

U-PC

Université Sorbonne
Paris Cité

PARIS
université
DIDEROT
PARIS 7

PHD THESIS

FROM UNIVERSITÉ SORBONNE PARIS CITÉ
PREPARED AT UNIVERSITÉ PARIS DIDEROT
ÉCOLE DOCTORALE DE GÉOGRAPHIE DE PARIS (ED 434)

UMR CNRS 8504 Géographie-cités / Équipe P.A.R.I.S.
UMR-T IFSTTAR 9403 LVMT

CARACTÉRISATION ET MODÉLISATION DE LA CO-ÉVOLUTION DES RÉSEAUX DE TRANSPORT ET DES TERRITOIRES

Presented by JUSTE RAIMBAULT

PhD Thesis in Geography

Under the supervision of ARNAUD BANOS and FLORENT LE NÉCHET

Presented and defended publicly at the Institut des Systèmes Complexes (Paris) on June 11th 2018, with a jury composed by :

Professeure, Université Paris 1 (Présidente du Jury)
Directeur de Recherche, CNRS (Rapporteur)
Professeure, Ecole Polytechnique de Montréal (Rapporteuse)
Chargé de Recherche, IFSTTAR (Examinateur)
Directrice de Recherche, IFSTTAR (Examinateuse)
Directeur de Recherche, CNRS (Directeur)
Maître de Conférences, Université Paris-Est (Directeur)



This work is licensed under a Creative Commons
Attribution-NonCommercial-ShareAlike 4.0 International
License.

JUSTE RAIMBAULT: *Caractérisation et modélisation de la co-évolution des réseaux de transport et des territoires*, Mémoire de Thèse de Doctorat, © 16 Février 2018

TITRE

Caractérisation et modélisation de la co-évolution des réseaux de transport et des territoires

RÉSUMÉ

L'identification d'effets structurants des infrastructures de transports sur la dynamique des territoires reste un défi scientifique ouvert. Cette question est une des facettes de recherches sur la complexité des dynamiques territoriales, au sein desquelles territoires et réseaux de transport seraient en co-évolution. L'objectif de cette thèse est de mettre à l'épreuve cette vision des interactions entre réseaux et territoires, autant sur le plan conceptuel que sur le plan empirique, en les intégrant au sein de modèles de simulation des systèmes territoriaux. La nature intrinsèquement pluri-disciplinaire de la question nous conduit à mener un travail d'épistémologie quantitative, qui permet de dresser une carte du paysage scientifique et une description des éléments communs et des spécificités des modèles traitant la co-évolution entre réseaux et territoires dans chaque discipline. Nous proposons ensuite une définition de la co-évolution, ainsi qu'une méthode de caractérisation empirique, basée sur une analyse de corrélations spatio-temporelles. Deux pistes complémentaires de modélisation, correspondant à des ontologies et des échelles différentes sont alors explorées. A l'échelle macroscopique, nous construisons une famille de modèles dans la lignée des modèles d'interaction au sein des systèmes de villes développés par la Théorie Evolutive des Villes (Pumain, 1997). Leur exploration montre qu'ils capturent effectivement des dynamiques de co-évolution, et leur calibration sur des données démographiques pour le système de villes français (1830-1999) quantifie l'évolution des processus d'interaction comme l'effet tunnel ou le rôle de la centralité. A l'échelle méso scopique, un modèle de morphogenèse capture la co-évolution de la forme urbaine et de la topologie du réseau. Il est calibré sur les indicateurs correspondants pour la forme et la topologie locales calculés pour l'ensemble de l'Europe. De multiples processus d'évolution du réseau s'avèrent être complémentaires pour reproduire la grande variété des configurations observées, au niveau des indicateurs ainsi que des interactions entre indicateurs. Ces résultats suggèrent de nouvelles pistes d'exploration des modèles urbains intégrant les dynamiques co-évolutives dans une perspective multi-échelles.

MOTS-CLEFS

Territoires ; Réseaux de Transport ; Co-évolution ; Morphogenèse ; Théorie Évolutive des Villes ; Épistémologie Quantitative ; Systèmes de Villes ; Morphologie Urbaine ; Grand Paris ; Delta de la Rivière des Perles

TITLE

Characterizing and modeling the co-evolution of transportation networks and territories

ABSTRACT

The identification of structuring effects of transportation infrastructure on territorial dynamics remains an open research problem. This issue is one of the aspects of approaches on complexity of territorial dynamics, within which territories and networks would be co-evolving. The aim of this thesis is to challenge this view on interactions between networks and territories, both at the conceptual and empirical level, by integrating them in simulation models of territorial systems. The intrinsically multidisciplinary nature of the question requires first to proceed to a quantitative epistemology analysis, that allow us to draw a map of the scientific landscape and to give a description of common features and specificities of models studying the co-evolution between network and territories within each discipline. We propose consequently a definition of co-evolution and an empirical method for its characterization, based on spatio-temporal correlation analysis. Two complementary modeling approaches, that correspond to different scales and ontologies, are then explored. At the macroscopic scale, we build a family of models inheriting from interaction models within system of cities, developed by the Evolutive Urban Theory (Pumain, 1997). Their exploration shows that they effectively capture co-evolutionary dynamics, and their calibration on demographic data for the French system of cities (1830-1999) quantifies the evolution of interaction processes such as the tunnel effect or the role of centrality. At the mesoscopic scale, a morphogenesis model captures the co-evolution of the urban form and of network topology. It is calibrated on corresponding indicators for local form and topology, computed for all Europe. Multiple network evolution processes are shown complementary to reproduce the large variety of observed configurations, at the level of indicators but also interactions between indicators. These results suggest new research directions for urban models integrating co-evolutive dynamics in a multi-scale perspective.

KEYWORDS

Territories; Transportation Networks; Co-evolution; Morphogenesis; Evolutive Urban Theory; Quantitative Epistemology; Systems of Cities; Urban Morphology; Greater Paris; Pearl River Delta

标题

建模交通网络和地域的共同演变

摘要

运输基础设施对领土体系结构效应存在的问题远未得到解决。这是复杂的地域动态的一个方面，其中领土和交通网络正在共同演变。这篇论文的目的是测试网络和地域之间的相互作用。它将在概念和经验上做到这一点，目的是将其整合到地域系统的模拟模型中。我们正在处理的问题本质上是多学科的。出于这个原因，我们首先进行量化的认识论分析。它可以绘制科学的景观图，并精确地描述每个学科不同模型的结构。我们制定了一个共同进化的定义，并开发了一个基于时空相关分析的经验表征方法。探索两个互补的建模轨道。它们对应于不同的本体和尺度。在宏观层面上，我们根据城市演变理论发展起来的城市体系内的相互作用模型发展了一个模型家族。他们的探索表明，他们实际上捕捉到共同演化的动力。他们对法国城市系统（1830-1999）的人口统计数据的校准量化了互动过程的演变。这些例如是隧道效应或网络中心性的影响。在介观尺度上，形态演化模型捕捉城市形态和网络拓扑的共同演化。根据整个欧洲计算的局部形态和拓扑结构的相应指标进行校准。网络演进的多个过程被考虑到：成本效益计划，潜在的突破，自组织。它们似乎是互补的，可以产生所有的真实配置。校准也是按照第二顺序进行的，也就是指标之间的相互作用，模型重现了现有情况的多样性。这些结果一方面表明了把城市演变理论与形式演变相结合的理论建构。另一方面，他们开辟了新一代城市模式的探索，这些模型将不得不整合多尺度协同进化动力学。

关键字

地域; 交通网络; 共同演变; 形态; 演变城市理论; 量化认识论; 城市系统; 城市形态; 大巴黎; 珠江三角洲

READING NOTES

This thesis was initially intended to be written in English. A first third and most of papers were, to be then adapted and translated into French, in order to fulfill an administrative constraint from another age. It has also been thought as a “Paper Thesis”, but the strong recommendations of CNU have rapidly swept this ambition. Therefore, the current version has gone through several transformations and “smoothing”, in order to give it a “classical” form, background and identity. We apologize in advance to the reader if translation or articulation issues remain and disturb the fluidity of the reading, since this English version was moreover fully translated again back from French.

The layout is designed to be narrow in order to allow the reader to write notes on the manuscript where he wants, on the digital or paper version: maybe the dream of all manuscript is to become interactive.

All the figures in main text are produced by the author, at the exception of Fig. 12 (source xkcd <https://xkcd.com/>) and two illustrations in the Frame 11. A large majority of figures are *directly* reproducible, i.e. can be obtained by executing the scripts. All source code, from models to the interpretation of results and to this proper writing, is available openly with all its atomic history (*commits*) on the repository of the project¹. All the datasets produced in that frame are open, and all data used are open or made open (in an aggregated way corresponding to the level of use by models in the case of a third-party closed database).

This memoir in itself has been proofread by the following readers (in alphabetical order): Arnaud Banos (AB), Clémentine Cottineau (CC), Florent Le Néchet (FL), Cinzia Losavio (CL), Sébastien Rey (SR), Hélène Serra (HS) in the spirit of an open review. By following the successive commits at <https://github.com/JusteRaimbault/ThesisMemoire>, the use of specific commands for the review remarks allows to track the full review process.

Names in Mandarin (cities, places, people, etc.) are transcribed using the *pinyin* system.

¹ at <https://github.com/JusteRaimbault/CityNetwork>

PUBLICATIONS

The following publications and communications contain most content of this thesis. Sources are precisely mentioned at the beginning of each chapter. Translations are ensured by the author when they are needed.

PUBLICATIONS

Raimbault, J. (2018). Indirect evidence of network effects in a system of cities. *Environment and Planning B: Urban Analytics and City Science*, 2399808318774335.

Raimbault, J. (2018). Calibration of a density-based model of urban morphogenesis. *PloS ONE, in revision*. arXiv preprint arXiv:1708.06743.

Raimbault, J. (2017). Exploration of an interdisciplinary scientific landscape. *Scientometrics, in revision*. arXiv preprint arXiv:1712.00805.

Raimbault, J. (2018). An urban morphogenesis model capturing interactions between networks and territories, *forthcoming in Mathematics of Urban Morphology. D'Acci L., ed. Springer Nature - Birkhäuser Mathematics*. arXiv preprint arXiv:1805.05195

Raimbault, J. (2018). Models for the co-evolution of cities and networks, *forthcoming in Handbook on Cities and Networks, Rozenblat C., Neal Z., eds. arXiv preprint arXiv:1804.09430*

Raimbault, J. (2018). Complexity, complexities and complex knowledge, *forthcoming in Theories and models of urbanization. Pumain D., ed. Springer Lecture Notes in Morphogenesis*.

Raimbault, J. (2018). Unveiling co-evolutionary patterns in systems of cities : systematic exploration of the SimpopNet model, *forthcoming in Theories and models of urbanization. Pumain D., ed. Springer Lecture Notes in Morphogenesis*.

Raimbault, J. (2018). Mutations of transportation networks in China, *forthcoming in Medium Project e-book. Aveline N., ed.*

Raimbault, J. (2017). An applied knowledge framework to study complex systems, *Complex Systems Design & Management* (pp.31-45).

Raimbault, J. (2017). Identification de causalités dans des données spatio-temporelles, *Spatial Analysis and GEOMatics 2017*.

Bergeaud, A., Potiron, Y., & Rimbault, J. (2017). Classifying patents based on their semantic content. *PloS one*, 12(4), e0176310.

Rimbault, J. (2017). A discrepancy-based framework to compare robustness between multi-attribute evaluations, *Complex Systems Design & Management* (pp. 141-154). Springer International Publishing.

Rimbault, J. (2017). Investigating the empirical existence of static user equilibrium. *Transportation Research Procedia*, 22, 450-458.

Rimbault, J. (2017). Models coupling urban growth and transportation network growth: An algorithmic systematic review approach. *Plurimondi*, (17).

Rimbault, J. (2016). Generation of correlated synthetic data, *Actes des Journées de Rochebrune 2016*.

WORKING PAPERS

Rimbault, J. (2018). Co-evolution and morphogenetic systems. *Rejected for Artificial Life 2018*. arXiv preprint arxiv:1803.11457.

Rimbault, J. & Bergeaud, A. (2018). A large scale analysis of fuel prices spatial variability. *Under review for Transportation Research Part A*. arXiv preprint arxiv:1706.07467.

Banos A., Chasset P.-O., Commenges A. Cottineau C., Pumain D. & Rimbault J. (2018). Where do you mean? A spatialised bibliometrics approach of a scientific journal production. *Under review for RSOS; desk rejected from Big Data and Society, Journal of Informetrics, Science Communication*.

Rimbault J., Cottineau C., Le Texier M., Le Nechet F. & Reuillon R. (2017). Space Matters: extending sensitivity analysis to initial spatial conditions in geosimulation models. *Under review for Computer, Environment and Urban Systems; rejected from Environment and Planning B*.

Antelope, C., Hubatsch, L., Rimbault, J., and Serna, J. M. (2016). An interdisciplinary approach to morphogenesis. *Working Paper, Santa Fe Institute CSSS 2016*.

COMMUNICATIONS

Multi-modeling the morphogenesis of transportation networks, *extended abstract forthcoming in Proceedings of ALife 2018, Tokyo, July 2018*.

Modeling Urban Morphogenesis: towards an integration of territories and networks, *GoPro 2017, Lyon, Dec. 2017*.

Modeling the Co-evolution of Urban Form and Transportation Networks, *Conference on Complex Systems 2017, Cancun, Sept. 2017.*

Rimbault J. & Baffi S. (2017). Structural Segregation: Assessing the impact of South African Apartheid on Underlying Dynamics of Interactions between Networks and Territories, *ECTQG 2017, York, Sept. 2017.*

Invisible Bridges ? Scientific landscapes around similar objects studied from Economics and Geography perspectives, *ECTQG 2017, York, Sept. 2017.*

Rimbault, J. & Bergeaud, A. (2017). The Cost of Transportation: Spatial Analysis of Fuel Prices in the US, *EWGT 2017, Budapest, Sept. 2017.*

Cottineau C., Rimbault J., Le Texier M., Le Néchet F. & Reuillon R. (2017). Initial spatial conditions in simulation models: the missing leg of sensitivity analyses?, *Geocomputation 2017, Leeds, Sept. 2017*

A macro-scale model of co-evolution for cities and transportation networks, *Medium International Conference, Guangzhou, June 2017.*

Losavio C. & Rimbault J. (2017). Agent-based Modeling of Migrant Workers Residential Dynamics within a Mega-city Region: the Case of Pearl River Delta, China, *Urban China Development International Conference, London, May 2017.*

Co-construire Modèles, Etudes Empiriques et Théories en Géographie Théorique et Quantitative: le cas des Interactions entre Réseaux et Territoires, *Treizièmes Rencontres de ThéoQuant, Besançon, May 2017.*

Un Cadre de Connaissances pour une Géographie Intégrée, *Journée des jeunes chercheurs de l'Institut de Géographie de Paris, Paris, April 2017.*

Towards a Theory of Co-evolutive Networked Territorial Systems: Insights from Transportation Governance Modeling in Pearl River Delta, China, *MEDIUM Seminar : Sustainable Development in Zhuhai, Guangzhou, Dec 2016.*

Models of growth for system of cities : Back to the simple, *Conference on Complex Systems 2016, Amsterdam, Sept. 2016.*

Rimbault J., Bergeaud A. and Potiron Y. (2016). Investigating Patterns of Technological Innovation. *Conference on Complex Systems 2016, Amsterdam, Sep 2016.*

For a Cautious Use of Big Data and Computation. *Royal Geographical Society - Annual Conference 2016 - Session : Geocomputation, the Next 20 Years (1), London, Aug 2016.*

Indirect Bibliometrics by Complex Network Analysis. *20e Anniversaire de Cybergeo, Paris, May 2016.*

Raimbault, J. & Serra, H. (2016). Game-based Tools as Media to Transmit Freshwater Ecology Concepts, *poster corner at SETAC 2016 (Nantes, May 2016).*

Le Néchet, F. & Raimbault, J. (2015). Modeling the emergence of metropolitan transport authority in a polycentric urban region, *ECTQG 2015, Bari, Sept. 2015).*

Hybrid Modeling of a Bike-Sharing Transportation System, *poster presented at ICCSS 2015, Helsinki, June 2015.*

Raimbault, J. & Gonzales, J. (2015). Application de la Morphogénèse de Réseaux Biologiques à la Conception Optimale d'Infrastructures de Transport, *poster presented at Rencontres du Labex Dynamite, Paris, May 2015.*

REMERCIEMENTS

Une grande partie des résultats obtenus dans cette thèse ont été calculés sur l'organisation virtuelle vo.complex-system.eu de l'European Grid Infrastructure (<http://www.egi.eu>). Je remercie l'European Grid Infrastructure et ses National Grid Initiatives (France-Grilles en particulier) pour fournir le support technique et l'infrastructure.

Ce travail de recherche a été mené dans le cadre du project MEDIUM (New pathways for sustainable urban development in China's MEDIUM sized-cities). Je remercie donc le Centre National de la Recherche Scientifique (CNRS) et l'UMR 8504 Géographie-cités pour leurs soutien ainsi que les partenaires de MEDIUM, en particulier la Sun-Yat-Sen University. Le projet MEDIUM a été cofinancé par l'Union européenne au titre de l'Action Extérieure de l'UE – Contrat de subvention ICI+/2014/348-005.

Je tiens à remercier Denise Pumain pour l'honneur qu'elle me fait d'accepter la présidence du Jury, les examinateurs Olivier Bonin et Anne Ruas pour accepter d'évaluer ce travail, et les rapporteurs Didier Josselin et Catherine Morency pour avoir assuré le conséquent travail de digérer chaque mot de ce manuscrit.

Un travail de thèse est à la fois improbable et évident. Improbable car on ne s'imaginait pas quelques années en arrière quel pourrait bien être le sujet avec lequel on devrait s'obséder pendant trois longues années. Improbable quand on regarde l'écart entre le projet initial et les cavités et les arêtes effilées qui ont finalement été explorées. Mais aussi évident quand on regarde ce même projet, et qu'on retrouve les graines de chacun des développements fondamentaux, suggérant une morphogenèse de la connaissance. Étrangement évident par un travail d'introspection : les métiers des rêves de mon enfance ont été conducteur de métro puis cartographe, peut-être n'est-ce pas une coïncidence si le cœur du sujet ici rassemble les transports et les territoires. Évident et improbable quand on contemple les futurs possibles, la métá-structure qui s'en dégage finalement. Autant un commencement qu'une fin, un moyen qu'une finalité, une trajectoire qu'une position, une fête qu'une tristesse, une poésie qu'un rapport technocratique. Je vais tenter de remercier ici tout ceux qui ont permis la concrétisation de cette complexité.

Ma profonde reconnaissance va naturellement à mes directeurs, qui ont rendu cette aventure possible et ont permis sa forme finale, par un

pilotage subtil du système complexe que formaient objets, modèles, idées. J'ai rencontré pour la première fois Arnaud Banos en octobre 2012 à l'ISC qu'il dirigeait, alors toujours rue Lhomond. C'était dans le cadre d'une supervision des *Open Problems* du PA Systèmes Complexes, et nous nous étions immisés avec mon collègue Jorge dans le monde du multi-échelle, de l'optimisation multi-objectif, des réseaux biologiques auto-organisés (projet dont l'implémentation originale a d'ailleurs été reprise ici). Ou plutôt jetés inconsciemment à l'eau au risque de se noyer, merci à Arnaud de nous avoir repêchés. Je garde un certain nombre de paradigmes fondamentaux qu'il nous avait transmis dès le premier contact avec la recherche. Cette bifurcation coïncide étrangement avec une autre plus personnelle, peut être ironiquement pour rappeler la place du sujet dont l'objectivité de la recherche ne fait aucun sens.

Ma première rencontre avec Florent Le Néchet a eu lieu en mars 2014, à la cafétéria des Ponts, pour discuter de ce projet de thèse. Naïvement, je lui présentais mon modèle RBD ainsi que des idées floues sur les ruptures de potentiel. Il a alors immédiatement donné de la profondeur au projet, en évoquant les Mega-city Regions, les nouveaux régimes urbains, la Chine : vision finalement prémonitoire (ou prophétie auto-réalisatrice ?). La richesse de ses idées n'a cessé d'irriguer ce travail mais aussi mes reflexions de manière plus générale. Sans lui, cette thèse n'aurait de géographie que le nom, et je lui suis fortement reconnaissant d'avoir été patient devant mes difficultés à appréhender les Sciences Humaines et Sociales.

Par ailleurs, même si Denise Pumain n'a pas officiellement dirigé cette thèse, son conseil a été d'une valeur inestimable, autant sur le plan thématique qu'épistémologique. Son intérêt pour les différents projets a été une source de motivation considérable, comme pour les nombreux projets futurs en perspective. Enfin, son soutien académique a été précieux.

Je remercie les acteurs académiques ayant accepté de mener des entretiens qui ont servi de matériau de recherche : Denise Pumain, Romain Reuillon, Clémentine Cottineau et Alain Bonnafous.

Le soutien technique a été crucial, je remercie l'équipe d'OpenMole et en particulier Romain Reuillon pour sa rapidité de réponse et de résolution des problèmes. Je remercie Maziyar Panahi pour le soutien technique sur Zebulon.

Je remercie les différents relecteurs de ce mémoire qui ont grandement contribué à le rendre lisible : Arnaud Banos, Clémentine Cottineau, Florent Le Néchet, Cinzia Losavio, Sébastien Rey, Hélène Serra. Je remercie également ceux avec qui les discussions ont été déterminantes dans la fin de la rédaction : Nicolas Coulombel, Hadrien Com-

menges, Caroline Gallez.

Cette trajectoire de recherche n'aurait pas été possible sans les personnes qui ont joué un rôle clé dans ma formation. Je tiens ainsi à remercier Paul Bourgine et Kashayar Pakdaman pour m'avoir introduit aux systèmes complexes, et l'ensemble de l'équipe pédagogique du Master pour la qualité de l'enseignement, en particulier René Dourrat pour son efficacité de formation à la recherche. Je remercie Eric Marandon pour la qualité scientifique et la stimulation intellectuelle pendant le stage chez L2. Je remercie également l'équipe du Département Ville, Environnement, Transport des Ponts, en particulier Nicolas Coulombel, Fabien Leurent, Zoi Cristoforou, Antoine Picon.

Le projet Medium a déjà été mentionné "officiellement", mais je me dois de remercier personnellement Natacha Aveline pour m'avoir donné l'opportunité d'y participer, Chenyi Shi et Ming pour leur aide précieuse à Zhuhai, Florent Resche-Rigon pour la supervision, Céline Rozenblat pour l'organisation de la session modélisation à la conférence Medium, et les participants Cinzia Losavio, Valentina Ansoize, Judith Audin, Yinghao Li pour les moments passés à Zhuhai.

Les écoles d'été ont également pris une place importante dans ma formation. Je remercie l'ensemble de l'équipe pédagogique et des participants de la SFI Complex Systems Summer School 2016 à Santa Fe et ceux de l'École d'été du Labex Dynamite 2014 à Florence.

Je remercie également mes co-auteurs et collaborateurs sur les différents projets reliés de près ou de loin à cette thèse :

- l'équipe SpaceMatters Clémentine Cottineau, Florent Le Néchet, Marion Le Texier, Romain Reuillon ;
- l'équipe CybergeoNetworks Arnaud Banos, Pierre-Olivier Chasset, Clémentine Cottineau, Hadrien Commenges, Denise Pumain ;
- l'équipe de PatentsMining Antonin Bergeaud et Yoann Potiron ;
- Antonin Bergeaud pour EnergyPrice ;
- Hélène Serra pour le projet de communication scientifique ;
- Cinzia Losavio pour les dynamiques migratoires en Chine ;
- Solène Baffi pour les dynamiques structurelles en Afrique du Sud ;
- l'équipe Morphogenesis Chenling Antelope, Lars Hubatsch, Jesus Mario Serna ;

- Florent Le Néchet pour Lutecia.

Je remercie Benjamin Carantino pour l'organisation conjointe de la session Eco-geo à ECTQG2017, et les participants invités Antonin Bergeaud, Clémentine Cottineau, Olivier Finance, Céline Rozenblat, Medhi Bida, Elfie Swerts, Denise Pumain, d'avoir accepté d'y participer.

Je remercie Céline Rozenblat, Luca D'Acci et Denise Pumain de m'avoir invité à rédiger divers chapitres d'ouvrage rendant compte de ce travail de thèse.

Apprendre c'est aussi apprendre à apprendre, et donc à enseigner. Je remercie les membres de l'équipe pédagogique de Paris 7 qui ont rendu cette expérience agréable, et pardonne ceux qui m'ont fait souffrir par psycho-rigidité. Dans les moments difficiles, la curiosité des élèves a été vraiment porteuse de sens, et je remercie tout ceux qui étaient motivés et qui ont aimé appréhender la multi-modélisation.

Les laboratoires qui m'ont accueilli ont joué un rôle déterminant dans la réussite de cette thèse (malgré les difficultés récurrentes de financement qui laissent pessimiste sur l'avenir de la recherche publique). Je remercie les différents membres de Géographie-cités (Oven Street et Olympe) et du LVMT qui ont rendu l'accueil toujours chaleureux. Je remercie en particulier parmi les doctorants Thibault Le Corre pour le soutien intellectuel et logistique, Julien Migozzi pour le soutien théâtral, Paul Gourdon pour le soutien poétique, Pierre-Olivier Chasset pour le soutien informatique, Daphnée Caillol pour le soutien qualitatif, Mathieu Pichon pour le soutien épistémologique, Anne-Cécile Ott pour le soutien pédagogique, Flora Hayat pour le soutien cartographique, Anaïs Dubreuil pour le soutien alpin, Laetita Verhaeghe pour le soutien territorial, Ryma Hachi pour le soutien réticulaire, Natalia Zdanowska pour le soutien sportif, Cinzia Losavio pour le soutien ethnographique, Eugenia Viana pour le soutien moral ; les anciens Solène Baffi, Brenda Le Bigot, Olivier Finance, Julie Gravier, Lucie Nahassia, Robin Cura, Etienne Toureille ; les titulaires Thomas Louail, Clémentine Cottineau, Paul Chapron, Hadrien Commenges, François Queroy pour les discussions stimulantes ; et tous ceux que j'oublie.

Je remercie Joris, Mario, Marius pour les expériences autant scientifiques que d'amitié, dédicace circulaire.

Je remercie Cinzia, Chenyi, Ming, Jing Jing, Xing et Meng pour l'expérience chinoise et la patience devant mes difficultés linguistiques.

La recherche c'est une vie et malheureusement souvent oublier sa vie, je suis extrêmement reconnaissant à mes amis qui m'ont permis d'en garder un semblant : Alexis, Emmanuel, les SFR, Antonin, Yoann, Maximilien, Simon, Arnaud, Hélène, Axel, Jonas, Nihal, Fabrice. Je remercie (partiellement seulement, pour la quantité de vie injectée dans ce travail en conséquence) celle qui m'a laissé rapidement seul avec ce démon de thèse, et celles et ceux qui m'ont permis de me sentir moins seul par moments. Enfin je remercie ma famille dont la présence a été indispensable.

★ ★

★

CONTENTS

Introduction	1
I FONDATIONS	19
1 INTERACTIONS BETWEEN NETWORKS AND TERRITORIES	23
1.1 Territories and networks	25
1.2 Transportation projects from Paris to Zhuhai	47
1.3 Fieldwork observations of interactions	65
2 MODÉLISER LES INTERACTIONS ENTRE RÉSEAUX ET TERRITOIRES	79
2.1 Modeling Interactions	81
2.2 An epistemological Approach	96
2.3 Systematic Review and Modelography	110
3 POSITIONING	127
3.1 Modeling, big data and intensive computing	129
3.2 Reproducibility	146
3.3 Epistemological positioning	165
II BRIQUES ÉLÉMENTAIRES	187
4 THÉORIE EVOLUTIVE URBAINE	197
4.1 Correlations between form of territories and network topology	202
4.2 Spatio-temporal causalities	218
4.3 Macroscopic growth model	236
5 MORPHOGENÈSE URBAINE	257
5.1 An Interdisciplinary Approach to Morphogenesis	259
5.2 Urban morphogenesis by aggregation-diffusion	274
5.3 Generation of correlated territorial configurations	290
III SYNTHÈSE	303
6 CO-EVOLUTION AT THE MACROSCOPIC SCALE	307
6.1 Exploring macroscopic models of co-evolution	309
6.2 Dynamical extension of the interaction model	319
7 CO-EVOLUTION AT THE MESOSCOPIC SCALE	339
7.1 Network growth models	340
7.2 Co-evolution at the mesoscopic scale	352
7.3 Co-evolution and governance	361
Conclusion et Ouverture	383
8 CONCLUSION ET OUVERTURE THÉORIQUE	387
8.1 Contributions and Perspectives	388
8.2 A geographical theory	399
8.3 An Applied Knowledge Framework	406
BIBLIOGRAPHY	439

IV ANNEXES	487
A INFORMATIONS SUPPLÉMENTAIRES	489
A.1 Fieldwork Elements	491
A.2 Quantitative Epistemology	498
A.3 Modelography	503
A.4 Static Correlations	511
A.5 Causality regimes	525
A.6 Aggregation-diffusion morphogenesis	529
A.7 Correlated Synthetic data	538
A.8 Exploration of the SimpopNet model	540
A.9 Macroscopic co-evolution model	544
A.10 Network generation heuristics	552
A.11 Co-evolution at the mesoscopic scale	555
A.12 Transportation system governance modeling	556
B DÉVELOPPEMENTS MÉTHODOLOGIQUES	565
B.1 An unified framework for stochastic models of urban growth	566
B.2 Sensitivity of Urban Scaling Laws to Spatial Extent	570
B.3 Generation of Correlated Synthetic Data	574
B.4 A Discrepancy-based Framework	577
B.5 Socio-technical Systems	592
B.6 Exploration of an Interdisciplinary Scientific Landscape	603
C DÉVELOPPEMENTS THÉMATIQUES	625
C.1 Road Network and Prices Drivers	627
C.2 Multi-scalar modeling of residential dynamics	641
C.3 Generation of Correlated Synthetic Data	648
C.4 CybergeoNetworks : a multi-dimensional and spatialized bibliometric	655
C.5 Classifying Patents Based on their Semantic Content	680
C.6 Bridges between Economics and Geography	703
C.7 Gamed-based tools as media to transmit freshwater ecology concepts	705
D DONNÉES	711
D.1 Grand Paris Traffic Data	711
D.2 Topological Road Network	711
D.3 Interviews	712
D.4 Synthetic Data and simulation results	713
E OUTILS	715
E.1 Softwares and Packages	716
E.2 Architecture and Sources for Algorithms and Models of Simulation	719
E.3 Tools and Workflow for an open Reproducible Research	724
F REFLEXIVE ANALYSIS	727
F.1 Hypernetwork analysis	727
F.2 Interaction between projects	730

LIST OF FIGURES

Figure 1	Successive transportation projects for Greater Paris	51
Figure 2	Impact of <i>Grand Paris Express</i> on accessibility	52
Figure 3	Empirical lagged correlations between accessibility gain and territorial variables	56
Figure 4	Accessibility gain induced by the Hong-Kong-Zhuhai-Macao gain	62
Figure 5	High speed network in China	69
Figure 6	TOD in Hong-Kong and Zhuhai	70
Figure 7	Systematic review algorithm workflow	100
Figure 8	Citation Network	105
Figure 9	107
Figure 10	Systematic Review	113
Figure 11	Coupling types	115
Figure 12	Naïve use of data mining and computation	139
Figure 13	Relative distances of phase diagrams to the reference	144
Figure 14	Examples of phase diagrams	144
Figure 15	Reproducibility and visualization	148
Figure 16	Web-application for traffic data	157
Figure 17	Spatial variability of shortest paths	158
Figure 18	Variability of travel time and distance	159
Figure 19	Temporal stability of maximal betweenness centrality	161
Figure 20	Spatial auto-correlations for relative travel speed	163
Figure 21	Spatial distribution of morphologies	205
Figure 22	Spatial distribution of network indicators	211
Figure 23	Spatial correlations between morphological indicators and network indicators	213
Figure 24	Variation of correlations with scale	214
Figure 25	Auto-regressive time-series	226
Figure 26	Correlation in the RBD model	229
Figure 27	Identification of interaction regimes	230
Figure 28	Evolution of network measures	233
Figure 29	Lagged correlations in South Africa	234
Figure 30	Time-series correlations	245
Figure 31	Model output	246
Figure 32	Evidence of network effects	247
Figure 33	Calibrating the Gravity Model	248
Figure 34	Calibrated parameters	249
Figure 35	Full model calibration	252
Figure 36	Example of generated territorial forms	280
Figure 37	Behavior of indicators	283
Figure 38	Randomness and frozen accidents	285

Figure 39	Model calibration	287
Figure 40	PSE exploration	288
Figure 41	Exploration of feasible space for correlations between urban morphology and network structure	295
Figure 42	Examples of generated coupled configurations	296
Figure 43	Behavior of the SimpopNet model	316
Figure 44	Correlations patterns in space and time	318
Figure 45	Schematic model representation	320
Figure 46	Temporal behavior of the co-evolution model	324
Figure 47	Aggregated behavior of the co-evolution model	325
Figure 48	Profiles of lagged correlations	328
Figure 49	Empirical lagged correlations for the French system of cities	332
Figure 50	Pareto fronts for the calibration on population and distance	334
Figure 51	Evolution of calibrated parameters	335
Figure 52	Example of application of the macroscopic model with a self-reinforcing network	336
Figure 53	Biological network generation example	344
Figure 54	Network examples for different generation heuristics	346
Figure 55	Feasible topological space	348
Figure 56	Comparison to real networks	349
Figure 57	Morphogenesis at the mesoscopic scale	353
Figure 58	Calibration of the morphogenesis model	356
Figure 59	Causality regimes for the morphogenesis model	359
Figure 60	Network topologies obtained for different levels of governance	372
Figure 61	Impact of co-evolution on accessibility in the Lutecia model	374
Figure 62	Application of Lutecia to Pearl River Delta	375
Figure 63	Calibration of the Lutecia model	377
Figure 64	Citation Network of main publications of Evolutive Urban Theory	412
Figure 65	Full network of knowledge domains	416

LIST OF TABLES

Table 1	Synthesis of the approach by scales	40
Table 2	Public transportation in Pearl River Delta	61
Table 3	Interaction processes between networks and territories	75
Table 4	Synthesis of modeling processes	94
Table 5	Stationary lexical proximities	101
Table 6	Description of citation communities	104
Table 7	Semantic communities	106
Table 8	Model type	118

Table 9	Explanation of models characteristics	121
Table 10	Synthesis of processes included in models	123
Table 11	Interrelations between morphological indicators and network indicators	216
Table 12	Model Parameters summary	242
Table 13	Empirical AIC values	254
Table 14	Summary of parameters of the morphogenesis model . . .	278
Table 15	Sensitivity to space of the SimpopNet model	315
Table 16	Summary of network growth parameters	345
Table 17	Summary of LUTECIA model parameters	370
Table 18	Processus taken into account in our models	392
Table 19	Behavior of models regarding co-evolution	396
Table 20	Illustration of Knowledge Framework Application	413
Table 29	List of words and its probability to belong to the topic . .	671

INTRODUCTION

INTRODUCTION

Would the fog machine on the plateau of Saclay be the only non temporal artefact in this metropolitan environment that still has to find its own identity ? Let's project us in 2100, in this south suburb of what still be Paris. Local transformations have indeed happened, but not in the expected way, the local climate being still fond of this well-known fog. However, the urban environment and the relation to the city are entirely conditioned by a proximity to heavy transportation lines: the disappearance of thermic transportation modes, then of all light vehicles through the technological failure of electric alternatives, have exacerbated the role of existing train or metro lines. Densities have progressively increased around stations to produce impressing tower compounds, whereas the peri-urban space became progressively empty. Concerning transportation infrastructures, they stayed quite at the identical after 2030, the few available resources being dedicated to their maintenance, and their extension became conjointly rapidly out of the political agendas. This plateau is therefore filled with abandoned buildings, since it still expects this line of the Grand Paris Express which finally would never have been realized. Nature progressively finds its way again.

This scenario for a low budget anticipation film has the advantage of revealing the existence of complex processes entangled at different space and time scales in the production of cities: the historical development of the railway network in the Parisian region conditioned the future evolutions, the RER B followed the old Ligne de Sceaux; the masterplan by DELOUVRIER for regional development and its incomplete realization are elements explaining the structure of the Parisian public transportation network which strongly condition urban development in our scenario; relocation processes within the metropolitan space, related to a more or less strong need for proximity or accessibility depending on transportation modes used, play they role in the urban evolution; in the case of the plateau of Saclay specific planning processes at different levels play a crucial role in the differentiation of the territory.

The list could be much further developed, since each approach brings its mature vision related to a scientific body of knowledge in different disciplines such as geography, urban economics, transportation. This anticipation scenario is enough to give a glance on the complexity of territorial systems we will study. Our aim here is to dive within this complexity, and more particularly to give an original viewpoint on the study of relations between transportation networks and territories. The choice of this positioning will be largely discussed in a thematic part, and we now concentrate on the originality of the point of view we will take.

ON GENERAL POSITIONING

The ambition of this thesis is to have no a priori ambition. Such an introduction, although seeming rash, contains at all levels the implicit logics behind our research process. At the first degree, we try as much as possible to take a exploratory and constructive approach, as much on theoretical and methodological domains than thematic domain, but also proto-methodological (tools applying the method) : if uni-dimensional or integrated ambitions should emerge, they would be conditioned by the arbitrary choice of a time sampling among the continuity of the dynamic that structures any research project. In the structural sense, the self-reference that underlines an apparent contradiction points out the central aspect of reflexivity in our constructive approach, as much in the sense of the recursion of theoretical apprals, than for application of tools and methods developed to the work itself, or in the sense of the co-construction of the different approaches and of the different thematic axis. The processus of knowledge production can this way be understood as a metaphor of studied processes. Finally, from a point of view closer to the interpretation, it suggests the intention of a delicate positioning linking a political positioning which necessity is intrinsic to humanities (for example here against the technocratic application of models, or for the development of tools for an Open Science) with a rigor of objectivity coming more from other fields used, position that impose an increased prudence.

SCIENTIFIC CONTEXT : PARADIGMS OF COMPLEXITY

To better introduce our subject, it is necessary to insist on the scientific context we are working in. This context is crucial both to understand the general epistemology underlying research questions, and to be aware of the variety of methods and tools used.

Contemporaneous science is progressively taking the shift of complexity in many fields that we will illustrate in the following, what implies an epistemological mutation to abandon strict reductionism² that failed in most of its synthesis attempts [Anderson, 1972]. [Arthur, 2015] recently recalled that a mutation of methods and paradigms was also at stake, through the increasing role of computational approaches replacing purely analytical techniques generally limited in their modeling and resolution scope. Capturing *emergent properties* in models of complex systems is one of the ways to understand the essence of these approaches.

² In a schematic way, reductionism consists in the epistemological positioning that systems are entirely understandable from the fundamental elements they are constituted of and from the laws driving their evolution. Superior levels have neither an autonomy nor irreducible causal powers.

These considerations are well known in Social Science and Humanities (both quantitative and qualitative), for which the complexity of studied agents and systems is one of the justifications of their existence: if humans were indeed particles, we could expect that most fields studying them would have never emerged, as thermodynamics would have solved most of social issues.³. They are however less known nor accepted in more “hard” sciences such as physics : [Laughlin, 2006] develops a view of physics at a similar position of a “frontier of knowledge” compared to other more recent fields that could appear as being still in their genesis. Most of knowledge concerns classical simple structures, whereas a large number of systems appear as *self-organized*, in the sense that the single microscopic laws are not enough to determine macroscopic properties unless system evolution is entirely simulated (more precisely this view can be taken as a definition of emergence on which we will come back later, and self-organized properties are indeed emergent). This corresponds to the first nightmare of Laplace’s Deamon developed in [Deffuant et al., 2015].

At the crossroads of epistemological positions, methods, and fields of applications, *Complexity Sciences* focus on the importance of emergence and self-organization in most of phenomena of the real world, which make it lie closer to a frontier of knowledge closer than we can imagine for classical disciplines [Laughlin, 2006]. These concepts are indeed not recent and had already been shown by [Anderson, 1972]. We can also interpret Cybernetics as a precursor of Complexity Sciences, by reading it as a bridge between technology and cognitive sciences [Wiener, 1948], and moreover by developing the notions of feedback and control.

Later, Synergetics [Haken, 1980] paved the way for a theoretical approach of collective phenomena in physics. Possible reasons for the recent growth of works claiming a complexity approach are numerous. The explosion of computing power is surely one of these because of the central role of numerical simulations [Varenne, 2010b]. They could also be related to epistemological progresses: introduction of the notion of perspectivism [Giere, 2010c], finer reflexions around the nature of models [Varenne and Silberstein, 2013]⁴.

The theoretical and empirical potentialities of such approaches play surely a role in their success⁵, as confirmed by the various domains of

³ Even if this affirmation can also be discussed, since classical physics also failed in their attempts to include irreversibility and evolutions of Complex Adaptive Systems as [Prigogine and Stengers, 1997] points out.

⁴ In that frame scientific and epistemological progresses can not be dissociated and can be seen as co-evolving, in the sense of a strong interdependency and a mutual adaptation

⁵ Although the adoption of new scientific practices may be strongly biased by imitation and lack of originality [Dirk, 1999], or in a more ambivalent way, by marketing strategies independent of knowledge strategies, as the fight for funds is becoming a huge obstacle for research [Bollen et al., 2014].

application (see [Newman, 2011] for a general survey), as for example Network Science [Barabasi, 2002]; Neuroscience [Koch and Laurent, 1999]; Social Sciences including Geography [Manson, 2001; Pumain, 1997]; Finance with econophysics approaches [Stanley et al., 1999]; Ecology [Grimm et al., 2005]. The Complex Systems Roadmap [Bourgine, Chavalarias, and al., 2009] proposes a double entry to studies on Complex Systems: an horizontal approach connecting fields of study with transversal questions on theoretical foundations of complexity and empirical common stylized facts, and a vertical approach to disciplines, with the aim at constructing integrated disciplines and corresponding multi-scale heterogeneous models. Interdisciplinarity is thus central in our scientific background.

INTERDISCIPLINARITY

We must further insist on the role of interdisciplinarity in the research positioning taken here. This is as much a work in Theoretical and Quantitative Geography than in Complex Systems Modeling, being finally both depending on the point of view taken by the reader. In that sense, we claim it to belong to *Complex Systems Science* that we aim at positioning as a proper discipline through this precise implementation⁶. There are risks of being read with mistrust or even defiance by scholars of various concerned disciplines, as recent examples of misunderstandings and conflicts have illustrated [Dupuy and Benguigui, 2015]. We need to recall the importance of BANOS' virtuous circle between disciplinarity and interdisciplinarity [Banos, 2013]. It must necessarily imply different scientific agents, and it is complicated for an agent to be positioned in the two branches; our scientific background will have to allow us to not be positioned only within *geographical disciplinarity* (even if it will simultaneously be a crucial component) but as much within Complex Systems (which is interdisciplinary, see 3.3 to go beyond the apparent contradiction), and our scientific and epistemological sensitivity leads us to do the same.

The scientific evolution of complexity sciences, that some see as a revolution [Colander, 2003], or even as *a new kind of science* [Wolfram, 2002], could indeed face intrinsic difficulties due to behaviors and a priori of researchers as human beings. More precisely, the need for interdisciplinarity which makes the strength of Complexity Science may be one of its greatest weaknesses, since the highly partitioned structure of the organization of science may have negative impacts on works involving different disciplines. We do not tackle the issue of over-publication, quantification, competition, which is more linked to a question of Open Science and its ethics, also of high importance but of an other nature. That barrier haunting us and that we might

⁶ An abstract level of reading of the work in its entirety will bring informations on knowledge production itself, as we will develop in 8.3.

struggle to triumph of, has as the most obvious symptom *cultural disciplinary differences*, and resulting opinion conflicts. The drama of scientific misunderstandings is that they can indeed totally annihilate progresses by interpreting as a falsification some works that answer a totally different question.

The recent example in economics of a work on top-income inequalities presented in [Aghion et al., 2015], which conclusions are presented as opposed to the ones obtained by [Piketty, 2013], is typical of this scheme. The latest focuses on the construction of long-time clean databases for income data and shows empirically a recent acceleration of income inequalities, his simple model aiming to link this stylized fact with the accumulation of capital has been criticized as oversimplified. On the other hand, [Aghion et al., 2015] show with econometric analyses that there indeed exist a causality link from innovation to top-income inequalities, the innovation however increasing social mobility, being thus also a driver of inequalities reductions. Therefore do they obtain divergent conclusion on the role of capitals in an economy, in particular on their ambiguous relation to innovation. But diverging *points of view* or *interpretations* do not imply a scientific incompatibility, and one could even imagine to try gathering both approaches in an unified framework and model, yielding possibly similar or different interpretations. Such an integrated approach will have chances to contain more information (depending on how coupling is done) and to be a scientific progress.

This thought experiment illustrates the potentialities and the necessity of interdisciplinarity. In an other but similar vein, [Holmes et al., 2017] reanalyses biological data from a 1943 experiment that claimed to rule out Lamarckian over Darwinian evolution processes, and show that the conclusions do not hold in the current context of data analysis (enormous advances in theoretical and processing techniques) and scientific context (with numerous other proofs today of Darwinian processes): this is a good example of a misunderstanding on the context and how conclusions strongly depend on both technical and thematic frameworks. We shall now briefly develop other examples to give an overview how conflicts between disciplines can be damaging.

As already mentioned, DUPUY and BENGUIGUI point out in [Dupuy and Benguigui, 2015] the fact that in the field of urban studies, have recently appeared open conflicts between classical heres of dicsciplines and new incomers, in particular physicists, even if their entry in this domain is not new. The availability of large datasets for new types of data (social networks, data from new information and communication technologies) have drawn an increased attention towards the study of objects traditionally studied by human sciences, as analytical and computational methods of statistical physics became applicable. Although these studies are generally presented as the con-

struction of a scientific approach to cities, discussing the scientific character of existing approach, the effective novelty of the results obtained and the discredit of “classical” approaches are discussable. To give a few examples, [Barthelemy et al., 2013] conclude that Paris has followed a transition during the Haussman period and it global planning operations, which are well-known facts for a long time in urban history and urban geography. [Chen, 2009] rediscovers that the gravity model can be improved by adding lags in interactions and theoretically derives the expression of the force of interaction between cities, without any thematic theoretical or thematical background. Similar examples could be multiplied, confirming the current discomfort between physicists and urban geographers. Significant benefices could results from a wise integration of disciplines [O’Sullivan and Manson, 2015] but the road seems to be still long.

Similar conflict can be found at the interface of relations between economics and geography: as [Marchionni, 2004] describes, the discipline of geographical economics, traditionally close to geography, has heavily criticized at its emergence the relatively recent approach of the *New Economic Geography*. This approach comes from economics and its purpose is to take space into account in classical economic methods. They have indeed not the same purposes and intentions, and the conflict appears as a complete misunderstanding when seen from an external eye. For example, the New Economic Geography will privilege explications that imply universal economic processes and independent of scales, whereas Geographical Economics will base its arguments on local particularities and the contingency of processes. Underlying epistemological assumptions are also very different, such as for example the relation to realism, the first being founded on an abstract realism which is not necessary concretely realistic (use of abstract processes), whereas the second will be more pragmatic. The extent in which these approaches are complementary or incompatible remains however an open question according to [Marchionni, 2004]. Similar disciplinary relations will be encountered in our work, such as between physics and geography. We furthermore illustrate this question in C.6 by an exploration of links between economics and geography from the point of view of modeling.

Disciplinary conflicts may also emerge under the form of a reject of novel methods by dominating currents. According to [Farmer and Foley, 2009], the operational failure of most classic economic approaches could be compensated by a broader use of agent-based modeling and simulation practices. The lack of analytical resolution which is inevitable for the study of most complex adaptive systems, seen to repel most of economists. However, [Barthelemy, 2016] insists on the exacerbated non-connection between numerous economic models and theories and empirical observation, at least in the field of urban economics. This could be a symptom of the disciplinary non-connection

evoked above. Still in economics, [Storper and Scott, 2009] also propose paradigms shifts for a return to the agent and an associated construction of *evidence-based theories*.

Quantitative finance can be instructive for our purpose and subject, through the similarities of its interdisciplinary kitchen with our domain (relations with physics and economics, fields more or less “rigorous”, etc.). In this domain coexist various fields of research having very few interactions between them. We can consider two example. On the one hand, statistics and econometrics are highly advanced in theoretical mathematics, using for example stochastic calculus and probability theory to obtain very refined estimators of parameters for a given model (see e.g. [Barndorff-Nielsen et al., 2011]). On the other hand, econophysics aims at studying empirical stylized facts and infer empirical laws to explain economic phenomena, for example the ones linked to complexity of financial markets [Stanley et al., 1999]. They include cascades leading to market crashes, fractal properties of asset signals, complex structure of correlation networks. Both have their advantages in a particular context and each would benefit from increased interactions between the fields.

These diverse examples caught in the wind give short illustrations of how crucial interdisciplinarity is and how it is difficult to achieve. Without being close to exaggerating, we could imagine all researchers complaining about bad or difficult experiences in interdisciplinarity, with a largely positive return in the rare cases of a success. We will in the following try to follow that narrow path, borrowing ideas, theories and methods from diverse disciplines, in the spirit of the construction of an integrated knowledge.

COMPLEXITY PARADIGMS IN GEOGRAPHY

Coming back to our introducing anecdote, we will focus on the study of a thematic object that will be territorial systems: at the microscopic scale, agents can indeed be seen as fundamental elements constituting the territory, which will emerge as a complex process at different scales. More generally, we propose to begin with sketching an overview of the role of complexity in geography. Geographers are naturally familiar with complexity, since the study of spatial interactions is one of their preferred object. The variety of fields in geography (geomorphology, physical geography, environmental geography, human geography, health geography, etc. to give a few) has certainly played a key role in the constitution of a subtle geographical thinking, which considers heterogeneous and multi-scalar processes.

[Pumain, 2003] gives a subjective history of the emergence of complexity paradigms in geography, that we synthesize here. Cybernetics yielded system theories such as the one used for first system dynamics models aiming at simulating the evolution of variables character-

izing a territory, under the form of coupled differential equations, as [Chamussy et al., 1984] illustrate for a model coupling population, employments and housing stock. Later, the shift towards concepts of self-organized criticality and self-organisation in physics lead to corresponding developments in geography, as [Sanders, 1992] which witnesses the application of concepts from synergetics to the dynamics of urban systems.

Finally, current paradigms of complex systems have been introduced through several relatively independent entries. We can exhibit among them concepts from fractals, cellular automata, *Scaling* concepts, and the evolutive urban theory. We briefly review these approaches below.

The study of the fractal nature of urban form was introduced by [Batty and Longley, 1994], has been later synthesized by [Batty and Longley, 1994] and had numerous application including more recent developments such as [Keersmaecker, Frankhauser, and Thomas, 2003] for analyzing the urban form or [Tannier et al., 2010] for the conception of sustainable urban planning.

The theory of *Scaling* has furthermore been imported from physics and biology (allometric relations) to explain urban scaling laws as universal properties linked to the type of activity: infrastructure and economies of scale (infralinear scaling) or resulting from a process of social interactions (supralinear scaling), and assumes cities as scaled versions of each other [Bettencourt et al., 2007]. We will not explicitly use these two approaches but they remain underlying in the paradigms we will use⁷.

Cellular automata, introduced in geography by TOBLER [Coulcelis, 1985], are an other entry of complex approaches for urban modeling. BATTY proposes a joint synthesis of it with agent-based models and fractals in [Batty, 2007]. This type of model will take a modest but not negligible place in our work.

An other incursion of complexity in geography was for the case of urban systems through the evolutive urban theory of PUMAIN. We will position more particularly within its heritage and will develop it with more details. In close relation with modeling from the beginning (the first Simpop model described in [Sanders et al., 1997] enters the theoretical framework of [Pumain, 1997]), this theory aims at understanding systems of cities as systems of co-evolving adaptive agents, interacting in many ways, with particular features emphasized such as the importance of the diffusion of innovations.

The series of Simpop models [Pumain, 2012a] was conceived to test various assumptions of the theory, such as the role of innovation diffusion processes in the organisation of the urban system. Thus,

⁷ For example, scaling laws have a privileged role in the application of the evolutive urban theory [Pumain et al., 2006].

different underlying regimes were revealed for systems of cities in Europe and in the United States [Bretagnolle and Pumain, 2010a].

At other time scales and in other contexts, the SimpopLocal model [Schmitt, 2014] aims at investigating the conditions for the emergence of hierarchical urban systems from disparate settlements. A minimal model (in the sense of sufficient and necessary parameters) has been isolated through to the use of intensive computation with the model exploration software OpenMole [Schmitt et al., 2015], what was a result impossible to obtain analytically for such a kind of complex model. The technical progresses of OpenMole [Reuillon, Leclaire, and Rey-Coyrehourcq, 2013] were done simultaneously with theoretical and empirical advancements.

Epistemological advances were also crucial to this framework, as [Rey-Coyrehourcq, 2015] develops, and new concepts such as incremental modeling [Cottineau, Chapron, and Reuillon, 2015] were discovered, with powerful concrete applications: [Cottineau, 2014] applies it on the soviet system of cities and isolates dominating socio-economic processes, by systematic testing of thematic assumptions and implementation functions. Directions for the development of such modeling and simulation practices in quantitative geography were recently introduced by [Banos, 2013]. He concludes with nine principles⁸, among which we can cite the importance of intensive exploration of computational models and the importance of heterogeneous models coupling, that are among other principles such as reproducibility at the center of the study of complex geographical systems from the point of view described before. We will be positioned mainly within the legacy of this line of research, working conjointly in the theoretical, empirical, epistemological and modeling aspects.

CITIES, SYSTEMS OF CITIES, TERRITORIES

We can enter now the heart of the matter to progressively construct the precise problematic which will enter the global context developed up to here. Our elementary geographical objects (in the sense of precursors in our theoretical genesis) will be the *City*, the *System of cities*, and the *Territory*, that we will now define.

A central element of socio-geographical systems is the *City* object, on which we position for a proper epistemological consistence. The question of the definition of the city has fostered numerous contributions. [Robic, 1982] shows for example that REYNAUD had already conceptualized the city as a central place of a geographical space, allowing aggregation and exchanges, theory that will be reformulated by CHRISTALLER as the *Central Place Theory*. This theoretical definition

⁸ Must it become the ten commandments ? RENÉ DOURSAT underlined the absence of the last Banos' commandment, the intrinsic essence of our enterprise may be linked to its pursuit.

is rejoined by the conception of PUMAIN which considers the city as a clearly identifiable spatial entity, constituted by social agents (that may be elementary or not) and of technical artifacts, and which is the incubator of social change and innovation [Pumain, 2010]. We will use this definition in our work. We must however keep in mind that the concrete definition of a city in terms of geographical entities and spatial extent is problematic: morphological definitions (i.e. based on the shape and the distribution of the built environment), functional definitions (based on the use of urban functions by agents, for example through area of dominating daily commuting), administrative definitions, etc., are partly orthogonal and more or less adapted to the problem studied [Guérois and Paulus, 2002]. Recently, several studies have shown the strong sensitivity of urban scaling laws⁹ to the delineation chosen for the estimation, leading sometimes to an inversion of expected qualitative properties (see for example [Arcaute et al., 2015]). Variations of estimated exponents as a function of parameters of the definition, as done by [Cottineau et al., 2015], can be interpreted as a more global property and a signature of the urban system.

This confirms the necessity to consider cities within their system, and the importance of the notion of *Urban System*¹⁰. An urban system can be considered as a set of cities in interaction, which dynamics will be more or less strongly coupled. [Berry, 1964] considers cities as "*systems within systems of cities*", insisting on the multi-scalar nature (in the sense of intricate scales with a certain level of autonomy)¹¹ and necessarily complex, conception which is adopted and extended by the evolutive urban theory previously detailed. The term of *System of Cities* will be used when we will be able to clearly identify cities as sub-systems, and we will use the term of urban system more generally (a city being itself an urban system).

Finally, underlying to the understanding of urban systems dynamics intervenes the notion of *Territory*. Polymorphic and corresponding to multiple visions, as we will develop deeply in 1.1, it can be simply defined in a preliminary way. The territory thus designates the spatial distribution of urban activities, of agents practicing or developing

⁹ Scaling laws consist in a statistical regularity which can be observed within a system of cities, linking for example a characteristic variable Y_i to the population P_i under the form of a power law $Y_i = Y_0 \cdot (P_i/P_0)^\alpha$.

¹⁰ Concerning the definition of a system, we can take it in all generality as a set of elements in interaction, presenting a certain structure determined by it, and which posses a certain level of autonomy in its environment. It can be a mainly ontological autonomy in the case of an open system, or a real autonomy in the case of a closed system.

¹¹ The definition of scale is ambiguous in geography, since according to [Hypergeo 2017], the scale designates simultaneously a spatial and/or temporal extent (scale of the map) and an abstract representation of "levels which make sense regarding a particular problem". As [Manson, 2008] indicates, scale is indeed placed within an epistemological continuum, from realistic conceptions to constructivist conceptions, and the ones making it correspond to intrinsic levels of self-organization of the system considered. We will position in a privileged way in this latest logic of complexity.

them, and of technical artifacts, including infrastructure, supporting them, and also the superstructure¹² which is associated to it¹³.

NETWORKS, INTERACTIONS AND CO-EVOLUTION

A fundamental characteristic of urban systems and territories is their simultaneous inscription in space and time, that is contained in spatio-temporal dynamics, at multiple scales. The notion of *process* in the sense of [Hypergeo 2017], i.e. a dynamical chain of facts with causal properties¹⁴, allows to capture relationships between components of these dynamics, and is thus an interesting approach for a partial understanding of such systems. Any partial understanding will be associated to the choice of *scales* and an *ontology* which corresponds to the specification of real objects studied¹⁵. We will now specify these abstract concepts, by introducing *networks*, their *interactions* with territories and their approach through *co-evolution*.

A particular ontology will hold our attention: within territories emerge *Physical Networks*, which can be understood according to [Dupuy, 1987] as the materialization of a set of potential connections between agents of the territory. The question of the implication of these networks and their dynamics in territorial dynamics, which we can synthesize as *interactions between networks and territories*, has been the subject of numerous technical and scientific debates, in particular in the case of transportation networks. We will come back on their nature and positioning in Chapter 1, but we can already take some of the underlying difficulties as a starting point for our questioning. One recurring aspect is the *myth of structuring effects*, suggested by [Offner, 1993] when criticizing an exaggerated use by planners and politics of a scientific concept which empirical basis are still discussed. The fundamental underlying question that we reformulate is the following: *to what extent is it possible to associate territorial dynamics to an evolution*

¹² We understand the superstructure in its marxist sense, i.e. the organizational structure and the ideas of a society, including political structures.

¹³ The link between the Territory and the City, or the System of Cities, will be also developed more deeply further when the concept will be constructed.

¹⁴ We will understand causality in the sense of circular causality in complex systems, which considers fostering cycles between phenomena, or more complex structures. Linear causality, i.e. a phenomenon driving another, is an idealized particular case of this. We will come back with more details on the notion of causality and on its different approaches by geographers in section 4.2.

¹⁵ More precisely, we use the definition of [Livet et al., 2010] which couples the ontological approach from the point of view of philosophy, i.e. “*the study of what can exist*”, and the one from computer science which consists in defining classes, objects and their relations which constitute the knowledge of a domain. This use of the notion of ontology naturally biases our research towards modeling paradigms, but we take the position (developed in more details later) to understand any scientific construction as a *model*, making the boundary between theory and models less relevant than for more classical visions. Any theory has to make choices on described objects, their relations, and the implicated processes, and contain thus an ontology in that sense.

of the transportation infrastructure ? We can ask the question reciprocally, and even generalize it: what are the processes capturing the interactions between these two objects ?

An approach allowing to consider the problem from an other angle is the notion of *co-evolution*, used in the evolutive urban theory to designate strongly coupled processes¹⁶ of evolution of cities as used by [Paulus, 2004], and applied to the relations between networks and cities by [Bretagnolle, 2009]¹⁷. This last work distinguishes a phase of “mutual adaptation” between networks and cities, corresponding to a dynamic in which causal effects can clearly be attributed to one on the development of the other (for example, new transportation lines answer to a growing demand induced by urban growth, or inversely urban growth is favored by a new connectivity to the network), from the phase of co-evolution, which is defined as a “strong interdependency” (p. 150) in which retroactions play a privileged role and “the dynamic of the system of cities is not anymore constrained by the development of transportation networks” (p. 170). These feedback loops and this mutual interdependency, seen in their dynamical perspective, correspond to circular causal relationships (in the sense given above) that are difficult to disentangle. We will take as preliminary definition of co-evolution between two components of a system *the existence of a strong coupling, corresponding generally to circular causal relationships*.

PROBLEMATIC

This framework allows to capture a certain degree of complexity, but however remains fuzzy or too general in its characterization, both theoretically and empirically. We will try here to challenge and to deepen this approach, to shed a light on its potential contributions for the understanding of interactions between networks and territories. The clarification on the one hand of what it means and on the other hand of its empirical existence will be a Gordian knot of our

¹⁶ We will use the term of *coupling* systems or processes to designate the constitution of a system including the coupled elements, through the emergence of new interactions or new elements. The definition of the nature and the strength of a coupling is an open question, and we will use the notion in an intuitive way, to designate a more or less high level of interdependency between coupled sub-systems.

¹⁷ [Paulus, 2004] directly transfers the biological concept of co-evolution (which consists in a strong interdependency between two species in their evolutionary trajectories, and which in fact corresponds to the existence of an *ecological niche* constituted by species as we will further develop in 8.2), and studies cities which “are in concurrence, imitate themselves, and cooperate”. This transfer remains fuzzy (on temporal scales implied, the status of objects which co-evolve) and finally not explored. Similar trajectories can not be enough to exhibit strong interdependencies as he states in conclusion, since these can be spurious. Furthermore, the transfer of concepts between disciplines is an operation on which one must remain cautious (we will illustrate this through the interdisciplinary study of morphogenesis, concept which is initially from biology, in Chapter 5).

approach. Our general problematic is thus decomposed into two complementary axis:

1. How to define and/or characterize co-evolution processes between transportation networks and territories ?
2. How to model these processes, at which scales and through which ontologies ?

The second aspect is a consequence of our scientific positioning, which postulates the use of modeling, and more particularly of simulation of models, as a fundamental tool for the knowledge of processes within complex systems.

GENERAL ORGANIZATION

We propose to answer to the above problematic through the following strategy. A first part will build the necessary foundations, by detailing definitions, studied concepts and objects, by sketching the scientific landscape gravitating around our question, and by refining the epistemological positioning. This part is composed by three chapters:

1. A first chapter develops the question of interactions between networks and territories, from a theoretical point of view but also by illustrating them by case studies and fieldwork elements. It allows to situate the notion of co-evolution both from a concrete and abstract point of view.
2. A second chapter aims in a similar way at clarifying the positioning regarding the modeling of co-evolution. The state of the art is completed by a mapping of concerned scientific disciplines and by a modelography, i.e. a classification and systematic decomposition of a corpus of models in order to understand the ontologies used and possible determinants of these.
3. A third chapter develops our epistemological positioning, which appears to have a considerable influence on modeling choices that will be taken in the following. We develop therein issues linked to modeling practices, to datamining and intensive computation, to reproducibility and open science, and more general epistemological considerations that are intrinsic to the systems studied.

From these complementary analyses emerge two thematic positioning that correspond to two modeling scales, that remain poorly explored for our particular question: the evolutive urban theory which induces a macroscopic modeling at the level of the system of cities, and urban morphogenesis which allows to consider the links between form and function at the mesoscopic scale. The second part will aim

thus at constructing elementary bricks from these approaches, which will be used in the following to construct models:

4. The fourth chapter deals with different aspects implied by the evolutive urban theory. The non-stationary character of processes in space is a crucial element, that we empirically demonstrate in a first section through the study of spatial correlations between urban form and road network topology for Europe and China. Then, the notion of circular causality is explored, and we develop a method allowing to isolate what we call *causality regimes*, i.e. typical configurations of interaction captured by lagged correlation patterns. It is tested on synthetic data and observed data in the case of South Africa, for which we demonstrate an effect of segregation policies on the interactions between networks and territories themselves. This first part of the chapter complements in an empirical way the characterization of co-evolution sketched in the first part. Finally, we construct a model of an urban system based on interactions between cities, which allows to indirectly demonstrate the existence of network effects.
5. The fifth chapter will deepen the notion of *morphogenesis*, by beginning with proposing a point of view consistent across disciplines using it, in order to exhibit a characterization based on the emergence of an architecture through causal circular relations between form and function. This precision will be crucial for the nature of models we will elaborate. A second section develops a simple model of urban growth taking into account the distribution of population alone, and capturing the contradictory forces of concentration and dispersion. We demonstrate its ability to reproduce existing urban forms using urban form data previously computed. It is then coupled in a sequential manner to a network generation model, what allows to exhibit a large spectrum of potentially generated correlations.

At this stage, we build in the third part from the foundations and with elementary bricks our fundamental construction, which consists in different models of co-evolution, that we differentiate according to the two approaches considered. Still within a logic of parallel and complementary approaches, we elaborate developments of the two previous chapters, in two chapters modeling co-evolution:

6. The sixth chapter develops a co-evolution model at the macroscopic scale. Firstly, we explore systematically the unique existing analog model. We then develop the model by extending the interaction model already introduced. Its systematic exploration reveals its ability to produce different regimes of co-evolution, some witnessing circular causalities. It is also calibrated on the

French system of cities on a long time period, on population and railway network data, which allows to infer indirect informations on implied processes.

7. The seventh chapter deals with urban morphogenesis models which capture co-evolution processes. The question of network generation heuristics is first tackled, by comparing the potentialities of diverse methods. In an approach of multi-modeling, these are then integrated in a family of morphogenesis models, which are calibrated on urban form and network topology indicators, at the first order (values of indicators) and at the second order (correlations matrices). We then sketch a more complex model, aiming at integrating governance processes in the growth of the transportation network. It is explored in a preliminary way.

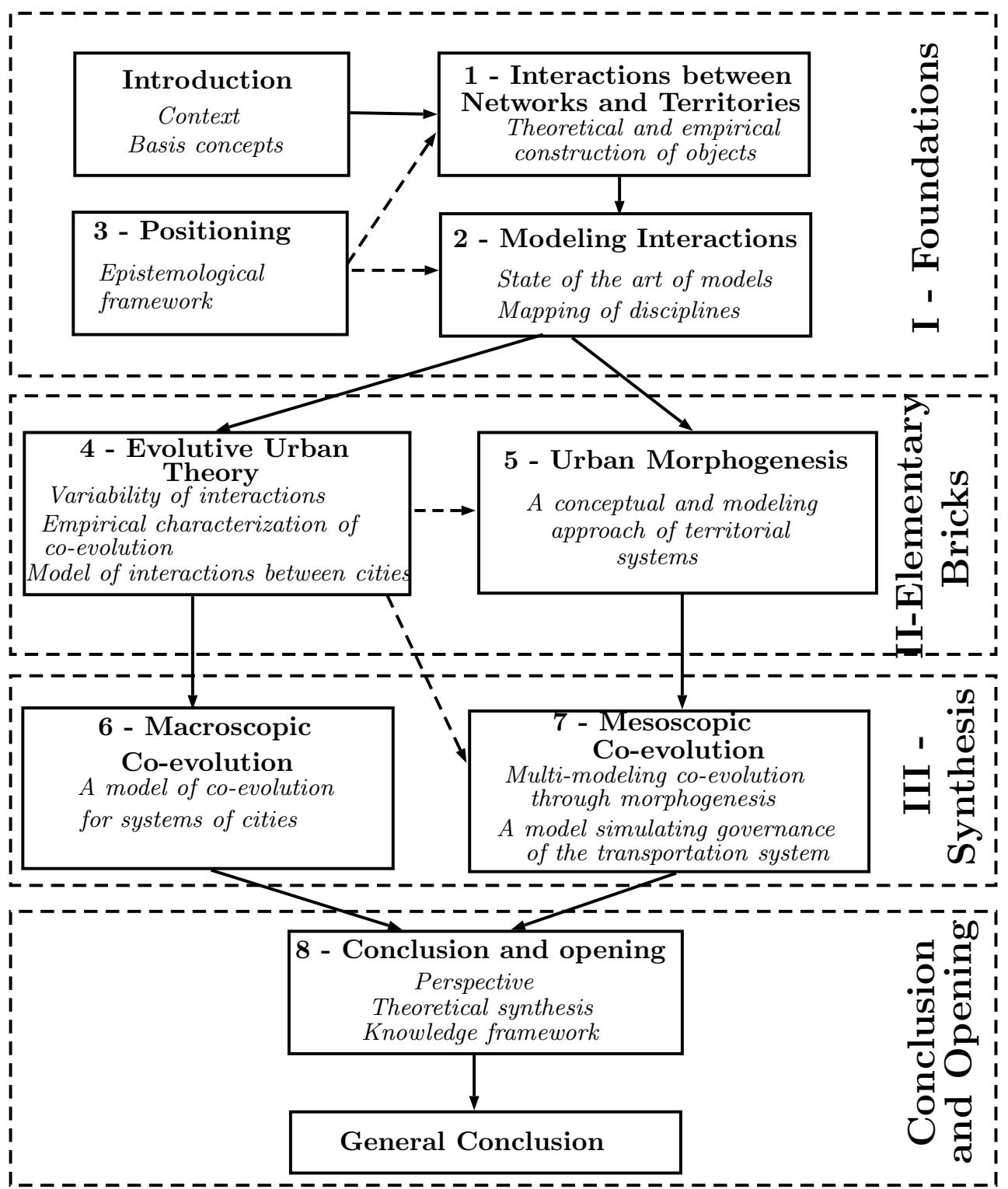
After having demonstrated the potentialities of our two approaches to capture some aspects of co-evolution and to inform corresponding processes, we finally proceed with an opening:

8. The eighth and last chapter consists in an theoretical and epistemological opening. We first draw a bilan of our contributions and put them into perspective. We then sketch a theoretical reconciliation of morphogenesis and the evolutive urban theory, in which co-evolution is central. This development could be the basis of a theory and multi-scalar models for co-evolution. We finally develop in a reflexive manner a knowledge framework for the study of complex systems, both product and precursor of all our work.

We summarize this organisation, and also direct or indirect dependencies between the different chapters, in the Frame 1 on the following page.

★ ★

★



FRAME 1: General organisation of the memoire. Full arrows give a direct dependency (logical chaining or extensions), dotted arrows an indirect dependency (reuse of data or methods).

Part I

FONDATIONS

This part builds the foundations of our work, by reconstructing the question in a theoretical way and through the illustration of case studies, and then by describing the scientific landscape of its existing approaches in modeling. We also develop our epistemological position with important practical implications.

INTRODUCTION OF PART I

A journey, discovering a city, new encounters, sharing ideas: as much processes which imply a cognitive generativity and a complex interaction between our representations, our actions, and the environment. The construction of a scientific knowledge does not escape these rules. We could then see in the studied object itself, let take the city and its agents, an allegory of the knowledge production process on the object. As Romain Duris which lands in l'Auberge Espagnole, and discovers these unknown streets that later we will have walked a hundred times, where we will have lived a thousand things: we land in a world of complementary concepts, approaches, points of view, on things that are not the same thing. This ontological discrepancy is indeed as much present in our representations of the urban space: Oven Street is one center of knowledge for the member of Géocités; it is the center of Paris, thus of France, thus of the World for the proud native of the 6th arrondissement ; it is the Saint-Germain market and globalized luxury shopping for the international tourist ; it is a piece of history for the student of Ecole des Ponts to which it reminds the era of Saint-Pères. Objects, concepts, understood and defined by multiple disciplines and agents that produce knowledge: do we finally designate the same thing ? How to benefit from this wealth of viewpoints, how to integrate the complexity allowed by this diversity ? To bring elements of answer requires a constructive, generative, and as much inclusive as possible approach. Choices are always more enlighten if we have a grasp on a maximum of alternatives. The trader living in his loft at the top of mid-levels and works in his close building between two rails, knows well Hong-Kong, but only one among its multiple faces, and it will be difficult to conceive the existence of a misery in Kwoloon, which inhabitants do not conceive the ephemeral but sometimes cyclic Hong-Kong of temporary workers from mainland, which them do not conceive the administrative and financial difficulties of migrants from Thailand or India, the whole picture being even less conceivable for a lost Parisian student. But it is indeed the loss, which in appropriate doses, will be source of a broader knowledge: ants establish their very precise optimizations from a walk that can be considered as random. Genetic algorithms, and even more biological evolution processes anchored in the physical, rely on a subtle compromise between order and disorder, between signal and noise, between stability and perturbations. To loose oneself to better find oneself makes the essence and the charm of the journey, let it be physical, conceptual, social. Finally, no possible comparison between orienteering in Le Caylar or Montagne de Bange to a rectilinear boredom in the Orléans forest.

This literary interlude raises fundamental issues induced by a demand of interdisciplinarity and the will to construct a complex inte-

grative knowledge. First, reflexivity and making a relation between a perspective taken with a certain number of other existing perspectives is necessary for its relevance. It is thus about constructing concepts in a solid way and to specify empirical references, in order to precise the problematic and its objectives *endogenously*. Secondly, the epistemological frame of the approach must be given. Above is indeed pictures a *perspectivist* approach, which is a particular epistemological positioning that we will detail here. Furthermore, the status of proofs is conditioned by the conception of methods and tools, which is particular in the case of simulation models.

This part respond to these constraints, by building the *foundations* necessary to the following of our work. In a relatively shifting terrain, these will have in some cases to be particularly deep for the global stability of the construction: this will for example be the case of the state of the art which will use techniques in quantitative epistemology. We recall that it is organized the following way:

1. The first chapter constructs concepts and objects from a theoretical point of view, and unveils a broad spectrum of possible approaches to interactions between transportation networks and territories.
2. The second chapter develops the different approaches in modeling interactions between networks and territories. It establishes the state of the art, structured by a typology previously obtained. It then describes the scientific landscape of concerned disciplines, and suggests the characteristics of models proper to each discipline and also possible determinants for it in a modegraphy.
3. The third chapter is relatively independent and precises our epistemological positioning. It allows in particular to situate the complexity which we aim at reaching, to specify what can be expected from a modeling approach, and to give a broader definition of the concept of co-evolution.

* * *

*

INTERACTIONS BETWEEN NETWORKS AND TERRITORIES

Networks and territories seem to be interlaced in complex causal relationships. In order to better understand notions of circular causalities within complex systems, and why these can lead to apparent paradoxes, the image given by DIDEROT in [Diderot, 1965] is enlightening: *"If you are embarrassed by the precedence of the chicken by the egg or of the egg by the chicken, it is because you are assuming that animals have always be the way they are now"*. By trying to naively tackle similar questions induced by our problematic previously introduced, causalities within geographical complex systems can be presented as a "chicken-and-egg" problem: if one effect seem to cause the other and reciprocally, is it possible and even relevant to try to isolate corresponding processes, if they are indeed part of a larger system which evolve at other scales ?

A reducing approach, which would consist in attributing systematic roles to one component or the other, is opposed to the idea suggested by DIDEROT which rejoins the one of *co-evolution*. One of the issues is thus to give an overview of interaction processes between networks and territories, in order to precise the definition of co-evolution, what will be after a similar work for modeling approaches, at the end of the first part.

This chapter must be read as the construction introducing our objects and positions of study, and will be completed by an exhaustive literature review on the precise subject of modeling interactions, which will be the object of chapter 2.

In a first section 1.1, we will precise the approach we take of the territory object, and to what extent it implies to consider transportation networks for the understanding of coupled dynamics. This allows to construct a framework which gives a definition of territorial systems, and which is particularly suited to our approach through co-evolution.

These abstract considerations will be illustrated by empirical case studies in the second section 1.2, chosen as very different to understand the underlying universality issues: the Greater Paris metropolitan area and Pearl River Delta in China.

Finally, in the last section 1.3, fieldwork observation elements obtained in China will precise and make more complex the construction of this theoretical and empirical framework.

★ ★

★

This chapter is fully unpublished.

1.1 TERRITORIES AND NETWORKS

We begin by constructing more precisely the concepts we will use. This construction helps to understand how the concepts of territory and network are rapidly in strong interaction, implying an ontological importance of interactions between corresponding objects. We will see that territories imply the existence of networks, but that reciprocally they are also influenced by them. A refined focus on properties of transportation networks allow to progressively a precise vision of *co-evolution*, that we will take up to there in its preliminary sense given before, i.e. the existence of circular causal relationships between transportation networks and territories.

1.1.1 Territories and networks, closely linked since their definition

Territories: an approach by systems of cities

The concept¹ of *territory*, that we introduced before through cities and systems of cities, will be central to our reasoning and must be depthen and enriched. In ecology, a territory corresponds to a spatial extent occupied by a group of agents or more generally an ecosystem [Tilman and Kareiva, 1997]. Territories of human societies imply supplementary dimensions, for example through the importance of their semiotic representations². These play a significant role in the emergence of social constructions, which genesis is profoundly linked to the one of urban systems. According to [Raffestin, 1987], the *Human Territoriality* is the “conjunction of a territorial process with an informational process”, what means that the physical occupation and exploitation of space by human societies can not be dissociated from the representations (cognitive and material) of these territorial processes, driving in return its further evolutions.

In other words, as soon as social constructions are implied in the constitution of human settlements, concrete and abstract social structures will play a role in the evolution of territories, and these two objects will be intimately binded. Examples of such links are for example the propagation of information and representations, political processes, or the conjunction or disjunction between lived and perceived territory. A territory is thus understood as a social structure organized in space, which includes its concrete abstract artifacts.

This approach of the territory rejoin the preliminary definition we took and reinforces it. The approach of RAFFESTIN insists on the role of cities as places of power (in the sense of a place gathering decision processes and of socio-economic control) and of wealth creation

¹ We will use the term *concept* for constructed knowledge, more than *notion*, which following [raffestin1978construits] is closer to an empirical information.

² In the sense signs marking the territory and their meaning, but also their representations, as maps for example.

through social and economical exchanges and interactions³. The city has however no existence without its hinterland, that can be interpreted as the *territory of a city*⁴. This correspondence sheds a light on all territories from the point of view of systems of cities, as developed by the evolutive urban theory [Pumain, 2010]. This theory interprets cities as complex self-organized systems, which act as mediators of social change: for example, innovation cycles initialize within cities and propagate between them (see C.5 for an empirical entry on the notion of innovation). It yield a vision of the territory as a space of flows, what will introduce the notion of network as we will see further. Cities are furthermore seen as competitive agents that co-evolve [Paulus, 2004], what already suggests the importance of co-evolution for territorial dynamics.

We have thus two complementary approaches of the territory that allow us to consider human territories structured by systems of cities⁵.

Moreover, a central aspect of human settlements that were studied in geography for a long time, and that relates directly to the concept of territory, is the one of *networks*. We will detail their definition and show how switching from one to the other is intrinsic to the approaches we take on these.

Definition of networks

A *network* must be understood in the broad sense of the establishment of relations between entities of a system, that can be seen as

-
- 3 An interaction will be taken in its broader meaning, as a reciprocal action of several entities one on the other. It can be physical, informational, transform the entities, etc. See [Morin, 1976] for a complete and complex construction of the concept, closely linked with the concept of organisation.
 - 4 Although an exact correspondance between territories and cities is probably only a simplification of reality, since territories can be entangled at different scales, along different dimensions. A reading through central places typical of CHRISTALLER [Banos et al., 2011] gives a conceptual idea of this correspondance. Functional definitions such as *Insee's urban areas*, that defines the area around a center above a critical size (10000 jobs) by the cities for which a minimal threshold of actives work in that center (40%) - see <https://www.insee.fr/fr/metadonnees/definition/c2070>, is a possible approach. The sensitivity of the properties of the urban system to these parameters is tested by [Cottineau et al., 2015]. The definition of the city is therefore intimately linked to the one of territories, and the definition of the urban system to the set of territories.
 - 5 These complementary views on the territory can also be enriched with an historical perspective. [Di Meo, 1998] gives an historical analysis of the different conceptions of space (that lead in particular to the lived space, the social space and the classical space of geography) and shows how their combination yields what RAFFESTIN describes as territories. [Giraut, 2008] recalls the different recent uses that have been done of the concept of territory, from cultural geography where it was used more as a scientific fashion, to geopolitics where it is a very specific term linked to governance structures, to uses where it is more an abstract concept, and highlights therein the interdisciplinary aspect of an object capturing a certain level of complexity of the systems studied.

abstract relations, links, interactions. [haggett1970network] postulates that the existence of a network is necessarily linked to the existence of flows⁶, and recalls the topological representation as a graph of any geographical system in which flows circulate between entities or places that are abstracted as nodes, linked by edges. Edges of the graph have then a *capacity*, which translate their ability to transport flows (that can be defined in a similar way as an *impedance*). The topological analysis already unveils a certain number of system properties, but [haggett1970network] precises the importance of the network spatialization, included in the properties of its nodes (localization) and of its links (localization, impedance), for the understanding of dynamics within the network (flows) or of the network itself (network growth). This specificity is recalled by [Barthelemy, 2011] which puts into perspective empirical domains that relate to spatial networks, some network growth models, and some models of processes within networks: for example, topological structures, or diffusion processes will be strongly constrained by the spatial dimension.

To study more thoroughly the concept of network by focusing on its strong interdependency with the concept of territory, we follow [Dupuy, 1987] which proposes elements for “a territorial theory of networks” inspired by the concrete case of an urban transportation network. This theory distinguishes *real networks*⁷ and *virtual networks*, that are themselves induced partly by the territorial configuration. Real networks are the materialization of virtual networks. More precisely, a territory is characterized by strong spatio-temporal discontinuities induced by the non-uniform distribution of agents and resources. These discontinuities naturally induce a network of potential interactions between the elements of the territorial system, namely agents and resources. [Dupuy, 1987] designates these potential interactions as *transactional projects*. These induce the notion of *potential of interaction*, i.e. a property of space from which the interactions derive⁸. For example nowadays people need to access the resource of employments, economic exchanges operate between different territories that can be more or less specialized in different types of production.

⁶ Flows are defined as a material exchange (people, goods, raw materials) or immaterial (information) between two entities.

⁷ Real networks include a category that can be described as concrete, material or physical networks - we will use these terms in an interchangeable manner in the following, to which transportation networks belong; other categories such as social networks are also real networks that we will not study.

⁸ Given any vectorial field of class C^1 on \mathbb{R}^3 , the HELMOLTZ theorem yields a vector potential and a scalar potential from which this field derives as a rotational and a gradient. It justifies in the particular case of such a viewpoint the correspondence between an interaction field between agents and a potential field.

From networks to real networks

In some cases, a potential network is materialized into a real network. The underlying question is then to determine if the potential field of territories is partly at the origin of this materialization, if it is totally independent, or if the dynamic of the two is strongly coupled, in other terms in co-evolution. The materialization will generally result of the combination of economic and geographical constraints with demand patterns, in a non-linear way. Such a process is not immediate, leading to strong non-stationarity and path-dependency effects⁹: the extension of an existing network will depend on previous configurations, and depending on involved time scales, the logic and even the nature of operators, i.e. agents participating to its production, may have evolved.

Examples of concrete trajectories can be quite varied: [Kasraian, Maat, and Wee, 2015] show for example, in the case of Randstad on long time, a first period during which the railway network has developed to follow urban development, whereas opposite effects has been more recently observed. At a urban scale on long time, the path-dependency is shown for Boston by [Block-Schachter, 2012] since the built environment and the distribution of population appear as highly dependant of past tramway lines even when they do not exist anymore: the way the transportation line changes the urban space acts on immediate dynamics but also on a longer time through reinforcement effects or because of the inertia of the built environment for example.

Therefore, the existence of a human territory necessarily imply the presence of abstract interaction networks, and concrete networks are crucial for the transport of people and ressources (including communication networks as information is a crucial ressource [Morin, 1976]), but the processes through which they are established are difficult to identify generally. Our ontological choice of positioning within DUPUY's theory, gives a privileged place to the relations between networks and territories, since it induces in the construction of the objects themselves a complex entanglement between these.

The status of the network in relation with the territory is moreover highly conditioned by the socio-economical and technological context. Following [Duranton, 1999], a factor influencing the form of pre-industrial cities was the performance of transportation networks. Technological progresses, leading to a decrease in transportation costs, have induced a regime change, what conducted to a preponderance of land markets in shaping cities (and thus a role of transportation network since they influence prices through accessibility), and more recently to the rising importance of telecommunication

⁹ Spatial non-stationarity consists in the dependency of the covariance structure of processes to space, whereas path-dependency corresponds to the fact that trajectories taken in the past strongly influence the current trajectories of the system.

networks what induced a “tyranny of proximity”, since a physical presence can not be replaced by virtual communications [Duranton, 1999].

This territorial approach to networks seems natural in geography, since networks are studied conjointly with geographical objects they connect, in opposition to theoretical works on complex networks which study them in a relatively disconnected way from their thematic background [Ducruet and Beauguitte, 2014].

Networks shaping territories ?

However networks are not only a material manifestation of territorial processes, but play their role in these processes since their evolution may influence the evolution of territories in return. Here comes an intrinsic difficulty: it is far from evident to attribute territorial mutations to an evolution of the network, and reciprocally the materialization of a network to precise territorial dynamics. Different exogenous factors are furthermore important, such as the price of energy or existing technologies in the case of the effect of the network on territories for example. In the case of *technical networks*, an other designation of concrete networks given in [Offner and Pumain, 1996], many examples of such feedbacks can be found: an increased accessibility may shape urban growth, or the interconnectivity of different transportation networks allows a significant extension of mobility ranges. At a smaller scale, changes in accessibility may induce relocalizations of different urban components. These retroactions of networks on territories does not necessarily act on concrete components: [Claval, 1987] shows that transportation and communication networks contribute to the collective representation of a territory by acting on the sentiment to belong to the territory, that can then play a crucial role in the emergence of a strongly coherent regional dynamic. We first develop with more details the possible influences of networks on territories.

The confusion on possible simple causal relationships has fed a scientific debate that is still active nowadays. The underlying question relies on more or less deterministic attributions of impacts to transportation infrastructures or to a new transportation mode on territorial transformations. Precursors of such a reasoning can be tracked back in the twenties: MCKENZIE, from the Chicago school, mentions in [Burgess, McKenzie, and Wirth, 1925] some “modifications of forms of transportation and communication as determining factors of growth and decline cycles [of territories]” (p. 69). Methodologies to identify what is then called *structuring effects* of transportation networks has been developed for planning in the seventies: [Bonafous and Plassard, 1974] situates the concept of structuring effect in the perspective of using the transportation offer as a planning tool (the alternatives are the development of an offer to answer to a congestion of the network, and the simultaneous development of asso-

ciated offer and planning). These authors identify from an empirical viewpoint direct effects of a novel offer on the behavior of agents, on transportation flows and possible inflexions on socio-economic trajectories of concerned territories. [Bonnafous, Plassard, and Soum, 1974] develop a method to identify such effects through the modification of the class of cities in a typology established a posteriori. More recently, [Bonnafous, 2014] recalls that the institution of *permanent observatories* for territories makes such analyses more robust, allowing a continuous monitoring of the territories that are the most concerned by the extent of a new infrastructure.

According to [Offner, 1993] which follows ideas already given by [Plassard, 1977] for example, a not reasoned and out-of-context use of these methods has then been developed by planners and politicians which generally used them to justify transportation projects in a technocratic manner: through the argument of a direct effect of a new infrastructure on local development (for example economic), politics are able to ask for subsidies and to legitimate their action in front of the people. [Offner, 1993] insists on the necessity of a critical positioning on these issues, recalling that there exists no scientific demonstration of an effect that would be systematic. A special issue of the journal *L'Espace Géographique* [Offner et al., 2014] on that debate recalled that on the one hand misconceptions and misuses were still greatly present in operational and planning communities, which can be explained for example by the need to justify public actions, and on the other hand that a scientific understanding of relations between networks and territories is still in construction. A. BONNAFOUS (interview on the 09/01/2018, see Appendix D.3) gives the current example of the project of the Seine-Nord-Europe canal¹⁰ as a transportation project for which traffic revisions were largely overestimated and that politics of concerned territories have largely instrumentalized.

An other concrete illustration in the actuality gives an idea of this instrumentalization: debates in July of 2017 concerning the opening of the *LGV Bretagne* and the *LGV Sud-Ouest* have shown the full ambiguity of positions, conceptions, imaginaries both of politics but also of the public: worries on the speculation on real estate in stations neighborhoods, questionings on daily mobility but also social mobility¹¹. The complexity and the reach of these subjects show well the

¹⁰ The canal project links the Oise at Compiègne to the Dunkerque-Escault canal in the north, see <https://www.canal-seine-nord-europe.fr/Projet>.

¹¹ See for example <http://www.liberation.fr/futurs/2017/07/02/immobilier-plus-de-parisiens-comment-les-bordelais-voient-l-arrivee-de-la-lgv-1580776>, or http://www.lemonde.fr/big-browser/article/2017/10/24/a-bordeaux-une-fronde-anti-parisiens-depuis-l-ouverture-de-la-ligne-a-grande-vitesse_5205282_4832693.html for an immediate reaction of diverse local actors, witnessing at least an impact on representations. For example, people in Bordeaux seem to fear the arrival of Parisians searching for cheaper housing and better living conditions, what could increase prices in the surroundings of the station.

difficulty of a systematic understanding of effects of transportation on territories.

An integrative approach: Territorial Systems

This overview as an introduction, from territories to networks, allows us thus to clarify our approach of territorial systems that will be underlying all the following. Taking into account diverse potential feedbacks of networks for the understanding of territories is suggested when coming back to the citation by Diderot that introduced the subject, in the sense that we must consider neither the network nor territories as independent systems that would influence themselves through one directional causal relations, but as strongly coupled components of a broader system, and thus being in a circular causal relationship. Depending on components and the scale that are considered, different manifestations of these will be observable, and there will exist some cases where there is apparently the influence of one on the other, other where influences are simultaneous, or moreover others where no relationship can be observed in a significant way.

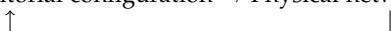
Since we have highlighted the role of networks in several aspects of territorial dynamics, we propose a definition of territorial systems that explicitly includes them. We consider a *Territorial System* as a *human territory that contains both interactions networks and real networks*. Real networks, and more particularly concrete networks¹², are an entire component of the system, influencing evolution processes, through multiple feedbacks with other components at many spatial and temporal scales.

The network is not necessarily a component in itself of the territory, but indeed of the *Territorial System* in our sense¹³. This view rejoins the positioning of [Dupuy, 1985] which introduces the territory as the “product of a dialectic” between territorial components and networks. We remark the semantic shortcut to designate components of the territorial system that are not the network and which interact with it, through the term of territory. These depend on ontologies and scales considered, as we will see in the following, and can span from microscopic agents to cities themselves. As we will also see in the following

¹² Which are as we previously saw materialized real networks.

¹³ This ontological choice is not innocent and reinforces the dialectic between networks and territories. Starting from the distant past where physical networks did not exist, the emergence of a human territory, that we assume equivalent to a network of interactions, induces the establishment of the complex diachronic dialectic between physical networks and human territories. We can thus read the genesis of a territorial system as a morinian loop [Morin, 1976], in which we enter by the initial territory and which then loops from the physical network to territorial components to produce the territorial system (thus the territory in most cases) in the following recursive way:

Initial territory → Territory = Territorial configuration → Physical network



(see 2.1), there exists some paradigms in which this simplification is not done, such as in the particular case of interactions between transportation and land-use where entities are specific. But it is done if we stay in a more general framework, as witnesses one of the reference works on the subject [Offner and Pumain, 1996]¹⁴. We will similarly postulate this semantic simplification, when designating by *interactions between networks and territories* or *co-evolution between networks and territories*, the interactions or the co-evolution between physical networks and components they connect, within a territorial system and thus a territory.

1.1.2 *Transportation networks, specific carriers of interactions*

We now precise the particular case of transportation networks and develop associated specific concepts that will play an important role in the precision of our problematic.

Characteristics and specificities of transportation networks

Central to the already evoked debates on structuring effects of networks, transportation networks play a significant role in the evolution of territories, but it is of course out of question to give them deterministic causal effects. We will generally use the term of transportation network to designate the functional entity allowing a movement of agents and resources within and between territories¹⁵. Even if other types of networks are also strongly implicated in the evolution of territorial systems (see for example the debates on the impact of communication networks on the localization of economic activities), transportation networks condition other types of networks (logistic, commercial exchanges, concrete social interactions to give a few examples) and are a privileged entry regarding patterns of territorial evolution, in particular in our contemporary societies for which transportation networks play a crucial role [Bavoux et al., 2005]. We will therefore focus in the following only on transportation networks.

The development of the French high speed rail network is an illustration of the role of transportation networks on policies of territorial development. Presented as a new era of railway transportation, it consisted in a top-down planning of totally novel lines, relatively independent through they two times higher speed, as [Zembri, 1997] puts it. High speed has been defended by political actors among other

¹⁴ When [Amar, 1985] proposes a conceptual model of network morphogenesis, he designates the territorial components as "The World", what does not solve the semantic issue. The choice to keep the term of territory, within the territory, suggests a recursivity, and thus a complexity in the generativity of the system [Morin, 1976]. The use of the concept of morphogenesis starting from chapter 5 suggests that this recursivity would not be spurious, but indeed intrinsic to the problem.

¹⁵ We designate thus simultaneously the infrastructure, but also its exploitation conditions, the rolling stock, the exploitation agents.

things as central for the development. The weak integration of these new networks with the existing network and with local territories is now understood as a structural weakness [Zembri, 1997] (i.e. that is a consequence of network structure such as it was planned in the *Scéma Directeur* of 1990), and negative impacts on some territories, such as the suppression of intermediate stops on classical lines used by the TGV, what contributes to an increase of the tunnel effect¹⁶ have been shown [Zembri, 2008]. A review done in [Bazin et al., 2011] confirms that no general conclusions on local effects of a connection to a high speed line could be drawn, although it keeps a strong place in imaginaries of politics¹⁷. The development of different high speed lines takes place in very different territorial contexts, and it is in any case difficult to interpret processes out of context: for example, the LGV Nord and LGV Est lines are situated within European scales that are broader than for the LGV Bretagne opened in July 2017¹⁸. The effects of the opening of a line can extend beyond the directly concerned territories: [L'Hostis, Leysens, and Liu, 2014] show through the use of indicators from *Time Geography*¹⁹ (measuring an available working time in the context of a return journey within the day) that the Tours-Bordeaux line has potential impacts in the North and East of France. These examples illustrate well the way transportation networks can have effects both directly and indirectly, positive or negative, at different scales, or no effect at all on territorial dynamics.

Processes depending on scales

The question of concerned temporal and spatial scales has until now been tackled only on a secondary plan compared to the concepts introduced. We propose now to integrate them to our reasoning in a structural way, i.e. guiding the developments of new concepts. Therefore, the concepts of *Mobility*, *Accessibility*²⁰, and *Structural Dynamics on long time*, correspond each to decreasing scales in time and space: intra-urban and daily, metropolitan and decennial, regional (in the

¹⁶ The tunnel effect designates the process of telescoping the territory traversed by the infrastructure, when it is not accessible from this territory.

¹⁷ But particular conclusions exist in some cases: for example a positive effect of the LGV Sud-Est on the touristic intensity in intermediate medium-sized cities such as Montbard or Beaune [Bonnafous, 1987]; or the positioning of Lille as an European metropolis in which the connexions to the LGV have played a role [Giblin-Delvallet, 2004].

¹⁸ The LGV Nord line links Paris to Lille then Calais (entirely opened in 1997), and is used for the link with London, Brussels, Amsterdam. The LGV Est line links Paris to Strasbourg (partially opened in 2007, fully in 2016) and allows to serve Luxembourg and Germany. The LGV Bretagne line, opened in 2017, is the branch of the LGV Ouest towards Rennes and its service is uniquely to Britanny [Zembri, 2010].

¹⁹ The *Time Geography*, introduced by the Swedish geographer HÄGERSTRAND, focuses mainly on trajectories of individuals in time and space, and of their implications in interactions with the environment [Chardonnel, 2007].

²⁰ The accessibility, as we will see, can be defined at different scales, but we will use this term in a privileged way for accessibility landscapes at the metropolitan scale.

broad and flexible sense of the range of a system of cities) and centennial. The correspondence we postulate here between time scales and spatial scales, far from being an evidence, will be shown during the development of each of these concepts. However, to take into account multiple scales is important, as shows [Rietveld, 1994] with a review of economic approaches to interactions, which insists on the difference between intra-urban and intra-regional: at a large scale, different methods (models or qualitative approaches) give very different results concerning the impact of the infrastructure stock, whereas at a small scale, the positive impact of the global stock on productivity can not a priori been discussed.

Transportation and mobility

The notion of mobility and all the associated approaches capture partly our questionings at a large scale. We will define mobility in a broad manner as a movement of territorial agents in space and time. It is related to use patterns of transportation networks. [Hall, 2005] introduces a theoretical framework that yields a typology of mobility practices. In particular, he shows a rapid decrease of the frequency of journeys with spatial range and duration, and thus that "micro-micro" patterns (for the daily temporal scale and the intra-urban spatial scale), that we designate as *daily mobility*, correspond to the most of journeys. It does not however mean an absence of link with other scales: on the one hand mobility patterns are very strongly conditioned by the distribution of activities as illustrate [Lee and Holme, 2015], but are on the other hand correlated to the social structure [Camarero and Oliva, 2008], that evolve both at time scales of a different magnitude (larger than a decade, thus at least one order in magnitude). Therefore, infrastructure and superstructure determine mobility practices, giving an important role to transportation networks in these.

Reciprocally, use patterns of transportation networks are the product of daily mobility dynamics, and they adapt to it, while inducing relocations of actives and employments: there exists a co-evolution between transportation and territorial components at the microscopic and mesoscopic scales, which are objects of study in themselves. For example, [Fusco, 2004] unveils an influence²¹ of mobility on the urban structure, whereas the offer in infrastructure and its properties have however simultaneous effects on mobility and on the urban structure. In the case of freeway networks, [Faivre, 2003] recalls the necessity to construct a framework going beyond the logic of structuring effects on long times, and exhibits also interactions at a large scale that are typical of mobility on which more systematic conclusions can be established, such as an evolution of mobility practices implying a different use of the transportation network. We have thus at a large scale

²¹ Which is interpreted as causal in the sense of bayesian networks.

a first strong interdependency between transportation networks and territories, a first scale of co-evolution.

It is important to keep in mind the strong contingency of concepts we use here. The co-construction of the concept of mobility with technical solutions that model it with an operational purpose, has been illustrated by [Commenges, 2013] for the French context, which reveals among other things an application of frameworks and methods imported from the United States which were not well adapted to the French context. This contingency means that even the choice of concepts depends of broader conditions than their direct utility, and suggests a global systemic insertion within the *Territorial System*.

Finally, we have to remark that our approach of mobility is necessarily in a way reductionist, and overshadows socio-economic problematics for example: following [Rémy, 2000] mobility is indeed a “virtual field”, i.e. it increases the potentialities offered to individuals, but in a way strongly dependent to the social class and to the socio-economic status. Indeed, mobility practices and political measures acting on transportation are closely linked and can lead to high socio-spatial inequalities in access to urban amenities [Gallez, 2015] (p. 236). Mobility practices will be indeed indirectly studied in an empirical preliminary study of traffic flows in 3.2, but we will not be able to treat of their socio-economic aspect: we must stay conscious that this aspect is not taken into account in our work.

Transportation and accessibility

The concept of *accessibility* is fundamental to our question, since it is positioned at the exact crossroad of networks and territories. Based on the ability to access a place through a transportation network (that can take into account the speed, the difficulty to travel), it is generally defined as a spatial interaction potential²² [Bavoux et al., 2005]. It was initially introduced in this form by [Hansen, 1959], with the aim to be applied to planning. Various formulations and formalizations of corresponding indicators have been proposed. It was shown that these enter the same theoretical frame. Indeed, [Weibull, 1976] develops an axiomatic approach to accessibility, i.e. proposing to characterize it starting from a minimal number of fundamental hypothesis (axioms). [Miller, 1999] takes the same frame and shows that it includes three classical ways to view accessibility. These are respectively the one based on *Time Geography* and constraints, the one on utility measures for the user, and the one on an average travel time. Corresponding measures are derived within an unified mathematical framework, what allows both a theoretical and operational link between approaches of the concept that are a priori different.

²² And often generalized as a *functional accessibility*, for example employments accessible to the actives in one place. Spatial interaction potentials that are expressed in gravity laws can also be understood in the same way.

We can first see to what extent accessibility patterns induce a evolution of the network. This concept is often used as a planning tool or as an explicative variable for the localization of agents, since it is for example a good indicator of the quantity of people concerned by a transportation project.

Recent debates on the planning of *Grand Paris Express* [Mangin, 2013], this new metropolitan transportation infrastructure planned for the next twenty years, has revealed the opposition between a vision of accessibility as necessary to open up disadvantaged territories, and a vision of accessibility as a driver of economic development for already dynamic areas, both being not necessarily compatible since they correspond to different transportation corridors. One was initially defended by the state in the perspective of competitive clusters, the other by the region in a perspective of territorial equity. These two logics answer naturally to different objective at various levels, and the chosen solution must be a compromise. We will come back on this precise example of the greater Paris in details in the following.

This example allows us to suggest an effect of patterns of potential on network evolution: even if this goes through complex social structures (we will also come back on this point in details further), there exists numerous situations where a growth of the transportation network (that can correspond to a topological evolution, i.e. the addition of a link, but also an evolution of link capacities) is directly or indirectly induced by a distribution of the accessibility [Zhang and Levinson, 2007]. This phenomenon can concern fundamental modifications of the networks or minor modifications: [Rouleau, 1985] studies the evolution on long times (from 1800 to 1980) of satellite villages around Paris that have progressively been integrated to its urban fabric and shows both a persistence of the roads and parcels frame, but also local evolutions answering to a logic of connectivity for example, while being part of a more complex evolution context (as in the case of Haussmann). We will designate this abstract process of an answer of the network to a connectivity demand as *potential breakdown*²³.

An other significant process is the impact of an evolution of accessibility through relocations on network use patterns, and more particularly congestion, inducing a modification of capacity (flow that can be carried by network links): this phenomenon is shown in the case of Beijing by [Yang, 2006], which unveils modification of the impedance (effective speed in the transportation network) up to 30%. This can be put in correspondence with processes linked to mobility, even if we are more within meso-meso scales here, i.e. an evolution of the network and relocations on time scales of the order of the decade (the

²³ In analogy with the phenomenon of *dielectric breakdown* which corresponds to the breakthrough of electrical current in a insulator when the difference of electrical potential is too high.

network being slower, of the order of two decades), and on spatial metropolitan scales²⁴.

Reciprocally, an evolution of the network implies an immediate re-configuration of the spatial distribution of accessibilities (in the sense of all existing approaches, since all take the network into account), and also potentially of territorial transformations on a longer time: we finally come back to the debate of structuring effects we already commented on. We have seen that accessibility co-evolves²⁵ with mobility practices, what suggests an effect at this scale. Concerning relocations and distribution of populations, there exists some cases where it is indeed possible to attribute some territorial dynamics to network growth, that we will develop in the following.

[Duranton and Turner, 2012] show thus at a medium time scale of 20 years for the United States, through the use of instrumental variables²⁶, that accessibility growth in a city causes the growth of employments. On a similar time scale, but at the spatial scale of the country for Sweden, [Johansson, 1993] show the local accessibility ("intra-regional") and global accessibility ("inter-regional") explains the growth of production and of the productivity of companies. [Kasraian et al., 2016] proceed to a systematic review of empirical studies of impacts at a medium term of transportation infrastructures, and show that an urban densification at the proximity of new infrastructures is highly probable, being residential in the case of a railway infrastructure and for employments and industrial activity in the case of a road infrastructure²⁷. Similarly, it is possible to show strong effects of the presence of infrastructures for particular types of land-use: [Nilsson and Smirnov, 2016] show it for example for fast foods in two cities in the United States, by showing statistically that the access to an important infrastructure induces a spatial aggregation of commerces.

The latest examples suggest the potential existence of effects of accessibility, and thus of the network, on territorial dynamics. In some cases, structuring effects are thus present. But these are always links

²⁴ Which correspond to spatial extents from 100 to 200km, but to various urban realities. A metropolis will be a city of importance in a system of cities at a small scale, and will be seen with its functional territory (for example Paris and a consequent part of Ile-de-France). The emergence of new metropolitan forms, such as *Mega-city-regions* which are composed by metropolis of comparable sizes, on a small spatial extent, and with strong interactions, makes this question of the scale more complicated. We will come back on these objects in 1.2.

²⁵ The concept applies a priori at different scales, what will be confirmed by the more precise definition we will take at the end of this first part.

²⁶ The method of instrumental variables aims at unveiling causal relations between an explicative and an explicated variable. The choice of a third variable, called the instrumental variable, must be done such that it influences only the explicative variable but not the explicated variable, in a sense an exogenous shock.

²⁷ The studies reviewed cover mainly the second half of the 20th century and Europe, the United States and East Asia. It is important to keep in mind that even if they are relatively general, conclusions must always be contextualized.

to the precise context and also to scales. This allows us to make the transition to concepts linked by dynamics of urban systems on long times.

Transportation and urban systems

The third conceptual entry on interactions between networks and territories, and which will be particularly linked to the idea of co-evolution, is the one of urban systems, at a small spatial scale and on long times. We will designate the concept by *structural dynamics of the urban system*.

The evolutive urban theory considers systems of cities as systems of systems at multiple scale, from the intra-urban microscopic level, to the macroscopic level of the whole system, through the mesoscopic level of the city [Pumain, 2008]. These systems are complex, dynamical, and adaptive: their components *co-evolve* and the system answers to internal or external perturbations by modifying its structure and its dynamics. We will largely develop the multiple implications of this approach all along our work, and retain here processes of interactions between cities. These interactions consist in material or informational exchanges, and the diffusion of innovation is therein a crucial component [Pumain, 2010]. These are necessarily carried by physical networks, and more particularly transportation networks. We expect thus from a theoretical point of view strong interdependencies between cities and transportation networks at these scales, i.e. a co-evolution.

From the empirical point of view, it has already been shown: [Bretagnolle, 2009] reveals an increasing correlation in time between urban hierarchy and the hierarchy of temporal accessibility for the French railway network (which is a priori clearer for this measure than for integrated measures of accessibility that are prone to auto-correlation as we will see in 4.2). This correlation is a witness of positive feedbacks between urban ranks and network centralities. Different regimes in space and times has been identified: for the evolution of the French railway network, a first phase of adaptation of the network to the existing urban configuration was followed by a phase of co-evolution, in the sense that causal relations became difficult to identify. The impact of the contraction of space-time by networks on patterns of growth potential had already been shown for Europe with an exploratory analysis in [Bretagnolle, Pumain, and Rozenblat, 1998].

Modeling results by [Bretagnolle and Pumain, 2010a], and more particularly the different parametrizations of the Simpop2 model²⁸,

²⁸ The generic structure of the Simpop2 model is the following [Pumain, 2008]: cities are characterized by their population and their wealth; they produce goods according to their economic profile; interactions between cities produce exchanges, determined by the offer and demand functions; populations evolve according to wealth after exchanges.

show that the evolution of the railway network in the United States has followed a rather different dynamic, without hierarchical diffusion, shaping locally urban growth in some cases. This particular context of conquest of a space empty of infrastructures implies a specific regime for the territorial system. Other contexts reveal different impacts of the network at short and long term: [Berger and Enflo, 2017] study the impact of the construction of the Swedish railway network on the growth of urban populations, from 1800 to 2010, and find an immediate causal effect of the accessibility increase on population growth, followed on long times of a strong inertia for population hierarchy. In each case, we indeed observe the existence of *structural dynamics* on long times, which correspond to the slow dynamics of the urban system structure, and witness in that sense of *structuring effects on long times* as [Pumain, 2014] puts it.

We must be careful to differentiate the latest from the structuring effects previously mentioned which are subject to debates. At the level of the urban system, it is relevant to globally follow trajectories that were possible, and locally the effect has necessarily a probabilistic aspect. Moreover, we insist on the role of path-dependency for trajectories of urban systems: for example the existence in France of a previous system of cities and network (postal roads) has strongly influenced the development of the railway network, or as [Berger and Enflo, 2017] showed for Sweden. The same way, [Chaudhuri, G. and Clarke, Keith C., 2015] highlight the importance of historical events in coupled dynamics of the road network and territories, historical shocks that can be seen as exogenous and inducing bifurcations of the system that accentuate the effect of path-dependency. Therefore, for these structural dynamics on long times, forecasting can difficultly be considered.

This third approach allowed us to unveil a complementary point of view on co-evolution, at an other scale.

Links between scales suggested by Scaling Laws

Our framework with successive scales, that yield a reasonable correspondence between spatial and temporal scales, and also to associate the corresponding concepts, does naturally not capture the full range of possible processes: these that would fundamentally be multi-scalar, for example by implying the emergence of their own intermediate level, are not evoked. These are important and we will come back to them below. First we propose to establish a conceptual link between scales by the intermediary of *scaling laws* (that we understand in the general sense given in introduction). This link aims in particular at going beyond a reductionist reading through the compartmentalization of scales.

Transportation networks are by essence hierarchical, this property depending on scales they are embedded in, and leading to the emer-

Table 1: **Synthesis of the approach by scales of interactions between transportation networks and territories.**
References give a possible theoretical frame for each scale.

Scale	Spatial scale	Temporal scale	Concept	Reference
Micro	Intra-urban (10km)	daily (1d)	Mobility practices	[Hall, 2005]
Meso	Metropolitan (100km)	Decade (10y)	Metropolitan reconfiguration	[Wegener and Fürst, 2004]
Macro	Regional (500km)	Century (100y)	Structural dynamic on long times	[Pumain, 1997]

gence of scaling laws for their properties. For example, [Louf, Roth, and Barthelemy, 2014] show empirical scaling properties for a consequent number of metropolitan areas across the world. Indeed, scaling laws reveal the presence of hierarchy within a system, as for size hierarchy for systems of cities expressed by Zipf's law [Nitsch, 2005] or other urban scaling laws [Arcaute et al., 2015; Bettencourt and Lobo, 2016], what suggests a particular structure for these systems. We can expect to find it again in interaction processes themselves. Transportation network topology follows such laws for the distribution of its local measures such as centrality [Samaniego and Moses, 2008], these being directly linked to accessibility patterns at different scales. Furthermore, network topology is among the factors inducing the hierarchy of use, since it influences congestion negative externalities, in relation with the spatial distribution of land-use [Tsekeris and Geroliminis, 2013]. Thus, considering scaling laws for transportation networks, and more generally for territorial systems, is first a signature of the complexity of these systems, and secondly yields an implicit link between scales.

Scales: a synthesis

To recall our framework by scales, we propose the Table 1. Designations and orders of magnitude of temporal and spatial scales are of course indicative, such as key concept that are indeed the ones that allowed us to enter these scales. We also give references that illustrate corresponding conceptual frameworks. This table will however be useful to keep in mind the typical scales to which we refer.

Processus: a synthesis

At this stage, we can already propose a preliminary of the interaction processes we introduced. A more exhaustive typology will be possible at the end of this chapter.

Thus, territorial components can act on networks by:

- Impact of mobility patterns on impedances and capacities

- Potential breakdown, emergence of centralities
- Hierarchical selection of accessibility
- Systemic structural effects and bifurcations

Reciprocally, processes where network properties act on territories include:

- Relocations induced by mobility constraints
- Land-use changes due to a transportation infrastructure
- Accessibility patterns induced by networks, that can induce relocations
- Interactions between territories carried by network, including the tunnel effect when these are telescoped

These different processes do not all have the same level of abstraction neither the same scales. We have furthermore hidden some processes already evoked, within which the coupling is stronger and for which the circularity is already present in the ontology, such as processes linked to planning. We will now detail these, what will allow us then to refine the list above and to present it as a typology after having enriched it with empirical studies.

1.1.3 *From interactions to co-evolution*

At this stage, we have identified processes of interaction between transportation networks and territories that play a significant role in the complexity of territorial systems. In the frame of our preliminary definition of a territorial system, this question can be reformulated as the study of networked territorial systems with an emphasis on the role of transportation networks. We have seen that the extent of spatial and temporal scales spans from daily mobility (micro-micro) to processes on long time in systems of cities (macro-macro), with the possibility of intermediate combinations. The precision of scales that are particularly relevant will be the subject of most of preliminaries (Part 1) and of foundations (Part 2), until chapter 5 that concludes foundations. We now extend this list and give concrete examples in terms of the complexity of interactions.

Importance of the geographical context

The contextualization of our question in a particular frame reveals the importance of taking into account the geographical context. The example of mountain territories, where constraints on resources and travel are stronger, shows the richness of possible situations when a generic frame is put in context of a particular case.

For example, on comparable French mountain territories, [Berne, 2008] shows that reactions to a same context of evolution of the transportation network can lead to very different territorial dynamics, some territories highly benefiting of the increased accessibility, others in the contrary becoming more closed. In the same frame, these possible opposed processes are scrutinized with more details by [Bernier, 2007], for which he proposes a typology based on the opening potential both of territorial dynamics and network dynamics: for example, a territory can exhibit rich opportunities to be attractive, such as touristic opportunities, but keep a low accessibility. Reciprocally, he gives the illustration of custom constraints that can impede the opening potential of a performant infrastructure.

Similarly to approaches considering systems of cities, [Torricelli, 2002] shows how it is possible in that context to establish a link between the nature of transportation flows and the local development of the urban system: cities in the mountains have first emerged as waypoints on paths to mountain passes, then have lost their importance when roads came into existence. The construction of railways gave them a new dynamic, through tourism and industry, and finally freeways has more recently inducted a loss of urban structure through peri-urbanization for example. Thus, structural dynamics on long time are particular, as a consequence of the geographical context.

Planification processes

As we already suggested, potential impacts of territorial dynamics on networks imply processes at different levels. This way, infrastructure projects are generally planned²⁹, in order to fulfil some objectives fixed generally by institutional actors. These objects bring progressively the concept of governance, but let first give some illustrations of planned projects.

The example of the failure in the planning of the Ciudad Real airport in Spain shows that the answer to a planned infrastructure is far from systematic. The explanations to it are probably a complex combination of diverse factors, difficult to disentangle. [Otamendi, Pastor, and Garcí, 2008] predicted before the opening of the airport a complex management due to the dimension of expected flows and proposed a suited model, but the order of magnitude of effective flows were closer to thousands than millions that were planned and the airport rapidly closed. It is complicated to know the reason of the failure, if it is an optimism of the regional level of polycentricity (the airport is halfway between Madrid and Seville), the lack of construc-

²⁹ We will use the term *planning* in general, territorial or urban, of an infrastructure project, when a project and its plan is willingly elaborated by a planning stakeholder, with an aim at transforming space according to some motivations depending on the stakeholder and on its interactions with other stakeholders.

tion of a train station on the high speed line, or just purely economical factors.

[Heddebaut and Ernecq, 2016]³⁰ show for the impact of infrastructures on the long term, in the case of the Channel tunnel³¹, through an analysis of investments and political actions in time, that the effect effectively observed for the Nord-Pas-de-Calais region such as a gain in centrality and in European visibility, are in strong distortion with the initial justifications of the project, and that the renewing of stakeholders implies that the project is not accompanied on the long time what makes its impact more uncertain. We rejoin the idea advocated by [Offner et al., 2014] according to which some “structure effects” effectively exist but that these can be observed on the long time in terms of the dynamic of the system for which a short time local vision does not make much sense. At the intra-urban scale, [Fritsch, 2007] takes the example of the tramway in Nantes to show, through a localized study of urban transformations in the neighborhood of a new line, that urban densification dynamics are far from what was expected from deciders and planners, i.e. a strong correspondence between the proximity to the line and a densification.

These examples confirm that the understanding of effects of territories on infrastructures imply to take into the concept of *governance*.

Governance

The development of a transportation network necessitate actors disposing of both concrete and economic capabilities to proceed to the construction, and furthermore having the legitimacy to lead this development. This must thus necessarily be actors of the social superstructure, that can be different levels of public governance, sometimes associated with private actors. The concept of *governance*, that we understand as the management of an organisation with common ressources with targets linked to the interest of the concerned community (these can be defined in different ways, for example in a *top-down* manner by governance actors or in a *bottom-up* manner by consulting the agents concerned with the decision), is then crucial to understand the evolution of transportation projects and thus of transportation networks. We will use the term of *territorial governance* when decision imply directly or indirectly components of territorial systems.

For example, [Offner, 2000] illustrates the difficulties posed by the deregulation of some networked public services concerning the territorial competences of authorities, and proposes the emergence of a

³⁰ The possible pun with the ambiguous title on the existence of the “Tunnel effect” recalls the effect through which an infrastructure traversing a territory has no interaction with it.

³¹ Put into service in 1994 between Calais in France and Folkestone in the United Kingdom, this railway underwater tunnel with a length of 50km establishes a physical link between the continent and the UK.

new local regulation for a new compromise between networks and territories.

Some aspects of territorial governance can have a significant impact on the development of transportation infrastructures. We can illustrate some for particular cases of the application of *urban models*³². [Deng and Liu, 2007] show in the case of Chinese cities that new directives in terms of housing can significantly deteriorate the performance of infrastructures, and that specific actions must be taken to anticipate these negative externalities. These concern in particular the dispositions in terms of *Transit Oriented Development* (TOD). TOD is a particular approach to urban planning that aims at articulating the development of public transportation and urban development. It can be understood as a voluntary co-evolution by developers (administrative authorities and/or planning authorities), in which the articulation is thought and planned. We will come back on TOD during empirical studies in the following.

These concepts are not new, since they were for example implicit in the planning of new towns in *Ile-de-France*, under a different form since these were strongly zoned (i.e. planned into relatively isolated mono-functional areas) and dependant on the automotive for some districts [Ostrowetsky, 2004]. [L'Hostis, Soulard, and Wulfhorst, 2012] give an example of an European project that has explored some implementations of TOD paradigms: planning details such as a quality of the network for active mobility modes at a short range are crucial for the concretization of principles. For example, [L'Hostis, Soulard, and Vulturescu, 2016] use a multi-criteria analysis³³ to understand determining factors in the selection of stations for the planned city, including urban density and access time to stations. [Liu and L'Hostis, 2014] show that even if some planning policies do not directly take a positioning as such, particularly in France, they exhibit very similar characteristics as shows the case of Lille.

The articulation between transportation and urban planning must often be operated in a strongly coupled manner to attain the expected objectives, even more when the project is specialized: [Larroque, Margairaz, and Zembri, 2002] recall the case of the SK metro in Noisy-le-Grand which unveils a case of a complete dependency of the functionality of the transport to local development. In order to serve a project of a office complex, a specific line with a lightweight equipment is constructed to make a link with the RER station of Mont-d'Est. The real estate project will fail whereas the line is inaugurated in 1993, it

³² In the sense of planning, i.e. conceptual generic schemas acting as a guide to the planification.

³³ In the frame of decision making for the planning of transportation infrastructures, multi-criteria analysis is an alternative to cost-benefit analysis (that compare projects by aggregating a generalized cost) which allows to take into account multiple dimensions, that are often contradictory (for example construction cost and robustness for a network), and obtain optimal solutions in the Pareto sense.

will be first regularly maintained and then abandoned without having never been opened to the public.

Therefore, governance processes, that manifest themselves in different ways, such as planning, or more particularly as TOD, play an important role in interactions between transportation networks and territories. These add up to our panorama, being of a particular type since they imply their own level of emergence and a strong autonomy.

Co-evolution of networks and territories

This progressive construction allowed us to highlight the complexity of interactions between networks and territories, what suggests the relevance of the particular ontology of *co-evolution* as we defined in introduction. [Levinson, 2011] insists on the difficulty of understanding the co-evolution between transport and land-use in terms of circular causalities, partly because of the different time scales implied, but also because of the heterogeneity of components. [Offner, 1993] uses the term of congruence, that can be understood as systemic dynamics implying correlations that can be spurious or not, that would be a preliminary vision of co-evolution.

The necessity to go past reducing approaches of structuring effects, together with the capture of the complexity of interactions between networks and territories through their co-evolution, is confirmed by the case of economic effects of high speed lines: [Blanquart and Koning, 2017] proceeds to a both empirical and theoretical review, including grey literature, of studies of this specific case, and concludes, beyond the direct effects linked to the construction on which there is a consensus, that proper effects on a long time seem to be random. This witnesses in fact complex local situations, a large number of conjunctural aspects playing a role in the production of effects, that can then not be attributed to transport only. This review confirms moreover the gap between political and technical narratives preceding transportation projects and the effective posterior analysis, revealed by [Bazin et al., 2010]. [Bazin, Beckerich, and Delaplace, 2007] conduct also a targeted study of the real estate market in Reims in anticipation to the arrival of the *TGV Est*. Through a diachronic analysis for each year between 1999 and 2005, for each district, of the real estate prices and the origin of buyers (locals or from the region of Paris), they conclude that only very localized operations can be directly linked to the *TGV*, the whole market following a global independent dynamic.

* * *

*

Thus, our constructive overview, broad and conceived as circular, of interactions between transportation networks and territories, confirms the relevance of the concept of *co-evolution* on the one hand, but suggests on the other hand a more thorough investigation and clarification for it.

We have therefore seen in this section that (i) the concept of territory naturally yields the concept of network; (ii) reciprocally, networks can transform territories, following different processes more or less established depending on scales; (iii) there exists a large number of cases and of particular processes for which the relation between networks and territories is imbricated, and for which we can use the term of *co-evolution*.

We will aim in the following section at studying more thoroughly in an empirical way various aspects evoked here, to put into perspective and refine the questions we aim at answering here.

★ ★

★

1.2 TRANSPORTATION PROJECTS FROM PARIS TO ZHUHAI

We develop in this section some geographical case studies at the metropolitan scale as we previously defined. We choose them to be very different to maximize the diversity of processes that can potentially be identified (since as we showed the geographical context is crucial). These are the Greater Paris metropolitan area, and the mega-city-region of Pearl River Delta in the South of China.

The objective of this section is to specify, precise, illustrate, enrich, the overview of co-evolution processes that we established in a general manner. Geography can not draw general conclusions, in the cases where these are relevant, without very precise and particular case studies. When applying a generic model to a set of territories, we will investigate the deviation to the model, that must then be explained through geographical reasoning, meaning a strong implication with the place in particular. Our approach is similar: if we can link several developed concepts to a case study, these will be necessarily enriched³⁴.

1.2.1 *Greater Paris: history and issues*

The Parisian region is a good illustration of the complexity of interactions between transportation networks and territories. The relevant time period for our question ranges from the end of the 19th century to nowadays. We propose, after a brief presentation of the context, to recall the history of the development of public transportation in *Ile-de-France*, which allows to reveals its articulations with urbanism, in particular the issues linked to transportation network planning. We will then study the present and future of *Grand Paris*, first concerning the emergence of a new governance structure at the level of the metropolitan area, and then the implied recent transportation projects, putting the example at the core of our problematic. We will finally make a more detailed incursion within an empirical analysis of relations between territorial variables and accessibility differentials for transportation projects, sketching some of the methodological developments we will develop in the following.

Context

The spatial context is the intermediate scale of a globally monocentric metropolitan area. Let precise this spatial structure. If the metropolis taken up to the *moyenne couronne* (i.e. the extent corresponding roughly to the central urban core with continuous built environment)

³⁴ And possibly connected through the transfer of the structure of the particular system to the structure of knowledge.

exhibits a certain level of polycentrism³⁵, in particular through the effect of new towns, which became important local employment centers [Berroir et al., 2005].

The role of different transportation infrastructures in the different economical dynamics in *Ile-de-France* is not trivial, as shows [Padeiro, 2013] which aims at statistically explicating employment growth between 1993 and 2008 in medium-sized and small communes in the Parisian region as a function of the proximity to an infrastructure: effects depends both on transportation mode (highway or airport) but also on the economic sector considered. Reciprocally, successive developments of transportation projects, generally operate in a discontinuous way in time. As we will detail in the following, they are linked to planning dynamics and governance processes that must be understood conjointly to territorial dynamics. The Parisian metropolis thus witnesses of complex relations between territories and networks.

Greater Paris transportation network

The history of the development of the transportation network of Parisian metropolitan area is recalled in [Larroque, Margairaz, and Zembri, 2002]. The French particularity with centralization lead to a particular structure for the railway network at the national scale, but also at the regional scale. The domination of Paris has indeed strongly shaped the structuration of the transportation network during the different historical periods during which it underwent significant evolutions. [Larroque, Margairaz, and Zembri, 2002] decompose the second half of the twentieth century in three periods.

Before 1975, the distribution of accessibility of actives to employments is clearly centralized and the center of Paris exhibits a strong congestion. The establishment of the RER network between 1975 and 1988 allows, thanks to the conjoint construction of *Villes Nouvelles*, an articulation between transportation and urbanism and a certain degree of polycentrism. [Larroque, Margairaz, and Zembri, 2002] however recall that realizations during this period show an increasing gap with the real demand for transportation. The period following 1988 until 2000, year of a political alternance, will mostly consist in the renewing of actors and the elaboration of new strategies, as witnesses the *Schéma Directeur* in 1994. Network developments during this period do not induce any major change in the spatial distribution of

³⁵ Polycentrism, by opposition to monocentrism, means that it is possible to identify different centers in an urban system. The way to define a center will depend on the scale and on the phenomena considered: it can for example be the existence of different employment poles of comparable size at the infra-metropolitan scale. The same way that the concept is polymorphic, the way to measure it quantitatively are multiple and complementary [Servais et al., 2004].

accessibility, despite the realization of the central interconnection for RER D, of the line 14 and of RER E.

The successive planning schemes lead to the SDRIF of 2013 [SDRIF, 2013]. They present early signs of the future network of the *Grand Paris Express*, of which a strong impact is expected in terms of territorial cohesion by favouring links between suburbs which are the most problematic in the current network. Furthermore, the plan is voluntary integrated, by densification around stations and an articulation between urban operations and new infrastructures. This aspect of network integration within territories and of territories by networks can be indeed observed in the public communication of the transportation organisation authority (former STIF, which became *Ile-de-France Mobilités*)³⁶. We therefore find again the importance of governance processes in the articulation between transportation networks and territories for the example of Ile-de-France in time.

Other processes already mentioned also manifest themselves, under different forms. For example, the role of path-dependency in trajectories of the territorial system is illustrated by [Larroque, Margairaz, and Zembri, 2002] which shows the inertia due to successive technical choices when they are successful: the initial choice of a metropolitan network within Paris' walls, the realization of the RER network, the tarification politic by areas for the *carte orange* at the end of the nineties, are different decisions in diverse domains but having each their significant part in the possible posterior developments. These authors also show how decisions concerning the public transportation network can induce, through a bad covering or performance of the public transportation network, the emergence of interaction processes where the couple use of the car and periurbanization³⁷ is favored, in a way similar to the *automobile city* described by [Newman and Kenworthy, 1996].

[Padeiro, 2009] recalls that the extension of metro lines to the close suburbs has always been restricted, reinforcing the role of Paris' city in the relation between the metropolitan territory and networks. Furthermore, he shows that urban polarizations (adaptation of the built environment and of the socio-economical composition) around stations beyond the limits of Paris are for their socio-economical part anterior dynamics that the arrival of the metro then accompanies: in that case, there is no structuring effect in the proper sense.

³⁶ See for example the actuality of the 4th October 2017 at <https://www.iledefrance-mobilites.fr/actualites/un-reseau-de-transports-qui-grandit/> which underlines that "With 29km of additional network length and the opening of 28 desserve points, territories are getting closer", witnessing the importance of accessibility for territories, notion which is furthermore fuzzy. Similar orientations in discourse can be found for the different projects of extension or construction of new lines.

³⁷ The periurban belongs to the new forms of urbanization, and consists in intermediate territories between the rural and the urban, benefiting from a good accessibility but exhibiting low densities and mostly individual dwellings.

Towards a metropolitan governance

To the metropolitan context previously described corresponds a complexity of the governance structure. In particular, current developments, both of the transportation network and of urban projects, coincide with the emergence of a new level of governance, an intermediary between *communes* and *départements* on one side, and the Region and the State on the other side. We can ask to what extent this emergence is linked to dynamics of interactions between territories, and how it will influence the interactions between territories and networks. [Gilli and Offner, 2009] propose in 2009 a diagnostic of the institutional situation of the Parisian region, and directions for a coupled approach between governance and planning. They highlight the early signs of the “establishment of a collective metropolitan actor”, which corresponds to the *métropole du Grand Paris* which will be inaugurated 7 years later, since the metropolitan council is put into place in the end of 2016.

The establishment of this new level of governance has been studied more recently still by [Gilli, 2014], which situates it within a broader socio-economical context and of other levels of governance (State, Region, *intercommunalités*). It allows him to sketch a territorial diagnosis which gives elements explaining its emergence: gaining retard in the domain of planification compared to its past dynamics, but also in the social domain given very high local socio-economical inequalities, the metropolis needs to reinvent itself, and this new dynamics naturally crystallize in the *Grand Paris*, what means that, as he concludes, “the future of Paris are its suburbs”. This initiative is made concrete by the convergence on the one hand of initiatives and the voluntarism of local politics, and on the other hand of a redefinition of the role of the State, wanted with a centralization until 2012 and freeing the stage to metropolitan governance with the political alternance in 2012. The projects launched and financing remain roughly the same: the project of the *Grand Paris Express* is a compromise between the solution wanted by the State and the one defended by the Region. Following [Desjardins, 2016], although the metropolitan governance structure has still today relatively no power, and although the negligence of the social aspect of metropolitan development is always highly present, these mutations however witness a deep structural change in the organisation of the region. We now detail the transportation project of the *Grand Paris Express*.

Project of the Grand Paris Express: towards a rebalancing of accessibilities ?

The metropolitan region of Paris is currently undergoing significant transformations, with the constitution of a metropolitan governance and new transportation infrastructures. The construction of a ring metro network allowing suburbs to suburbs links answers to an an-

Légende

- Arc Express - proche (2007)
- Arc Express - éloigné (2007)
- Grand Paris Express (2011)
- Existant

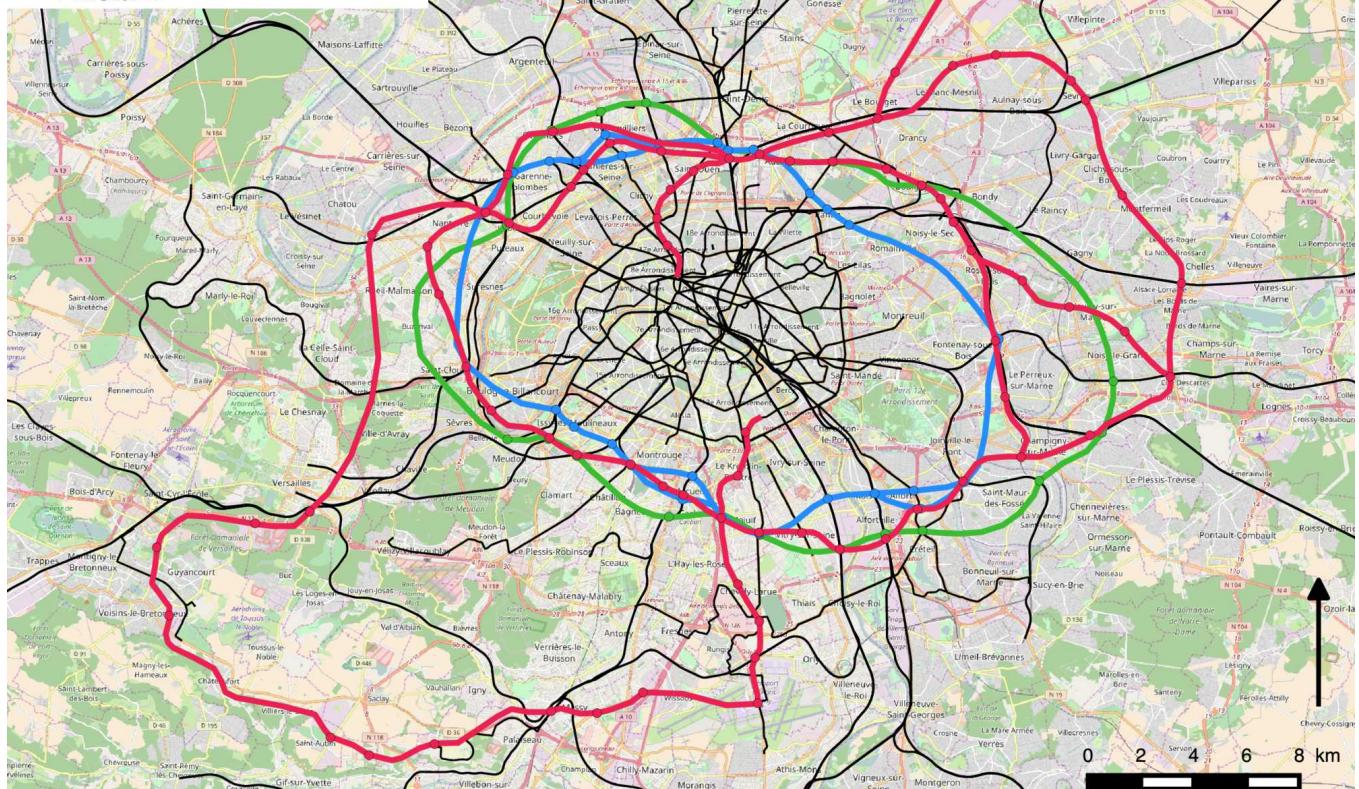


Figure 1: Successive transportation network projects for the Grand Paris metropolitan area. We show the two alternatives for the *Arc Express* project elaborated by the Region, and the *Grand Paris Express* (GPE) advocated by the State. The *Réseau du Grand Paris*, a precursor for GPE, is not shown here for visibility reasons because of its proximity with it. The source of the map background, given to situate the lines, is OpenStreetMap.

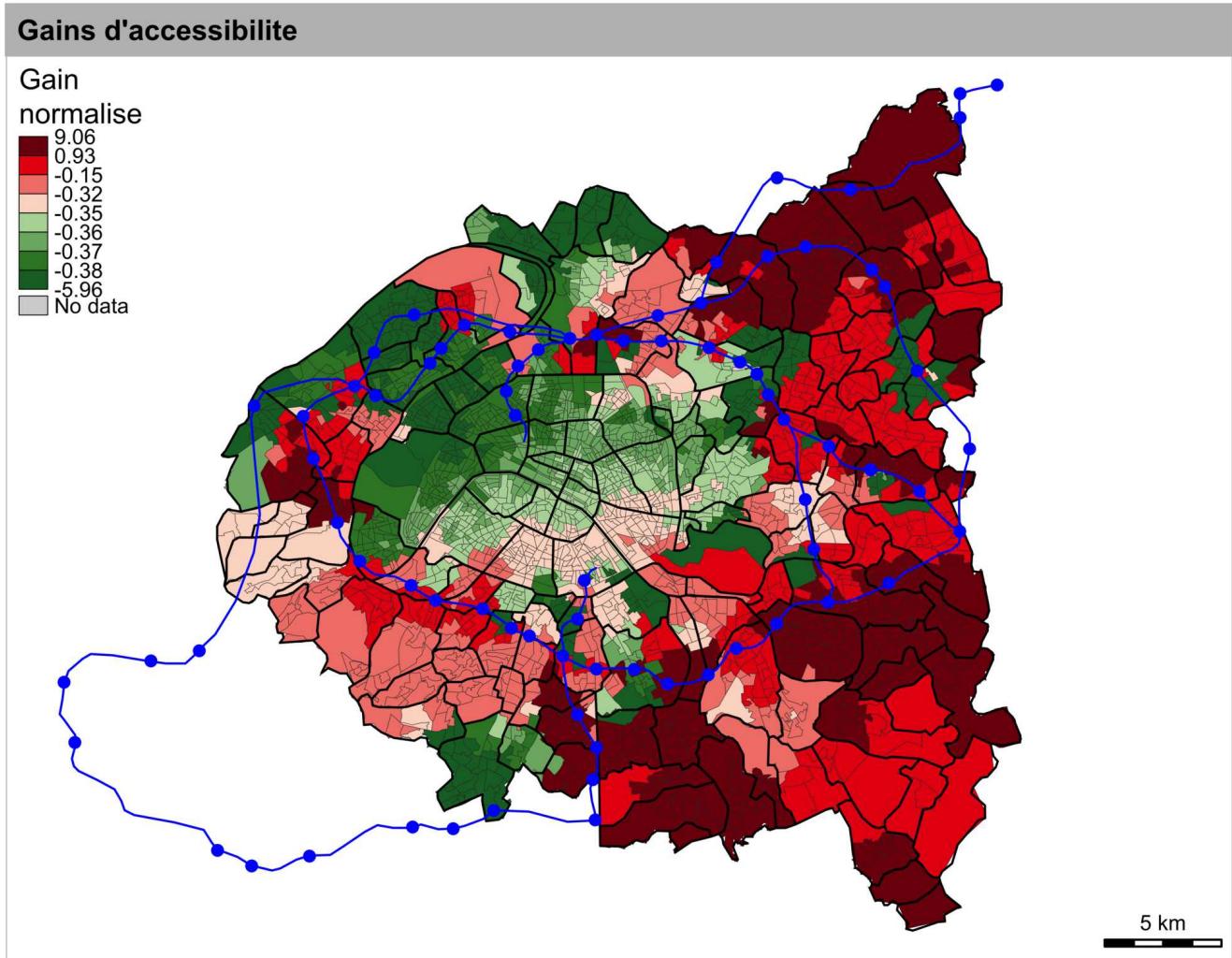


Figure 2: **Impact of GPE lines on temporal accessibility.** The map gives, for the *départements de petite couronne* and Paris (75, 92, 93, 94) the temporal accessibility gains, defined for each Iris (elementary infra-communal statistical unit) as the average travel time with public transport to all centroids of other communes, weighted by destination population. The gain is computed as the accessibility difference with and without *Grand Paris Express*. We show a normalized gain, i.e. centered (with a null average) and reduced (unit standard deviation). In blue, the lines and new stations of GPE. We observe the strongest gains mostly in the East, in consistence with the existing literature such as [Beaucire and Drevelle, 2013]. The territorial imprints of RER lines (A in the West, D and B in North, B in the South) exhibit relatively low gains since they are already very accessible.

cient need, and lead to several proposals on which the State and the Region have been in conflict around 2010 [Desjardins, 2010]. The *Arc Express* project [STIF, 2010], advocated by the Region and more focused on territorial equity, can be contrasted with initial proposals for a *Réseau du Grand Paris* aimed at linking “excellence clusters” despite a potential tunnel effect. The solution finally adopted (see the last *Schéma Directeur* [SDRIF, 2013]) is a compromise and allows a rebalancing of accessibility between the west and the east [Beaucire and Drevelle, 2013]. The Fig. 1 maps the different projects.

The immediate impacts of a new transportation infrastructure in terms of accessibility, i.e. of the transformation of the spatial distribution of different accessibilities, generally occur for much larger territories than the areas in which the line and its stations are constructed: accessibility patterns are a consequence of topological properties of the network and these are strongly discontinuous as a function of graph structure. We can illustrate the case of *Grand Paris Express* lines and of their direct impact on regional accessibility. We map in Fig. 2 the temporal accessibility gains allowed by the *Grand Paris Express* for metropolitan *départements* (75, 92, 93 and 94). The temporal accessibility is computed for each Iris i in the following way: with P_j the populations of *communes*, t_0 a parameter giving the typical commuting duration (that we fix at one hour [Zahavi and Talvitie, 1980]), t_{ij} the travel time with public transport between the centroid of i and the one of commune j , we take a weighted average defined by

$$Z_i = \sum_j \left(\frac{P_j}{\sum_k P_k} \right) \cdot \exp(-t_{ij}/t_0)$$

This expression indeed allows to have an accessibility potential, and the weighting by population should remove some bias due to potentially negligible trajectories as a proportion of total travels. We recall that this is a normative accessibility in the sense of [Páez, Scott, and Morency, 2012] since the gravity parameter is fixed in a stylized way.

We observe, in accordance with the analysis by [Beaucire and Drevelle, 2013], a rebalancing of accessibility differentials between the East and the West. At an equal distance of the center, accessibility is lower for Seine-Saint-Denis and Val-de-Marne than for Hauts-de-Seine, i.e. that these *départements* have potentially more difficulties to access the rest of the metropolis. The map of average time gains also exhibits the highest gains for this two *départements*. Some *communes* that are socially and economically disadvantaged as Aulnay benefit from the highest time gains. The line 16 indeed allows a significant opening up of the North-east of Seine-Saint-Denis [Desjardins, 2016]. The creation of links from suburbs to suburbs is a crucial aspect of this opening up and is conceived as a motor of the emergence of new centralities,

towards an always more polycentric metropolis, in the inheritance of the planning policy of *Villes Nouvelles*, in order to obtain not neighboring suburbs anymore but districts that are a full part of Greater Paris. The effects can remain however mitigated depending on the areas: [L'horty and Sari, 2013] show that the *Grand Paris Express* will induce a direct access to a larger number of employments for a significant number of unemployed within the *Petite Couronne*, but that inequalities with *Grande Couronne* will increase and that there exists some risks of dropping out for far away *communes* with a low accessibility.

One of the crucial issues for the construction of Greater Paris is to stay careful on not obtaining a metropolis with multiple separated levels, and to exploit the increased connectivity at different scales (international, national, regional, metropolitan) in order to reduce territorial inequalities instead of increasing them³⁸. The novel network seems to contribute to this dynamic, under the condition of a coordinated territorial development, allowing the realization of immediate accessibility gains in terms of territorial transformations. There exists no method that can forecast it in a deterministic way as we already developed. It is however possible to retrospectively analyze from an empirical point of view the couplings between territorial variables and network variables, in order to quantitatively unveil co-evolution phenomena. We propose now to illustrate this approach.

Linking territorial dynamics and construction of the Grand Paris Express

One of the aims of our work in the following will be to empirically clarify situations in which strongly coupled dynamics linked to our problematic can be exhibited, and then through models to isolate processes and conditions allowing one or the other situation. We propose to deepen the illustration of GPE, while introducing a potential approach to link a territorial dynamic with the one of the anticipated network.

Various aspects of territories are concerned by interactions with networks. In previous empirical studies, no socio-economic attributes of populations inhabiting the territory nor economic values for land and real estate was considered. Both are however crucial elements of territorial dynamics and are extensively studied in fields such as territorial analysis or urban economics : for example, [Homocianu, 2009] studies households residential choices to understand land-use transportation interactions. We propose here to use a database of Real Estate transactions for Parisian region on the last 20 years, with 2 years temporal granularity and exact spatial coordinates. [Guérois and Le

³⁸ We recall that an unequal distribution of agents and resources will generate differences in potential larger than a uniform distribution, these can then be linked to the evolution of the network.

Goix, 2009] used it for example to obtain typologies of spatial dynamics of the Parisian real estate market.

This more precise study can be understood as a research of early warnings of network potential breakdowns: indeed, if intrinsic territorial dynamics anticipate the arrival of a new public transportation station, the implications will be much different to the case where it will then drive these variables after its construction. The interpretation in terms of “structuring effects” will indeed be significantly different. We apply here the method of spatio-temporal causalities developed in 4.2. We propose to study the relations between the accessibility differential for each project, and variables linked to land (real estate transactions) and socio-economical, in order to see if it is possible to capture a link between accessibility differentials and differentials in territorial variables. Indeed, the links between new lines and real estate value evolution are sometimes dramatic [Damm et al., 1980].

Data for real estate transactions are provided by the BIENS database (*Chambre des Notaires d'Ile de France*, proprietary database). The number of transactions that can be used after cleaning is 862360, distributed across all IRIS areas (basic census units in France), for a temporal span covering the years 2003 to 2012 included. The data at the IRIS level for population and income (median income and Gini index) come from INSEE. Network data have been vectorialized from projects maps (see figure ?? for the different projects). Travel times are computed by public transportation only, with standard values for average speeds of different modes³⁹.

The travel time matrix is computed from all the centroids of IRIS to all the centroids of *Communes* (above aggregation level). These are linked to the network with abstract connectors to the closest station, with a speed of 50km.h⁻¹ (travel by car). Analysis are implemented in R [R Core Team, 2015] and all data, source code and results are available on an open git repository⁴⁰.

We compute for each project, the accessibility differentials ΔT_i in average travel time from each IRIS, in comparison with the network without the project. Average travel time accessibility is defined as $T_i = \sum_k \exp -t_{ik}/t_0$ with k *Communes*, t_{ik} travel time, and t_0 a decay parameter. We do not weight here by the population of destination communes, on the contrary to the accessibility Z_i we used previously, to ensure we do not capture any auto-correlation for population or correlations between population and the territorial vari-

³⁹ That we take as the following: RER 60km.h⁻¹, Transilien 100km.h⁻¹, Metro 30km.h⁻¹, Tramway 20km.h⁻¹.

⁴⁰ At

<https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/SpatioTempCausality/GrandParis>. Data for the BIENS database are given only at the aggregated level of IRIS and for price and mortgage variables, for contractual reasons closing the database.

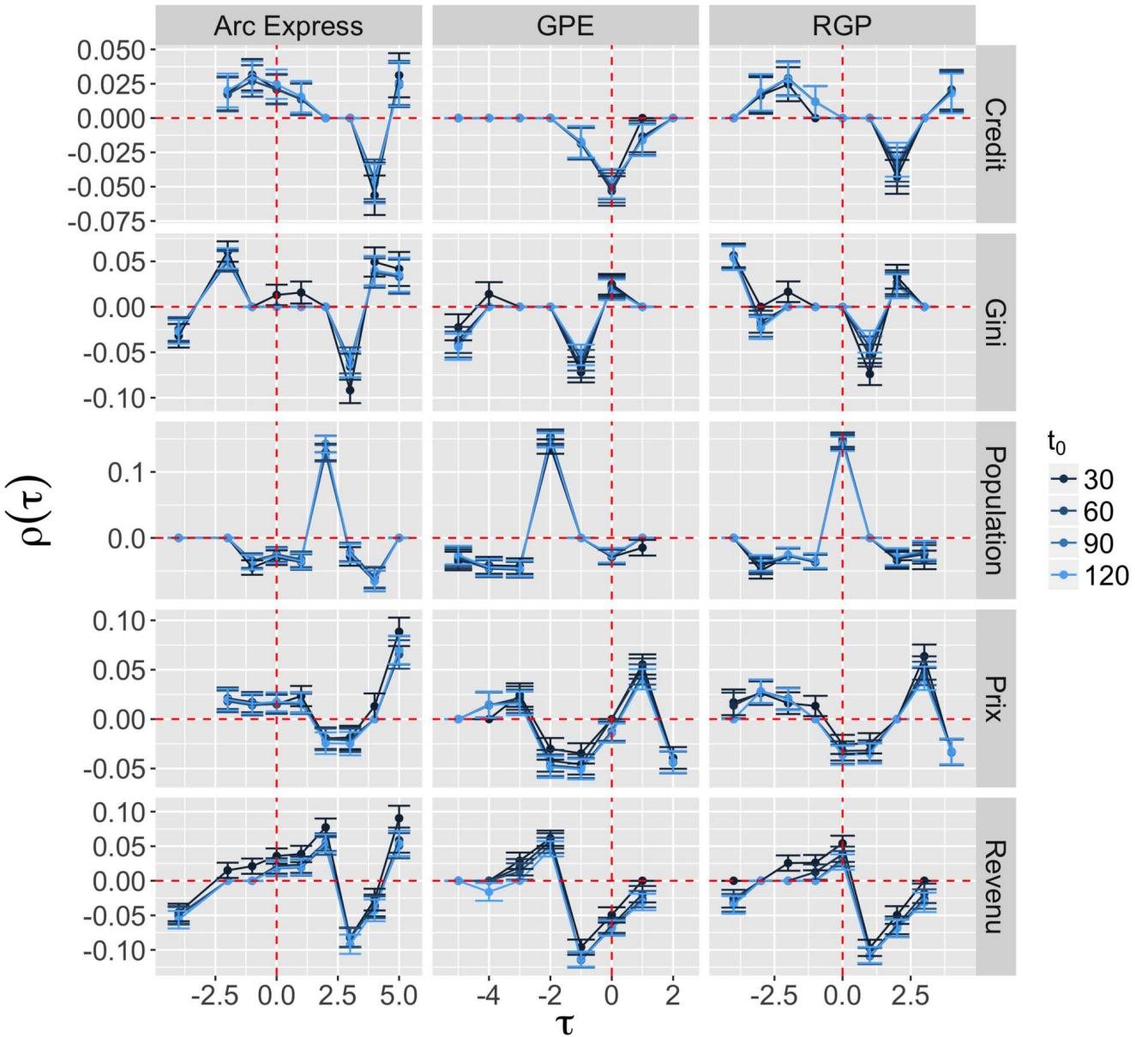


Figure 3: Empirical lagged correlations between accessibility differentials and territorial variables. Plots show the value of the lagged correlation between differentials of accessibility $\rho(\tau)$ as a function of the lag τ , in terms of average travel time ΔT_i , for each project (in columns: Arc Express, Grand Paris Express (GPE), Réseau du Grand Paris (RGP)) and the differential of the different socio-economic and real estate variables ΔY_i (in rows: values of real estate mortgages (Credit), Average price of real estate transactions (Price), Median income (Income), Gini index for incomes (Gini), Population), for different values of the decay parameter t_0 . Error bars give the 95% confidence interval. Dotted red lines are a reading guide: they allow horizontally to check if correlations are significant, and vertically to check the value of the optimal lag. For example, an interpretation of the first row suggests that the older projects have caused a decrease in granted real estate mortgages in Iris which accessibility had a positive growth, and that these variables are synchronized for GPE.

ables we study. To each project is associated a date⁴¹, corresponding roughly to the mature announcement of the project, what remains a bit arbitrary as it is difficult on the one hand to determine precisely as a planning project does not emerge from nothing in one day, and on the other hand it may correspond to different realities of learning about the project by economic agents (we do therefore the limiting but necessary assumption of a diffusion of information for the majority of agents in a time smaller than a year).

The link between accessibility differentials and variations of territorial variables is done through the study of lagged correlations. This method will be developed in details in 4.2, but we do not need to enter into technical details here. The idea is the following: if two variables exhibit a strong correlation at a given temporal lag, there is a weak notion of causality, and the variation of the upstream variable may be at the origin of the ones of the variable which is not lagged in time (we use the term weak, since it is of course always possible that correlations are spurious).

We study the lagged correlations of ΔT_i with the variations ΔY_{ij} of the following socio-economic variables: population, median income, Gini index for income, average price of real estate transactions and average value of real estate mortgages. Correlation is estimated by lagging accessibility, i.e. by estimating $\rho[\Delta T_i(t - \tau), \Delta Y_i(t)]$. A Fisher test is done for each estimation and the value is set to 0 if it is not significant ($p < 0.05$ in a classical manner). The study with generalized accessibility in the sense of Hansen [Hansen, 1959] (weighted by populations at destination, or with populations at the origin and employments at destination) has also been conducted but is less interesting as it has a very low sensitivity to the mobility component (network and decay) compared to the variables themselves. It informs therefore only on relations between these and is not presented here.

We show in figure ?? the results for all networks and variables. The interpretation can be done the following way: for a variable and a given project, the curve $\rho(\tau)$ can exhibit maxima for a value $\tau_m > 0$ or $\tau_m < 0$. This maximal correlation corresponds to a lag giving a “maximal synchronization” between the two variables, and the sign of the lag gives the sense of causality between the two variables.

It is first remarkable to note the presence of significant effects (in the sens of significant correlations and a 95% confidence interval which does not contains 0) for all variables. Lower values for the parameter t_0 give correlations higher in absolute value, unveiling a possible higher importance of local accessibility on territorial dynamics. The behavior of population shows a clearly detached peak corresponding to 2008, what suggests an impact of the older project *Arc Express* on population growth. Under this assumption, the effect of other projects would then be spurious from their proximity in the

⁴¹ 2006 for *Arc Express*, 2008 for *Réseau du Grand Paris* and 2010 for *Grand Paris Express*

most important branches. It would imply that areas where they are fundamentally different such as *Plateau de Saclay* are less sensitive to transportation projects, what would confirm the artificial planned aspect of the development of this territory.

Concerning income, we observe a similar behavior but in a negative way, what would imply a decrease of wealth linked to the increase of accessibility, however accompanied by a decrease of inequalities since the Gini coefficient also presents a negative correlation in positive lags. Finally, real estate prices are as expected driven by the potential arrival of new networks, suggesting a temporal speculation bubble. We demonstrate thus the existence of complex lagged correlation links, that we call causalities in this sense, between territorial dynamics and anticipated dynamics of networks. A finer understanding of implied processes is beyond the scope of this preliminary study and would imply for example qualitative fieldwork or targeted case studies.

This study suggests potential effects of the modification of accessibility due to Greater Paris projects, since some effects that were revealed can be linked to planning policing that also anticipate the new network. We thus suggest an effective existence of processes implying an effect of the network on territories, since most optimal lags are positive.

1.2.2 *Pearl River Delta*

We now switch the geographical region, the urban structure, and the time period, in order to describe an other relevant case study in China. The extended Parisian region can be read as a consistent entity⁴²: it would be a *mega-city region*, concept that we will now define and develop for the particular instance of Peral River Delta.

New urban regimes and mega-city regions

The notion of megalopolis has been introduced by [Gottmann, 1961] to designate the emergence of urban agglomerates at a scale that did not exist before. It is at the origin of the concept of *Mega-city Region* (MCR) which was consecrated by [Hall and Pain, 2006]. For the European case, they unveil assemblies of metropolis that are strongly connected regarding mobility flows, connections between companies, which form what they call polycentric *Mega-city Regions* (for example Randstad in Netherlands, the Rhine-Rhur region in Germany). Their characteristics are a certain geographical proximity of centers, a strong integration through flows, and a certain level of polycentrism.

⁴² [Gilli, 2005] recalls the importance of the hinterland of Bassin Parisien and the importance of not considering the hypercenter in an isolated way, and thus considerate the MCR which includes a certain number of important urban centers at one hour of Paris: Chartres, Orléans, Rouen, Reims and Lille thanks to High Speed Lines.

It consists in an urban form that did not exist before, which emergence seems linked to globalization processes.

This concept is even more relevant with the recent emergence of new types of urbanization, in particular through the accelerated urbanization in countries with a strong economic growth and undergoing a very rapid mutation such as China [Swerts and Denis, 2015].

The second case that we develop here enters this category: Pearl River Delta (PRD) is one of the classical illustrations of the structure of a strongly polycentric MCR. Historically initially only composed by Guangzhou, the development of Hong-Kong and the establishment of Special Economic Zones (SEZ) in the context of opening policies by DENG XIAOPING, lead to an extremely rapid development of Shenzhen, and in a less proportion of Zhuhai⁴³. Guangdong province in which PRD is fully located has currently the highest regional GDP within China, and the MCR contains a population of around 60 millions (estimations strongly fluctuate depending on the definition of the MCR which is taken, and the inclusion of the floating population). The phenomenon of migrations from rural areas is highly present in the region and a city such as Dongguan has for example based its economy on factories employing these migrant workers.

Governance of the mega-city region

[Ye, 2014] analyzes the actions of metropolitan governance at the scale of centers of the MCR, and more particularly how municipalities of Guangzhou and Foshan have progressively increased their cooperation to form an integrated metropolitan area, what can thus strongly influence the development of transportation for example and allowing the construction of a connected network. A strong tension between bottom-up processes, and a state control which is relatively strong in China, which originates from the Central State, to the province government and local government, has allowed the emergence of such a structure. The competition with other cities in the MCR remains strong, and the logic of integration (in the sense of articulation between the different centers, of interactions and of flows between these) of the MCR is only partly guided by the region. The particular nature of SEZ of Shenzhen and Zhuhai, linked to the privileged relations with the Special Administrative Zones of Hong-Kong and Macao, which have returned to the Popular Republic only at the end of the last millennium and keep a certain level of independence in terms of governance, complicates even more the relations between actors within the region. The issue of a correspondence between some levels of governance and urban processes is a tricky one: [Liao and

43 Shenzhen and Zhuhai were among the first Special Economic Zones, created in 1979 to attract foreign investments in these areas with flexible economic rules. The development model of Zhuhai was different of Shenzhen, since heavy industry was forbidden.

Gaudin, 2017] interprets the progressive transfers of economic initiatives from the central power to local authorities as a form of a multi-level governance.

Transportation Governance

In the frame of transportation within the MCR, there is no specific authority at this scale for the organization of transportation (but indeed entities at the level of the State, of the province and of municipalities), and each municipality manages independently the local network, whereas the connections between cities are ensured by the national train network. This leads to particular situations in which some areas have a very low accessibility, with a very strong heterogeneity locally. Therefore, the southern part of the city of Guangzhou which constitutes a direct access to the sea, is geographically closer to the center of Zhongshan, but a direct link by public transport is difficult to imagine, whereas the area is well linked to the center of Guangzhou by the metro line. A similar situation can be observed at the terminus of line 11 in Shenzhen, for the neighbor district of Dongguan, the latest having a very low accessibility by public transport⁴⁴. This situation could however be transitory, given the infrastructures already being built and the ones planned on a longer term: the Shenzhen metro, which covers today 285km, is planned to reach 30 lines and a length of around 1100km⁴⁵ in 2030 as declared by the official plan of the city [Shenzhen Planning Bureau, 2016]. It is clear that these developments mostly follow an existing urban development, a crucial issue is the voluntarism and the capacity to contain urban sprawl and to structure future developments around this new network, in the spirit of a voluntary integration between urbanism and transportation of the type *Transit Oriented Development* that we introduced before. Different final stations will be connected to the Dongguan metro, and new intercity lines will structure the longer range mobility, what will make the Delta a relatively well integrated in terms of public transport in a close temporal horizon. To have an idea of the development of the network in the coming years, the Table 2 gives the size of the planned networks in the different cities for 2030.

⁴⁴ See the map 4 for locations, the map giving also the accessibility with the road network.

⁴⁵ For comparison, the Transilien network has a length around 1300km with RER lines included, what could make them comparable, but one must keep in mind that Ile-de-France has a surface of 12000km² against 2000km² for Shenzhen. This implies for Shenzhen a much higher transport density, corresponding to high urban density areas, such that the plan anticipates 70% of commuting by metro at the horizon 2030.

Table 2: Public transportation in Pearl River Delta. We give populations in 2010 taken from [Guangdong Province, 2013]. Network lengths are taken from the different planning documents for the Guangzhou metro [Guangzhou Metro, 2016], the Shenzhen metro [Shenzhen Planning Bureau, 2016] and the Dongguan metro [Dongguan Metro, 2017], and for the Zhuhai tramway [Zhuhai Tramway, 2016]. Zhongshan is not included since it exploits a BRT system but no heavy infrastructure.

Ville	Population	Réseau 2016	Réseau 2030
Guangzhou - Foshan	18.9 Mio	390km	800km
Shenzhen	10.4Mio	286km	1124km
Dongguan	8.2Mio	38km	195km
Zhuhai (Tramway)	1.6Mio	10km	173km

Impact of the Zhuhai-Hong-Kong-Macao bridge

A major transportation infrastructure project in the region is the bridge-tunnel closing the mouth of the Delta, linking Zhuhai and Macao to Hong-Kong (HZMB). The length of the crossing is 36.5km, what makes it an exceptional infrastructure [Hussain et al., 2011]. The opening to traffic was delayed of several years and is finally planned for 2018⁴⁶. [Zhou, 2016] shows that the expected changes in accessibility patterns for the West of the Delta are relatively strong, and these can potentially induce strong bifurcations in the trajectories of cities. The necessity of the project is advocated by the different stakeholders of the project (Guangdong province, Hong-Kong Special Administrative Region, Macao Special Administrative Region) using an argumentation of a strong economic benefit in the frame of opening policies, and also through a social benefit for the West in particular. For example, Zhuhai is positioned as a new pivot between Hong-Kong and the West. The balancing of accessibility, in the sense of a diminution of spatial accessibility inequalities, operates however only for the private car transportation mode, what conducts to question its potential impacts: on the one hand the access to automotive remains reserved to a part of the population only, on the other hand the negative impacts of congestion can rapidly moderate the accessibility gains. These accessibility gains are mapped following the same method as previously, and shown with accessibility Z_i itself in Fig. 4.

The medium and long term impacts of the bridge are difficult to estimate. [Wu et al., 2012] finds patterns similar to the ones we estimate, i.e. a significant benefit for Zhuhai (and Hong-Kong that we did not take into account), and also immediate effects of traffic modification and economic impacts due to the toll or the increase of tourism. They mostly postulate the position of Zhuhai-Macao as a new pivot in the region. Even if it can be directly verified in terms of centrality and accessibility, it is not evident that this new position will influence par-

⁴⁶ See the official website at <http://www.hzmb.org/cn/default.asp>.

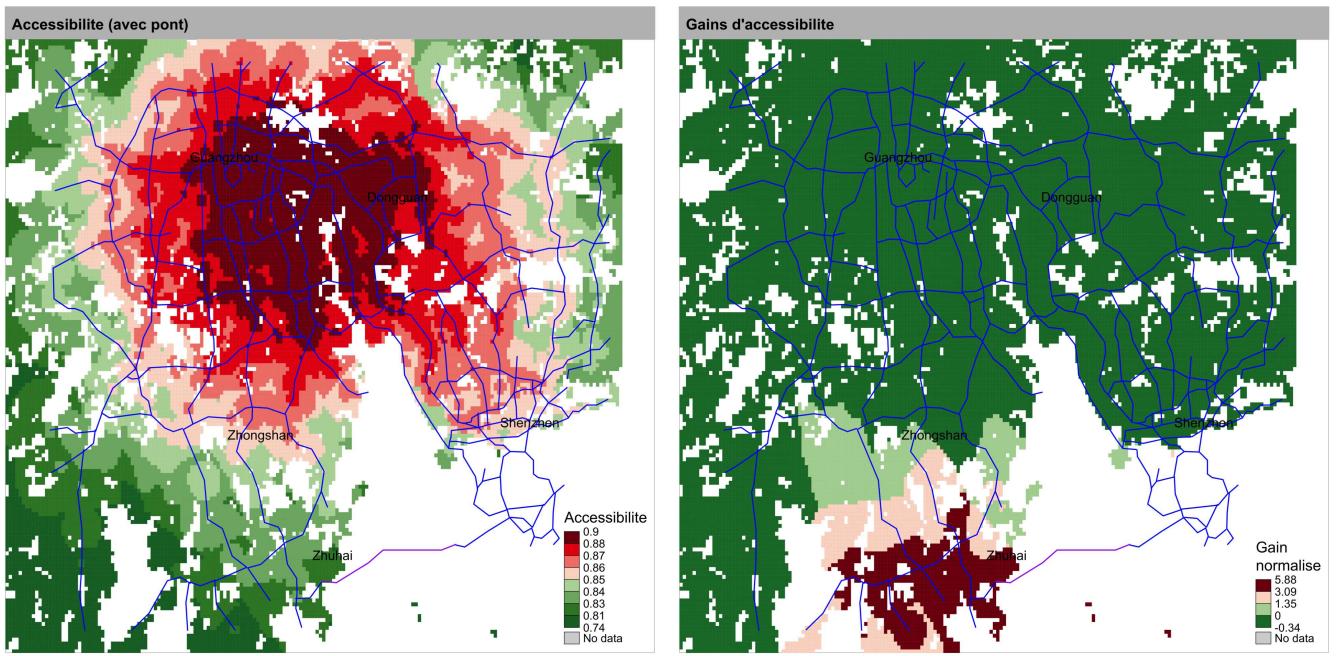


Figure 4: Accessibility gain induced by the HZMB in Pearl River Delta, for the territory of mainland China. (Left) Accessibility to population Z_i ; (Right) Normalized accessibility gains. The population of Hong-Kong is taken into account in destination points. The highway network (2017) is mapped in blue and the new link of the bridge in purple.

ticularly the socio-economic trajectory of Zhuhai. An increased particular political accompaniment implying an increased collaboration between Hong-Kong, Zhuhai and Macao will be important [Zhou, 2016]. Immediate economic effects are expected, as an increase of Zhuhai residents working in Hong-Kong (Zhuhai inhabitants are the only ones in the region to benefit of a special card allowing them to regularly visit the Special Administrative Areas⁴⁷), but cases showing the contrary, such as investments from Hong-Kong towards the West of the Delta, have no reason to be systematic: the first case extends the already existing dynamic with Macao, the second is mostly to be constructed. Thus, this example is a typical case of our general problematic.

Perspectives

A direction of exploration through modeling consists in considering the problem differently and to try to understand the dynamics of the metropolitan system in an intergated way, i.e. as a territorial system in our sense, in which the strong coupling between territory and network is operated through a proper ontology for governance entities. It will be the object of section 7.3.

⁴⁷ Source: fieldwork on 06/11/2016 with C. Losavio (see A.1).

This second shorter study has allowed us to emphasize a fundamentally different governance structure, but the same idea of a considerable transportation project which deeply modifies accessibility patterns. The expectations of actors regarding territorial mutations potentially induced are comparable in the sense that a high expectation is put in the project.

1.2.3 Comparability of case studies

We have studied here two cases of metropolitan development and infrastructure projects in their frame. The possibility of transfer of urban models (such as TOD), in the sense of the applicability of generic frameworks to different geographical contexts, is generally difficult. The synthesis of empirical conclusions obtained from very diverse case studies is also difficult.

The East-Asian particularity has already been shown for the economic structure, and how it can not be interpreted in a simple way by a separation of microscopic and macroscopic processes as some quick and ideologically oriented readings may have done, such as the approach of the World Bank [Amsden, 1994]. The comparability of urban systems is an open question at the core of issues for the Evolutive Urban Theory. It is linked to the ergodic character of these systems: the ergodicity assumption postulates that the trajectory of a city in time captures the set of possible urban states, and also that different cities are different manifestations of the same stochastic process at different periods. In that case, an ensemble of cities would allow to understand their temporal trajectories. It is intuitively not the case, and urban systems would rather be non-ergodic [Pumain, 2012b]. Empirically, this non-correspondence between global statistics and individual dynamics of cities is shown for traffic data by [Depersin and Barthelemy, 2017]. Thus we will have to remain cautious for the generalization of conclusions, as much as empirical as theoretical, or obtained through modeling.

★ ★

*

We have thus seen in this section, from two very different case studies, but having the common feature to exhibit significant transportation infrastructure projects, that the immediate impacts of these in terms of accessibility can be consequent, but that it is complicated to associate these gains to possible future mutations. We begin to foresee the difficulty to characterize co-evolution.

We will in the next section even more diversify our examples, from fieldwork observations, and thus from a more subjective and complementary point of view.

★ ★

★

1.3 FIELDWORK OBSERVATIONS OF INTERACTIONS

This section proposes to illustrate the issue of interactions between transportation networks and territories, and more particularly their complexity and the diversity of possible situations already perceptible in a qualitative way (and also subjective in a second time) at the microscopic scale, through concrete fieldwork examples. The geographical subject is Pearl River Delta, in Guangdong province, that we already described before, and more particularly mostly the city of Zhuhai. The objective is to enrich our repertory with concrete situations, to understand if these can be associated to the generic processes we have already exhibited, or if others can be observed at the scales of observation.

We assume the term of *Geographical Fieldwork*, with all knowledge of epistemological debates its use can raise. Indeed, we extract observations from places that were experimented, in the context of a given problematic [Rettaillé, 2010]. Our approach will also highlight the role of representations, underlined as a type of fieldwork in itself by [Lefort, 2012], when we will give a subjective view.

In the frame of the European project Medium⁴⁸, aiming at an interdisciplinary approach of sustainability for Chinese cities by concentrating on medium-sized cities⁴⁹, this city was chosen as a case study. When the source is not explicitly precised, observations come from fieldwork, for which narrative reports are available in Appendix A.1. The format of narrative reports is “on-the-fly” following the recommendations of [Goffman, 1989] for taking notes in an immersive fieldwork in particular, whereas the voluntary subjective position rejoins [Ball, 1990] which recalls the importance of reflexivity in order to draw rigorous conclusions from qualitative fieldwork observations of which the researcher is a part in itself⁵⁰.

⁴⁸ The Medium project, which establishes a partnership between European and Chinese universities, is entitled “*New pathways for sustainable urban development in China's medium-sized cities*”. It aims at studying sustainability through an interdisciplinary and multidimensional viewpoint, in the case of rapidly growing urban areas. Three medium-sized Chinese cities were chosen as a case study. See <http://mediumcities-china.org/> for more information.

⁴⁹ The definition of medium-sized cities considered for the project is broader than the official statistical definition of the Chinese government, and covers cities from 1 to 10 millions of inhabitants.

⁵⁰ The consideration of the researcher as a *subject* in relation with its object of study does not imply in our case a feedback of the researcher on the system because of its size in the case of a transportation network at the scale of the city, and indeed a conditioning of observations by a subjectivity of which we must detach in the posterior exploitation of the observation material, but which ignoring can only increase the biases.

1.3.1 *Development of a transportation network*

The objective of fieldwork is thus to observe the multiple facets and layers of a complex public transport system which is always transforming, its links with observable urban operations, and to what extent these witness of interaction processes between networks and territories. The spatial extent of observations spans on Zhuhai as an illustration of local transportation but also punctually on other regions in China. These observations have their proper logic in comparison to the modeling of transportation networks or data analysis, such as accessibility studies or interaction models between land-use and transportation, that will be done in the following. Indeed, these fail generally in capturing aspects at a large scale, which are often directly linked to the user, and which can become crucial regarding the effective use of the network. For example, multi-modality⁵¹ can be in practice made efficient through the emergence of self-organized informal transportation modes, or the establishment of new modes such as bike-sharing, what solves the “last-mile problem” [Liu, Jia, and Cheng, 2012], which seems to be often neglected in the planning of newly developed areas in China. On the contrary, practical details such as tickets reservation or check-in delays at boarding can considerably influence use patterns.

Several trips on the Chinese territory were made to observe the concrete manifestations of the high speed network development. Since 2008, China has established the larger HSR network in the world from scratch, which has a great success and which lines are currently saturated. It answers primary demand patterns in terms of city size, showing that it was planned such that the network answers to territorial dynamics. Its high usage shows the impact of network on mobility, what is a possible precursor of territorial mutations.

To show to what extent territories can influence the development of network in diverse ways, we can take a particular example, linked to the development of tourism, which corresponds to a particular dimension taken into account in planning. Thus, the line between Guangzhou and Guiyang (North-West axis which is precursor of the future direct link Guangzhou-Chengdu) have witnessed the opening of stations specifically for the development of tourism, such as Yangshuo in Guangxi, which number of visits has then strongly increased (see maps in Appendix A.1). One year after the opening of the station, the main road link with the city is still under construction, showing that the different networks react differently to constraints at different

⁵¹ Multi-modality consists in the combination of different transportation modes: road, train, metropolitan, tramway, bus, peaceful modes, etc., in a mobility pattern. A multimodal transportation system consists in the superposition of modal layers, and these can be more or less well articulated for the production of optimal routes following multiple objectives (cost, time, generalized cost, comfort, etc.) which themselves depend on the user, and of the mobility pattern.

levels. A higher number of trains stops on week-ends - more than one each hour, are full more than two weeks in advance. New mobility patterns can be induced by this new offer, as illustrate the interview of an inhabitant of Guangzhou done in Yangshuo, which came for a short week-end with her colleagues, in the context of a "team-building" trip financed by her startup in information technology. These new mobility practices are shown in a second interview of an inhabitant of Beijing met at Emeishan, sent by her company in Industrial Design for a short stay in Chengdu for a training in a local subsidiary. The company prefers the high speed train, and it recently increased the mobility practices for its employees.

A similar strategy can be observed concerning the connection of touristic destinations for the line Chengdu-Emeishan. The principal objective of this line is for now to serve the highly frequented touristic destinations of Emeishan and Leshan. However, the missing link between Leshan and Guiyang is already well advanced in its construction and will complete the direct link between Guangzhou and Chengdu. This reveals diachronic and complementary dynamics of network development following properties of territories. This line is a part of the structuring skeleton of the "8+8" recently reformulated by the central government⁵², and the traversed territories expect a lot from it as shows [LU et al., 2012] for the city of Yibin halfway between Chengdu and Guiyang.

We also observe joint mutations of the railway network and of the city. We illustrate thus in Fig. 5 the insertion of the HSR in its territories. Direct effects of the network are linked to the development of totally new districts in the neighborhood of new stations, sometimes in an approach of type "*Transit Oriented Development*" (TOD)⁵³ - we will come back to it with more details. Furthermore, more subtle indirect effects are suggested by clues such as the promotion of operations through advertisement. It shows the socio-economic expectations regarding the network and the local agents which have to contribute to its success: advertisements claiming the merits of high speed, and the selling of apartments in the associated real estate operations. This dynamic seems to contribute to the construction of a "middle class" and of the role it has to play in the dynamism of territories [Rocca, 2008]⁵⁴. The insertion of lines in territories seems in some case to be forced, as shows the Yangshuo station which exploits the tourism opportunity offered by the passage of the line in a low populated area

⁵² It corresponds to the general plan for future high speed lines, recently actualized to include 8 North-South parallels and 8 East-West others, completing the 4+4 already realized.

⁵³ As we defined in 1.1, this planning paradigm aims at articulating the development of an heavy transportation infrastructure with urbanization, typically through a densification around stations.

⁵⁴ Construction which is, as JEAN-LOUIS ROCCA emphasizes, as much concrete since it depends on objective realities, as imaginary in the academic and political discourse, which construct the object simultaneously to its study or use.

but which is very attractive by its landscapes, or the new real estate operations in Zhuhai which are not very accessible because of their price.

Finally, it is important to remark the network development answers simultaneously to different types of territorial contexts. Branches of the new high speed network with a short range, such as the line Guangzhou-Zhuhai, can be seen as being at the intermediary between a long range service and a proximity regional transport, depending on the modularity of serving patterns. This line is thus placed within long range urban interactions (the service Zhuhai-Guiyang being for example ensured) and within interactions in the mega-city region, most of the service being trains to Guangzhou. To this can be added the classical train network which keeps a certain role in territorial interactions: some connections require the use of both networks and of urban transportation, such as the link between Zhuhai and Hong-Kong, experimented through terrestrial transportation modes only⁵⁵.

1.3.2 Implementing TOD: contrasted illustrations

The local urban network and real estate development operations are planned in close conjunction with the new train network : Zhuhai new tramway, of which a single line is today open and in test, aims at participating to a “Transit-oriented development” (TOD) approach of Urban Development which aims at promoting the use of public transport and a city with less cars, as claimed for example by the High-Tech Zone planning committee in charge of the development around Zhuhai North station. Observing the surroundings of Tangjia station, also constructed in the same spirit, the anti-urban atmosphere and unpractical setting can lead to question the effectiveness of the approach and wonder if it is not more a kind of self-fulfilling prophecy, as suggested by the advertisements for new real estate to sell highlighting the role of the train line.

Other field observations, such as in Hong-Kong new territories, witness of an efficient and well achieved TOD, with the smart combination of heavy transit and local light rail, together with high urban densities around stations.

These observations recall the complexity of urban trajectories coupled with network development, and how one must be careful before drawing any general conclusion from particular cases.

⁵⁵ Following the Hato Typhoon on 23/08/2017, maritime links with the center of Hong-Kong and the international airport has been interrupted for a significant part of the delta, and has been reopened for Zhuhai in the beginning of November 2017 only.



Figure 5: Local manifestations of the mutations induced by the new high speed network. (*Top Left*) High speed station of Tangjia, in Zhuhai city. The monumental advertisement for a real estate operation praises the merits of the proximity to the network, which is also used as an argument for higher prices; (*Top Right*) High speed line in Zhuhai, deserted bus stop and real estate project being realized in a difficultly accessible area: this urban fringe is in direct contact with the rural environment on the other side of the line, and eccentric from the city; (*Bottom Left*) Yangshuo station on the Guangzhou-Guiyang line, which principal function is the development of this touristic destination which bases most of its economy on that field; (*Bottom Right*) Advertisement for high speed in Sichuan, at the station of the international Chengdu airport on the line to Leshan and Emeishan. The train departs from the futurist city to fly over the countryside, recalling the tunnel effect of territories telescoped by high speed.

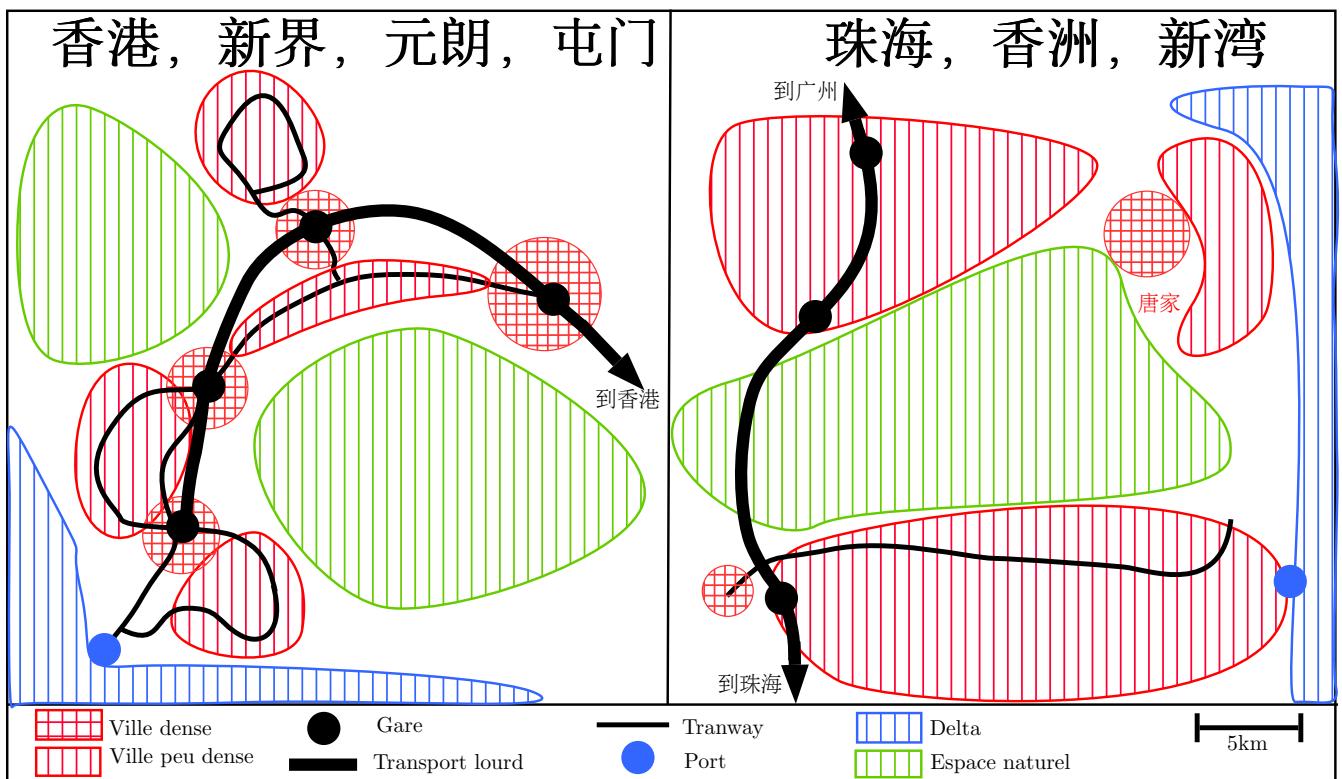


Figure 6: Comparative analysis of two implementations of TOD in PRD.

1.3.3 *Floating Observation*

The devil is in the details, and transportation systems are a typical embodiment of this image. What some will see as detail contains the majority of information for others. Logically trapped in a scientific information bubble, despite all the positions developed in introduction, we must stay aware of the nature and the range of knowledge produced here. What can be called detail in our case, for an accessibility study of a transportation network for example, such as subjective views of users or social relations inducted by the situations consequent to the dynamics of the transportation system, will be at the center of the questioning for some viewpoints in anthropology or sociology. Such knowledge, that could surely find a way in our work, is out of scope because of the lack of long-time fieldwork. We propose here however to sketch such a qualitative entry, to suggest directions for future research and better understand processes in a concrete way.

RESULTS Nos séquences d'observation de terrain ont eu lieu d'une part en Chine, majoritairement dans le Guangdong à Zhuhai, lors de sessions dédiées. Les observations s'étendent entre le 10/10/2016 et le 23/01/2017 ainsi qu'entre le 08/06/2017 et le 01/09/2017. Le mode de transport majoritaire est le bus de ville, suivi par le train régional, puis le train à grande vitesse et le ferry ; la portée des déplacements correspondent à celle des modes. Les compte-rendus détaillés, écrits à la volée de manière subjective et édités *a posteriori* le moins possible, comme expliqué précédemment, sont disponibles en Annexe A.1. Les observations pour la région parisienne sont quasi-quotidiennes et non consignées ; celles-ci ont eu lieu en plus grande partie sur la ligne 4 du métro et sur la ligne A du RER entre février 2016 et octobre 2016, sur la ligne R du Transilien et la ligne A du RER entre novembre 2016 et septembre 2017 puis entre février 2017 et mai 2017, puis sur la ligne 9 et la ligne 4 entre septembre 2017 et octobre 2017.

Les deux synthèses d'observation flottante pour chacune des régions, matériaux produit à partir des notes brutes, sont présentées dans les encadrés ci-dessus. Celles-ci illustrent entre autres par des exemples subjectifs certaines instances d'interactions entre réseaux et territoires, majoritairement aux échelles microscopique et mesoscopique, pour des processus touchant à la mobilité. La subjectivité et l'interprétation permet aussi d'extrapoler sur des processus à plus petite échelle, en terme d'accessibilité par exemple. Ceux-ci ne peuvent toutefois être pris plus que comme une illustration et introduction thématique. Par une prise de recul, nous proposons de lister certains enseignements qui peuvent être tirés de cette expérience à un niveau

Le ciel est gris et les visages fermés, ce Soleil du Nord n'a bien de lumière que le nom. L'initié ne saura s'y tromper et ressentira au fond de lui-même cette banale routine d'un aller-retour quotidien en RER. Il ne cherchera ni à maudire les planifications successives dont les stratifications temporelles ont laissé décanter cette organisation territoriale incongrue, ni à se prendre à rêver d'une trajectoire de vie alternative puisque choisir c'est un peu mourir et qu'il ne se sent pas une âme de Phoenix aujourd'hui. Peut être que la beauté de la ville est finalement dans ces tensions qui la façonnent à tous les niveaux et dans tous les domaines, ces paradoxes qui deviennent cadre de vie au point d'asséner quotidiennement une vérité. Cette philosophie de couloir de métro, le francilien en fait son cheval de bataille car après tout s'il vit en ville il doit bien la connaître. Encore un rail cassé sur le A, "tout cela est mal géré, et ce réseau est mal conçu" vocifère un utilisateur journalier, s'improvisant expert en planification ; d'autres plus patients prennent leur mal en patience mais se présentent tout aussi connasseurs d'une illusoire vision d'ensemble d'un territoire aux multiples visages. Ces usagers *sont* pourtant le système, de manière concrète à leur échelle d'espace et de temps, par induction et émergence aux échelles supérieures. La fourmi est supposée ne pas avoir conscience de l'intelligence collective dont elle est une des composantes fondamentales. Ils n'ont de la même manière que peu de perception de l'auto-désorganisation dont ils sont la source, peut-être la cause, et qui très sûrement subissent les désagréments de ses dynamiques. Se laisser flotter dans les transports franciliens est une expérience intemporelle. Presque thérapeutique parfois, quand l'un commence à perdre son optimisme quant à l'intérêt d'une vie urbaine, une excursion aléatoire en métro rappelle rapidement la richesse et la diversité qui sont un des plus grand succès des villes. C'est cette variété apparente de profils que le chercheur retiendra principalement de ces errements, et il gardera à l'esprit qu'il n'existe pas d'échelle où un traitement spécifique de chaque objet géographique n'est pas nécessaire : en quelques stations sur la ligne 4 le profil socio-économique des quartiers change profondément et souvent sans transition au moins trois fois, comme sur la ligne 13 nord où les motifs horaires soulignent d'autant plus de dures réalités socio-économiques qui sont en fait géographiques dans cet *espace produit* de la métropole. Lorsqu'il s'agit de modéliser, prendre en compte les limites de toute tentative de généralisation est d'autant plus cruciale comme chaque modèle est un équilibre fragile entre spécificité et généralité.

FRAME 2: A floating observation experiment in the Paris region.

de synthèse élevé, en contraste avec l'aspect subjectif et spécifique du produit de l'expérience. Ils sont les suivants :

1. La complexité du système de transport et en conséquence de son intégration avec l'urbanisme dans le système territorial, peut avoir des conséquences divergentes en termes de performance finale, et par exemple de soutenabilité. Dans le cas Chinois, l'auto-organisation et l'adaptabilité locale sont des atouts de la performance locale des nouvelles gares, tandis qu'en France la complexité semble être source de freins et finalement d'externalités négatives⁵⁶.

⁵⁶ Cet effet étant par ailleurs nécessairement en interdépendance forte avec les propriétés culturelles, qui est en fait une composante fondamentale des territoires.

Le trajet sera long. La perturbation choisie est la simulation de l'événement malencontreux, “ 我的护照丢了，我得去法国的领事馆在广州 ”, c'est-à-dire la perte de son passeport, qui oblige à prendre les transports pour se rendre au consulat. Celle-ci en Chine est assurément malencontreuse, puisque l'intégralité des trajets interurbains y est conditionnée. Traverser la mega-région urbaine du sud vers le nord pour rejoindre Guangzhou dans cette situation relève du défi. De bus urbain en bus urbain, des terminus plus ou moins bien articulés. Un village traditionnel factice est sorti de terre pour faire le bonheur des touristes, non loin de la maison natale de Zhongshan, peu crédible vu l'accessibilité. Des contrastes saisissants et un paysage très hétérogène, des enclaves de pauvreté dans des zones nouvellement prisées. Les relocalisations plus ou moins volontaires vers les franges façonnent un nouveau paysage d'inégalité géographique que l'on connaît déjà bien en Europe. À l'image de cet embouteillage continu, la réinvention de la ville déjà bien avancée ici se doit de faire des choix cruciaux pour être l'exemple d'une trajectoire durable. Une résilience impressionnante des usagers à une perturbation majeure, une capacité d'auto-organisation locale rendant fonctionnels des aménagements qui auraient pu ne pas l'être : de Shenzhen, Baoan à Zhuhai, Tangjia ou à Zhongshan, Xiaolan, la flotte de moto-taxis informels sauve l'accessibilité locale, comme me le confirme Jingzi habitant le sud de Zhongshan et étudiant au nord de Zhuhai et pour qui le train est une solution de mobilité même pas envisagée. Du tramway au BRT, choix et compromis équivalents ? Le premier étonne plus les nouveaux usagers. Peut-être aussi un argument percutant pour valoriser le complexe spécialement conçu autour du terminus. Les choix locaux sont d'autant plus différenciables qu'il est difficile de passer d'une zone à l'autre. Bloqué non loin de Guangzhou, le pont est fermé, le métro est en face mais impossible de le rejoindre. Juste le temps pour se rabattre sur la gare de Xiaolan et retour à la case départ, défi bien loin d'être réalisé. Observer l'adaptabilité ne suffit pas à la développer ? Des pratiques de mobilité très vite adaptées par les usagers : des trains à grande vitesse bondés en toute heure de la semaine, semble-t-il pour des motifs très divers. Un développement territorial apparent, des impacts à moyen terme qu'on peut parier non discutables. Si la structure est intégrée et flexible, discuter d'effets structurants devient une tautologie puisque la trajectoire du système urbain devient alors l'aspect plus ou moins contrôlable, selon les échelles de temps et d'espace.

FRAME 3: A floating observation in Guangdong, Zhuhai.

2. L'adaptabilité des territoires, dont l'une des composantes est par exemple la vitesse de mutation des pratiques de mobilité et reliée à l'adaptabilité, semble également très sensible aux particularités géographiques.
3. La question des échelles de temps et d'espace observables, ce qui conditionnera partiellement celles qu'on peut modéliser, est ambiguë dans l'observation, comme le témoigne l'observation conjointe de la mobilité et de manifestation de motifs d'accessibilité.
4. La comparabilité des cas et des situations géographiques est, dans notre cas, mais a priori plus généralement, un point épineux auquel il n'existe pas de solution idéale. Le compromis entre généralité et particularité est alors déterminant dans la construction d'une théorie et de modèles géographiques. Cette conclu-

sion tirée sur des études empiriques devrait s'appliquer aussi aux modèles, mais dans quelle mesure il s'agit d'une question ouverte.

Ces considérations participeront à l'orientation des postures ontologiques et épistémologiques que nous prendrons par la suite.

★ ★

★

Table 3: Interaction processes between networks and territories.

	Réseaux → Territoires	Territoires → Réseaux	Réseaux ↔ Territoires
Micro	Motifs de mobilité	Congestion du réseau ; Externalités négatives	Mobilité et structure sociale
Meso	Relocalisations ; Effets locaux des infrastructures	Rupture de potentiel	Planification métropolitaine ; TOD
Macro	Interactions entre villes ; Effet tunnel	Différenciation hiérarchique de l'accessibilité	Planification à grande échelle ; Dynamique structurelle ; Bifurcations

SYNTHESIS OF STUDIED PROCESSES

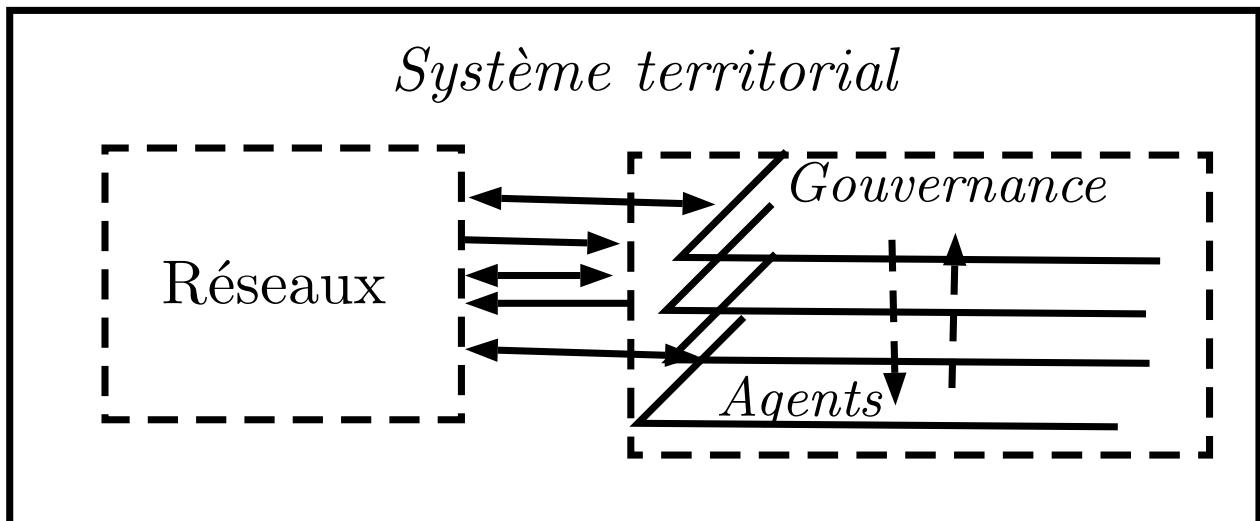
Nous concluons ce chapitre introductif par une synthèse et une mise en perspective des processus d'interaction identifiés par l'analyse théorique, empirique et la littérature. Celle-ci permettra de situer les revues des entreprises de modélisation auxquelles nous procéderons dans le chapitre 2, puis pourra être comparée à celle que nous établirons dans le cas des modèles.

A view from scales

Une première entrée pour synthétiser les processus abordés consiste à les considérer par échelle. Nous avons vu qu'une lecture multi-échelle était pertinente, et que celle-ci permettait globalement de dégager des échelles spatiales et temporelles caractéristiques : microscopique, mesoscopique et macroscopique, avec une assez bonne correspondance des échelles spatiales et temporelles. Cette typologie est bien sûr réduite, puisqu'elle simplifie la classe des processus qui pourraient sortir de ces correspondances, par exemple une mobilité à grande échelle, ou une bifurcation du système urbain qui se manifeste rapidement. De même, les processus eux-mêmes multi-échelles (la gouvernance du Grand Paris en est une bonne illustration, mobilisant des niveaux de gouvernance et des enjeux territoriaux à différentes échelles) sont pris en compte de manière simplifiée. L'axe complémentaire à celui des échelles se base sur les "effets et causes" : bien que nous restions toujours dans le cadre d'une causalité complexe comme présenté en introduction, nous avons mis en évidence des processus pour lesquels il est possible d'identifier un précurseur parmi le réseau ou le territoire (nous les noterons alors A → B), d'autres sont intrinsèquement complexes et contiennent déjà des causalités circulaires (par exemple dans le cas des processus de gouvernance), nous les noterons Réseaux ↔ Territoires. Le tableau de synthèse est alors donné en Table 3.

A hierarchical view

Une deuxième entrée privilégie le rôle des *acteurs*, c'est-à-dire des agents qui font le territoire. En effet, les problématique liées à la mobilité concernent les agents microscopiques, celles liées à l'accessibilité des acteurs urbains et économiques, celles liées à la planification des acteurs de gouvernance. Cet aspect peut être résumé par le schéma en Encadré 4.



Dans ce schéma, on identifie les acteurs territoriaux au sein du système territorial, qui se déclinent schématiquement sur deux échelles : les agents à l'échelle microscopique qui seront centraux pour les processus de mobilité, et les acteurs de gouvernance à des échelles supérieures, qui mènent les processus de gouvernance. Ils interagissent entre eux de manière complexe, et sont séparés ici conceptuellement par les pointillés d'autres aspects du territoire avec lesquels ils sont aussi couplés fortement.

Cette entrée peut être mise en perspective avec le cadre conceptuel de [Le Néchet, 2010], qui étudie les liens entre forme urbaine et pratiques de mobilité dans des contextes métropolitains. Celui-ci comprend le système urbain comme un couplage fort entre système de localisation, système d'activités et système de transport, en précisant l'influence des agents demandeurs (agents micro-économiques) et des agents aménageurs (agents de gouvernance) sur chaque système. Le système de transport correspond à nos réseaux et les deux autres systèmes à un aspect des agents territoriaux, qui contiennent aussi les agents précisés dans ce cadre. Ce parallèle reste à nuancer lorsqu'on change d'échelle : à celle du système de villes, lorsque les agents sont les villes, le système de localisation n'a plus de sens, puisque celui-ci

est adapté à une échelle au plus métropolitaine, et surtout aux ontologies correspondantes.

Cette double entrée de lecture des processus d'interaction entre réseaux et territoires conditionnera d'une part la revue de littérature des modèles faite en chapitre 2, et sera d'autre part complétée et précisée à l'issue de celui-ci.

★ ★

★

CHAPTER CONCLUSION

Les territoires interagissent de manière complexe avec les réseaux, en particulier ceux de transport, comme montré par les nombreux exemples empiriques ou les constructions théoriques passés en revue. À différentes échelles temporelles typiques (le jour, la décennie et le siècle), correspondent des échelles spatiales (urbaine, métropolitaine et système de villes), ainsi que des processus (mobilité, accessibilité et relocalisations, effets systémiques structurels et bifurcations). Les situations concrètes témoignent de réalités locales déclinées avec différentes nuances, et des processus portant ces processus abstraits avec différents rôles et interactions entre eux.

Nous avons dans une première section clarifié cette notion d'interaction entre réseaux de transports et territoires en construisant un cadre théorique qui permet de les considérer comme des composantes du système territorial dans son ensemble. Nous avons alors suggéré une approche par la co-évolution pour tenir compte de cette complexité. Afin de mieux cerner ces notions sur des exemples géographiques concrets, nous avons développé en 1.2 deux cas d'étude métropolitain d'actualité, et souligné les certitudes en termes d'impact d'accessibilité pour des projets majeurs d'infrastructures qui s'accompagnent systématiquement d'incertitude en terme de trajectoire du système à plus long terme. Enfin, nous proposons en 1.3 une excursion par des éléments de terrain dans le Guangdong, Chine.

À ce stade, ayant introduit l'objet d'étude thématique, nous proposons de nous intéresser plus particulièrement aux approches impliquant une modélisation, faisant le choix d'un rôle fondamental du *modèle* (sur lequel nous reviendrons plus en détails par la suite) dans la production de connaissance.

* * *

*

2

MODÉLISER LES INTERACTIONS ENTRE RÉSEAUX ET TERRITOIRES

La littérature empirique et thématique, ainsi que les cas d'études développés précédemment, semblent converger vers un consensus sur la complexité des relations entre réseaux de transport et territoires. Dans certaines configurations et à certaines échelles, il est possible de mettre en valeur des relations circulaires causales entre dynamiques territoriales et dynamiques des réseaux de transports. Nous désignons leur existence par le concept de *co-évolution*. Il semble difficile d'introduire des explications simples ou systématiques de ces dynamiques, comme le rappelle par exemple les débats autour des effets structurants des infrastructures [Offner, 1993].

Par ailleurs, les multiples situations géographiques suggèrent une forte dépendance au contexte, donnant une pertinence au travail de terrain et aux études ciblées. Or l'explication géographique et la compréhension des processus est très vite limitée dans cette approche, et intervient un besoin d'un certain niveau de généralisation. C'est sur un tel point que la théorie évolutive des villes se concentre en particulier, puisqu'elle permet de combiner des schémas et modèles généraux aux particularités géographiques. Au contraire, certaines théories issues de la physique s'appliquant à l'étude des systèmes urbains [West, 2017] peuvent être plus difficiles à accepter pour les géographes de par leur positionnement d'universalité qui est à l'opposé de leurs épistémologies habituelles.

Dans tous les cas, le *medium* qui permet de gagner en généralité sur les processus et structures des systèmes est toujours le modèle. Comme le rappelle J.P. MARCHAND¹, “*notre génération a compris qu'il y avait une co-évolution, la votre cherche à la comprendre*”, ce qui appuie le pouvoir de compréhension apporté par la modélisation et la simulation que nous jugeons être encore aujourd’hui à très fort potentiel de développement.

Sans développer pour le moment les nombreuses fonctions que peut avoir un modèle, nous nous baserons sur la position de BANOS qui soutient que “modéliser c'est apprendre”, et suivant notre positionnement dans une science des systèmes complexes suggéré en introduction, nous ferons ainsi de la *modélisation des interactions entre réseaux et territoires* notre principal sujet d'étude, outil, objet². Ce chapitre doit être pris comme un “état de l’art” des démarches de

¹ Communication personnelle, Mai 2017.

² Même si dans une relecture à la lumière de 8.3 de ce positionnement n'a pas de sens puisque notre démarche contenait déjà des modèles à partir du moment où elle était scientifique.

modélisation des interactions entre réseaux et territoires. Il vise en particulier à capturer différentes dimensions des connaissances : pour cela, nous mobiliserons des analyses en épistémologie quantitative.

Dans une première section 2.1, nous passons en revue de manière interdisciplinaire les modèles pouvant être concernés, même de loin, sans a priori d'échelle temporelle ou spatiale, d'ontologies, de structure, ou de contexte d'application. Cet aperçu est possible par les entrées disciplinaires diverses révélées au chapitre précédent : par exemple géographie, géographie des transports, planification. Cet aperçu suggère des structures de connaissances assez indépendantes et des disciplines ne communiquant que rarement.

Nous procédons dans 2.2 à une revue systématique algorithmique, qui correspond à une reconstruction par exploration itérative d'un paysage scientifique. Ses résultats tendent à confirmer ce cloisonnement. L'étude est complétée par une analyse de réseau multi-couches, combinant réseau de citations et réseau sémantique issu d'analyse textuelle, qui permet de mieux cerner les relations entre disciplines, leur champs lexicaux et leur motifs d'interdisciplinarité.

Cette étude permet la constitution d'un corpus utilisé pour la modélographie (typologie de modèles) et la métá-analyse (caractérisation de cette typologie) effectuée en dernière section 2.3. Celle-ci dissèque la nature d'un certain nombre de modèles et la relie au contexte disciplinaire, ce qui pose les bases et le cadre précis des efforts de modélisation qui seront développés par la suite.

★ ★

★

Ce chapitre est inédit pour sa première section ; reprend dans sa deuxième section le texte traduit de [Raimbault, 2017d], puis pour sa deuxième partie la méthodologie introduite par [Raimbault, 2016c] et développée dans [Raimbault, 2017] ainsi que les outils de [Bergeaud, Potiron, and Raimbault, 2017a] ; et enfin est inédit pour sa dernière partie.

2.1 MODELING INTERACTIONS

2.1.1 Modeling in Quantitative Geography

History

Modeling has in Theoretical and Quantitative Geography (TQG) a privileged role. [Cuyala, 2014] proposes an analysis of the spatio-temporal development of French speaking TQG scientific movement and underlines the emergence of the discipline as the combination between quantitative analysis (e.g. spatial analysis or modeling and simulation practices) and theoretical constructions. This dynamic can be tracked back to the end of the seventies, and is closely linked to the growing use and appropriation of mathematical tools [Pumain and Robic, 2002]. The integration of these two components allows to construct theories from empirical stylized facts, which then produce theoretical hypothesis that can be tested on empirical data. This approach is born under the influence of the *New Geography* in Anglo-Saxon countries and in Sweden.

Concerning urban modeling in itself, other fields than geography have proposed simulation models approximatively at the same period. For example, the LOWRY model, developed by [Lowry, 1964] with the objective to be applied directly to the Pittsburgh metropolitan region, assumes a system of equations for the localization of actives and employments in different areas. This model has been a cornerstone of urban modeling, since as shows [Goldner, 1971] it already had less than ten years after a broad heritage of conceptual and operational developments³. Relatively similar models are still largely used nowadays.

Simulation of models and intensive computation

A broad history of the genesis of models of simulation in geography is done by [Rey-Coyrehourcq, 2015] with a particular emphasis on the notion of validation of models (we will come back on the role of these aspects in our work in 3). The use of computation resources for the simulation of models is anterior to the introduction of current paradigms of complexity, coming back for example to FORRESTER, a computer scientist pioneer in spatial economics models inspired by cybernetics⁴. With the increase of computational capabilities, epistemological transformations have also occurred, with the emergence of explicative models as experimental tools. REY compares the dy-

³ [Goldner, 1971] makes the hypothesis that this success is due to the combination of three factors: a possibility of an immediate operational application, a causal structure of the model easy to grasp (actives relocate depending on employments), and a flexible frame that can be extended or adapted.

⁴ Which was, together with the systemic trend, precursors of current paradigms of complexity as we already developed.

namism of seventies when computation centers were opened to geographers to the current democratization of High Performance Computing⁵. Today, this ease of use is in particular exemplified by grid computing with a transparent use, i.e. without the need for advanced technical skills related to mechanisms of computation distribution. This way, [Schmitt et al., 2015] givef an exemple of the possibilities offered in terms of model validation and calibration, reducing the computational time from 30 years to one week - these techniques will play a crucial role in the results we will obtain in the following. This evolution is also accompanied by an evolution of modeling practices [Banos, 2013] and techniques [Chérel, Cottineau, and Reuillon, 2015].

Modeling, and in particular computational models of simulation, is seen by many as a fundamental building brick of knowledge: [Livet et al., 2010] recalls the combination of empirical, conceptual (theoretical) and modeling domains, with constructive feedbacks between each domain. A model can be an exploration tool to test assumptions, an empirical tool to validate a theory against datasets, an explicative tool to reveal causalities and internal processes of a system, a constructive tool to iteratively build a theory jointly with associated models. These are examples among others: [Varenne, 2010b] proposes a classification of diverse functions of a model. We will consider modeling as a fundamental instrument of knowledge on processes within systems, and more particularly in our case within complex adaptive systems. We recall thus that our research question will focus on *models which ontology is mainly composed by interactions between transportation networks and territories*.

2.1.2 Modeling networks and territories

We develop now an overview of different approaches modeling interactions between networks and territories. First of all, we need to notice a high contingency of scientific constructions underlying these. Indeed, according to [Bretagnolle, Paulus, and Pumain, 2002], the “*ideas of specialists in planning aimed to give definitions of city systems, since 1830, are closely linked to the historical transformations of communication networks*”. The historical context (and consequently the socio-economical and technological contexts) conditions strongly the formulated theories. It implies that ontologies and corresponding models addressed by geographers and planners are closely linked to their current historical preoccupations, thus necessarily limited in scope and/or operational purpose. In a perspectivist vision of science [Giere,

⁵ The development of the first urban simulation models coincides with the opening of the first computation centers to social sciences and humanities, as recalls also PUMAIN (interview on 31/03/2017, see Appendix D.3) for example for the implementation of the ALLEN entropy model.

2010c], such boundaries are the essence of the scientific enterprise, and as we will argue in chapter 8 their combination and coupling in the case of models is generally a source of knowledge.

The entry we take here to sketch an overview of models is complementary to the one taken in chapter 1, by declining them through their main object (i.e. the relations Network → Territory, Territory → Network and Territory ↔ Network)⁶.

The reference frame for scales is also the one introduced in chapter 1, knowing that we do not consider the microscopic scales by choice to discard daily mobility. We have therefore roughly mesoscopic and macroscopic temporal and spatial scales.

We have seen that the correspondence to temporal and spatial scales is not systematic (see the provisional double entry typology for processes). On the contrary, the correspondence to fields of study and types of stakeholders is more systematic. This literature review is thus done following the latest logic.

Territories

The main current dealing with the modeling of the influence of transportation networks on territories lies in the field of planning, at medium temporal and spatial scales (the scales of metropolitan accessibility we developed before). Models in geography at other scales, such as the Simpop models already described [Pumain, 2012a], do not include a particular ontology for transportation networks, and even if they include networks between cities as carriers of exchanges, they do not allow to study in particular the relations between networks and territories. We will come back later on extensions that are relevant for our question. First, let recall the context of models closer to planning studies.

LUTI MODELS These approaches are generally named as *models of the interaction between land-use and transportation* (*LUTI*, for *Land-Use Transport Interaction*). Land-use generally means the spatial distribution of territorial activities, generally classified into more or less precise typologies (for example housing, industry, tertiary, natural space). These works can be difficult to apprehend as they relate to different scientific disciplines⁷. Their general principle is to model and simu-

⁶ We recall the meaning of this notation introduced in chapter 1: a direct arrow corresponds to processes that we can relatively univocally attribute to the origin, whereas a reciprocal arrow assumes the intrinsic existence of reciprocal interaction, generally in coincidence with the emergence of entities playing a role in these.

⁷ We make here the choice to gather numerous approaches having the common characteristic to principally model the evolution of land-use, on medium temporal and spatial scales. The unity and the relative positioning of these approaches covering from economics to planning, remain an open question that to the best of our knowledge has never been frontally tackled. The work done in 2.2 introduces elements of answer through an approach in quantitative epistemology.

late the evolution of the spatial distribution of activities, taking transportation networks as a context and significant drivers of localizations. To understand the underlying conceptual frame to most approaches, the Frame 5 sums up the one given by [Wegener and Fürst, 2004]⁸.

For example, from the point of view of urban economics, propositions for such models have existed for a relatively long time: [Putman, 1975] recalls the frame of urban economics in which main components are employments, demography and transportation, and reviews economic models of localization that relate to the LOWRY model already mentioned.

[Wegener and Fürst, 2004] give more recently a state of the art of empirical studies and in modeling on this type of approach of interactions between land-use and transport. The theoretical positioning is closer of disciplines such as transportation socio-economics and planning (see the disciplinary landscapes described in 2.2). [Wegener and Fürst, 2004] compare and classify seventeen models, among which no one includes an endogenous evolution of the transportation network on relatively short time scales for simulations (of the order of the decade). We find again indeed the correspondance with typically mesoscopic scales previously established. A complementary review is done by [Chang, 2006], broadening the context with the inclusion of more general classes of models, such as spatial interactions models (which contain traffic assignment and four steps models), planning models based on operational research (optimization of locations of different activities, generally homes and employments), the microscopic models of random utility, and models of the real estate market.

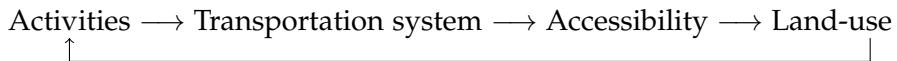
In order to give a better intuition of the logic underlying some Luti models, we detail in Frame 6 and in Frame 7 the structures, the ontologies, and assumptions of two models developed in the specific case of Ile-de-France (allowing on the one hand a comparison between both and on the other hand echoing the thematic development of 1.2). Even for very close ontologies (real estate prices, households localizations), we see the variety of possible assumptions and of issues raised by the models.

OPERATIONAL MODELS The variety of possible models has lead to operational comparisons [Paulley and Webster, 1991; Wegener, Mackett, and Simmonds, 1991].

More recently, the respective advantages of static and dynamic modeling was investigated in [Kryvobokov et al., 2013].

⁸ A more general frame that we already developed, that allows to bridge it with our frame, is the one given by [Le Néchet, 2010], which situates the triad Transportation system/Localization system/Activities system within the relation with agents: agents creating demand, agents building the city, external factors.

[Wegener and Fürst, 2004] introduces a general theoretical and empirical frame for land-use transport interaction models. The four concepts included are land-use, localization of activities, the transportation system and the distribution of accessibility. A cycle of circular effects are summed up in the following loop:



The transportation system is assumed with a *fixed infrastructure*, i.e. effects of the distribution of activities are effects on the *use* of the transportation system (and thus link to *mobility* in our more general frame): modal choice, frequency of trips, length of travels.

The theoretically expected effects are classified according to the direction of the relation (*Land-use*→*Transport* or *Transport*→*Land-use*, and a loop *Transport*→*Transport* that is not taken into account in our case), and according to the acting factor (residential density, of employments, localization, accessibility, transportation costs) and also by the aspect that is modified (length and frequency of trips, modal choice, densities, localizations). We can for example take:

- *Land-use*→*Transport*: a minimal residential density is necessary for the efficiency of public transportation, a concentration of employments implies longer trips, larger cities have a greater proportion of the modal part of public transportation.
- *Transport*→*Land-use*: a high accessibility implies higher prices and an increased development of residential housing, companies locate for a better accessibility to transportation at a larger scale.
- *Transport*→*Transport*: places with a good accessibility will produce more and longer trips, modal choice and transportation cost are highly correlated.

These theoretical effects are then compared to empirical observations, which for most of them give the way processes are implemented. Some are not observed in practice, whereas most converge with theoretical expectations.

Comment 1: An unisclar framework ? This framework takes schematically into account two main scales, the scale of daily mobility and the scale of the localization of activities. Knowing that in practice mobility behaviors are generally taken into account as average flows, it often reduces to a unique mesoscopic scale. All in all, it does not allow to take into account dynamics on longer time scales, that would include the evolution of the transportation network infrastructure or structural dynamics of systems of cities on long time periods.

Comment 2: A systematic view of structuring effects ? Furthermore, critics of the rhetoric of structuring effects may find in this framework its strong presence, since direct effects of accessibility on land-use and then the localization of activities are assumed here. These critics can be undermined by observing that these are theoretical expected effects, and that the framework is put into perspective of empirical effects indeed observed. We will however always take it with caution, by situating it in terms of context and scales.

FRAME 5: Conceptual framework of land-use transport interactions according to [Wegener and Fürst, 2004].

The Pirandello®model^a is presented in [Delons, Coulombel, and Leurent, 2008] as one of the first attempts to develop an operational Luti model in France. The model is based on four fundamental economic processes: the real estate market and the dwellings offer, the residential mobility of households, the attribution of travel destinations, the model choice. The model is static, i.e. computes n equilibrium for spatial distributions of actives and employments, and also for transportation flows. The fundamental processes taken into account and their implementation are the following:

- Residential choices of households are based on a utility function taking into account (i) a comfort term as a Cobb-Douglas of housing surface and income, corrected by a linear preference for individual dwellings; (ii) an accessibility term based on generalized cost (aggregation of transportation cost and time, with a value of time); (iii) the dwelling price and the local tax as a function of the housing surface; (iv) a fixed effect by income and by area; and (v) a random term assumed to follow a Gumbel law. Location probabilities for an income group are then given by a discrete choice model given this utility.
- The housing prices are formed following a scaling law of population.
- A local bidding mechanism answers to the demand previously obtained, as a function of an exogenous dwelling offer.
- Companies locate by maximizing their profit, function of the productivity (Cobb-Douglas in the salary and the accessibility) and the real estate price, under the constraint of a fixed spatial distribution of the number of employments, of the office surface, and of the total production of the region.
- Transportation is taken into account through a four steps model, which distributes model choices and destination choices with a discrete choice model, and flows are assigned according to a Wardrop equilibrium (see 3.2), what allows to adjust the values of accessibility given a spatial distribution of activities.

The mechanism to combine these different processes to obtain a global equilibrium is detailed by [Kryvobokov et al., 2013], and consists in the establishment of three sub-equilibriums at different scales: transportation flows (giving costs) on a short term, location and real estate prices on the middle term, land prices and available terrains (fixed in an exogenous way for all the modeled period).

Commentary: Equilibrium, operational model and calibration. A certain number of remarks can be done concerning this model, the most important for our approach are: (i) the equilibrium assumption can be a powerful tool to understand the structure of the attractors of the system, but has no empirical foundation, and even less for the coupling of equilibriums at different scales; (ii) thus, the operational nature of the model can be discussed, since the study of the impact of scenarios on the movements of attractors can difficultly allow to infer on local dynamics of the system; and (iii) sub-models are calibrated more or less rigorously and relatively separately, but the conditions of a calibration by decomposition are an open question still not well explored and linked to the nature of model coupling. In our sense, such a micro-based model would in any case be in better consistence with a philosophy of dynamical generative modeling and parsimony (see 3.1).

^a The origin of the name is not given, but strongly suggests the influence of its original creators V. PIROU and J. DELONS.
FRAME 6: **The Pirandello model.**

The Nendum2D model, described in details by [Viguié, Hallegatte, and Rozenberg, 2014], is focused on the localization of actives and their interaction with land rent and real estate promoters: it is a model inspired by the Fujita-Ogawa model [Fujita and Ogawa, 1982], inheriting from the literature in Urban Economics.

The processes included in the model are, with each its own time scale fixed by a parameter:

- Households make a compromise between housing surface and available budget without transportation costs and rent, following a Cobb-Douglas function for the corresponding utility. This process induces a dynamic for housing surface as a function of the distance to the center.
- They relocate in order to have an expected utility larger than the average.
- Rents evolve to maximize the occupation or in response to an external demand.
- New buildings are built by promoters that aim at maximizing their profits.

This model is dynamical and simulates the evolution of these different variables in space (the formulation above is monocentric, a polycentric extension and one taking into account an exogenous distribution of employments exist) and time. Its spatial scale is metropolitan, and the time scale can range from a medium scale (decade) to longer time-periods (century), knowing that the latest has a low credibility since it keeps static numerous other components of the urban system.

Comment: extension of ontologies. The coupling of Nendum with a model for traffic assignment, the Modus model^a, aims at including the feedback of congestion in the transportation system on costs, and thus on the localization and on the urban structure. Fundamental questions arise from the first coupling experiments:

- Is the masterplan *Schéma Directeur* really useful, since it seems to only accompany already existing dynamics ? In other words, *is the governance process endogenous* ? Does the Sdrif in fact capture an intrinsic dynamic on a longer time ?
- The coupling of models raises in itself technical difficulties, for communication between modules already implemented in different languages and for convergence of the coupled model in a reasonable number of iterations.
- It furthermore raises ontological difficulties: each model includes opposite mechanisms for the same ontology (aggregation effect against congestion effect for the distribution of population). The question is then if a specific coupling ontology is necessary (for example with specific equations integrating these contradictory effects), to allow on the one hand a better convergence, on the other hand a better ontological consistency.

^a In the frame of the current research project ANR VITE! (see <http://www.agence-nationale-recherche.fr/Projet-ANR-14-CE22-0013>).

These techniques operate also at small scales and consider at most land-use evolution. [Iacono, Levinson, and El-Geneidy, 2008] covers a similar scope with a further emphasis on cellular automata models of land-use change and agent-based models. These type of models are still largely developed and used today, as for example [Delons, Coulombel, and Leurent, 2008] which is used for Parisian metropolitan region. The short-term range of application and their operational character makes them useful for planning, what is far from our pre-occupation to obtain explicative models for geographical processes.

PERSPECTIVES FOR LUTI MODELS

Network Growth

Passons à présent au paradigme “opposé”, centré sur l’évolution du réseau. Il peut sembler incongru de considérer un réseau variable en négligeant les variations du territoire, au regard de l’aperçu de certains des mécanismes potentiels d’évolution revus précédemment (rupture de potentiel, auto-renforcements, planification du réseau) qui se produisent à des échelles de temps majoritairement plus longues que les évolutions territoriales. On verra ici qu’il n’y a pas de paradoxe, vu que (i) soit la modélisation s’intéresse à l’évolution des *propriétés du réseau*, à une courte échelle (micro) pour des processus de congestion, de capacité, de tarification, principalement d’un point de vue économique ; (ii) soit les composantes territoriales jouant en effet sur le réseau sont stables au échelles longues considérés.

Network growth can be used to design modeling enterprises that aim to endogenously explain growth of transportation networks, generally from a bottom-up point of view, i.e. by exhibiting local rules that would allow to reproduce network growth over long time scales (generally the road network).

[Xie and Levinson, 2009b] develops a broad review on network growth modeling extending to other fields: transportation geography early developed empirical-based models but which did concentrate on topology reproduction rather than on mechanisms according to [Xie and Levinson, 2009b]; statistical models on case studies provide mitigated conclusions on causal relations between offer and demand; economists have studied infrastructure provision from both microscopic and macroscopic point of views, generally non-spatial; network science has provided toy-models of network growth based on structural and topological rules rather on rules inspired from real processes.

ECONOMICS Economists have proposed such models: [Zhang and Levinson, 2007] reviews transportation economics literature on network growth within an endogenous growth theory [Aghion et al., 1998], recalling the three main features studied by economists on that

subject that are road pricing, infrastructure investment and ownership regime, and describes an analytical model combining the three. An other approach not mentioned that we will develop further is biologically inspired network design. We first give some example of economic-based and geometrical-based network growth modeling attempts. [Yerra and Levinson, 2005] shows through a reinforcement economic model including investment rule based on traffic assignment that local rules are enough to make hierarchy of roads emerge for a fixed land-use.

PHYSICS A very similar model in [Louf, Jensen, and Barthelemy, 2013] with simpler cost-benefits obtains the same conclusion.

Whereas these models based on processes focus on reproducing macroscopic patterns of networks (typically scaling), geometrical optimization models aim to ressemble topologically real networks. [Barthelemy and Flammini, 2008] proposes a model based on local energy optimization but it stays very abstract and unvalidated. The morphogenesis model given in [Courtat, Gloaguen, and Douady, 2011] using local potential and connectivity rules, even if not calibrated, seems to reproduce more reasonably real street patterns. Very close work is done in [Rui et al., 2013].

BIOLOGICAL NETWORKS Finally, an interesting and original approach to network growth are biological networks. These belong to the field of morphogenetic engineering pioneered by DOURSAT that aim to design artificial complex system inspired from natural complex systems and in which a control of emerging properties is possible [Doursat, Sayama, and Michel, 2012]. *Physarum Machines*, that are models of a self-organized mould (slime mould) have been shown to provide efficient bottom-up solution to computationally heavy problems such as routing problems [Tero, Kobayashi, and Nakagaki, 2006] or NP-complete navigation problems such as the Travelling Salesman Problem [Zhu et al., 2013a]. It has been shown to produce networks with Pareto-efficient cost-robustness properties [Tero et al., 2010], relatively close in shape to real networks (under certain conditions, see [Adamatzky and Jones, 2010]). This type of models can be of interest for us since auto-reinforcement mechanisms based on flows are analog to mechanisms of link reinforcement in transportation economics.

PROCEDURAL MODELING Other tentatives [De Leon, Felsen, and Wilensky, 2007; Yamins, Rasmussen, and Fogel, 2003] are closer to procedural modeling [Lechner et al., 2004; Watson et al., 2008] and therefore not of interest in our purpose as they can difficultly be used as explicative models.

2.1.3 Modeling co-evolution

We can now switch to models that integrate dynamically the paradigm Territory ↔ Network, which as we recall assumes that the conditioning of one by the other can not be identified. The ontologies used, as we will see, often couple⁹ network elements with territorial components, but this positioning is not necessary and some elements may be hybrid (for example a governance structure for the transportation network may simultaneously belong to both aspects). In our reading of models, these different specifications will naturally arise.

We will broadly designate by model of co-evolution simulation models that include a coupling of urban growth dynamics and transportation network growth dynamics. These are relatively rare, and for most of them still at the stage of stylized models. The efforts being relatively sparse and in very different domains, there is not much unity in these approaches, beside the abstraction of the assumption of an interdependency between networks and territorial characteristics in time. We propose to review them still through the prism of scales.

Microscopic and mesoscopic scales

GEOMETRICAL MODELS [Achibet et al., 2014] describes a co-evolution model at a very large scale (scale of the building), in which evolution of both network and buildings are ruled by a same agent, influenced differently by network topology and population density, and that can be understood as an agent of urban development. The model allows to simulate an auto-organized urban extension and to produce district configurations. Even if it strongly couples territorial components (buildings) and the road network, described results do not imply any conclusion on the processes of co-evolution themselves.

A generalization of the geometrical local optimization model described before is developed in [Barthelemy and Flammini, 2009]. It aims at capturing the co-evolution of network topology with the density of its nodes. The localization of new nodes is simultaneously influenced by density and centrality, yielding the looping of the strong coupling. More precisely, the global behavior of the model is the same, as the network extension behavior. Centers then localize following a utility function that is a linear combination of average betweenness centrality in a neighborhood and of the opposite of density (dispersion due to higher price as a function of density). This utility is used to compute the probability of localization of new centers following a discrete choices model. The model allows to show that the influence of centrality reinforces aggregation phenomena (in particular through an analytical resolution on a one-dimensional version of the

⁹ We recall the definition of model coupling, which corresponds to the one of system or process coupling given in introduction: it is the construction of a model that is simultaneously the extension of each initial model.

model), and furthermore reproduces exponentially decreasing density profiles (Clarcke's law) which are observed empirically.

[Ding et al., 2017] introduce a model of co-evolution between different layers of the transportation network, and show the existence of an optimal coupling parameter in terms of inequalities for the centrality in network conception: if the road network is assimilated at a fine granularity to a population distribution, this model can be compared with the precedent model of co-evolution between the transportation network and the territory.

ECONOMIC MODELS [Levinson, Xie, and Zhu, 2007] take an economic approach, which is richer from the point of view of network development processes implied, similar to a four step model (i.e. including the generation of origin-destination flows and the assignment of traffic in the network) including travel cost and congestion, coupled with a road investment module simulating toll revenues for constructing agents, and a land-use evolution module updating actives and employments through discrete choice modeling. The exploration experiments show that co-evolving network and land uses lead to positive feedbacks reinforcing hierarchies. These are however far from satisfying, since network topology does not evolve as only capacities and flows change within the network, what implies that more complex mechanisms (such as the planning of new infrastructures) on longer time scales are not taken into account. [Li et al., 2016] have recently extended this model by adding endogenous real estate prices and an optimization heuristic with a genetic algorithm for deciding agents.

From an other point of view, [Levinson and Chen, 2005] is also presented as a model of co-evolution, but corresponds more to a predictive model based on Markov chains, and thus closer to a statistical analysis than a simulation model based on these processes. [Rui and Ban, 2011] describe a model in which the coupling between land-use and network topology is done with a weak paradigm, land-use and accessibility having no feedback on network topology, the land-use model being conditioned to the growth of the autonomous network.

CELLULAR AUTOMATONS A simple hybrid model explored and applied to a stylized planning example of the functionnal distribution of a new district in [Raimbault, Banos, and Doursat, 2014], relies on mechanisms of accessibility to urban activities for the growth of settlements with a network adapting to the urban shape. The rules for network growth are too simple to capture more elaborated processes than just a simple systematic connection (such as potential breakdown for example), but the model produces at a large scale a broad range of urban shapes reproducing typical patterns of human settlements. This model is inspired by [Moreno, Badariotti, and Banos,

2012] for its core mechanisms but yield a much broader generation of forms by taking into account urban functions.

At these relatively large scales, spanning from the urban to the metropolitan scale, mechanisms of population localization influenced by accessibility coupled to mechanisms of network growth optimizing some particular functions seem to be the rule for this kind of models: in the same way, [Wu et al., 2017] couple a cellular automaton for population diffusion to a network optimizing local cost that depends on the geometry and on population distribution.

Models answering to more remote questions can furthermore be linked to our problem: for example, in a conceptual way, a certain form of strong coupling is also used in [Bigotte et al., 2010] which by an approach of operational research propose a network design algorithm to optimize the accessibility to amenities, taking into account both network hierarchy and the hierarchy of connected centers.

This way, co-evolution models at the microscopic and mesoscopic scales globally have the following structure: (i) processes of localization or relocalization of activities (actives, buildings) influenced by their own distribution and network characteristics; (ii) network evolution, that can be topological or not, answering to very diverse rules: local optimization, fixed rules, planning by deciding agents. This diversity suggests the necessity to take into account the superposition of multiple processes ruling network evolution.

Urban systems modeling

At a macroscopic scale, co-evolution can be taken into account in models of urban systems. [Baptiste, 1999] propose to couple an urban growth model based on migrations (introduced by the application of synergetics to systems of cities by [Sanders, 1992]) with a mechanism of self-reinforcement of capacities for the road network without topological modification. More precisely, the general principles of the model are the following.

- Attractivity and repulsion indicators allow for each city to determine emigration and immigration rates and to make populations evolve.
- Network topology is fixed in time, but capacities of links evolve. The rule is an increase in capacity when the flow becomes greater given a fixed parameter threshold during a given number of iterations. Flows are affected with a gravity model of interaction between cities.

The last version of this model is presented by [Baptiste, 2010]. General conclusions that can be obtained from this work are that this coupling yield a hierarchical configuration¹⁰ and that the addition of the

¹⁰ But we also know that simpler models, only a preferential for example, allow to reproduce this stylized fact. The model must have as an objective to answer to broader

network produces a less hierarchical space, allowing medium-sized cities to benefit from the feedback of the transportation network.

The model proposed by [Blumenfeld-Lieberthal and Portugali, 2010] can be seen as a bridge between the mesoscopic scale and the approaches of urban systems, since it simulates migrations between cities and network growth induced by potential breakdown when detours are too large. In the continuity of Simpop models for systems of cities, [Schmitt, 2014] describes the SimpopNet model which aims at precisely integrating co-evolution processes in systems of cities on long time scales, typically via rules for hierarchical network development as a function of the dynamics of cities, coupled with these that depends on network topology. Unfortunately the model was not explored nor further studied, and furthermore stayed at a toy-level. [Cottineau, 2014] proposes an endogenous transportation network growth as the last building brick of the Marius modeling framework, but it stays at a conceptual level since this brick has not been specified nor implemented yet. To the best of our knowledge, there exists no model which is empirical or applied to a concrete case based on an approach of co-evolution by urban systems from the point of view of the evolutive urban theory.

We can see well the opposition to epistemological principles of economic geography: [Fujita, Krugman, and Mori, 1999] introduce for example an evolutionary model able to reproduce an urban hierarchy and an organization typical of central place theory [Banos et al., 2011], but that still relies on the notion of successive equilibria, and moreover considers a “Krugman-like” model, i.e. a one dimensional and isotropic space, in which agents are homogeneously distributed¹¹. This approach can be instructive on economic processes in themselves but more difficultly on geographical processes, since these imply the embedding of economic processes in the geographical space which spatial particularities not taken into account in this approach are crucial. Our work will focus on demonstrating to what extent this structure of space can be important and also explicative, since networks, and even more physical networks induce spatio-temporal processes that are path-dependent and thus sensitive to local singularities and prone to bifurcations induced by the combination of these with processes at other scales (for example the centrality inducing a flow).

At the macroscopic scale, existing models are based on the evolution of agents (generally cities) as a consequence of their interactions,

questions, such as the fine understanding of co-evolution processes, what is not done here. However, one of its operational objectives is otherwise fulfilled, through the application to France and the study of the impact of a high speed line project, recalling the multiple possible functions of a model (see 3.1).

¹¹ The absence of a real space is not an issue in this economic approach that aims at understanding processes out of their context. In our case, the structure of the geographical space is not separable, and indeed at the core of the issues we are interested in.

Table 4: Synthesis of modeling processes.

Type	Classe	Echelle Temporelle	Echelle Spatiale	Fonction	Résultats	Paradigmes
Réseaux → Territoires	LUTI	Moyenne	Mesoscopique	Planification, Prédiction	Simulation de l'usage du sol	Économie urbaine
Territoires → Réseaux	Économie des Réseaux	Moyenne	Mesoscopique	Explication	Rôle de processus économiques	Économie, Gouvernance
Réseaux	Croissance géométrique	Longue	Meso ou Macro	Explication	Reproduction de formes stylisées	Modèles de Simulation, Optimisation locale
	Réseaux biologiques	Longue	Mesoscopique	Optimisation	Production de réseaux optimaux	Réseau auto-organisé
Territoires ↔ Réseaux	Économie des Réseaux	Moyenne	Mesoscopique	Explication	Effets de renforcement	Économie
	Croissance géométrique	Longue ou NA	Micro, Meso ou Macro	Explication	Reproduction de formes stylisées	Modèles de Simulation, Optimisation locale
	Systèmes Urbains	Moyenne, Longue	Macroscopique	Explication, prospection	Faits stylisés	Géographie complexe

carried by the network, whereas the evolution of the network can follow different rules: self-reinforcement, potential breakdown. The general structure is globally the same than at larger scales, but ontologies stay fundamentally different.

Synthesis

Il est essentiel à ce stade de s'oser à une synthèse et une mise en perspective de l'ensemble des modèles que nous avons passé en revue, puisque même si celle-ci sera nécessairement réductrice et simplificatrice, elle donne les fondations pour les analyses qui suivront.

Nous synthétisons les grands types de modèles que nous avons passé en revue dans le tableau suivant, en les classant par type (relation entre réseaux et territoires), par classe (grandes classes correspondant à la stratification de la revue), et en précisant les échelles temporelle et spatiales concernées, les fonctions, le type de résultats obtenus, les paradigmes utilisés. Celle-ci est donnée en Table 4.

An unthought coevolution ?

Le déséquilibre entre la dernière section rendant compte des modèles intégrant effectivement une dynamique fortement couplée (et possiblement une co-évolution) et les précédentes interroge : les modèles intégrant la co-évolution sont-ils si marginaux ? Est-il alors possible d'expliquer cette marginalité ?

L'objet des deux sections qui suivent sera de proposer des éléments de réponse à ces questions par des analyses épistémologiques en accroissant la connaissance des champs concernés et des modèles correspondants.

* * *

*

Nous avons ainsi donné dans cette section un aperçu large des modèles s'intéressant aux interactions entre réseaux de transport et territoires, incluant les modèles de co-évolution. Nous commençons donc à entrevoir une précision de la définition du concept de co-évolution dans ce cadre.

Nous proposons dans la section suivante de dresser une cartographie plus systématique de ce paysage scientifique, afin de renforcer le point de vue épistémologique et mieux situer la position que nous prendrons et les modèles que nous introduirons par la suite.

* * *

*

2.2 AN EPISTEMOLOGICAL APPROACH

A corollary of the thematic background introduced in chapter 1 is the need of an understanding of involved disciplines themselves to be able to build integrated heterogeneous models. The potentialities of couplings and integrations are greatly determined by existing approaches and corresponding gaps.

Diverses hypothèses peuvent être avancées pour tenter d'expliquer l'absence d'investigation des modèles de co-évolution :

- Suivant [Commenges, 2013], les acteurs scientifiques et opérationnels qui seraient concerné par l'application pratique de tels modèles se verrait remplacés par ces mêmes modèles et donc n'ont aucune incitation à les développer (explication sociologique).
- Les différentes disciplines qui détiennent les diverses composantes nécessaires à de tels modèles sont cloisonnées et ont des motivations divergentes (explication épistémologique).
- La construction de tels modèles comporte des difficultés intrinsèques rendant leur développement décourageant et pas parfaitement maîtrisé actuellement.

Nous n'aurons pas les moyens d'explorer la première hypothèse (ou plutôt elle demanderait un sujet à part entière, impliquant entre autres entretiens sociologiques). La troisième est soit une tautologie soit indémontrable, à-la-Church dirait-on, et l'ensemble de notre travail permettra d'y apporter des pistes de réponse. La deuxième par contre est comme nous allons le voir plus à notre portée.

This implies an advanced epistemological study in each field, that we propose to tackle in a systematic and quantitative way.

We describe and explore first a systematic review exploration algorithm, that retrieve corpuses of references through iterative semantic extraction. We describe then briefly possible extended bibliometrics by presenting an external example of application. We finally suggest possible development directions towards unsupervised data and text-mining.

Commençons par situer le contexte des analyses en *épistémologie quantitative*¹² que nous proposons de mener.

2.2.1 Quantitative epistemology

The possible methods for quantitative insights into epistemology are numerous. A good illustration of the variety of approaches is given

¹² Nous proposons d'utiliser ce terme pour des travaux à la croisée de la bibliométrie et de la scientométrie, des sciences cognitives, de l'épistémologie et des systèmes complexes, à l'image de l'*Epistémologie Appliquée* développée jusqu'en 2011 par le laboratoire CREA.

by network analysis. Using citation network features, a good predicting power for citation patterns is for example obtained by [Newman, 2014]. Co-authorship networks can also be used for predictive models (Sarigöl et al., 2014). A multilayer network approach was proposed in [Omodei, De Domenico, and Arenas, 2017], using bipartites networks of papers and scholars, in order to produce measures of interdisciplinarity using generalized centrality measures. Disciplines can be stratified into layers to reveal communities between them and therein collaboration patterns (Battiston et al., 2016). Keyword networks are used in other fields such as economics of innovation: for example, [Choi and Hwang, 2014] proposes a method to identify technological opportunities by detecting important keywords from the point of view of topological measures. In a similar manner, [Shibata et al., 2008] uses topological analysis of the citation network to detect emerging research fronts.

Systematic reviews

Literature review is a crucial preliminary step for any scientific work and its quality and extent may have a dramatic impact on research quality. Systematic review techniques have been developed, from qualitative review to quantitative meta-analyses allowing to produce new results by combining existing studies [Rucker, 2012]. Ignoring some references can even be considered as a scientific mistake in the context of emerging information systems [Lissack, 2013]. We aim to take advantage of such techniques to tackle our issue. Indeed, observing the form of the bibliography obtained in previous section raises some hypothesis. It is clear that all components are present for co-evolutive models to exist but different concerns and objectives seem to stop it. As it was shown by [Commenges, 2013] for the concept of mobility, for which a “small world of actors” relatively closed invented a notion ad hoc, using models without accurate knowledge of a more general scientific context, we could be in an analog case for the type of models we are interested in. Restricted interactions between scientific fields working on the same objects but with different purposes, backgrounds and at different scales, could be at the origin of the relative absence of co-evolving models. While most of studies in bibliometrics rely on citation networks [Newman, 2014] or co-autorship networks [Sarigöl et al., 2014], we propose to use a less explored paradigm based on text-mining introduced in [Chavalarias and Cointet, 2013], that obtain a dynamic mapping of scientific disciplines based on their semantic content. For our question, it has a particular interest, as we want to understand content structure of researches on the subject. We propose to apply an algorithmic method described in the following. The algorithm proceeds by iterations to obtain a stabilized corpus from initial keywords, reconstructing scientific semantic landscape around a particular subject.

Interdisciplinarity

The development of interdisciplinary approaches is increasingly necessary for most of disciplines, both for further knowledge discovery but also societal impact of discoveries, as it was recently coined by the special issue of *Nature* (*Nature*, 2015). [Banos, 2013] suggests that the development of such approaches must occur within a subtle spiral between and inside disciplines. An other way to understand this phenomenon is to understand it as the emergence of vertically integrated fields conjointly with horizontal questions as detailed in the Complex Systems roadmap ([Bourgne, Chavaliaras, and al., 2009]).

There are naturally multiple views on what is exactly interdisciplinarity (many other terms such as trans-disciplinarity, cross-disciplinarity also exist) and it actually depends on involved domains : recent hybrid disciplines (see e.g. the ones underlined by [Bais, 2010] such as astro-biology) are a good illustration of the case where entanglement is strong and new discoveries are vertically deep, whereas more loose fields such as “urbanism”, which have no precise definition and where integration is by essence horizontal, are an other illustration of how transversal knowledge can be produced. Interaction between disciplines are not always smooth, as shows the misunderstandings when urban issues were recently introduced to physicists as [Dupuy and Benguigui, 2015] recalls.

These concerns are part of an understanding of processes of knowledge production, i.e. the *Knowledge of the knowledge* as [Morin, 1986] puts it, in which evidence-based perspectives, involving quantitative approaches, play an important role. These paradigms can be understood as a *quantitative epistemology*. Quantitative measures of interdisciplinarity would therefore be part of a multidimensional approach of the study of science that is in a way “beyond bibliometrics” (Cronin and Sugimoto, 2014). The focus of this paper is positioned within this stream of research. We first review existing approaches to the measure of interdisciplinarity.

Definitions of interdisciplinarity itself and indicators to measure it have already been tackled by a large body of literature. [Huutoniemi et al., 2010] recall the difference between *multidisciplinary* (an aggregate of works from different disciplines) and *interdisciplinary* (implying a certain level of integration) approaches. They construct a qualitative framework to classify types of interdisciplinarity, and for example distinguish empirical, theoretical and methodological interdisciplinaries. The multidimensional aspect of interdisciplinarity is confirmed even within a specific field such as literature (Austin et al., 1996). A first way to quantify interdisciplinarity of a set of publications is to look at the proportion of disciplines outside a main discipline in which they are published, as [Rinia, Leeuwen, and Raan, 2002] do for the evaluation of projects in physics, complementary with judgement of experts. [Porter et al., 2007] designate this measure

as *specialization*, and compares it with a measure of *integration*, given by the spread of citations done by a paper within the different Subject Categories (classification of the Web of Knowledge), which is also called the *Rao-Stirling* index. [Larivière and Gingras, 2010] uses it on a Web of Science corpus to show the existence of an optimal intermediate level of interdisciplinarity for the citation impact within a five year window. A similar work is done in (Larivière and Gingras, 2014), focusing on the evolution of measures on a long time range. The influence of missing data on this index is studied by [Moreno, Auzinger, and Werthner, 2016], providing an extended framework taking into account uncertainty. The use of networks has also been proposed : [Porter and Rafols, 2009] combine the integration index with a mapping technique which consists in visualisation of synthetic networks constructed by co-citations between disciplines. [Leydesdorff, 2007] shows that the betweenness centrality is a relevant indicator of interdisciplinarity, when considering appropriate citation neighborhood.

2.2.2 Algorithmic Systematic Review

A broad bibliographical study suggests a scarcity of quantitative models of simulation integrating both network and urban growth. This absence may be due to diverging interests of concerned disciplines, resulting in a lack of communication. We propose to proceed to an algorithmic systematic review to give quantitative elements of answer to this question. A formal iterative algorithm to retrieve corpuses of references from initial keywords, based on text-mining, is developed and implemented. We study its convergence properties and do a sensitivity analysis. We then apply it on queries representative of the specific question, for which results tend to confirm the assumption of disciplines compartmentalization.

Tandis que la majorité des études en bibliométrie se reposent sur les réseaux de citations [Newman, 2014] ou les réseaux de co-auteurs [Sarigöl et al., 2014], nous proposons d'utiliser un paradigme moins exploré, basé sur l'analyse textuelle, introduit par [Chavalarias and Cointet, 2013], qui produit une cartographie dynamique des disciplines scientifiques en se basant sur leur contenu sémantique. Nous prenons le parti d'une appréhension de la diversité des domaines, introduite en 2.1, par cette information supplémentaire du paysage scientifique. Les méthodes que nous introduisons sont particulièrement adaptées pour notre étude puisque nous voulons comprendre la structure du contenu des recherches sur le sujet.

L'algorithme procède par itérations pour obtenir un corpus stabilisé à partir de mots-clés initiaux, reconstruisant l'horizon sémantique scientifique autour d'un sujet donné. La description formelle de l'algorithme est détaillée en Annexe A.2, avec les détails de son implémentation et des analyses de sensibilité. Sa logique est donnée

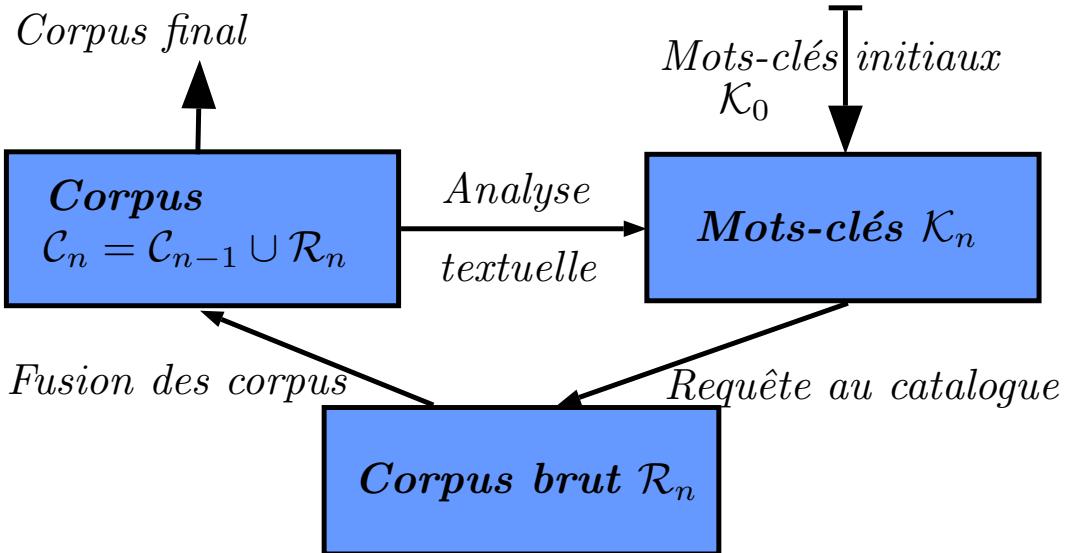


Figure 7: Global workflow of the algorithm, including implementation details: catalog request is done through Mendeley API; final state of corpuses are RIS files.

par le schéma en Fig. 7 : étant donné un ensemble de mots-clés de départ que l'on rassemble en une unique requête, on récolte des travaux qui en traitent, dont on extrait de nouveaux mots-clés pour itérer en boucle jusqu'à convergence éventuelle.

We start from five different initial requests that were manually extracted from the various domains identified in the bibliography (that are “city system network”, “land use transport interaction”, “network urban modeling”, “population density transport”, “transportation network urban growth”). We take the weakest assumption on parameter $N_k = 100$, as it should less constrain reached domains. After having constructed corpuses, we study their lexical distances as an indicator to answer our initial question. Large distances would go in the direction of the assumption made above, i.e. that discipline self-centering may be at the origin of the lack of interest for co-evolutive models. We show in Table 5 values of relative lexical proximity, that appear to be significantly low, confirming this assumption.

The disturbing absence of models simulating the co-evolution of transportation networks and urban land-use, confirmed through a state-of-the-art covering many domain, may be due to the absence of communication between scientific disciplines studying different aspects of that problems. We have proposed an algorithmic method to give elements of answers through text-mining-based corpus extraction. First numerical results seem to confirm the assumption. However, such a quantitative analysis should not be considered alone, but rather come as a back-up for qualitative studies that will be the object of further work, such as the one lead in [Commenges, 2013], in which

Table 5: Symmetric matrix of lexical proximities between final corpuses, defined as the sum of overall final keywords co-occurrences between corpuses, normalized by number of final keywords (100). We obtain very low values, confirming that corpuses are significantly far. Size of final corpuses is given as W .

Corpuses	1	2	3	4	5
1 ($W=3789$)	1	0	0.0719	0.0078	0.0724
2 ($W=5180$)	0	1	0.0338	0	0.0125
3 ($W=3757$)	0.0719	0.0338	1	0.0100	0.1729
4 ($W=3551$)	0.0078	0	0.0100	1	0.0333
5 ($W=8338$)	0.0724	0.0125	0.1729	0.0333	1

questionnaires with historical actors of modeling provide highly relevant information.

2.2.3 *Indirect Bibliometrics*

As described before, semantic analysis of final corpus does not contain all the information on disciplinary compartmentation nor on patterns of propagation of scientific knowledge as the ones contained in citation networks for example. Furthermore, data collection in the previous algorithm is subject to convergence towards self-consistent themes because of the proper structure of the method. It may give more insight about scientific social patterns of ontological choices in modeling to study communities in broader networks, that would more correspond to disciplines (or sub-disciplines depending on granularity level). We propose to reconstruct disciplines around our thematic, to obtain a more precise view of interdisciplinarity and the scientific landscape on our subject.

Context

The approach developed here couples citation network exploration and analysis with text-mining, aiming at mapping the scientific landscape in the neighborhood of a particular corpus. The context is particularly interesting for the methodology developed. First of all, the subject studied is very broad and by essence interdisciplinary. Secondly, bibliographical data is difficult to obtain, raising the concern of how the perception of a scientific landscape may be shaped by actors of the dissemination and thus far from objective, making technical solutions as the ones consequently developed here crucial tools for an open and neutral science.

Our approach combine semantic communities analysis (as done in [Palchykov et al., 2016] for papers in physics but with keyword extraction ; [Gurciullo et al., 2015] analyses semantic networks of political debates) with citation network to extract e.g. interdisciplinarity measures. Our contribution differs from the previous works quantifying interdisciplinarity as it does not assume predefined domains nor classification of the considered papers, but reconstructs from the bottom-up the fields with the endogenous semantic information. [Nichols, 2014] already introduced a close approach, using Latent Dirichlet Allocation topic modeling to characterize interdisciplinarity of awards in particular sciences. [Larivière and Gingras, 2014] quantifies interdisciplinarity over a long time range by looking at the field of references of publications.

Database Construction

Our approach imposes some requirements on the dataset used, namely:

- (i) cover a certain neighborhood of the studied journal in the citation network in order to have a consistent view on the scientific landscape;
- (ii) have at least a textual description for each node. For these to be met, we need to gather and compile data from heterogeneous sources, using therefore a specific application, which general architecture is synthesized in Fig. ???. For the sake of simplicity, we will denote by *reference* any standard scientific production that can be cited by another (journal paper, book, book chapter, conference paper, communication, etc.) and contains basic records (title, abstract, authors, publication year). We will work in the following on networks of references. Note that one significant contribution of this paper is the construction of such an hybrid dataset from heterogeneous sources, and the development of associated tools that can be reused and further developed for similar purposes.

INITIAL CORPUS Notre corpus initial est construit à partir de l'état de l'art établi en 2.1. Sa composition complète est donnée en Annexe A.2. Il s'agit de 7 références "phares" identifiées pour chacune des disciplines abordées précédemment. Le but ici n'est pas d'être exhaustif (cela le sera en 2.3), mais de construire une description du voisinage des domaines qui nous concernent. Celui-ci est pris de taille raisonnable (conduisant à un réseau final traitable sans méthode spécifique concernant la taille des données), mais les méthodes utilisées ici ont été développées sur des données massives, pour les brevets par exemple [Bergeaud, Potiron, and Raimbault, 2017a], et comme il le sera en Annexe F à l'ensemble de notre bibliographie.

C (FL) : police

CITATION DATA Citation data is collected from [Google Scholar](#), that is the only source for incoming citations [Noruzi, 2005] in our

case as the journal is poorly indexed in other databases¹³. We are aware of the possible biases using this single source (see e.g. [Bohannon, 2014])¹⁴, but these critics are more directed towards search results than citation counts. We retrieve that way two sub-corpora: references *citing Cybergeo* and references *citing the ones cited* by Cybergeo. At this stage, the full corpus contains around $4 \cdot 10^5$ references.

TEXT DATA A textual description for all references is necessary for a complete semantic analysis. We use for this another source of data, that is the online catalog of *Mendeley* reference manager software [Mendeley, 2015]. It provides a free API allowing to get various records under a structured format. Although not complete, the catalog provides a reasonable coverage (over 55%), yielding a final corpus with full abstracts of size $2.1 \cdot 10^5$, which structure is recalled in Fig. ??

Results

CITATION NETWORK PROPERTIES Des statistiques basiques pour le réseau de citation donnent déjà des informations intéressantes. Le réseau a un degré moyen de $\bar{d} = 2.53$ et une densité de $\gamma = 0.0013^{15}$. Le degré entrant moyen (qui peut être interprété comme un facteur d'impact stationnaire) est de 1.26, ce qui est relativement élevé pour des sciences humaines. Il est important de noter sa connexité faible, ce qui signifie que les domaines initiaux ne sont pas en isolation totale : les références initiales sont partagées à un degré minimal par les différents domaines. Nous travaillons sur la suite sur le sous-réseau des noeuds comprenant au moins deux liens, pour extraire le cœur de la structure du réseau et se débarrasser de l'effet "grappe". De plus, le réseau est nécessairement complet entre ces noeuds puisqu'on est remonté au deuxième niveau.

Nous procédons pour le réseau de citation à une détection de communautés par l'algorithme de Louvain, sur le réseau non-dirigé correspondant. L'algorithme fournit 13 communautés, de modularité dirigée 0.66¹⁶, extrêmement significative en comparaison à une estimation par bootstrap de la même mesure sur le graphe aléatoirement rebranché qui donne une modularité de 0.0005 ± 0.0051 sur $N = 100$ répétitions. Les communautés font sens de manière thématique, puisqu'on retrouve pour les plus grosses les domaines présentés dans la Table 6.

¹³ or was just added as in the case of *Web of Science*, indexing *Cybergeo* since May 2016

¹⁴ or <http://iscpif.fr/blog/2016/02/the-strange-arithmetic-of-google-scholars>

¹⁵ Pour référence, [Batagelj, 2003] présente les caractéristiques de 11 réseaux scientifiques de domaines divers et de taille allant de 40 à 8851 noeuds, et rapporte des densités variant de $3.3 \cdot 10^{-4}$ à 0.038, avec une médiane à 0.003, proche de celle de notre réseau.

¹⁶ La modularité est une mesure du "niveau de clustering" d'une partition d'un réseau en classes. L'algorithme de Louvain construit les communautés par optimisation gourmande de la modularité.

Table 6: Description and size of citation communities.

Domaine	Taille (% de noeuds)
LUTI	18%
Géographie Urbaine et des Transports	16%
Planification des infrastructures	12%
Planification intégrée - TOD	6%
Réseaux Spatiaux	17%
Etudes d'accessibilité	18%

Les appellations sont à regard d'expert *a posteriori*, selon les grands domaines dégagés dans la revue de littérature en 2.1¹⁷.

La Fig. 8 montre le réseau de citation et permet de visualiser les relations entre ces domaines. Il est intéressant d'observer que les travaux des économistes et des physiciens dans le domaine tombent dans la même catégorie d'étude des *Spatial Networks*. En effet, la littérature citée par les physiciens comporte souvent plus d'ouvrage en économie qu'en géographie, tandis que les économistes utilisent des techniques d'analyse de réseau. Ensuite, le planning, l'accessibilité, les LUTI et le TOD sont très proches mais se distinguent dans leur spécificités : le fait qu'ils apparaissent dans des communautés séparées témoigne d'un certain niveau de cloisonnement. Ceux-ci font le pont entre les approches réseaux spatiaux et les approches géographiques, qui comportent une partie importante de sciences politiques par exemple. Les liens entre physique et géographie restent très faibles. Ce panorama dépend bien sûr du corpus initial, mais nous permet de mieux comprendre le contexte de celui-ci dans son environnement disciplinaire.

SEMANTIC COMMUNITIES L'extraction des mots-clés est faite suivant une heuristique inspirée de [Chavalarias and Cointet, 2013]. La description complète de la méthode et de son implémentation est donnée en Annexe B.6. Elle se base sur les relations au second ordre entre les entités sémantiques, qui sont des *n-grams*, c'est-à-dire des mots-clés multiples pouvant avoir une longueur jusqu'à 3. Celles-ci sont estimées via la matrice de co-occurrence, dont les propriétés statistiques fournissent une mesure de déviation à des co-occurrences uniformes, qui est utilisée pour juger la pertinence des mots-clés. Sélectionnant un nombre fixe de mots-clés pertinents $K_W = 10000$, nous pouvons ensuite construire un réseau pondéré par les co-occurrences.

¹⁷ On note que cette dénomination est bien exogène et nécessairement subjective. Comme développé plus loin pour le réseau sémantique, il n'existe pas de technique simple pour une désignation endogène. Il faut garder cet aspect en tête pour la mise en perspective des interprétations et conclusions.

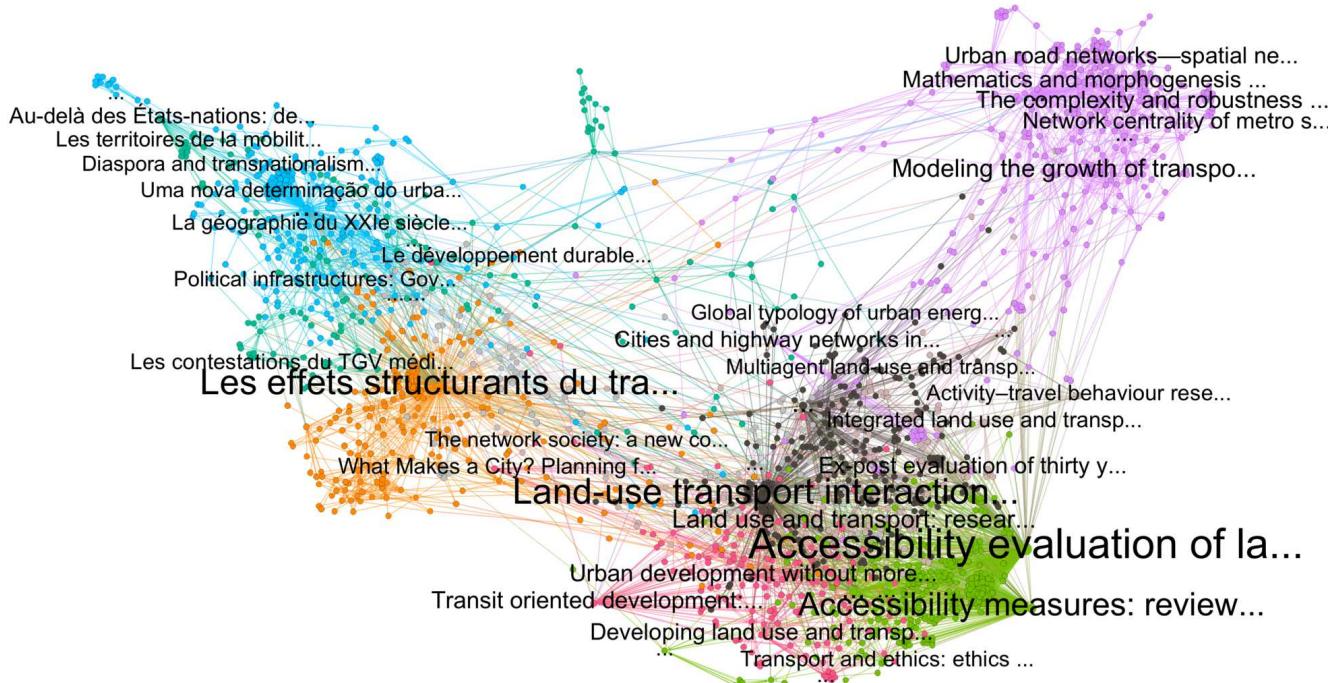


Figure 8: Citation Network

The topology of raw networks does not allow the extraction of clear communities, in particular because of the presence of hubs that correspond to frequent terms common to many fields (e.g. model, space). We assume these highest degree terms do not carry specific information on particular classes and can be thus filtered given a maximal degree threshold k_{\max} . Similarly, edge with small weight are considered as noise and filtered according to a minimal edge weight threshold θ_w . Keywords are preliminary filtered by a document frequency window $[f_{\min}, f_{\max}]$ which is slightly different from network filtering and complementary. A sensitivity analysis of resulting network topology to these parameters is presented in Fig. ???. We choose parameter values that maximize modularity under the constraint of a community number and size distribution of same magnitude as technological classes. This multi-objective optimization does not have a unique solution as objectives are somehow contradictory, and a compromise point must be chosen.

We then retrieve communities in the semantic network (using standard Louvain algorithm, with the optimized filtering parameters).

Table 7: Disciplines/domains/fields reconstructed from community detection in the semantic network

Name	Size	Weight	Keywords
Networks	820	13.57%	social network, spatial network, resili
Policy	700	11.8%	actor, decision-mak, societi
Socio-economic	793	11.6%	neighborhood, incom, live
High Speed Rail	476	7.14%	high-spe, corridor, hsr
French Geography	210	6.08%	système, développement, territoire
Education	374	5.43%	school, student, collabor
Climate Change	411	5.42%	mitig, carbon, consumpt
Remote Sensing	405	4.65%	classif, detect, cover
Sustainable Transport	370	4.38%	sustain urban, travel demand, activity-bas
Traffic	368	4.23%	traffic congest, cbd, capit
Maritime Networks	402	4.2%	govern model, seaport, port author
Environment	289	3.79%	ecosystem servic, regul, settlement
Accessibility	260	3.23%	access measur, transport access, urban growth
Agent-based Modeling	192	3.18%	agent-bas, spread, heterogen
Transportation planning	192	3.18%	transport project, option, cba
Mobility Data Mining	168	2.49%	human mobil, movement, mobil phone
Health Geography	196	2.49%	healthcar, inequ, exclus
Freight and Logistics	239	2.06%	freight transport, citi logist, modal
Spanish Geography	106	1.26%	movilidad urbana, criteria, para
Measuring	166	1.0%	score, sampl, metric

communities correspond to well-defined scientific fields (and/or domains, approaches). An expert validation allow us to give names to these, a more complicated naming procedure would eventually be possible (as in [Yang et al., 2000] for the case of patents where a chi-square test on distribution of documents in classes), but we prefer to stick here to a certain level of supervision. Table ?? summarizes the communities

MEASURES OF INTERDISCIPLINARITY Distribution of keywords within communities provides an article-level interdisciplinarity. Combination of citation and semantic layers in the hyper-network provide second order interdisciplinarity measures, that we don't use here because of the modest size of the citation network. More precisely, a reference can be viewed as a probability vector on semantic classes.

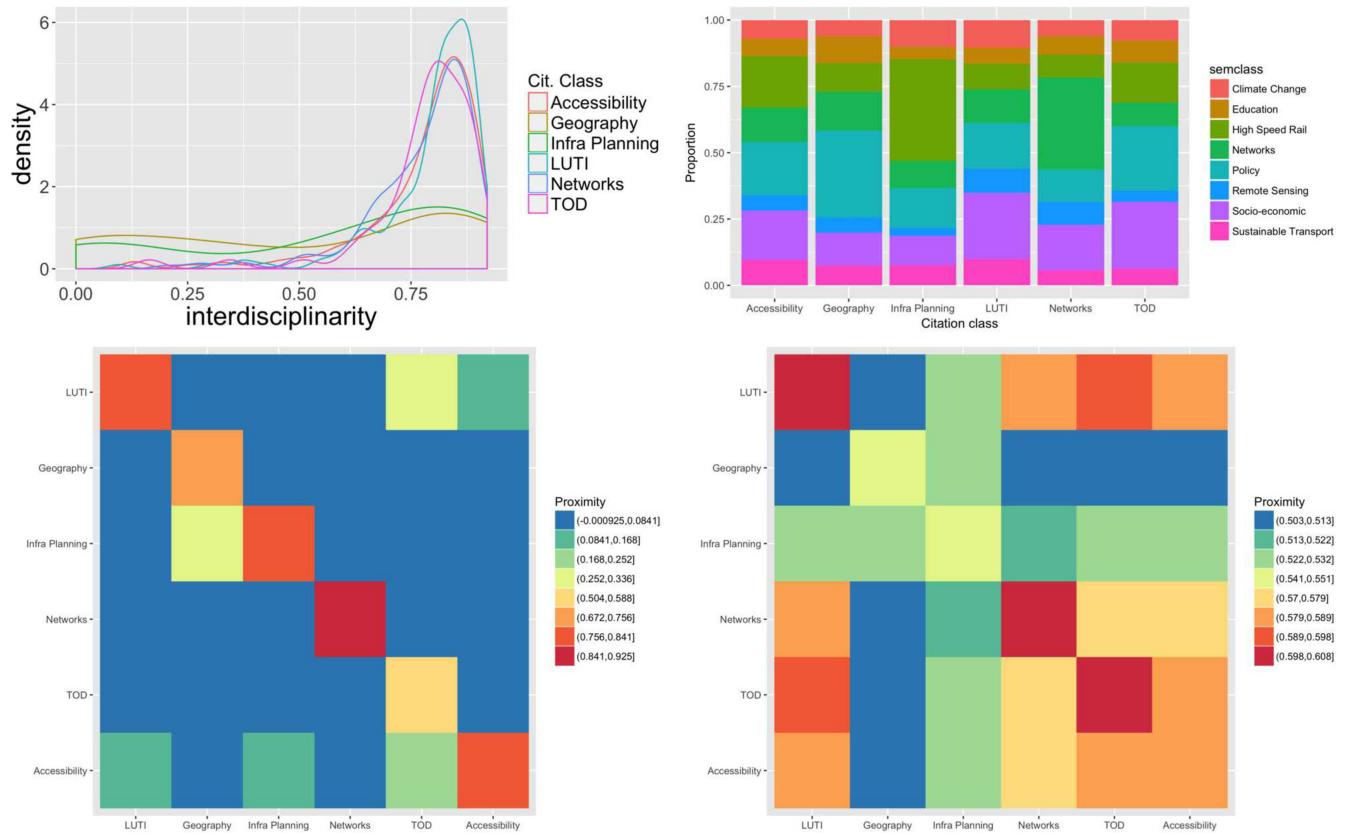


Figure 9:

Discussion

Donnons brièvement des directions d'extension de l'analyse que nous venons de mener ainsi que des implications pour le positionnement épistémologique de notre travail.

Towards modeling purpose and context automatic extraction

A possible direction to strengthen our quantitative epistemological analysis would be to work on full textes related to the modeling of interaction between networks and territories, with the aim to automatically extract thematics within articles. The idea would be to perform some kind of automatized modelography, with possible features to be extracted that would be ontologies, model architecture or structures, scales, or even typical parameter values. It is not clear to what degree structure of models can be extracted from their description in papers and it surely depends on the discipline considered. For example in a framed field such as transportation planning, using a pre-defined ontology (in the sense of dictionary) and a fuzzy grammar could be efficient to extract information as the discipline is relatively formatted. In theoretical and quantitative geography, beyond the barrier of language, information organisation is surely less subject to unsupervised data-mining because of the more literary nature of the discipline : synonyms and figures of speech are generally the norm in good level human sciences writing, fuzzing a possible generic structure of knowledge description.

Reflexivity

The methodology developed here is particularly interesting since it is reflexive, i.e. it can be used on our work itself. Therefore, an other application will be the reflexivity of our thesis : we attend to proceed to similar analysis on our proper bibliography (and possibly its evolution, available via git history), to understand our patterns of knowledge, possible gaps or unveil unexpected developments. The detailed development is done in Appendix F.

* * *

*

Cette section nous a ainsi permis de dresser un paysage des disciplines en relation avec notre problématique, et des relations entre ces disciplines, en termes de citations mais aussi de niveau d'interdisciplinarité.

La section suivante se positionnera dans une démarche voisine, mais avec un but plus marqué d'exhaustivité en termes de modéli-

sation des interactions : nous procéderons ainsi à une revue systématique et à une modélographie, afin de renforcer la typologie des modèles obtenue en section 2.1.

★ ★

★

2.3 SYSTEMATIC REVIEW AND MODELOGRAPHY

Tandis que les études menées précédemment proposaient de construire un horizon global de l'organisation des disciplines s'intéressant à notre question, nous proposons à présent une étude plus ciblée des caractéristiques de modèles existants. Nous proposons pour cela dans un premier temps une revue systématique, c'est-à-dire la construction d'un corpus plus précis répondant à certaines contraintes, suivie d'une métá-analyse, c'est-à-dire une tentative d'explication de certaines caractéristiques des modèles par des modèles statistiques.

2.3.1 *Systematic Review and Meta-analysis*

Les revues systématiques classiques ont majoritairement lieu dans des domaines où une recherche très ciblée, même par titre d'article, fournira un certain nombre d'études étudiant quasiment la même question : typiquement en évaluation thérapeutique, où des études standardisées d'une même molécule varient uniquement par taille des effectifs et modalités statistiques (groupe de contrôle, placebo, niveau d'aveugle). Dans ce cas la construction du corpus est d'une part aisée par l'existence de bases spécialisées permettant des recherches très ciblées, et d'autre part par la possibilité de procéder à des analyses statistiques supplémentaires pour croiser les différentes études (par exemple métá-analyse par réseau, voir [Rucker, 2012]). Dans notre cas, l'exercice est bien plus aléatoire pour les raisons exposées dans les deux sections précédentes : les objets sont hybrides, les problématiques diverses, et les disciplines variées. Les différents points soulevés par la suite auront souvent autant de valeur thématique que de valeur méthodologique, suggérant des points cruciaux lors de la réalisation d'une telle revue systématique hybride.

Nous proposons une méthodologie hybride couplant les deux méthodologies développées précédemment avec une procédure plus classique de revue systématique. Nous souhaitons à la fois une représentativité de l'ensemble des disciplines que l'on a découvertes, mais aussi un bruit limité dans les références prises en compte pour la modélographie. Pour cela, nous combinons le corpus obtenu précédemment et un corpus constitué par requêtes de mots-clés, de manière similaire à [Tahamtan and Bornmann, 2018]. Le protocole est donc le suivant :

1. Partant du corpus de citation isolé en 2.2.3, nous isolons un nombre de mots-clés pertinents, en sélectionnant les 5% de liens ayant le plus fort poids (seuil arbitraire), puis parmi les noeuds correspondants ceux ayant un degré supérieur au quantile à 0.8 de leur classe sémantique respective. Le premier filtrage permet de se concentrer sur le "coeur" des disciplines observées, et le second de ne pas biaiser par la taille sans perdre la structure globale, les classes étant relativement équilibrées. Un exa-

men manuel permet de supprimer les mots-clés clairement non-pertinents (télédétection, tourisme, réseaux sociaux, ...), ce qui conduit à un corpus de $K = 115$ mots-clés (K est endogène ici).

2. Pour chaque mot-clé, nous effectuons automatiquement une requête au catalogue (scholar) en y ajoutant `model*`, d'un nombre fixé $n = 20$ de références. L'ajout du terme est nécessaire pour obtenir des références pertinentes, après test sur des échantillons.
3. Le corpus potentiel composé des références obtenues, ainsi que des références composant le réseaux de citation, est revu manuellement (passage en revue des titres) pour assurer une pertinence au regard de l'état de l'art de 2.1, fournissant le corpus préliminaire de taille $N_p = 297$.
4. Ce corpus est alors inspecté pour les résumés et textes complets si nécessaire. On sélectionne les articles mettant en place une démarche de modélisation, hors modèles conceptuels. Les références sont classifiées et caractérisées selon des critères décrits ci-dessous. Nous obtenons alors un corpus final de taille $N_f = 145$, sur lequel des analyses quantitatives sont possibles.

La méthode est résumée en Fig. 10, avec les valeurs des paramètres et la taille des corpus successifs. Cet exercice permet tout d'abord un certain nombre de points méthodologiques, dont la connaissance pourra être un atout pour mener des revues systématiques hybrides similaires :

- Les biais de catalogue semblent inévitables. Nous reposons sur l'hypothèse que l'utilisation de Scholar permet un échantillonnage uniforme au regard des erreurs ou biais de catalogage. Le développement futur d'outils ouverts de catalogage et de cartographie, permettant un effort contributif pour une connaissance plus précise de domaines étendus et de leurs interfaces, sera un enjeu crucial de la fiabilité de ce genre de méthodes (voir B.6).
- La disponibilité des textes complets est particulièrement un problème pour une revue si large, vu la multiplicité des éditeurs. L'existence de moyens d'émancipation de la science ouverte comme Sci-hub¹⁸ permet d'effectivement accéder à l'ensemble des textes. En écho au débat sur le bras de fer récent avec les éditeurs concernant l'exclusivité de la fouille de textes complets, il parait de plus en plus évident qu'une science ouverte réflexive est totalement antagoniste au modèle actuel de l'édition. Nous espérons également une évolution rapide des pratiques sur ce point.

¹⁸ <http://sci-hub.cc/>

- Les revues, et en fait les éditeurs, semblent influencer différemment le référencement, augmentant potentiellement le biais de requête. La littérature grise ainsi que les preprints sont pris en compte différemment selon les champs.
- Le passage en revue manuel des grands corpora permet de pas louper des “poids lourds” qui auraient pu être omis en amont [Lissack, 2013]. La question de la mesure dans laquelle on peut s’attendre d’être au courant de la manière la plus exhaustive des découvertes récentes liées au sujet étudié évolue très probablement vu l’augmentation de la quantité totale de littérature produite et la fragmentation des domaines pour certains toujours plus pointus [Bastian, Glasziou, and Chalmers, 2010]. Rejoignant les points précédents, on peut supposer que des outils d’aide à l’analyse systématique permettront de garder cet objectif raisonnable.
- Les résultats de la revue automatique sont sensiblement différents des domaines dessinés dans la revue classique : certaines associations conceptuelles, notamment l’inclusion des modèles de croissance de réseaux, ne sont pas naturelles et existent peu dans le paysage scientifique comme nous l’avons montré précédemment.

D’autre part, l’opération de construction du corpus permet déjà de tirer des observations thématiques intéressantes en elles-mêmes.

- Les articles sélectionnés supposent une clarification de ce qui est entendu par “modèle”. Nous donnons en 8.3 une définition très large s’appliquant à l’ensemble des perspectives scientifiques. Notre selection ici ne retient pas les modèles conceptuels par exemple, notre critère de choix étant que le modèle doit inclure un aspect numérique ou de simulation.
- Un certain nombre de références consistent en des revues, ce qui revient à un groupe de modèles ayant des caractéristiques similaires. Nous pourrions compliquer la méthode en retranscrivant chaque revue ou meta-analyse, ou en pondérant par le nombre d’article correspondant les enregistrements des caractéristiques correspondants. Nous faisons le choix d’ignorer ces revues, ce qui reste cohérent de manière thématique en restant dans l’hypothèse d’échantillonnage uniforme.
- Une première clarification du cadre thématique est opérée, puisque nous ne sélectionnons pas les études liées uniquement au trafic et à la mobilité (ce choix étant aussi lié aux résultats obtenus en ??), à l’urban design pur, au modèles de flux piétons, au fret, à l’écologie, aux aspects techniques du transport, pour donner

quelques exemples, même si ces sujets peuvent dans une vue extrême être considérés comme liés aux interactions entre réseaux et territoires.

- De la même façon, des domaines annexes comme le tourisme, les aspects sociaux de l'accès aux transports, l'anthropologie, n'ont pas été pris en compte.
- On observe une forte fréquence des études liées au Trains à Grande Vitesse (HSR), rappelant la non-dissociabilité des aspects politiques de la planification et des directions de recherche en transports.

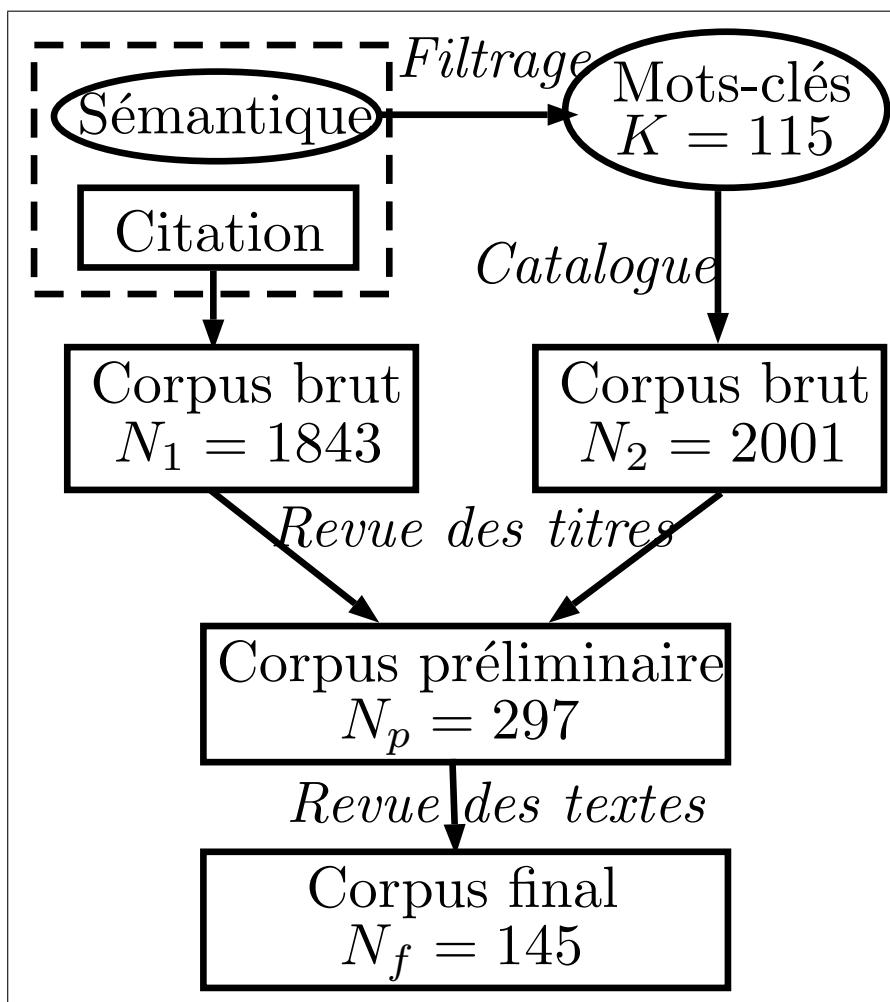


Figure 10:

2.3.2 Modelography

Nous passons à présent à une analyse mixte basée sur ce corpus, inspirée par les résultats des sections précédentes notamment.

ment pour la classification. Elle a pour but d'extraire et de décomposer précisément les ontologies, échelles et processus, puis d'étudier des liens possibles entre ces caractéristiques des modèles et le contexte dans lequel ils ont été introduits. Il s'agit ainsi de la métanalyse en quelque sorte, que nous désignerons ici par modélographie. Pour ne pas froisser les puristes, il ne s'agit en effet pas d'une métanalyse à proprement parler car nous ne combinons pas des analyses proches pour extrapoler des résultats potentiels d'échantillons plus grands. Notre démarche est proche de celle de [Cottineau, 2017] qui rassemble les références ayant étudié quantitativement la loi de Zipf pour les villes, puis lie les caractéristiques des études aux méthodes utilisées et hypothèses formulées.

La première partie consiste en l'extraction des caractéristiques des modèles. Automatiser ce travail constituerait un projet de recherche en lui-même, comme nous développons en discussion ci-dessous, mais nous sommes convaincu de la pertinence d'affiner de telles techniques (voir 8.3.3) dans le cadre d'un développement de disciplines intégrées. Le temps étant autant l'ennemi que l'allié de la recherche, nous nous concentrerons ici sur une extraction manuelle qui se voudra plus fine qu'une tentative peu convaincante de fouille de données. Nous extrayons des modèles les caractéristiques suivantes :

- quelle est la force du couplage¹⁹ entre les ontologies territoriales et celles du réseau, autrement dit s'agit-il d'un modèle de coévolution. Nous classerons pour cela en catégories suivant la représentation de la figure 11 : {territory ; network ; weak ; coevolution}, qui résulte de l'analyse de la littérature en 2.1 ;
- échelle de temps maximale ;
- échelle d'espace maximale ;
- hypothèses d'équilibre ;
- domaine “*a priori*”, déterminé par l'origine des auteurs et domaine de la revue ;
- méthodologie utilisée (modèles statistiques, système d'équations, multi-agent, automate cellulaire, recherche opérationnelle, simulation etc.) ;
- cas d'étude (ville, métropole, région ou pays) s'il y a lieu.

¹⁹ Il n'existe à notre connaissance pas d'approche générique du couplage des modèles n'étant pas liée à un formalisme particulier. Nous prendrons l'approche donnée en introduction, en distinguant ici un couplage faible comme un couplage séquentiel (sorties du premier modèle deviennent entrées du second) d'un couplage fort dynamique où l'évolution est interdépendante à chaque instant (soit par une détermination réciproque soit par une ontologie commune).

Nous collectons également de manière indicative, mais sans objectif d'objectivité ni d'exhaustivité, le "sujet" de l'étude (c'est-à-dire la question thématique dominante) ainsi que les "processus" inclus dans le modèle. Une extraction exacte des processus reste hypothétique, d'une part conditionnée à une définition rigoureuse et prenant en compte différents niveaux d'abstraction, de complexité, ou d'échelle, d'autre part dépendant de moyens techniques hors de portée de cette étude modeste. Nous commenterons ceux-ci de manière indicative sans les inclure dans les études systématiques.

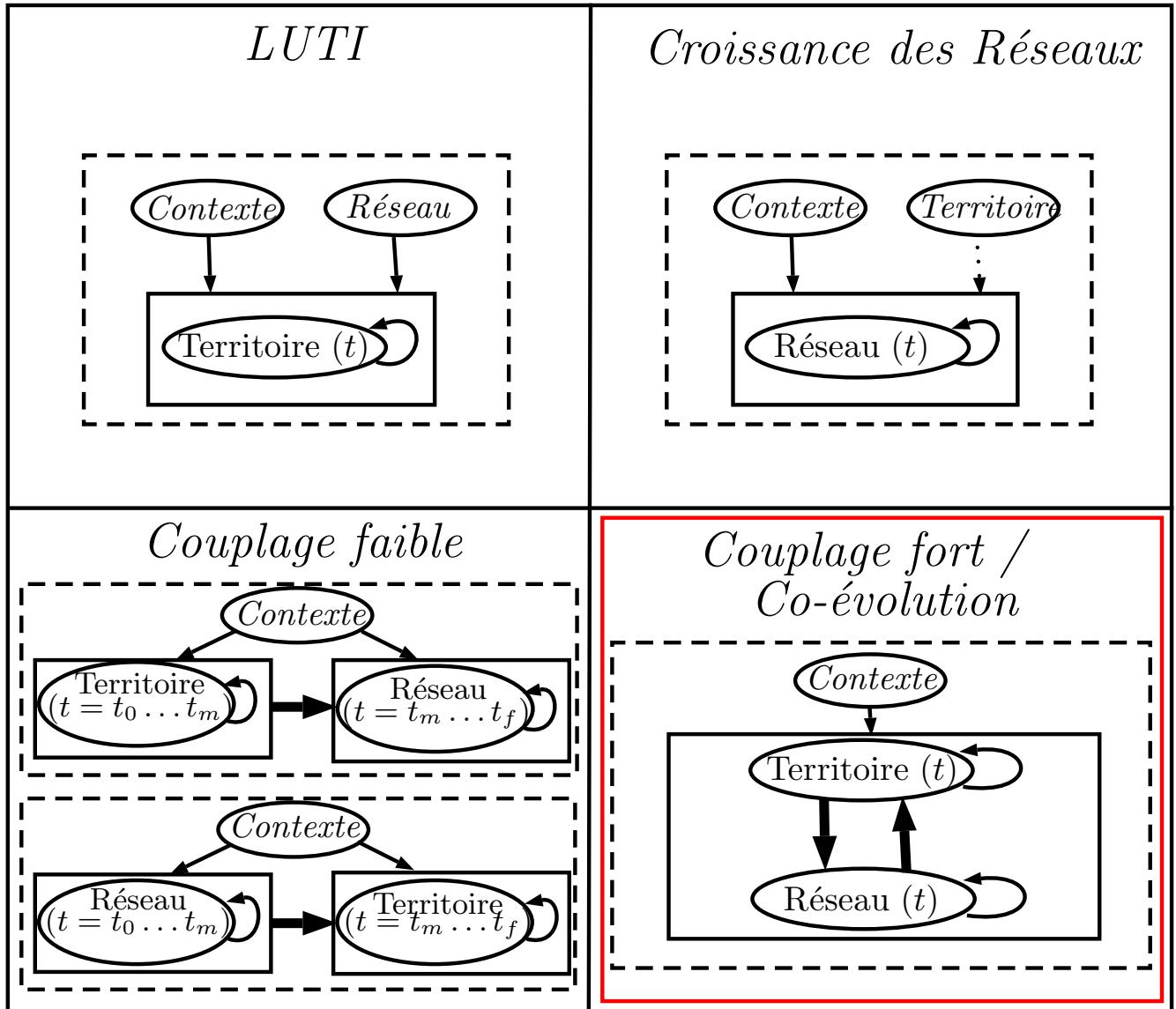


Figure 11: Schematic representation of the distinction between different types of models coupling networks and territories

Nous confondons également échelle, portée et dans un sens résolution pour ne pas rendre plus confus l'extraction. Même s'il serait pertinent de différencier lorsque un élément n'a pas lieu d'être pour un modèle (NA) de lorsque celui-ci est mal défini par son auteur, cette tâche apparaît sujette à subjectivité et nous fusionnons les deux modalités. Nous ajoutons aux caractéristiques ci-dessus les variables suivantes :

- domaine de citation (le cas échéant, c'est-à-dire pour les références initialement présentes dans le réseau de citation, i.e. 55% des références) ;
- domaine sémantique, défini par le domaine pour lequel le document a la plus grande probabilité ;
- indice d'interdisciplinarité.

Les domaines sémantiques et la mesure d'interdisciplinarité ont été recalculés pour ce corpus par collecte des mots-clés, puis extraction selon la méthode décrite en 2.2, avec $K_W = 1000$, $\theta_w = 15$ et $k_{max} = 500$. On obtient des communautés plus ciblées et plutôt représentatives de la thématique et des méthodes : Transit-oriented development (tod), Hedonic models (hedonic), Planification des infrastructures (infra planning), High-speed rail (hsr), Réseaux (networks), Réseaux complexes (complex networks), Bus rapid transit (brt).

Un "bon choix" de caractéristiques pour classer les modèles est un peu le problème du choix des *features* en apprentissage statistique : si on est en supervisé, c'est-à-dire qu'on veut obtenir une bonne prédiction de classe fixée a priori (ou une bonne modularité de la classification obtenue par rapport à la classification fixée), on pourra sélectionner les caractéristiques optimisant cette prédiction. On discriminera ainsi les modèles que l'on connaît et que l'on juge différents. Si l'on veut extraire une structure endogène sans a priori (classification non supervisée), la question est différente. Nous testerons pour cela en second temps une technique de regression qui permet d'éviter l'overfitting et faire de la selection de caractéristiques (forêts aléatoires).

Processes and Case studies

Concernant l'existence d'un cas d'étude et sa localisation, 26% des études n'en présentent pas, correspondant à un modèle abstrait ou modèle jouet (la quasi totalité des études en physique tombant dans ce cas). Ensuite, elles sont réparties à travers le monde, avec toutefois une surreprésentation des Pays-bas avec 6.9%. Les processus inclus sont trop variés (en fait autant que les ontologies des disciplines concernées) pour faire l'objet d'une typologie, mais nous noterons la domination de la notion d'accessibilité (65% des études), puis des

processus très variés allant de processus de marché immobilier pour les études hédoniques, aux relocalisations d'actifs et d'emplois pour les Luti, ou aux investissements d'infrastructure de réseau. Nous observons des processus abstraits géométriques de croissance de réseau, correspondant aux travaux des physiciens. La maintenance du réseau apparaît dans une étude, ainsi que l'histoire politique. Les processus abstraits d'agglomération et dispersion sont aussi le cœur de quelques études. Les interactions entre villes sont minoritaires, les approches de type systèmes de villes étant noyées dans les études d'accessibilité. Les questions de gouvernance et de régulation ressortent aussi, plutôt dans le cas de planification d'infrastructure et de modèle d'évaluation de démarches TOD, mais sont aussi minoritaires. On retiendra que chaque domaine puis chaque étude introduit ses propres processus quasi-spécifiques à chaque cas.

Corpus Characteristics

Les domaines "a priori" (i.e. jugés, ou plutôt préjugés sur la revue ou l'appartenance des auteurs), sont relativement équilibrés pour les disciplines majoritaires déjà identifiées : 17.9% Transportation, 20.0% Planning, 30.3% Economics, 19.3% Geography, 8.3% physics, le reste minoritaire se répartissant entre environnement, informatique, ingénierie et biologie. Concernant les poids des domaines sémantiques significatifs, le TOD domine avec 27.6% des documents, suivi par les réseaux (20.7%), les modèles hédoniques (11.0%), la planification des infrastructures (5.5%) et le HSR (2.8%). Les tables de contingences montrent que le Planning ne fait quasiment que du TOD, la physique uniquement des réseaux, la géographie se répartit équitablement entre réseaux et TOD (le second correspondant aux articles typés "aménagement", qui ont été classés en géographie car dans des revues de géographie) ainsi qu'une plus faible part en HSR, enfin l'économie est la plus variée entre hédonique, planning, réseaux et TOD. Cette interdisciplinarité n'apparaît cependant que pour les classes extraites pour la probabilité majoritaire, puisque les indices d'interdisciplinarité moyens par discipline ont des valeurs équivalentes (de 0.62 à 0.65), hormis la physique significativement plus basse à 0.56 ce qui confirme son statut de "nouveau venu" ayant une profondeur thématique plus faible.

Studied models

Il est intéressant pour notre question de répondre à la question "qui fait quoi?", c'est-à-dire quelles types de modèles sont mobilisés par les différentes disciplines. Nous donnons en Table 8 la table de contingence du type de modèle en fonction des disciplines a priori, de la classe de citation et de la classe sémantique. On constate les approches fortement couplées, les plus proches de ce qu'on considère

Table 8: Model types.

Discipline	economics	geography	physics	planning	transportation
network	5	3	12	1	4
strong	4	3	0	0	2
territory	35	22	0	28	20
Semantic	hedonic	hsr	infra ning	plan- ning	networks tod
network	1	0	0	14	2
strong	0	0	0	5	1
territory	15	4	8	11	37
Citation	accessibility	geography	infra ning	Plan- ning	LUTI networks TOD
network	0	0	0	0	24 0
strong	0	0	0	2	5 0
territory	13	1	6	18	2 3

comme des modèles de co-évolution, sont majoritairement contenues dans le vocabulaire des réseaux, ce qui est confirmé par leur positionnement en terme de citation, mais que les disciplines concernées sont variées. La majorité des études s'intéresse au territoire uniquement, le déséquilibre le plus fort étant pour les études sémantiquement liées au TOD et à l'hédonique. La physique est encore limitée en s'intéressant exclusivement aux réseaux.

Studied scales

Pour répondre ensuite à la question du comment, on peut regarder les échelles de temps et d'espace typiques des modèles. La planification et les transports se concentrent à des petites échelles spatiales, métropolitain ou local, l'économie également avec une forte représentation du local via les études hédoniques, et une étendue un peu plus grande avec l'existence d'études au niveau régional et quelques une du pays (études de panel généralement). Encore une fois, la physique se retrouve limitée avec l'ensemble de ses contributions à une échelle fixe, métropolitaine (pas forcément claire ni bien spécifiée dans les articles d'ailleurs puisqu'il s'agit de modèles jouets dont les contours thématiques peuvent être très flous). La géographie est relativement bien équilibrée, de l'échelle métropolitaine à l'échelle continentale. Le schéma pour les échelles de temps est globalement similaire. Les méthodes utilisées sont fortement corrélées à la discipline : un test du χ^2 donne une statistique de 169, très significatif

avec $p = 0.04$. De même, l'échelle d'espace l'est mais de manière moindre ($\chi^2 = 50$, $p = 0.08$).

Classical Regressions

Nous étudions à présent l'influence de divers facteurs sur les caractéristiques des modèles par des régressions linéaires simples. Dans une démarche de multi-modélisation, nous proposons de tester l'ensemble des modèles possibles pour expliquer chacune des variables à partir des autres. Le nombre d'observations pour lesquelles toutes les variables sont renseignées est très faible, il s'agit de prendre en compte le nombre d'observations utilisées pour ajuster chaque modèle. D'autre part, les performances du modèle peuvent être caractérisées par des objectifs complémentaires. Suivant [Igel, 2005], nous appliquons une optimisation multi-objectif, pour maximiser simultanément la variance expliquée (R^2 ajusté dans notre cas) et l'information capturée (Critère d'information d'Akaike corrigé AICc²⁰). Celle-ci est effectuée conditionnellement au fait d'avoir le nombre d'observations $N > 50$ (seuil fixé au regard de la distribution de N sur l'ensemble des modèles). La procédure d'optimisation est détaillée en Annexe A.2 pour chaque variable. L'échelle de temps et l'interdisciplinarité présentent des compromis difficiles à départager, et nous ajustons les deux candidats. Les autres variables présentent des solutions dominantes et nous n'ajustons qu'un seul modèle.

Les résultats complets des régressions sont donnés en Table ???. Les échelles temporelle et d'espace, ainsi que l'année, sont les variables les mieux expliquées au sens de la variance. L'échelle de temps est influencée très significativement par le type de modèle : territoire qui diminue celle-ci, ou couplage fort qui l'augmente. Le fait d'être en physique influe également significativement, et élargit la portée temporelle des modèles. Au contraire, les approches d'ingénierie (souvent design optimal d'un réseau de transport) correspondent à une courte durée.

Pour l'échelle d'espace, le fait d'être en géographie a une forte influence sur la portée spatiale des modèles : en effet, les études régionales et à l'échelle du système de villes sont bien l'apanage de la géographie. L'appartenance au domaine du transport augmente aussi faiblement la portée spatiale (voir significativité dans les regressions complètes en Annexe A.2). Aucune autre variable n'a une influence significative.

Le niveau d'interdisciplinarité est bien expliqué par l'année, qui l'influence de manière négative, ce qui confirme une augmentation des spécialisations scientifiques dans le temps. Les études économétriques

²⁰ L'AIC est une mesure du gain d'information entre deux modèles, et permet d'éviter l'ajustement abusif par un nombre trop grand de paramètres. L'AICc est une version prenant en compte la taille de l'échantillon, la mesure variant significativement pour les petits échantillons.

des modèles hédoniques apparaissent très spécialisées. Enfin, l'année de publication est expliquée significativement et positivement par le type territoire et par le fait d'être en transports, ce qui signifierait une recrudescence récente d'un profil particulier d'études. Un examen du corpus suggère qu'il s'agit des études sur la grande vitesse, apparaissant comme une mode scientifique récente.

Random Forest Regressions

Nous concluons cette étude par des régressions et classification par forêts aléatoires, qui sont une méthode très flexible permettant de dégager une structure d'un jeu de données [Liaw and Wiener, 2002]. Pour compléter les analyses précédentes, nous proposons de l'utiliser pour déterminer les importances relatives des variables pour différents aspects. Nous utilisons à chaque fois des forêts de taille 100000, une taille de noeud de 1 et un nombre de variable échantillonnée en \sqrt{p} pour la classification et $p/3$ pour la régression lorsque p est le nombre total de variables. Pour classifier le type de modèle, nous comparons les effets de la discipline, de la classe sémantique et de la classe de citation. Cette dernière est la plus importante avec une mesure relative de 45%, tandis que la discipline compte pour 31% et le sémantique pour 23%. Ainsi, le cloisonnement disciplinaire se retrouve, tandis que le sémantique et donc en partie les ontologies, est le plus ouvert. Cela nous encourage dans notre démarche de sortir de ce cloisonnement. Lorsqu'on applique une regression de forêt sur l'interdisciplinarité, toujours avec ces trois variables, on constate qu'elles expliquent 7.6% de la variance totale, ce qui est relativement faible, témoignant d'une disparité de sémantique sur l'ensemble du corpus indépendamment des différentes classifications. Dans ce cas, la variable la plus importante est la discipline (39%) suivie par le sémantique (31%) et la citation (29%), ce qui confirme que le journal visé conditionne fortement le comportement de langage employé. Cela nous alerte sur le danger de perte de richesse sémantique lorsqu'on s'adresse à un public particulier. Ainsi, nous avons pu dégager certaines structures et régularités des modèles nous concernant, qui seront riches d'enseignements lors de la construction de nos modèles.

2.3.3 *Discussion*

Further Developments

Further work may consist in the production of an automatic synthesis of this meta-analysis, from a modular modeling point of view, combined with a refined purpose and scale classification. Modular modeling consists in the integration of heterogeneous processes and implementation of processes in order to extract the set of mechanisms giving the best fit to empirical data [Cottineau, Chapron, and Reuil-

Table 9: Explanation of models characteristics.

	<i>Variable expliquée:</i>					
	TEMPSCALE (1)	SPATSCALE (2)	INTERDISC (3)	INTERDISC (4)	YEAR (5)	YEAR (6)
YEAR	0.674			-0.004*	-0.002*	
TYPEstrong		100.271***			-0.026	
TYPEterritory	-38.933***	-14.988			0.044	10.898***
TEMPSCALE			-5.179	-0.0003		0.035
FMETHODeq						-6.224
FMETHODmap						4.747
FMETHODro						6.128
FMETHODsem						1.009
FMETHODsim						5.153
FMETHODstat						-0.357
DISCIPLINEengineering	-52.107*	-9.609	-154.461	0.144		13.486
DISCIPLINEenvironment	17.110	17.886	-5.878	0.092		-3.668
DISCIPLINEgeography	3.640	9.126	1,445.457***	0.036		1.121
DISCIPLINEphysics	46.879*	77.897***	292.559	-0.103		3.392
DISCIPLINEplanning	1.304	4.553	-143.554	-0.047		-2.850
DISCIPLINEtransportation	-14.718	8.753	568.329	0.062		5.503*
INTERDISC	2.357					-12.876
SEMCOMcomplex networks					-0.217	
SEMCOMhedonic				-0.179	-0.184*	-5.769
SEMCOMhsr				-0.100	-0.122	6.135
SEMCOMinfra planning				-0.032	-0.096	-4.123
SEMCOMnetworks				-0.038	-0.107	4.711
SEMCOMtod				-0.105	-0.152	-1.653
Constant	-1,305.126	22.103*	235.357	8.962**	5.531**	2,004.945***
Observations	64	94	94	64	98	64
R ²	0.385	0.393	0.100	0.314	0.155	0.510
R ² ajusté	0.282	0.336	0.027	0.136	0.068	0.281

Note:

*p<0.1; **p<0.05; ***p<0.01

lon, 2015]. We can thus classify models described here according to their building bricks in terms of processes implemented and thus identify possible coupling potentialities.

Lessons for Modeling

Nous pouvons résumer les points principaux issus de cette métanalyse qui joueront sur notre attitude et nos choix de modélisation. Tout d'abord, la présence interdisciplinaire des approches effectuant un couplage fort confirme notre besoin de faire des ponts et de coupler les approches, et confirme également rétrospectivement les conclusions de 2.2 sur les conséquences du cloisonnement des disciplines en terme de modèles formulés. Ensuite, l'importance du vocabulaire des réseaux dans une grande partie des modèles nous poussera à confirmer cet ancrage. La spécificité des approches TOD et d'accessibilité, assez proches des modèles LUTI, seront secondaires pour nous. La portée restreinte des travaux issus de la physique, confirmée par la majorité des critères étudiés, nous pousse à nous méfier de ces travaux et de l'absence de sens thématique aux modèles. La richesse des échelles temporelles et spatiales couvertes par les modèles géographiques et économiques nous confirme l'importance de varier celles-ci dans nos modèles, idéalement de parvenir à des modèles multi-échelles. Enfin, les importances relatives des variables de classification sur le type de modèle vont également dans le sens de ponts interdisciplinaires pour croiser les ontologies.

* * *

*

Table 10: Synthesis of processes included in models.

	Réseaux → Territoires	Territoires → Réseaux	Réseaux ↔ Territoires
Micro	Economie : marché immobilier, relocalisation, marché de l'emploi	NA	Informatique : croissance spontanée
	Planification : régulations, développement		
Meso	Economie : marché immobilier, coût du transport, aménités	Economie : croissance du réseau, offre et demande	Economie : investissements, relocalisations, offre et demande, planification du réseau
	Géographie : usage du sol, centralité, étalement urbain, effets de réseau	Transports : investissements, niveau de gouvernance	Géographie : usage du sol, croissance du réseau, diffusion de population
	Planification/transports : accessibilité, usage du sol, relocalisation, marché immobilier	Physique : corrélations topologiques, hiérarchie, congestion, optimisation locale, maintenance du réseau	
Macro	Economie : croissance économique, marché, usage du sol, agglomération, dispersion, compétition	Economie : interactions entre villes, investissements	Economie : offre et demande
	Géographie : accessibilité, interaction entre villes, relocalisation, histoire politique	Géographie : interactions entre villes, rupture de potentiel	Transports : couverture du réseau
	Transports : accessibilité, marché immobilier	Transports : planification de réseau	

SYNTHESIS OF MODELED PROCESSES

Nous proposons de synthétiser les processus pris en compte par les modèles parcourus lors de la modélographie, afin de procéder à un effort similaire à celui concluant l'approche thématique du chapitre 1. Nous ne pouvons ni avoir une vision exhaustive (comme déjà précisé lors de la description de la méthodologie de la modélographie) ni rendre compte avec grande précision de chaque modèle en détail, puisque quasiment chacun est unique dans son ontologie. L'exercice de synthèse permet ainsi de s'extraire de ces limites et prendre un certain recul, et avoir ainsi un aperçu sur les *processus modélisés*²¹.

La table 10 propose cette synthèse à partir des 145 articles issus de la modélographie et pour lesquels une classification de type était

²¹ En gardant en tête les choix de sélection, qui emmènent par exemple à ne pas avoir les processus de mobilité dans cette synthèse.

possible, c'est-à-dire qu'il existait un modèle rentrant dans la typologie développée en 2.1. Être complètement exhaustif relèverait d'une opération de métamodélisation interdisciplinaire qui est bien hors de la portée de notre travail²², et la liste donnée ici est indicative.

Nous retrouvons les correspondances entre disciplines, échelles et types de modèles obtenues dans la modélographie en 2.3. Nous retirons les enseignements principaux suivants, en écho au tableau de synthèse obtenu en fin du Chapitre 1 (Table 3) :

1. La dichotomie des ontologies et des processus pris en compte entre les échelles et entre les types est autant manifeste ici dans les modèles que dans les processus en eux-même²³. Nous postulons qu'il existe bien des processus différents aux différentes échelles, et nous prendrons le parti d'étudier différentes échelles.
2. Le cloisonnement des disciplines démontré en 2.2 se retrouve qualitativement dans cette synthèse : il est évident qu'elles divergent originellement dans leurs différentes épistémologies fondatrices. Nous tacherons d'intégrer des paradigmes de différentes disciplines, tout en prenant en compte les limites imposées par les principes de modélisation que nous présenterons en 3.1 (par exemple, la parcimonie des modèles limite nécessairement l'intégration d'ontologies hétérogènes).
3. Un décalage important entre cette synthèse et celle des processus est la quasi absence ici de modèles intégrant des processus de gouvernance. Il s'agira d'une piste à explorer.
4. Au contraire, une très bonne correspondance s'établit entre les modèles géographiques des systèmes urbains et les positionnements théoriques de la théorie évolutive des villes. Cette adéquation, plus difficile à retrouver pour l'ensemble des autres approches revues, nous suggère également de suivre cette piste.

²² Il s'agirait pour cela d'avoir des correspondances entre les ontologies, sans lesquelles on se retrouverait avec au moins autant de processus que de modèles, même au sein d'une discipline. Il n'en existe à notre connaissance pas entre deux disciplines seulement. Une piste pour une approche formelle est donnée en B.5.

²³ Puisqu'on a plus détaillé cette étude, elle paraît même plus forte aussi, une plus grande précision permettant alors de séparer des catégories abstraites.

CHAPTER CONCLUSION

Les processus que nous cherchons à modéliser étant multi-scalaires, hybrides et hétérogènes, les angles d'approches et questionnements possibles sont nécessairement extrêmement variés, complémentaires et riches. Il pourrait s'agir d'une caractéristique fondamentale des systèmes socio-techniques, que PUMAIN formule dans [Pumain, 2005] comme "une nouvelle mesure de complexité", qui serait liée aux nombre de point de vue nécessaires pour appréhender un système à un niveau donné d'exhaustivité. Cette idée rejoint la position de *perspectivisme appliqué* que B.5 formalise et qui est implicitement présente dans l'investigation des relations entre économie et géographie développée en C.6. Ainsi, la modélisation des interactions entre réseaux et territoires peut être reliées à un ensemble très large de disciplines et d'approches revues en section 2.1.

Afin de mieux comprendre le paysage scientifique environnant, et quantifier les rôles ou poids relatifs de chacune, nous avons procédé à une série d'analyse en épistémologie quantitative en 2.2. Une première analyse préliminaire basée sur une revue systématique algorithmique suggère un certain cloisonnement des domaines. Cette conclusion est confirmée par l'analyse d'hyperréseau couplant réseau de citations et réseau sémantique, qui permet également de dessiner plus finement les contours disciplinaires, à la fois sur leur relations directes (citations) mais aussi leur proximité scientifique pour les termes et méthodes utilisées. On peut alors utiliser le corpus constitué et cette connaissance des domaines pour une revue systématique semi-automatique en 2.3, qui permet de constituer un corpus de travaux traitant directement du sujet, qui est ensuite inspecté intégralement, permettant de lier caractéristique des modèles au différents domaines. Nous avons alors à ce stade une idée assez précise de ce qui se fait, pourquoi et comment.

L'enjeu reste de déterminer les pertinences relatives de certaines approches ou ontologies, ce qui sera le but des deux chapitres de la deuxième partie. Nous concluons d'abord cette première partie par un chapitre de discussion 3, éclairant des points nécessaires à clarifier avant une entrée dans le vif du sujet.



3

POSITIONING

Toute activité de recherche serait, selon certains acteurs de celle-ci, nécessairement politisée, de par pour commencer le choix de ses objets. Ainsi, RIPOLL alerte contre l'illusion d'une recherche objective et les dangers de la technocratie [Ripoll, 2017]. Nous ne rentrerons pas dans ces débats bien trop vastes pour être traités même en un chapitre, puisqu'il rejoignent des thèmes de sciences politiques, d'éthique, de philosophie, liés par exemple à la gouvernance scientifique, à l'insertion de la science dans la société, à la responsabilité scientifique.

Il est clair que même des sujets a priori intrinsèquement objectifs, comme la physique des particules et des hautes énergies, ont des implications regardant d'une part les choix de leur financements et les externalités associées (par exemple, l'existence du CERN a largement contribué au développement du calcul distribué), mais d'autre part aussi les applications potentielles des découvertes qui peuvent avoir des répercussions sociales considérables. En biologie, l'éthique est au cœur des principes fondateurs des disciplines, comme en témoignent les débats soulevés par l'émergence de la biologie synthétique [Gutmann, 2011]. Les tenants d'approche prudentes dans celle-ci se recoupent avec la biologie intégrative, or les sciences intégratives défendues par PAUL BOURGINE, mises en oeuvre par l'intermédiaire du campus digital Unesco CS-DC¹, ont typiquement la responsabilité sociale et l'implication citoyenne au cœur de leur cercle vertueux. En sciences humaines et sociales, comme les recherches interagissent avec les objets étudiés (en quelque sorte l'idée des *interactive kind* de HACKING [Hacking, 1999]), les implications politiques et sociales de la recherche sont bien évidemment indiscutables.

Nous nous placerons ici à un niveau épistémologique, c'est-à-dire à des réflexions sur la nature et le contenu des connaissances scientifiques au sens large, c'est-à-dire co-construites et validées au sein d'une communauté imposant certains critères de scientificité [Morin, 1991], bien sûr évolutifs puisque nous nous positionnerons pour la systématisation de certains. Mais donc, même en restant à ce niveau, des prises de positions sont nécessaires, celles-ci pouvant être épistémologiques, méthodologiques, thématiques. Ces dernières ont déjà été ébauchées dans les deux chapitres précédents par les choix des objets d'étude, des problématiques, et seront renforcées à mesure de la progression.

¹ <https://www.cs-dc.org/>

Nous proposons ainsi ici un exercice relativement original mais que nous jugeons nécessaire pour une lecture plus fluide de la suite. Il consiste en le développement précis de certains positionnements qui ont une influence particulière dans notre démarche de recherche.

Dans une première section (3.1), nous précisons notre position au regard des modèles de simulation. Après avoir détaillé les fonctions que nous prêterons aux modèles, nous argumentons sous forme d'essai pour un usage raisonné des données massives et du calcul intensif, et illustrons notre positionnement par rapport à l'exploration des modèles par une étude de cas méthodologique pour l'exploration de la sensibilité des modèles aux conditions initiales.

Dans une deuxième section (3.2), nous développons des exemples pour illustrer le besoin et la difficulté de reproductibilité, ainsi que les liens avec des nouveaux outils pouvant la favoriser mais aussi la mettre en danger. Nous illustrons la question d'ouverture des données et d'exploration interactive par une étude de cas empirique des flux de trafic en Ile-de-France.

Enfin, la dernière section (3.3) explicite modestement des positions épistémologiques, notamment concernant le courant dans lequel nous nous plaçons, la complexité des objets en sciences sociales, et la nature de la complexité de manière générale.

Le lecteur très familier avec les "commandements" de BANOS [Banos, 2013] pourra trouver dans les deux premières sections des illustrations pratiques originales de ceux-ci, notre positionnement étant principalement dans leur lignée.

* * *

*

This chapter is composed of various works. The first section is novel for its two first parts, and for its last part describes ideas presented as [Cottineau et al., 2017]. The second section relates in its first part the theoretical content of [Rimbault, 2016a], and corresponds to [Rimbault, 2017b] for the empirical illustration. The third section follows for its first part the epistemological foundations of [Rimbault, 2017e] which were then depthen by [Rimbault, 2017c], follows a part of [Rimbault, 2018] for its second part and uses [Rimbault, 2018] for its last part.

3.1 MODELING, BIG DATA AND INTENSIVE COMPUTING

We now develop our positioning regarding issues linked to the use of modeling, of massive data and of intensive computing, what also induces by extension some comments on model exploration methods. It is not evident to what extent these new possibilities are necessarily accompanied of deep epistemological mutations, and we show on the contrary that their use necessitates more than ever a dialog with theory. Implicitly, this position foreshadows the epistemological frame for the study of complex systems of which we give the context in 3.3 and that we formalize in opening 8.3.

The points developed here cover some crucial issues linked to modeling enterprises, and can be of an epistemological, theoretical or practical nature. We will first try to answer the question of why modeling. We will then give our position on more technical issues linked to the use of emerging computing resources and new data. Finally, the last point is methodological, and both illustrates the first two points and introduces a new method to explore models.

3.1.1 Why modeling ?

We first develop the role of modeling in our process of scientific production. Models have in appearance diverse roles depending on disciplines: a model in physics results of a theory, allows to confront it with experiments and has to be validated through its predictive powers with strong requirements, whereas in computational social science one often settles for the reproduction of general stylized facts. A statistical model will be composed of assumptions on relations between variables and on the statistical distribution of an error term, and values of coefficients obtained will be interpreted even if the goodness-of-fit measure is very low. The aim here is thus to precise in which spirit our modeling approaches will be placed², what are their mechanisms and objectives.

Functions of models

As we just saw, the term of *model* has multiple meanings, and implies different realities, practices, uses (we can assume a proper ontology to models which become real objects, at least when they are implemented). A way to propose a sort of typology for models is to proceed to a typology of their functions, as does [Varenne, 2017], based on the study of diverse disciplines (biology, geography, social sciences). This classification is to the best of our knowledge the most

² If this work may appear as redundant, laborious and superficial to someone used to geosimulation models, it is crucial in our logic of disciplinary opening, in order first to avoid any misunderstanding on the status of results, and secondly to foster a dialog in the case of very different uses of models.

exhaustive existing. VARENNE thus distinguishes five broad classes of model functions³, which are in an increasing order according to their integration to a social practice :

1. Function of perception and observation: make accessible an object which can not be observed through perception (physical model of a molecule), allow experiments, a memorization, the reading and visualization of data.
2. Function of intelligibility: description of patterns, precision of ontologies, conception through prediction, explanation and comprehension of processes⁴.
3. Function of assistance to theorization: formulation, interpretation, illustration of a theory, internal consistency test (do deductive schemes induce model simulation results that are contradictory or consistent ?), applicability, computability (in the case of numerical schemes allowing to approximate the solutions of equations), co-computability (coupling of theories and models).
4. Function of social communication: scientific communication, consultation, action with actors (*stakeholders*⁵).
5. Function of decision-making: informing decision-making, action, self-fulfilling action in an abstract system (pricing models in finance).

It is clear that each discipline will have its own relation to these different functions, that some will be privileged, and others not accessible or without relevance for the object studied or the questions asked. In physics for example, the aspects of theory validation and of the existence of predictive models with a very high precision are at the heart of the discipline; whereas entire branches of social science such as urban planning for example are focused on models for communication and decision-making. Regarding this, we must not neglect the nature of social science for economics and stay doubtful of predictive aims of some modeling experiments⁶.

³ The broad classes of functions are declined into precise classes which form 21 classes. We do not detail them here, but give a synthesis describing the broad classes.

⁴ The comprehension is more general than explanation, since it assumes a reconstruction of the system structure and a deductive use, i.e. a projection and generation of the system considered in the psychological structure considering it [Morin, 1980].

⁵ We do not develop this aspect at all, but we recall that *stakeholder workshops* are one of the structuring axis of the Medium project we described in 1.3. Even if the percolation with the axis focusing on the analysis and modeling of urban systems dynamics in which our work is situated is not explicit, they implicitly operate in exchanges between perspectives, and the cohabitation within a project foreshadows more integrated future perspectives.

⁶ Even in finance at high frequencies, at which signals would be reasonably closer to physical systems than macro-econometric series for example as witnesses the appropriation of these problems by physicists, the predictability remains questionable and in any case limited [Campbell and Thompson, 2007].

This classification of functions can be found implicitly in modeling reasons developed outside any typology by [Epstein, 2008]: he insists on refuting the preconceived idea that models would only be used for prediction, and introduces diverse reasons, among which we can find intelligibility functions (explanation, uncover dynamics, reveal complexity or simplicity), of sustaining a theory (discover new questions, highlight uncertainties, suggest analogies), of informing decision-making (real-time crisis solutions, finding optimization compromises), and of communication (educate the public, train practitioners).

Within this frame of functional classification of models, our work will mainly use the following functions:

- Descriptive models and pattern extraction: these will be the diverse empirical analyses aiming at establishing stylized facts on co-evolution processes for given case studies.
- Models with an explanation and comprehension goal: models simulating territorial dynamics that we will construct, with the objective of integrating co-evolution processes, will have the principal goal of explaining stylized facts linked to some processes (for example: variations of a given parameter corresponding to a given process explain a given stylized fact), and ideally the *comprehension* of systems⁷.
- Models to test a theory: internal validation, i.e. consistence of model behavior regarding stylized facts implied by the theory, and external, in the sense of a more or less performant reproduction of dynamics for case studies considered in the frame of a theory; or more generally to answer a precise question or assumption.

Generative modeling

The *type*⁸ of models that we will mainly use in our work is related to *generative modeling*, in the sense given by [Epstein, 2006] in its manifest

⁷ Indeed, the boundary between explanation and comprehension is fuzzy and subjective. It is possible to consider that there already exists a certain level of comprehension when a model with a certain level of internal and ontological consistence, in relation with reasonable and relatively autonomous theoretical assumptions, allows to draw conclusions on global dynamics of the system considered.

⁸ In the functional perspective, structures, contents and processes, i.e. the nature of models in themselves (what corresponds to the nature and principles of models evoked but not classified by VARENNE), are given as illustrating examples, but a given function is not restricted to a given model (although reciprocally some models are not able to fulfill some functions). There does not exist to the best of our knowledge a general typology of models by *type*, that we could then define in terms of a typology of relations with other knowledge domains (see 8.3): for example a model using a given methodology, privileging a given tool, a particular or privileged use of data, etc. In any case, existing typologies or classifications of models are associated to literature reviews and synthesis that are proper to each discipline: for example,

for *generative social sciences*. The fundamental principle is to propose to explain macroscopic regularities as emerging from interactions between microscopic entities, by simulating the evolution of the system in a generative way⁹. This paradigm can be linked to the paradigm of *Pattern Oriented Modeling* in Ecology [Grimm et al., 2005], which aims at explaining through the bottom-up production of patterns¹⁰. Agent-based models, i.e. models implying a certain number of heterogeneous agents that are relatively autonomous and simulating their interactions, are a way to achieve it.

Knowledge through Modeling

Ainsi, nos modèles seront principalement à visée de compréhension (même s'ils n'atteignent pas l'objectif et restent au niveau d'une explication). Nous procéderons dans certains cas à des calibrations fines sur données observées, mais celles-ci n'auront à aucun moment l'objectif de prédiction. Ces calibrations serviront à extrapoler des paramètres et apprendre indirectement sur les processus modélisés, et le modèle est ainsi bien un instrument de *connaissance indirecte*.

Cette connaissance des processus est permise par l'utilisation de la simulation comme un laboratoire virtuel permettant le test d'hypothèses formulées à partir d'une théorie ou issues de faits stylisés empiriques : c'est exactement ce type de paradigme que construit [Pumain and Reuillon, 2017d], qui insiste sur (i) le besoin de parcimonie dans les modèles ; (ii) le besoin de multiples modèles (multi-modélisation) ; et (iii) le rôle de l'exploration extensive des modèles, pour y parvenir sans tomber dans le piège de l'équifinalité¹¹. Ainsi, l'établissement des *Calibration Profiles* du modèle SimpopLocal [Reuillon et al., 2015] permettent d'établir des conditions nécessaires et suffisantes pour reproduire un motif donné, et donc par exemple de déclarer indirectement un processus nécessaire ou non pour produire un fait stylisé.

[Harvey, 1969] (p. 157) proposes a general typology which remains however inspired from and limited to geography. Conditions for interdisciplinary typologies are an open question, which exploration is largely out of reach of our work.

⁹ Keeping in mind that the ability to generate is of course a necessary but not sufficient component of explanation, as illustrate the debate on this subject around the works of EPSTEIN synthesized by [Rey-Coyrehourcq, 2015] (p. 154).

¹⁰ Indeed, POM aims at reproducing by the model through simulation, i.e. *generates*, patterns expected at several scales, constituting a virtual laboratory in which assumptions can be tested. Furthermore, EPSTEIN's generativity is based on similar paradigms for explanation, implying models with a progressive complexity and which allow the test of assumptions, by isolating mechanisms sufficient to reproduce macroscopic patterns.

¹¹ L'équifinalité correspond à la possibilité pour un système d'atteindre un point de son espace des phases par des trajectoires différentes, c'est-à-dire dans notre cas des motifs macroscopiques pouvant être générés par différents processus microscopiques. Ce concept était déjà formulé dans la théorie générale des systèmes [Von Bertalanffy, 1972]. Il pose problème aux notions de causalité, et remet en cause des explications de causalité "directe" au niveau macroscopique - nous y reviendrons plus particulièrement en 4.2.

Ainsi, nous prendrons ici ce parti de l'utilisation des modèles (de simulation principalement), tout en gardant à l'esprit que celui-ci ne répond que partiellement aux challenges fondamentaux de la modélisation urbaine donnés par [Perez, Banos, and Pettit, 2016], notamment la capture de la complexité et de la multidimensionalité des systèmes urbains ainsi que la possibilité de générer des scénarii futurs possibles (ce qui est différent de la prédition), mais pas la question des modèles de planification urbaine, pouvant par exemple être participatifs et impliquant les *stakeholders*¹².

How to explore a model of simulation

Afin d'éviter au maximum le "bricolage" concernant l'ensemble des étapes de la genèse d'un modèle, de sa spécification, sa conception, son utilisation à son exploration, décrit par [Kotelnikova-Weiler and Le Néchet, 2017], nous proposons de nous fixer un protocole pour la partie d'exploration des modèles. Plus généralement, il existe des protocoles généraux comme celui introduit par [Grimm et al., 2014] pour accompagner l'ensemble de la démarche de modélisation. Nous considérons l'étape d'exploration et creusons celle-ci plus en détails. Nous nous plaçons dans le cadre fixé ci-dessus d'un modèle de simulation, majoritairement à visée de compréhension.

Le protocole simplifié est issu directement de la philosophie et de la structure d'OpenMole. On peut se référer par exemple à [Reuillon, Leclaire, and Rey-Coyrehourcq, 2013] pour les principes fondamentaux, la documentation en ligne¹³ pour un aperçu global des méthodes disponibles et de leur articulation dans un cadre standard, et [Pumain and Reuillon, 2017a] pour une contextualisation des différentes méthodes. Ces travaux¹⁴ ont apporté un nombre considérable d'innovations à la fois méthodologiques, techniques, thématiques et théoriques. La philosophie d'OpenMole s'articule autour de trois axes (voir entretien avec R. REUILLOU, Annexe D.3) : le modèle comme "boîte noire" à explorer (i.e. méthodes indépendantes du modèle), utilisation de méthodes avancées d'exploration, accès transparent aux environnements de calcul intensif. Ces différentes composantes sont en interdépendance forte, et permettent un changement de paradigme dans l'utilisation des modèles de simulation : utilisation de multimodélisation, c'est-à-dire structure variable du modèle [Cottineau et al., 2015], changement de la nature des questions posées au modèle (par exemple détermination complète de l'espace faisable [Chérel,

¹² Le rôle de la visée d'application des modèles est lié à la fois à une sensibilité disciplinaire, comme le domaine des Luti [Wegener and Fürst, 2004] qui l'est bien plus que celui de la géographie théorique et quantitative, mais aussi à une sensibilité "culturelle", comme l'illustre [Batty, 2013b] qui montre une branche de la géographie anglo-saxonne plus proche des applications concrètes.

¹³ Disponible à <https://next.openmole.org/Models.html>.

¹⁴ La majorité ayant été réalisés dans le cadre interdisciplinaire de l'ERC Geodiversity.

Cottineau, and Reuillon, 2015]), tout cela permis par l'utilisation du calcul intensif [Schmitt et al., 2015].

Nous considérons un modèle de simulation comme un algorithme produisant des sorties à partir de données et de paramètres en entrée. Dans ce cadre, nous proposons dans un cas idéal l'ensemble des étapes suivantes qui devraient être nécessaire pour une utilisation robuste des modèles de simulation.

1. Identification des mécanismes principaux et des paramètres cruciaux associés, possiblement des métaparamètres (ici compris comme paramètre générant la configuration initiale du modèle), ainsi que de leur domaine thématique ; identification des indicateurs pour évaluer la performance ou le comportement du modèle.
2. Évaluation des variations stochastiques : grand nombre de répétitions pour un nombre raisonnable de paramètres, établissement du nombre de répétitions nécessaire pour atteindre un certain niveau de convergence statistique.
3. Évaluation de la sensibilité aux meta-paramètres, suivant la méthodologie innovante développée par la suite¹⁵.
4. Exploration brutale pour une première analyse de sensibilité, si possible évaluation statistique des relations entre paramètres et indicateurs de sortie.
5. Calibration, exploration algorithmique ciblée par l'utilisation d'algorithmes spécifiques (*Calibration Profile, Pattern Space Exploration*)¹⁶
6. Retours sur le modèle, extension et nouvelles briques de multimodélisation, retours sur les faits stylisés et la théorie.

Le cas échéant, certaines étapes n'ont pas lieu d'être, par exemple l'évaluation de la stochasticité dans le cas d'un modèle déterministe. De même, les étapes prendront plus ou moins d'importance selon la nature de la question posée : la calibration ne sera pas pertinente dans le cas de modèles complètement synthétiques, tandis qu'une exploration systématique d'un grand nombre de paramètres ne sera pas forcément nécessaire dans le cas d'un modèle qui a pour but d'être calibré sur des données.

¹⁵ Un exemple de cette méthodologie consistant à la génération de données synthétiques sera utilisé dans la suite de cette section ; une description formelle de la méthode est donnée en B.3 et un autre exemple d'application en C.3.

¹⁶ Nous ne pratiquerons quasiment pas ce dernier point, trouvant suffisamment de réponses à nos questions avec les points précédents.

Link between modeling and Open Science

The last methodological point which we need to emphasize is the relation between the workflow we introduce and model exploration workflows.

The ideas of multi-modeling and extensive model exploration are nothing from new as Openshaw already advocated for “model-crunching” in [Openshaw, 1983], but their effective use only begins to emerge thanks to the apparition of new methods and tools together with an explosion of computation capabilities: [Cottineau, Rey, and Reuillon, 2016] claims for a renewed approach on multi-modeling. Coupling models as we do answers to similar questions. In that stream of research, the model exploration platform OpenMole [Reuillon, Leclaire, and Rey-Coyrehourcq, 2013] allows to embed any model as a black-box, write modifiable exploration workflow using advanced methodologies such as genetic algorithms and distribute transparently the computation on large scale computation infrastructures such as clusters or computation grids. In our case, the workflow tool is a powerful way to embed both the sensitivity analysis and the meta-sensitivity analysis, and allow to couple any generator with any model in a straightforward way as soon as the model can take its spatial configuration as input or from an input file.

Synthesis

Résumons brièvement les idées à garder à l'esprit à la suite de ce survol rapide d'enjeux cruciaux liés à la modélisation.

1. Les modèles peuvent avoir un grand nombre de fonctions [Varenne, 2017], parmi lesquelles nous utiliserons fondamentalement : extraction d'information et de motifs, explication et compréhension, vérification et construction des théories.
2. Nous nous placerons majoritairement dans le paradigme de la *modélisation générative*, dans un souci de parcimonie et de modèles multiples avec des protocoles d'exploration extensive appropriés [Pumain and Reuillon, 2017d].
3. Cette façon de modéliser à la fois suppose et participe à une démarche de science ouverte [Fecher and Friesike, 2014].

Dans ce contexte, nous proposons de développer à présent certains enjeux particulièrement important pour notre question de manière plus précise.

3.1.2 For a cautious use of big data and computation

The so-called *big data revolution* resides as much in the availability of large datasets of novel and various types as in the always increasing available computational power. Although the *computational shift*

([Arthur, 2015]) is central for a science aware of complexity and is undeniably the basis of future modeling practices in geography as [Banos, 2013] points out, we argue that both *data deluge* and *computational potentialities* are dangerous if not framed into a proper theoretical and formal framework. The first may bias research directions towards available datasets (as e.g. numerous twitter mobility studies) with the risk to disconnect from a theoretical background, whereas the second may overshadow preliminaries analytical resolutions essential for a consistent use of simulations. We argue that the conditions for most of results in this thesis are indeed the ones endangered by incautious big-data enthusiasm, concluding that a main challenge for future Geocomputation is a wise integration of novel practices within the existing body of knowledge.

Increase in computing power

The computational power available seems to follow an exponential trend, as some kind of Moore's law. Both effective Moore's law for hardware, and improvement of softwares and algorithms, combined with a democratization of access to large scale simulation facilities, makes always more and more CPU time available for the social scientist (and to the scientist in general but this shift happened quite before in other fields, as e.g. CERN is leading in cloud computing and grid computation). About 10 years ago, [Gleyze, 2005] concluded that network analysis, for the case of Parisian public transportation network, was "limited by computation". Today most of these analyses would be quickly done on a personal computer with appropriated software and coding: [Lagesse, 2015] is a witness of such a progress, introducing new indicators with a higher computational complexity, computed on larger networks. The same parallel can be done for the Simpop models: the first Simpop models at the beginning of the millenium [Sanders et al., 1997] were "calibrated" by hand, whereas [Cottineau et al., 2015] calibrates the multi-modeling Marius model and [Schmitt et al., 2015] calibrates very precisely the SimpopLocal model, both on grid with billions of simulations. A last example, the field of Space Syntax, witnessed a long path and tremendous progresses from its theoretical origins [Hillier and Hanson, 1989] to recent large-scale applications [Hillier, 2016].

A data deluge ?

Concerning the new and "big" data available, it is clear that always larger dataset are available and always newer type of data are available. Numerous examples of fields of application can be given. For example, mobility can now be studied from various entries, such as new data from smart transportation systems [O'brien, Cheshire, and Batty, 2014], from social networks [Frank et al., 2014], or other more

exotic data such as mobile phone data [De Nadai et al., 2016]. In an other spirit, the opening of “classic” datasets (such as city dashboards, open data government initiatives) should allow ever more meta-analyses. New ways to do research and produce data are also raising, towards more interactive and crowd-sourced initiatives. For example, [Cottineau, 2017] describes a web-application aimed at presenting a meta-analysis of Zipf’s law across numerous datasets, but in particular features an upload option, where the user can upload its own dataset and add it to the meta-analysis. Other applications allow interactive exploration of scientific literature for a better knowledge of a complex scientific landscape, as [Chasset et al., 2016] does.

On induced dangers

As always the picture is naturally not as bright as it seems to be at first sight, and the green grass that we try to go eating in the neighbor’s field quickly turns into a sad reality. Indeed, the purpose and motivation are fuzzy and one can get lost. Some examples speaks for themselves. [Barthelemy et al., 2013] introduces a new dataset and rather new methods to quantify road network evolution, but the results, on which the authors seem to be astonished, are that a transition occurred in Paris at the Haussmann period. Any historian of urbanism would be puzzled by the exact purpose of the paper, as in the end a vague and bizarre feeling of reinventing the wheel floats in the air. The use of computation can also be exaggerated, and in the case of agent-based modeling it can be illustrated by the example of [Axtell, 2016], for which the aim at simulating the system at scale 1:1 seems to be far from initial motivations and justifications for agent-based modeling, and may even give arguments to mainstream economists that denigrate easily ABMS. Other anecdotes raise worries: [[robin_cura_2014_11415](#)] is a web application that wastes computational ressources to simulate Gaussian distributions for a Gibrat model in order to compute their mean and variance, that are input parameters of the model. It basically checks the Central Limit Theorem, which is a priori well accepted among most scientists. Otherwise, the full distribution given by a Gibrat model is theoretically known as it was fully solved e.g. by [Gabaix, 1999]. Recently on the French speaking diffusion list *Geotamtam*, a sudden rush around *Pokemon Go* data seemed to answer more to an urgent unexplained need to exploit this new data source before anyone else rather than an elaborated theoretical construction. Simple existing accurate datasets, such as historical cities population (for France the Pumain-INED database for example), are far from being fully exploited and it may be more important to focus on these already existing classic data. One must also be aware of the possible misleading applications of some results: [Louail et al., 2017] makes a very good analysis of potential redistribution of bank card transactions within a city, but pushes the results as possible ba-

sis for social equity policy recommendation by acting on mobility, forgetting that urban form and function are coupled in a complex way and that moving transactions from one place to the other involves far more complex processes than policies.

For a cautious use

Our main claim here is that the computational shift and simulation practices will be central in geography, but may also be dangerous, for the reasons illustrated above, i.e. that data deluge may impose research subjects and elude theory, and that computation may elude model construction and solving. A stronger link is required between computational practices, computer science, mathematics, statistics and theoretical geography. Theoretical and Quantitative Geography is at the center of this dynamic, as it was its initial purpose that seems forgotten in some cases. It implies the need for elaborated theories integrated with conscious simulation practices. In other words we can answer complementary naive questions that have however to be tackled one and for once. If a theory-free quantitative geography would be possible, the answer if naturally no as it is close to the trap of black-box data-mining analysis. Whatever is done in that case, the results will have a very poor explanatory power, as they can exhibit relations but not reconstruct processes. On an other hand, the possibility of a purely computational quantitative geography is a dangerous vision: even gaining three orders of magnitudes in computational power does not solve the dimensionality curse. In our work here, without theory, we would not know which objects, measures and properties to look at (e.g. multi-scale and dynamical nature of processes), and without analytics, it would be sometimes difficult to draw conclusions from empirical analysis. Nothing is really new here but this position has to be stated and stood up, precisely because our work will use this kind of tools, trying to advance on a thin and fragile edge, with the void of the unfunded theoretical charlatanism on one side and the abyss of the technocratic blind drowning in foolish amounts of data. More than ever we need simple but powerful and funded theories à-la-Occam [Batty, 2016], to allow a wise integration of new techniques into existing knowledge.

3.1.3 *Extend sensitivity analyses*

Context

When evaluating data-driven models, or even more simple partially data-driven models involving simplified parametrization, an unavoidable issue is the lack of control on “underlying system parameters” (what is a ill-defined notion but should be seen in our sense as parameters governing system dynamics). Indeed, a statistics extracted from

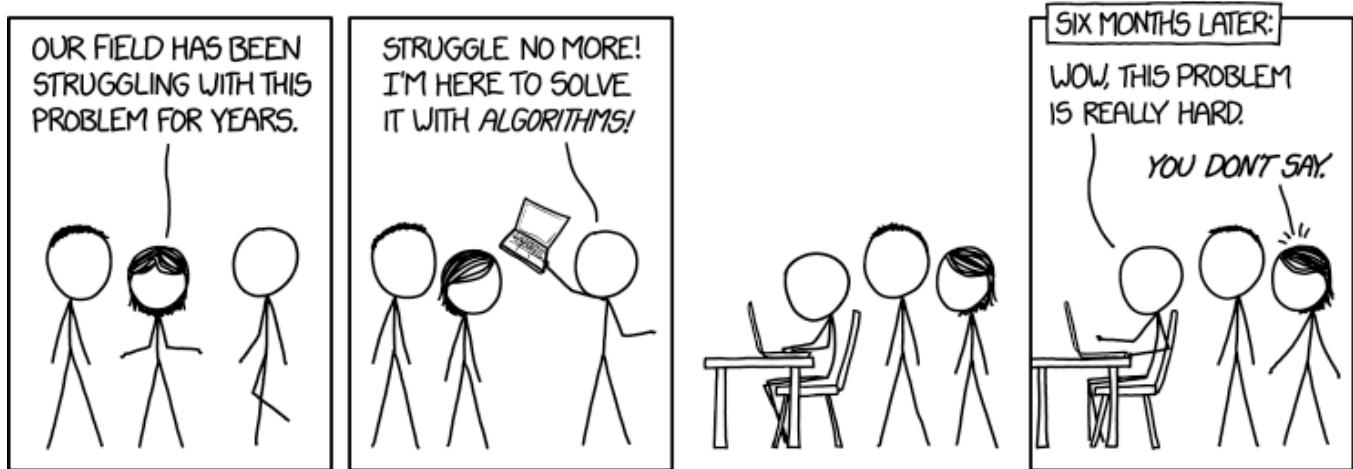


Figure 12: On naive use of data mining and intensive computation

running the model on enough different datasets can become strongly biased by the presence of confounding in the underlying real data, as it is impossible to know if result is due to processes the model tries to translate or to a hidden structure common to all data.

RATIONALE Although simulation models of geographical systems in general and agent-based models in particular represent a fantastic opportunity to explore socio-spatial behaviours and to test a variety of scenarios for public policy, the validity of generative models is uncertain until their results are proven robust. Sensitivity analysis usually include the analysis of the effect of stochasticity on the variability of results, as well as the effects of small parameter changes. However, initial spatial conditions are usually taken for granted in geographical models, thus leaving completely unexplored the effect of spatial arrangements on the interaction of agents and of their interactions with the environment. In this part, we present a method to assess the effect of initial spatial conditions on simulation models, using a systematic generator controlled by meta-parameter to create density grids used in spatial simulation models. We show, with the example of a very classical agent-based model (Sugarscape model of ressource allocation) that the effect of space in simulation is significant, and sometimes even larger than parameters themselves. We do so using high performance computing in a very simple and straightforward open-source workflow. The benefits of this approach are various but include for example the knowledge of model behavior in an extended frame, the possibility of statistical control when regressing model outputs, or a finer exploration of model derivatives than with a direct approach.

THE ROLE OF SPATIO-TEMPORAL PATH DEPENDENCIES Spatio-temporal path dependency is one of the main reasons making our approach relevant. Indeed, a crucial aspect of most spatio-temporal complex systems is their non-ergodicity ([Pumain, 2012b]) (the property that cross-sectional samples in space are not equivalent to samples in time to compute statistics such as averages), what witnesses generally strong spatio-temporal path-dependencies in their trajectories. Similar to what Gell-Mann calls *frozen accidents* in any complex system [Gell-Mann, 1995], a given configuration contains clues on past bifurcations, that can have had dramatic effects on the state of the system. Temporal and cumulative effects have been considered in various geographical subfields and at various geographical scales, including transportation and urban economics, urban geography, and interregional migrations[White, 1977],[White, 1978], [Allen and Sanglier, 1979],[Wilson, 1981],[D., Sanders, and Saint-Julien, 1989],[Allen and Sanglier, 1981],[Weidlich and Haag, 1988],[Portugali, 2000],[Wilson, 2002],[Batty, 2007],[Aziz-Alaoui and Bertelle, 2009]. Less studied is the impact of the spatial setting on models dynamics and potential bifurcations.

The example of transportation networks is a good illustration, as their spatial shape and hierarchy is strongly influenced by past investment decisions, technical choices, or political decisions sometimes not rational ([Zembri, 2010]). Some aggregated indicators will not take into account positions and trajectories of each agent (such as segregation in the Schelling model) but others, as in the case of spatial patterns of accessibility in a system of cities, fully capture the path-dependency and may therefore be highly dependent of the initial spatial configuration. It is not clear for example what shifted the economical and political capital of France from Lyon to Paris in the early Middle Age, some assumptions being the reconfiguration of trade patterns from South to North of Europe and thus an increased centrality for Paris due to its spatial position: the bifurcation induced by socio-economic and political factors took a deep significance with worldwide repercussions until today when magnified by the spatial configuration.

PREVIOUS ATTEMPTS IN THE LITERATURE The effect of the spatial configuration on area-based attributes of human behaviours has been largely discussed in geostatistics, mainly since the exposure of the Modifiable Areal Unit Problem (MAUP) [Openshaw, 1984],[Fotheringham and Wong, 1991]. Recently, [Kwan, 2012] claims for a careful examination of what she coins the uncertain geographic context problem (UGCoP), that is of the spatial configuration of geographical units even if the size and delineation of the area are the same. On the contrary, the scarcity of these considerations in the geographic simulation model literature questions the generalisation of their results, as

it has for instance been showed in the case of LUTI models [Thomas et al., 2018], of diffusion processes using ABM [Le Texier and Caruso, 2017].

Methods

In this section, we detail the method developed to analyse the sensitivity of simulation models to initial spatial conditions. In addition to the usual protocol, which consists of running a model μ with various values of its parameters and relating these variations of values to the variations in the simulation results, we here introduce a spatial generator, which itself is determined by parameters and produces sets of spatial initial conditions. Initial spatial conditions are clustered to represent types of spaces ex-ante (for example: moonocentric or polycentric density grids), and the sensitivity analysis of the model is now run against μ parameters as well as spatial parameters or spatial types. It allows the sensitivity analysis to produce qualitative conclusions regarding the influence of spatial distribution on the outputs of simulation models, alongside the classic variation of parameter values.

SPATIAL GENERATOR Our spatial generator applies an urban morphogenesis model developed and explored in 5.2. To present it in a nutshell, grids are generated through an iterative process which adds a quantity N (population) at each time step, allocating it through preferential attachment characterised by its strength of attraction α . This first growth process is then smoothed n times using a diffusion process of strength β . Grids are thus generated from the combination of the values of these four meta-parameters α , β , n and N , in addition to the random seed. To ease our exploration, only the distribution of density is allowed to vary rather than the size of the grid, which we fix to a 50x50 square environment of 100,000 units (cf. figure ??).

COMPARING PHASE DIAGRAMS In order to test for the influence of spatial initial conditions, we need a systematic method to compare phase diagrams. Indeed, we have as many phase diagrams than we have spatial grids, what makes a qualitative visual comparison not realistic. A solution is to use systematic quantitative procedures. Several potential methods could be used: for example in the case of the Schelling model, an anisotropic spatial segregation index (giving the number of clusters found and in which region in the parameter spaces they are roughly situated) would differentiate strong *meta phase transitions* (phase transitions in the space of meta parameters). The use of metrics comparing spatial distributions, such as the Earth Movers Distance which is used for example in Computer Vision to compare probability distributions [Rubner, Tomasi, and Guibas, 2000], or the comparison of aggregated transition matrices of the

dynamic associated to the potential described by each distribution, would also be potential tools. Map comparison methods, popular in environmental sciences, provide numeral tools to compare two dimensional fields [Visser and De Nijs, 2006]. To compare a spatial field evolving in time, elaborated methods such as Empirical Orthogonal Functions that isolates temporal from spatial variations, would be applicable in our case by taking time as a parameter dimension, but these have been shown to perform similarly to direct visual inspection when averaged over a crowdsourcing [Koch and Stisen, 2017]. To keep it simple and as such methodological considerations are auxiliary to the main purpose of this paper, we propose an intuitive measure corresponding to the share of between-diagrams variability relative to their internal variability. More formally, the distance is given by

$$d_r(\alpha_1, \alpha_2) = 2 \cdot \frac{d(f_{\vec{\alpha}_1}, f_{\vec{\alpha}_2})^2}{\text{Var}[f_{\vec{\alpha}_1}] + \text{Var}[f_{\vec{\alpha}_2}]} \quad (1)$$

where $\alpha \mapsto [\vec{x} \mapsto f_{\vec{\alpha}}(\vec{x})]$ is the operator giving phase diagrams with \vec{x} parameters and $\vec{\alpha}$ meta-parameters, and d is a distance between probability distributions that can be taken for example as basic L₂ distance or the Earth's Mover Distance. For each values $\vec{\alpha}_i$, the phase diagram is seen as a random spatial field, facilitating the definition of variances and distance.

Results

Sugarscape is a model of resource extraction which simulates the unequal distribution of wealth within a heterogenous population ([Epstein and Axtell, 1996]). Agents of different vision scopes and different metabolisms harvest a self-regenerating resource available heterogeneously in the initial landscape, they settle and collect this resource, which leads some of them to survive and others to perish. The main parameters of this model are the number of agents, their minimal and maximal resource. In addition, we are interested in testing the impact of the spatial distribution of the resource in this project, using the spatial generator. The outcome of the model is measured as a phase diagram of an index of inequality for ressource distribution (Gini index). We extend the implementation with agents wealth distribution of [Li and Wilensky, 2009].

For the exploration, 2,500,000 simulations (1000 parameter points x 50 density grids x 50 replications) allow us to show that the model is more sensitive to space than to its other parameters, both qualitatively and quantitatively: the amplitude of variations across density grids is larger than the amplitude in each phase diagram, and the behavior of phase diagram is qualitatively different in different regions of the morphological space. More precisely, we explore a

grid of a basic parameter space of the model, which three dimensions are the population of agents $P \in [10; 510]$, the minimal initial agent resource $s_- \in [10; 100]$ and the maximal initial agent resource $s_+ \in [110; 200]$. Each parameter is binned into 10 values, giving 1000 parameter points. We run 50 repetitions for each configuration, what yield reasonable convergence properties. The initial spatial configuration varies across 50 different grids, generated by sampling meta-parameters for the generator in a LHS. We did not use the clustered grids to test the flexibility of our framework, which is demonstrated in this case by a direct sequential coupling of the generator and the model. We measure the distance of all 3-dimensional phase diagrams to the reference phase diagram computed on the default model setup (see Fig. ?? for its morphological positioning regarded generated grids), using equation 1 with the L₂ distance to ensure direct interpretability. Indeed, it gives in that case the average squared distance between corresponding points of the phase diagrams, relative to the average of the variance of each. Therefore, values greater than 1 will mean that inter-diagram variability is more important than intra-diagram variability.

We obtain a very strong sensitivity to initial conditions, as the distribution of the relative distance to reference across grids ranges from 0.09 to 2.98 with a median of 1.52 and an average of 1.30. It means that in average, the model is more sensitive to meta-parameters than to parameters, and the relation variation can reach a factor of 3. We plot in Fig. 13 their distribution in a morphological space. The reduced morphological space is obtained by computing 4 raw indicators of urban form, namely Moran index, average distance, rank-size slope and entropy (see [LeNechet2015] for precise definition and contextualization), and by reducing the dimension with a principal component analysis for which we keep the first two components (92% of cumulated variance). The first measures a “level of sprawl” and of scattering, whereas the second measures aggregation.¹⁷ We find that grids producing the highest deviations are the ones with a low level of sprawl and a high aggregation. It is confirmed by the behavior as a function of meta-parameters, as high values of α also yield high distance. In terms of model processes, it shows that congestion mechanisms induce rapidly higher levels of inequality.

We now check the sensitivity in terms of qualitative behavior of phase diagrams. We show the phase diagrams for two very opposite morphologies in term of sprawling, but controlling for aggregation with the same PC2 value. These correspond to the green and blue frames in Fig. 13. The behaviors are rather stable for varying s_+ , what means that the poorest agents have a determinant role in trajectories. The two examples have not only a very distant baseline

¹⁷ We have $PC1 = 0.76 \cdot \text{distance} + 0.60 \cdot \text{entropy} + 0.03 \cdot \text{moran} + 0.24 \cdot \text{slope}$ and $PC2 = -0.26 \cdot \text{distance} + 0.18 \cdot \text{entropy} + 0.91 \cdot \text{moran} + 0.26 \cdot \text{slope}$.

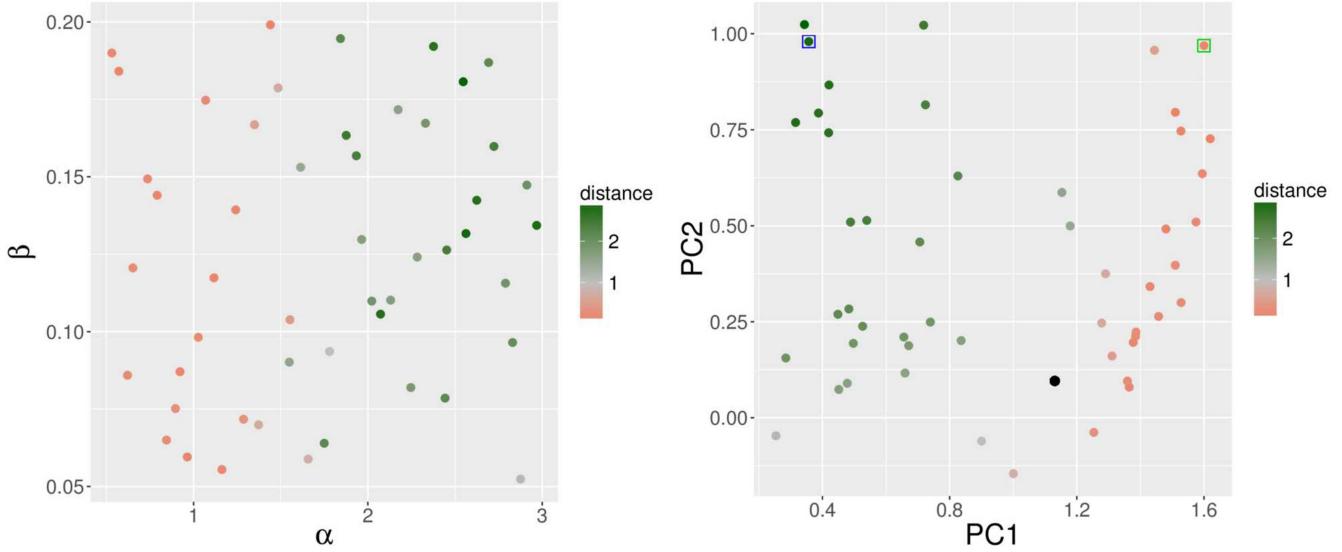


Figure 13: Relative distances of phase diagrams to the reference across grids. (Top) Relative distance as a function of meta-parameters α (strength of preferential attachment) and diffusion (β , strength of diffusion process). (Bottom) Relative distance as a function of two first principal components of the morphological space (see text). Red point correspond to the reference spatial configuration. Green frame and blue frame give respectively the first and second particular phase diagrams shown in Fig. 14.

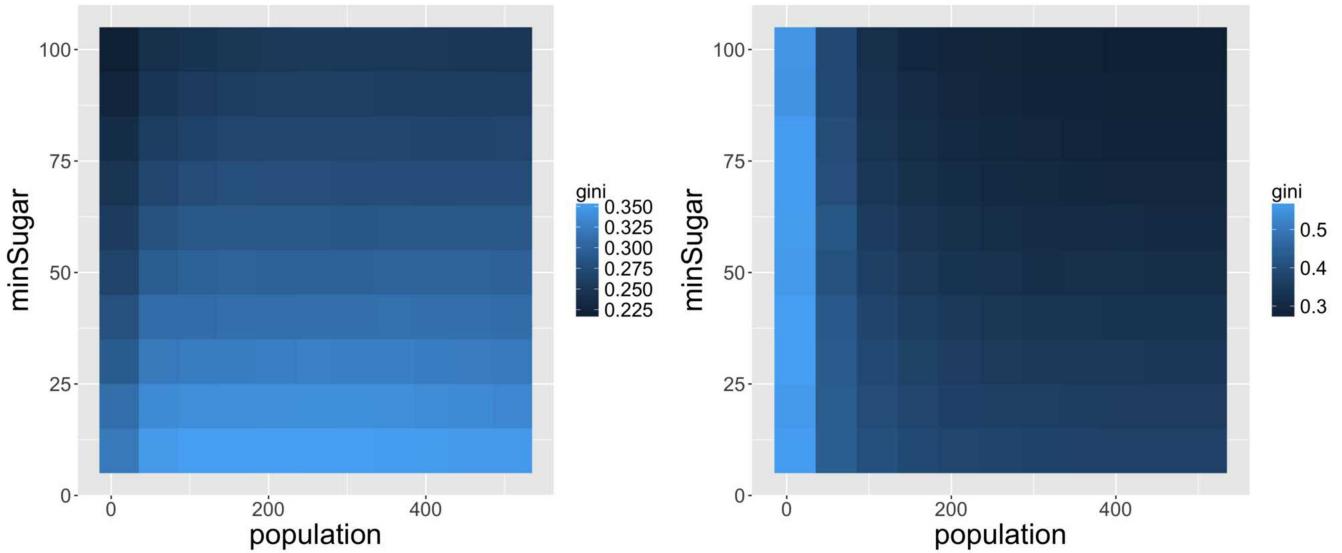


Figure 14: Examples of phase diagrams. We show two dimensional phase diagrams on (P, s_-) , both at fixed $s_+ = 110$. (Left) Green frame, obtained with $\alpha = 0.79$, $n = 2$, $\beta = 0.14$, $N = 157$; (Right) Blue frame, obtained with $\alpha = 2.56$, $n = 3$, $\beta = 0.13$, $N = 128$.

inequality (the ceil of the first 0.35 is roughly the floor of the second 0.3), but their qualitative behavior is also radically opposite: the sprawled configuration gives inequalities decreasing as population decreases and decreasing as minimal wealth increases, whereas the concentrated one gives inequalities strongly increasing as population decreases and also decreasing with minimal wealth but significantly only for large population values. The process is thus completely inverted, what would have significant impacts if one tried to schematize policies from this model. This second example confirms thus the importance of sensitivity of simulation models to the initial spatial conditions.

* * *

*

Nous avons vu dans cette section comment nous positionner par rapport à l'usage des modèles de simulation, et plus généralement par rapport au calcul intensif. Nous avons vu revenir de manière récurrente dans les problématiques abordées la question de l'ouverture des pratiques scientifiques.

Nous proposons dans la section suivante d'en détailler un aspect, celle de la reproductibilité, qui en est à la fois une composante mais aussi un produit : simultanément produit et producteur, elle permet une plus grande ouverture et est réciproquement encouragée par les pratiques d'ouverture.

* * *

*

3.2 REPRODUCIBILITY

The strength of science comes from the cumulative and collective nature of research, as progresses are made as Newton said “standing on the shoulder of giants”, meaning that the scientific enterprise at a given time relies on all the work done before and that advances would not be possible without constructing on it. It includes development of new theories, but also extension, testing or falsifiability of previous ones.

The effective practice of reproducibility seems to be increasing [Stodden, 2010] and technical means to achieve it are always more developed (as e.g. ways to make data openly available, or to be transparent on the research process such as git [Ram, 2013], or to integrate document creation and data analysis such as knitr [Xie, 2013]), at least in the field of modeling and simulation. However, the devil is indeed in the details and obstacles judged at first sight as minor become rapidly a burden for reproducing and using results obtained in some previous researches. We describe two cases studies where models of simulation are apparently highly reproducible but unveil as puzzles on which research-time balance is significantly under zero, in the sense that trying to exploit their results may cost more time than developing from scratch similar models.

3.2.1 *Explicitation, documentation and implementation of models*

On the Need to Explicit the Model

A current myth (to which we ourselves struggle to escape indeed) is that providing entire source code and data will be a sufficient condition for reproducibility. It will work if the objective is to produce exactly same plots or statistical analysis, assuming that code provided is the one which was indeed used to produce the given results. It is however not the nature of reproducibility. First, results must be as much implementation-independent as possible for clear robustness purposes. Then, in relation with the precedent point, one of the purposes of reproducibility is the reuse of methods or results as basis or modules for further research (what includes implementation in another language or adaptation of the method), in the sense that reproducibility is not replicability as it must be adaptable [Drummond, 2009].

Our first case study fits exactly that scheme, as it was undoubtedly aimed to be shared with and used by the community since it is a model of simulation provided with the Agent-Based simulation platform NetLogo [Wilensky, 1999]. The model is also available online [De Leon, Felsen, and Wilensky, 2007] and is presented as a tool to simulate socio-economic dynamics of low-income residents in a city based on a synthetic urban environment, generated to be close

in stylized facts from the real town of Tijuana, Mexico. Beside providing the source code, the model appears to be poorly documented in the literature or in comments and description of the implementation. Comments made thereafter are based on the study of the urban morphogenesis part of the model (setup for the “residential dynamics” component) as it is our global context of study. In the frame of that study, source code was modified and commented, which last version is available on the repository of the project¹⁸.

RIGOROUS FORMALIZATION An obvious part of model construction is its rigorous formalization in a formal framework distinct from source code. There is of course no universal language to formulate it [Banos, 2013], and many possibilities are offered by various fields (e.g. UML, DEVS, pure mathematical formulation). No paper nor documentation is provided with the model, apart from the embedded NetLogo documentation, that only thematically describes in natural language the ideas behind each step without developing more and provides information about role of different elements of the interface.

This formulation is a key for it to be understood, reproduced and adapted; but it also avoids implementation biases such as

- Architecturally dangerous elements: in the model, world context is a torus and agents may “jump” in the euclidian representation, what is not acceptable for a 2D projection of real world. To avoid that, many tricky tests and functions were used, including unadvised practices (e.g. dead of agents based on position to avoid them jumping).
- Lack of internal consistence: the example of the patch variable `land-value` used to represent different geographical quantities at different steps of the model (morphogenesis and residential dynamics), what becomes an internal inconsistence when both steps are coupled when option `city-growth?` is activated.
- Coding errors: in an untyped language such as NetLogo, mixing types may conduct to unexpected runtime errors, what is the case of the patch variable `transport` in the model (although no error occurs in most of run configurations from the interface, what is more dangerous as the developer thinks implementation is secure). Such problems should be avoided if implementation is done from an exact formal description of the model.

TRANSPARENT IMPLEMENTATION A totally transparent implementation is expected, including ergonomics in architecture and coding, but ...

¹⁸ at <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Reproduction/UrbanSuite>

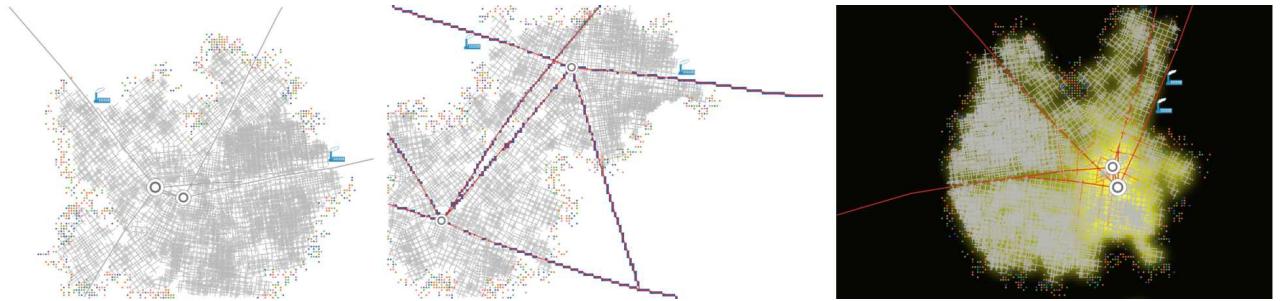


Figure 15: Example of simple improvement in visualization that can help understanding mechanisms implied in the model. *Left:* Example of original output ; *Middle:* Visualization of main roads (in red) and underlying patches attribution, suggesting possible implementation bias in the use of discretized trace of roads to track their positions ; *Right:* Visualization of land values using a more readable color gradient. This step confirms the hypothesis, through the form of value distribution, that the morphogenesis step is an unnecessary detour to generate a random field for which simple diffusion method should provide similar results, as detailed in the paragraph on implementation.

EXPECTED MODEL BEHAVIOR Whatever the definition, a model can not be reduced to its formulation and/or implementation, as expected model behavior or model usage can be viewed as being part of the model itself. In the frame of GIERE’s perspectivism [Giere, 2010c], the definition of model includes the purpose of use but also the agent who aims to use it. Therefore a minimal explication of model behavior and exploration of parameter roles is highly advised to decrease chances of misuses or misinterpretations of it. It includes simple runtime charts that are immediate on the NetLogo platform, but also indicators computations to evaluate outputs of the model. It can also be improved visualizations during runtime and model exploration, such as showed in Fig. ??.

On the Need of Exactitude in Model Implementation

Possible divergences between model description in a paper and the effectively implemented processes may have grave consequences on the final reproducibility. The road network growth model given in [Barthelemy and Flammini, 2008] is one example of such a discrepancy. A strict implementation of model mechanisms provide slightly different results than the one presented in the paper, and as source code is not provided we need to test different hypotheses on possible mechanisms added by the programmer (that seems to be a connexion rule to intersections under a certain distance threshold). Lessons that could be possibly drawn from this examples are

- the necessity of providing source code
- the necessity of providing architecture description along with code (if model description is in a langage too far from archi-

tectural specifications) in order to identify possible implementation biaises

- the necessity of performing and detailing explicitly model explorations, that would in that case have helped to identify the implementation bias.

Making the last point mandatory may ensure a limited risk of scientific falsification as it is generally more complicated to fake false exploration results than to effectively explore the model. One could imagine an experiment to test the general behavior of a subset of the scientific community regarding reproducibility, that would consist in the writing of a false modeling paper in the spirit of [Zilsel, 2015], in which opposite results to the effective results of a given model are provided, without providing model implementation. A first bunch of test would be to test the acceptance of a clearly non-reproducible paper in diverse journals, possibly with a control on textual elements (using or not “buzz-words” associated to the journal, etc.). Depending on results, a second experiment may be tested with providing open source code for model implementation but still with false results, to verify if reviewers effectively try to reproduce results when they ask for the code (in reasonable computational power limits of course, HPC being not currently broadly available in Humanities).

Interactive Exploration and Production of Results

L’usage d’applications interactives pour la fouille de données a des avantages non discutables, tel qu’une familiarisation avec la structure des données par une vue d’ensemble qui serait beaucoup plus laborieuse voire impossible autrement. C’est la même idée sous-jacente qui justifie l’interactivité pour l’exploration préliminaire des modèles multi-agents intégrée à des plateformes comme NetLogo [Wilensky, 1999] ou Gamma [Drogoul et al., 2013]. Un objectif similaire est implicite dans [Rey-Coyrehourcq, 2015], c’est-à-dire une intégration complète de l’exploration fine des modèles et de la production des graphes de sortie ainsi que leur exploration interactive. Comme le rappelle ROMAIN REUILLOU (Entretien du 11/04/2017, voir D.3), la plateforme OpenMole qui devait accueillir cette couche supplémentaire était à ses débuts à l’époque et ne l’est toujours pas aujourd’hui, puisque l’état de l’art de telles pratiques est en pleine construction et bouleversements réguliers [Holzinger, Dehmer, and Jurisica, 2014].

Des difficultés au regard de la reproductibilité, qui nous concernent particulièrement ici, sont récurrentes et loin d’être résolues. En effet, il faut bien situer la position de ces outils et méthodes comme une aide cognitive préliminaire¹⁹, mais peu souvent comme permettant la production de résultats finaux : lorsque les paramètres ou

¹⁹ Que nous ne jugeons pas superficielle puisque nous les mobilisons au moins deux fois par la suite, voir ci-dessous ainsi que C.1.

dimensions se multiplient, l'export d'un graphe est bien souvent déconnecté de l'information complète ayant conduit à sa production. De la même manière, l'utilisation de notebooks intégrés tel Jupyter, permettant d'intégrer analyses et rédaction du compte-rendu, peut devenir dangereux car on peut justement revenir sur un script, tester différentes valeurs d'un paramètre, et perdre les valeurs qui avaient produit un graphe donné. L'utilisation de versioning peut être une solution partielle mais souvent lourde.

Dans l'idéal, tout logiciel interactif permettant l'export de résultats devrait en même temps exporter un script ou une description exacte et utilisable permettant d'arriver exactement à ce point à partir des données brutes. La plupart des applications d'exploration interactives de données spatio-temporelles sont à ce regard relativement immatures scientifiquement, car même dans le cas où elles sont totalement honnêtes et transparentes sur les analyses présentées à l'utilisateur, ce qui n'est malheureusement pas la règle, les tâtonnements d'exploration progressive ne sont pas reproductibles et la méthode d'extraction de caractéristiques est ainsi relativement aléatoire. En poussant le raisonnement, leur utilisation révélerait plutôt l'aveu d'une faiblesse d'un manque de méthodes systématiques accompagnant la découverte de motifs dans des données spatio-temporelles complexes de manière efficace.

Par un plaidoyer visionnaire, BANOS avait déjà mis en garde contre "les dangers de la jungle" des données dans [Banos, 2001], quand il souligne très justement que l'exploration interactive doit nécessairement se doubler d'indicateurs locaux adaptés, mais surtout d'outils d'exploration automatisés et de critère d'évaluation des choix faits et des motifs découverts par l'utilisateur. Nous revenons encore à l'idée d'une plateforme intégrée dont OpenMole pourrait être un précurseur. La combinaison des capacités cognitives humaines au traitement machine, notamment pour des problèmes de vision par ordinateur, ouvre des possibilités de découvertes inédites, encore plus via une utilisation collective comme en témoigne le Galaxy Zoo [Radick et al., 2010]²⁰. Les résultats d'un crowdsourcing de la cognition humaine peuvent rivaliser avec les techniques automatiques les plus avancées comme le montre [Koch and Stisen, 2017] pour l'exemple de la comparaison de cartes spatiales.

Ces possibilités ne doivent cependant pas être sur-estimées ou utilisées à mauvais escient, et les questions d'intégration efficiente homme-machine sont d'ailleurs totalement ouvertes. Dans le domaine de la visualisation de l'information géographique, [Pfaender, 2009] introduit une sémiologie spécifique visant à favoriser l'exploration de grands jeux de données hétérogènes, et l'expérimente sur une application

²⁰ Le principe rejoint celui de *citizen science*, en faisant participer des volontaires hors de la communauté scientifique à des tâches requérant cognition mais pas de connaissances scientifique : la classification d'images, dans le but d'entraîner des algorithmes supervisés, est l'exemple initial du Galaxy Zoo pour la forme des galaxies.

spécifique : il s'agit d'une avancée considérable vers une plateforme intégrée et une exploration interactive saine et reproductible, les directions d'exploration répondant à des modèles basés sur les sciences cognitives.

Enfin, le rôle de l'interactivité dans la communication et la vulgarisation scientifiques est exploré par l'Annexe C.7, qui suggère la mise en place de jeux, notamment un jeu informatique interactif, pour faciliter la transmission de concepts scientifiques au public. cela nous montre que le développement de ces pratiques innovantes dépasse largement le seul cadre de l'analyse de données.

Application

Again, reproducibility and transparency is a non-negotiable feature of contemporaneous science, along with Open practices and Open Access. Too much examples (see a very recent one in experimental economics [Camerer et al., 2016]) show in various disciplines the lack of reproducibility of experiments, that is a falsification of previous results or a result in itself. Falsification is a costly practice, and even if necessary [Chavaliarias et al., 2005], could be made more efficient through more transparency and direct reproducibility, increase therein the global workflow of science. We develop in parallel of this thesis various tools aimed to ease reproducibility, for which an overview is given in appendix E.3.

Le développement et la systématisation de standards et de bonnes pratiques, de manière conjointe sur les différentes problématiques évoquées, est une condition nécessaire à une rigueur scientifique qui devrait être uniforme au travers de l'ensemble des disciplines existantes. Nous construisons par exemple des outils facilitant le flot de production scientifique, ceux-ci étant détaillés en Appendice E.3. Par exemple, pour les sciences computationnelles, on a déjà évoqué les potentialités de l'utilisation de git qui s'étendent en fait sans contrainte de disciplines ni de types de recherche si les bonnes adaptations sont introduites. Le suivi précis de l'ensemble des étapes d'un projet, gardé en historique offrant la possibilité de revenir à n'importe laquelle à tout moment, mais aussi de travailler de façon collaborative, plus ou moins parallèlement selon les besoins en utilisant les branches, est un exemple de service fourni par cet outil. Un exemple de bonnes pratiques d'utilisation est donné par [Perez-Riverol et al., 2016].

Plus généralement, les sciences computationnelles nécessitent l'adoption de certains standards et pratiques pour assurer une bonne reproductibilité, et ceux-ci restent majoritairement à développer : [Wilson et al., 2017] donne des premières pistes. Concernant la qualité des données, de nombreux efforts sont faits pour introduire des cadres de standardisation des données : par exemple [Veiga et al., 2017] décrit un cadre conceptuel visant à guider la résolution de problème récur-

rent liés à la qualité des données de biodiversité (comme par exemple évaluer des mesures jugeant de l'usage possible d'un jeu de données pour un problème donné). De nouvelles perspectives s'ouvrent pour des futurs cadres de traitement de données intrinsèquement ouverts et reproductibles, avec le développement de nouvelles techniques comme le *blockchain*²¹, comme proposé par [Furlanello et al., 2017].

3.2.2 *Open Data*

L'accès aux données est également un point crucial pour la reproductibilité, et sans nous y attarder car cela impliquerait des développements sur la définition, la philosophie, le droit des données etc. qui sont des sujets de recherche en eux-mêmes, nous donnons des perspectives sur les opportunités offertes par une ouverture systématique des données en recherche. En géographie, les *data paper* sont une pratique inexistante, et la règle est plutôt de garder la main jalousement sur un jeu produit, capitalisant sur le fait d'être le seul à y avoir accès²².

Il est évident que la qualité et quantité des connaissances produites sera nécessairement plus grande si un jeu de données est publiquement ouvert, puisqu'au moins la même chose sera obtenue, et on peut s'attendre à une prise en main par d'autres domaines, d'autres méthodes, et donc à une plus grande richesse²³.

La fermeture induira plutôt des effets négatifs, comme par exemple du temps perdu à recoder une base vectorielle donnée uniquement sous forme de carte dans un article. L'argument du temps passé comme justification à la fermeture est absurde, puisqu'au contraire, en voyant les données comme une composante à part entière de la connaissance (voir le cadre de connaissances en 8.3), le temps passé doit impliquer plus de citations, donc plus d'utilisation, ce qui passe nécessairement par l'ouverture pour des données. De même, quelle

²¹ Le *blockchain* consiste en la distribution d'un graphe de transactions entre utilisateurs, celles-ci étant validées (dans le cadre historique classique de type *proof-of-work*) par la résolution de problèmes cryptographiques inverses par force brute, par des agents appelés mineurs, essentiels à la robustesse de l'écosystème.

²² Il n'existe à notre connaissance pas de travail quantifiant la proportion de données ouvertes sur l'ensemble des données produites en géographie. Cela pourrait être l'objet d'un travail d'épistémologie quantitative appliquant des techniques similaires à celles développées en chapitre 2. La difficulté à trouver des données ouvertes, comparée à la fréquence des publications dans les domaines concernés, suggère une validité au moins qualitative de ce fait.

²³ Il est possible d'argumenter que le système de production scientifique est complexe, et qu'une monétarisation, compétition ou privatisation accrue de la recherche peut faire partie d'un écosystème de recherche dont les sorties pourront être jugées de qualité selon les indicateurs choisis. Ces considérations sont pertinentes, mais hors de notre portée puisque relevant d'un travail en anthropologie et sociologie des sciences. Nous postulons ici ce principe, et le considérons comme une position scientifique subjective.

logique, sinon la même absurde de propriété des connaissances, pousse les géographes à insérer un copyright sur l'ensemble de leurs cartes mais aussi leurs figures, jusqu'à un copyright pour un simple histogramme qui s'en serait bien passé si on avait pu l'interroger, honnête de simplicité ?

L'expérience d'évaluation d'articles nous induit à réellement nous inquiéter sur la valeur donnée à l'ouverture des données par les auteurs : au bout d'une dizaine d'articles, incluant des journaux affichant comme priorité et pré-requis l'ouverture totale des données et modèles, dont un seul est seulement partiellement ouvert et l'ensemble des autres implique de croire sur parole les résultats présentés (alors qu'un des but de la revue est de contourner les biais cognitifs qu'un ou des humains ont forcément par une validation croisée qui doit se faire sur les résultats bruts et non des interprétations contenant ces biais), il est difficile de croire que des mutations profondes des pratiques ne sont pas nécessaires.

Mais en suivant l'adage de Framasoft²⁴, "la route est longue mais la voie est libre", les perspectives sont nombreuses pour une évolution dont la lenteur n'est pas inéluctable. Le journal Cybergéo, pionnier des pratiques d'ouverture en sciences sociales (première revue entièrement électronique, première revue à lancer une rubrique de *model papers*), lance en 2017 une rubrique *data papers*²⁵ visant à inciter le développement du partage de données et de l'ouverture en géographie.

Il reste des zones grises sur lesquelles il est impossible aujourd'hui d'avoir des perspectives, notamment le droit des données. Nous avons un exemple dans les analyses que nous développerons : les données bibliographiques sont obtenues au prix d'une guerre de blocage par Google et un effort technique considérable pour la gagner (voir 2.2 et B.6).

L'ouverture implique un engagement qui fait résolument partie de nos positionnements. C'est la même idée qui soutient la construction de l'application CybergéoNetworks²⁶, qui couple les outils présentés en 2.2 avec d'autres approches complémentaires d'analyse de corpus, dans le but d'encourager la réflexivité scientifique, et de mettre cet outil ouvert à la disposition d'éditeurs indépendants, pour s'émanciper de la nouvelle main mise des géants de l'édition qui à la recherche d'un nouveau modèle pour sécuriser leur profits parient sur la vente de méta-contenu et de son analyse. Heureusement, la récente loi numérique en France a gagné le bras de fer contre leur revendication d'un droit exclusif sur la fouille de texte complets.

²⁴ Réseau pour la promotion du logiciel libre, <https://framasoft.org/>

²⁵ Dont l'index est disponible à <https://cybergeo.revues.org/28545>. Le premier article est [Swerts, 2017], que nous utilisons d'ailleurs en 7.3.

²⁶ Dont la démarche et le contexte sont détaillés en Annexe C.4. Elle est disponible en ligne à <http://shiny.parisgeo.cnrs.fr/CybergéoNetworks>.

3.2.3 *Illustration by an empirical study*

Nous proposons à présent de développer un exemple concret d'étude empirique illustrant les derniers points relevés ci-dessus et nous permettant une entrée progressive dans notre problématique. Dans le cas du trafic routier en Ile-de-France, nous menons une collecte d'un jeu de données là où il n'existe pas de source ouverte. Nous mettons également en place une application permettant son exploration interactive.

Nous avons développé en 1.1 le concept de mobilité quotidienne comme jouant un rôle clé dans les processus d'interaction entre réseaux de transport et territoires, à une échelle que nous avons désignée par microscopique. Il est de plus candidat à la mobilisation de dynamiques co-évolutives, comme le suggère l'effet des localisations sur la congestion et réciproquement.

Ici, la mobilité sera captée par le flux de trafic, et la co-évolution s'opère entre propriétés du réseau (congestion) et localisation des agents. Nous nous intéresserons plus particulièrement à l'équilibre hypothétique des flux de trafic, répondant indirectement à des problématiques que nous détaillons ci-dessous.

Context

Traffic Modeling has been extensively studied since seminal work by [Wardrop, 1952] : economical and technical elements at stake justify the need for a fine understanding of mechanisms ruling traffic flows at different scales. Many approaches with different purposes coexist today, of which we can cite dynamical micro-simulation models, generally opposed to equilibrium-based techniques. Whereas the validity of micro-based models has been largely discussed and their application often questioned, the literature is relatively poor on empirical studies assessing the stationary equilibrium assumption in the Static User Equilibrium (SUE) framework. Various more realistic developments have been documented in the literature, such as Dynamic Stochastic User Equilibrium (DSUE) (see e.g. a description by [Han, 2003]). An intermediate between static and stochastic frameworks is the Restricted Stochastic User Equilibrium, for which route choice sets are constrained to be realistic ([Rasmussen et al., 2015]). Extensions that incorporate user behavior with choice models have more recently been proposed, such as [Zhang, Mahmassani, and Lu, 2013] taking into account both the influence of road pricing and congestion on user choice with a Probit model. Relaxations of other restricting assumptions such as pure user utility maximization have been also introduced, such as the Boundedly Rational User Equilibrium described by [Mahmassani and Chang, 1987]. In this framework, user have a range of satisfying utilities and equilibrium is achieved when all users are satisfied. It produces more complex features such as the

existence of multiple equilibria, and allows to account for specific stylized facts such as irreversible network change as developed by [Guo and Liu, 2011]. Other models for traffic assignment, inspired from other fields have also recently been proposed : in [Puzis et al., 2013], an extended definition of betweenness centrality combining linearly free-flow betweenness with travel-time weighted betweenness yield a high correlation with effective traffic flows, acting thus as a traffic assignment model. It provides direct practical applications such as the optimization of traffic monitors spatial distribution.

Despite all these developments, some studies and real-world applications still rely on Static User Equilibrium. Parisian region e.g. uses a static model (MODUS) for traffic management and planning purposes. [Leurent and Boujnah, 2014] introduce a static model of traffic flow including parking cruising and parking lot choice: it is legitimate to ask, specifically at such small scales, if the stationary distribution of flows is a reality. An example of empirical investigation of classical assumptions is given in [Zhu and Levinson, 2015], in which revealed route choices are studied. Their conclusions question “Wardrop’s first principle” implying that users choose among a well-known set of alternatives. In the same spirit, we investigate the possible existence of the equilibrium in practice. More precisely, SUE assumes a stationary distribution of flows over the whole network. This assumption stays valid in the case of local stationarity, as soon as time scale for parameter evolution is considerably greater than typical time scales for travel. The second case which is more plausible and furthermore compatible with dynamical theoretical frameworks, is here tested empirically.

The rest of the paper is organized as follows : data collection procedure and dataset are described ; we present then an interactive application for the interactive exploration of the dataset aimed to give intuitive insights into data patterns ; we present then results of various quantitative analyses that give convergent evidence for the non-stationarity of traffic flows ; we finally discuss implications of these results and possible developments.

Dataset

DATASET CONSTRUCTION We propose to work on the case study of Parisian Metropolitan Region. An open dataset was constructed for highway links within the region, collecting public real-time open data for travel times (available at www.sytadin.fr). As stated by [Bouteiller and Berjoan, 2013], the availability of open datasets for transportation is far to be the rule, and we contribute thus to a data opening by the construction of our dataset. Our data collection procedure consists in the following simple steps, executed each two minutes by a python script :

- fetch raw webpage giving traffic information

- parse html code to retrieve traffic links id and their corresponding travel time
- insert all links in a sqlite database with the current timestamp.

The automatized data collection script continues to enrich the database as time passes, allowing future extensions of this work on a larger dataset and a potential reuse by scientists or planners. The latest version of the dataset is available online (sqlite format) under a Creative Commons License²⁷.

DATA SUMMARY A time granularity of 2 minutes was obtained for a three months period (February 2016 to April 2016 included). Spatial granularity is in average 10km, as travel times are provided for major links. The dataset contains 101 links. Raw data we use is effective travel time, from which we can construct travel speed and relative travel speed, defined as the ratio between optimal travel time (travel time without congestion, taken as minimal travel times on all time steps) and effective travel time. Congestion is constructed by inversion of a simple BPR function with exponent 1, i.e. we take $c_i = 1 - \frac{t_{i,\min}}{t_i}$ with t_i travel time in link i and $t_{i,\min}$ minimal travel time.

Analysis of traffic flow patterns

VISUALIZATION OF SPATIO-TEMPORAL CONGESTION PATTERNS
As our approach is fully empirical, a good knowledge of existing patterns for traffic variables, and in particular of their spatio-temporal variations, is essential to guide any quantitative analysis. Taking inspiration from an empirical model validation literature, more precisely Pattern-oriented Modeling techniques introduced by [Grimm et al., 2005], we are interested in macroscopic patterns at given temporal and spatial scales: the same way stylized facts are in that approach extracted from a system before trying to model it, we need to explore interactively data in space and time to find relevant patterns and associated scales. We implemented therefore an interactive web-application for data exploration using R packages shiny and leaflet²⁸. It allows dynamical visualization of congestion among the whole network or in a particular area when zoomed in. The application is accessible online at <http://shiny.parisgeo.cnrs.fr/transportation>. A screenshot of the interface is presented in Figure ???. Main conclusion from interactive data exploration is that strong spatial and temporal heterogeneity is the rule. The temporal pattern recurring most often, peak and off-peak hours is on a non-negligible proportion of days

²⁷ at http://37.187.242.99/files/public/sytadin_latest.sqlite3

²⁸ source code for the application and analyses is available on project open repository at
<https://github.com/JusteRaimbault/TransportationEquilibrium>

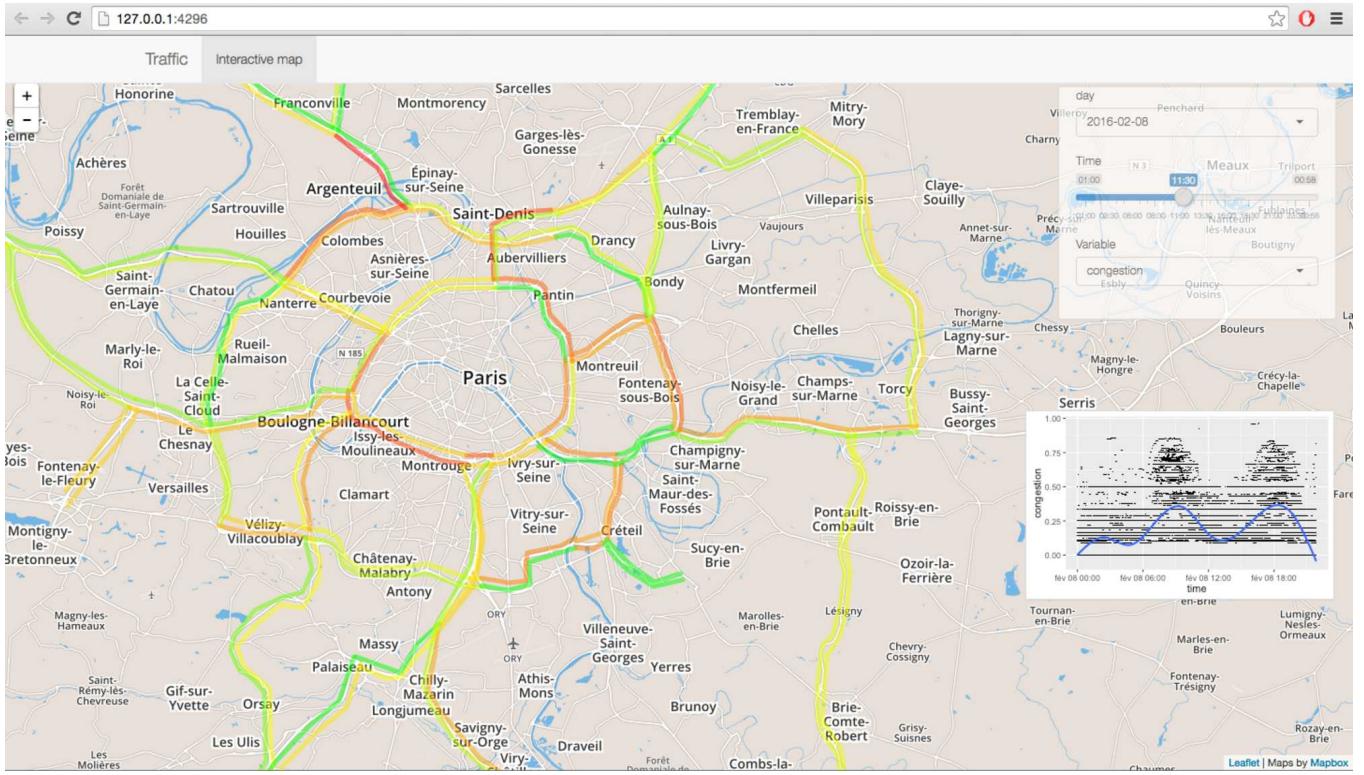


Figure 16: Capture of the web-application to explore spatio-temporal traffic data for Parisian region. It is possible to select date and time (precision of 15min on one month, reduced from initial dataset for performance purposes). A plot summarizes congestion patterns on the current day.

perturbed. In a first approximation, non-peak hours may be approximated by a local stationary distribution of flows, whereas peaks are too narrow to allow the validation of the equilibrium assumption. Spatially we can observe that no spatial pattern is clearly emerging. It means that in case of a validity of static user equilibrium, meta-parameters ruling its establishment must vary at time scales smaller than one day. We argue that traffic system must in contrary be far-from-equilibrium, especially during peak hours when critical phase transitions occur at the origin of traffic jams.

SPATIO-TEMPORAL VARIABILITY OF TRAVEL PATH Following interactive exploration of data, we propose to quantify the spatial variability of congestion patterns to validate or invalidate the intuition that if equilibrium does exist in time, it is strongly dependent on space and localized. The variability in time and space of travel-time shortest paths is a first way to investigate flow stationarities from a game-theoretic point of view. Indeed, the static User Equilibrium is the stationary distribution of flows under which no user can improve its travel time by changing its route. A strong spatial variability of shortest paths at short time scales is thus evidence of non-stationarity,

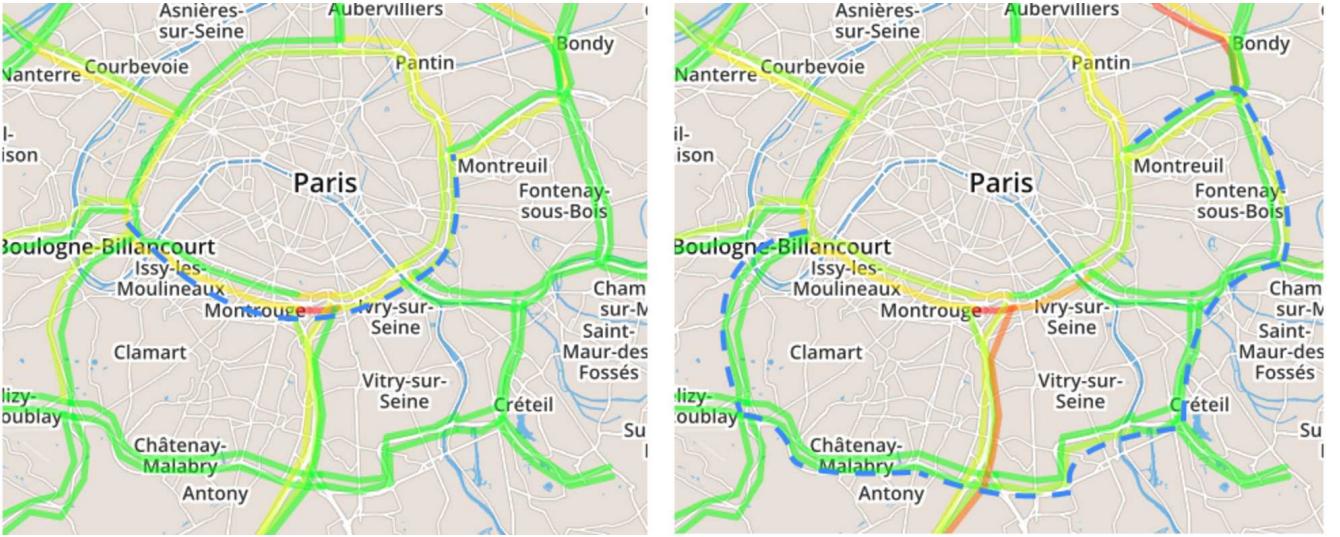


Figure 17: Spatial variability of travel-time shortest path (shortest path trajectory in dotted blue). In an interval of only 10 minutes, between 11/02/2016 00:06 (left) and 11/02/2016 00:16 (right), the shortest path between *Porte d'Auteuil* (West) and *Porte de Bagnolet* (East), increases in effective distance of $\simeq 37\text{km}$ (with an increase in travel time of only 6min), due to a strong disruption on the ring of Paris.

since a similar user will take a few time after a totally different route and not contribute to the same flow as a previous user. Such a variability is indeed observed on a non-negligible number of paths on each day of the dataset. We show in Figure 17 an example of extreme spatial variation of shortest path for a particular Origin-Destination pair.

The systematic exploration of travel time variability across the whole dataset, and associated travel distance, confirms, as described in Figure 3, that travel time absolute variability has often high values of its maximum across OD pairs, up to 25 minutes with a temporal local mean around 10min. Corresponding spatial variability produces detours up to 35km.

STABILITY OF NETWORK MEASURES The variability of potential trajectories observed in the previous section can be confirmed by studying the variability of network properties. In particular, network topological measures capture global patterns of a transportation network. Centrality and node connectivity measures are classical indicators in transportation network description as recalled in [Bavoux et al., 2005]. The transportation literature has developed elaborated and operational network measures, such as network robustness measures to identify critical links and measure overall network resilience to disruptions (an example among many is the Network Trip Robustness index introduced in [Sullivan et al., 2010]).

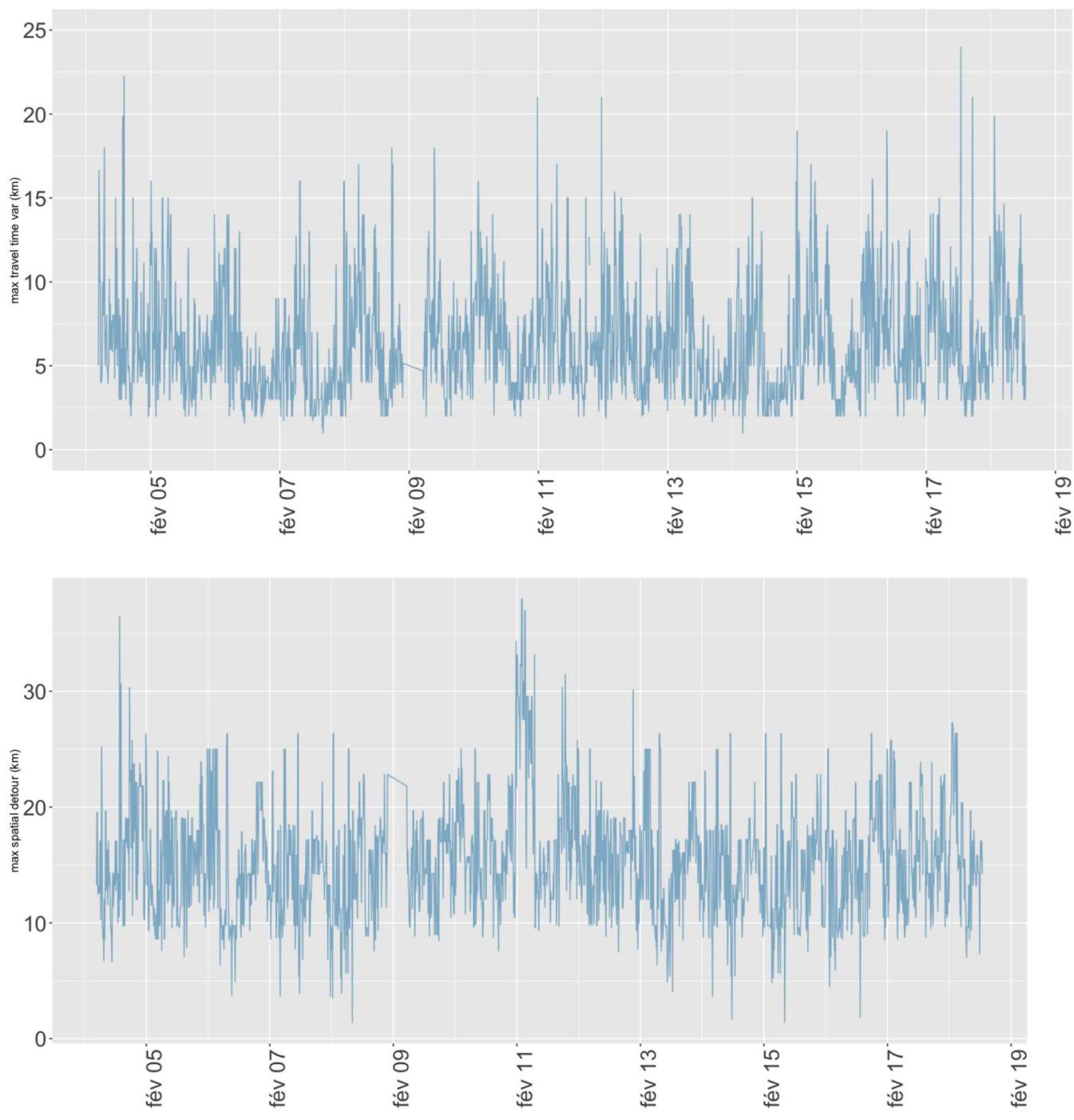


Figure 18: Travel time (top) in min and corresponding travel distance (bottom) maximal variability on a two weeks sample. We plot the maximal on all OD pairs of the absolute variability between two consecutive time steps. Peak hours imply a high time travel variability up to 25 minutes and a path length variability up to 35km.

More precisely, we study the betweenness centrality of the transportation network, defined for a node as the number of shortest paths going through the node, i.e. by the equation

$$b_i = \frac{1}{N(N-1)} \cdot \sum_{o \neq d \in V} \mathbb{1}_{i \in p(o \rightarrow d)} \quad (2)$$

where V is the set of network vertices of size N , and $p(o \rightarrow d)$ is the set of nodes on the shortest path between vertices o and d (the shortest path being computed with effective travel times). This index is more relevant to our purpose than other measures of centrality such as closeness centrality that does not include potential congestion as betweenness centrality does.

We show in Figure 4 the relative absolute variation of maximal betweenness centrality for the same time window than previous empirical indicators. More precisely we plot the value of

$$\Delta b(t) = \frac{|\max_i(b_i(t + \Delta t)) - \max_i(b_i(t))|}{\max_i(b_i(t))} \quad (3)$$

where Δt is the time step of the dataset (the smallest time window on which we can capture variability). This absolute relative variation has a direct meaning : a variation of 20% (which is attained a significant number of times as shown in Fig. ??) means that in case of a negative variation, at least this proportion of potential travels have changed route and the local potential congestion has decrease of the same proportion. In the case of a positive variation, a single node has captured at least 20% of travels. Under the assumption (that we do not try to verify in this work and assume to be also not verified as shown by [Zhu and Levinson, 2015], but that we use as a tool to give an idea of the concrete meaning of betweenness variability) that users rationally take the shortest path and assuming that a majority of travels are realized such a variation in centrality imply a similar variation in effective flows, leading to the conclusion that they can not be stationary in time (at least at a scale larger than Δt) nor in space.

SPATIAL HETEROGENEITY OF EQUILIBRIUM To obtain a different insight into spatial variability of congestion patterns, we propose to use an index of spatial autocorrelation, the Moran index (defined e.g. in [Tsai, 2005]). More generally used in spatial analysis with diverse applications from the study of urban form to the quantification of segregation, it can be applied to any spatial variable. It allows to establish neighborhood relations and unveils spatial local consistence of an equilibrium if applied on localized traffic variable. At a given point in space, local autocorrelation for variable c is computed by

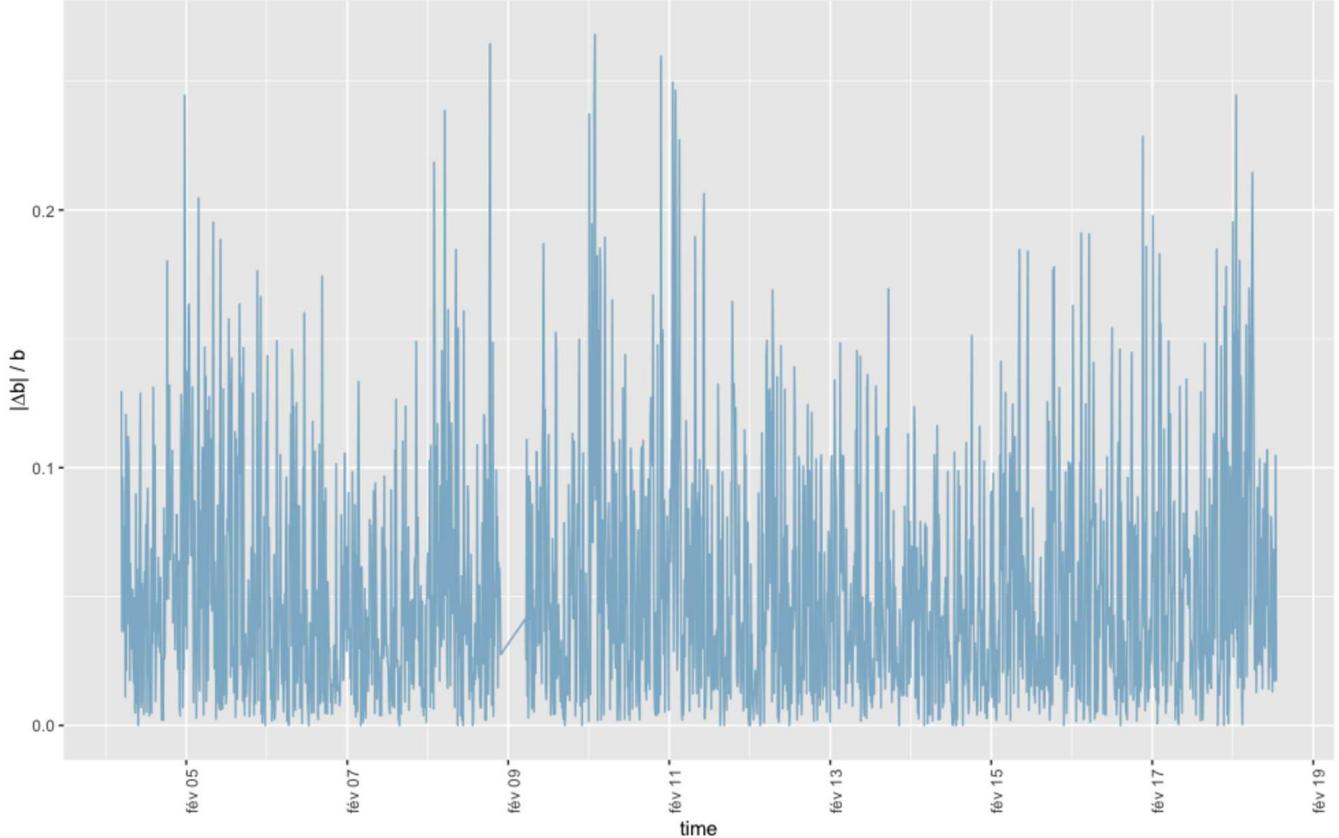


Figure 19: Temporal stability of maximal betweenness centrality. We plot in time the normalized derivative of maximal betweenness centrality, that expresses its relative variations at each time step. The maximal value up to 25% correspond to very strong network disruption on the concerned link, as it means that at least this proportion of travelers assumed to take this link in previous conditions should take a totally different path.

$$\rho_i = \frac{1}{K} \cdot \sum_{i \neq j} w_{ij} \cdot (c_i - \bar{c})(c_j - \bar{c}) \quad (4)$$

where K is a normalization constant equal to the sum of spatial weights times variable variance and \bar{c} is variable mean. In our case, we take spatial weights of the form $w_{ij} = \exp\left(\frac{-d_{ij}}{d_0}\right)$ with d_0 typical decay distance and compute the autocorrelation of link congestion localized at link center. We capture therefore spatial correlations within a radius of same order than decay distance around the point i . The mean on all points yields spatial autocorrelation index I . A stationarity in flows should yield some temporal stability of the index.

Figure 20 presents temporal evolution of spatial autocorrelation for congestion. As expected, we have a strong decrease of autocorrelation with distance decay parameter, for both amplitude and temporal average. The high temporal variability implies short time scales for potential stationarity windows. When comparing with congestion (fitted to plot scale for readability) for 1km decay, we observe that high correlations coincide with off-peak hours, whereas peaks involve vanishing correlations. Our interpretation, combined with the observed variability of spatial patterns, is that peak hours correspond to chaotic behaviour of the system, as jams can emerge in any link: correlation thus vanishes as feasible phase space for a chaotic dynamical system is filled by trajectories in an uniform way what is equivalent to apparently independent random relative speeds.

We have described an empirical study aimed at a simple but from our point of view necessary investigation of the existence of the static user equilibrium, more precisely of its stationarity in space and time on a metropolitan highway network. We constructed by data collection a traffic congestion dataset for the highway network of Greater Paris on 3 months with two minutes temporal granularity. The interactive exploration of the dataset with a web application allowing spatio-temporal data visualization helped to guide quantitative studies. Spatio-temporal variability of shortest paths and of network topology, in particular betweenness centrality, revealed that stationarity assumptions do not hold in general, what was confirmed by the study of spatial autocorrelation of network congestion. We suggest that our findings highlight a general need of higher connections between theoretical and empirical studies, as our work can discard misunderstandings on the theoretical static user equilibrium framework and guide the choice of potential applications.

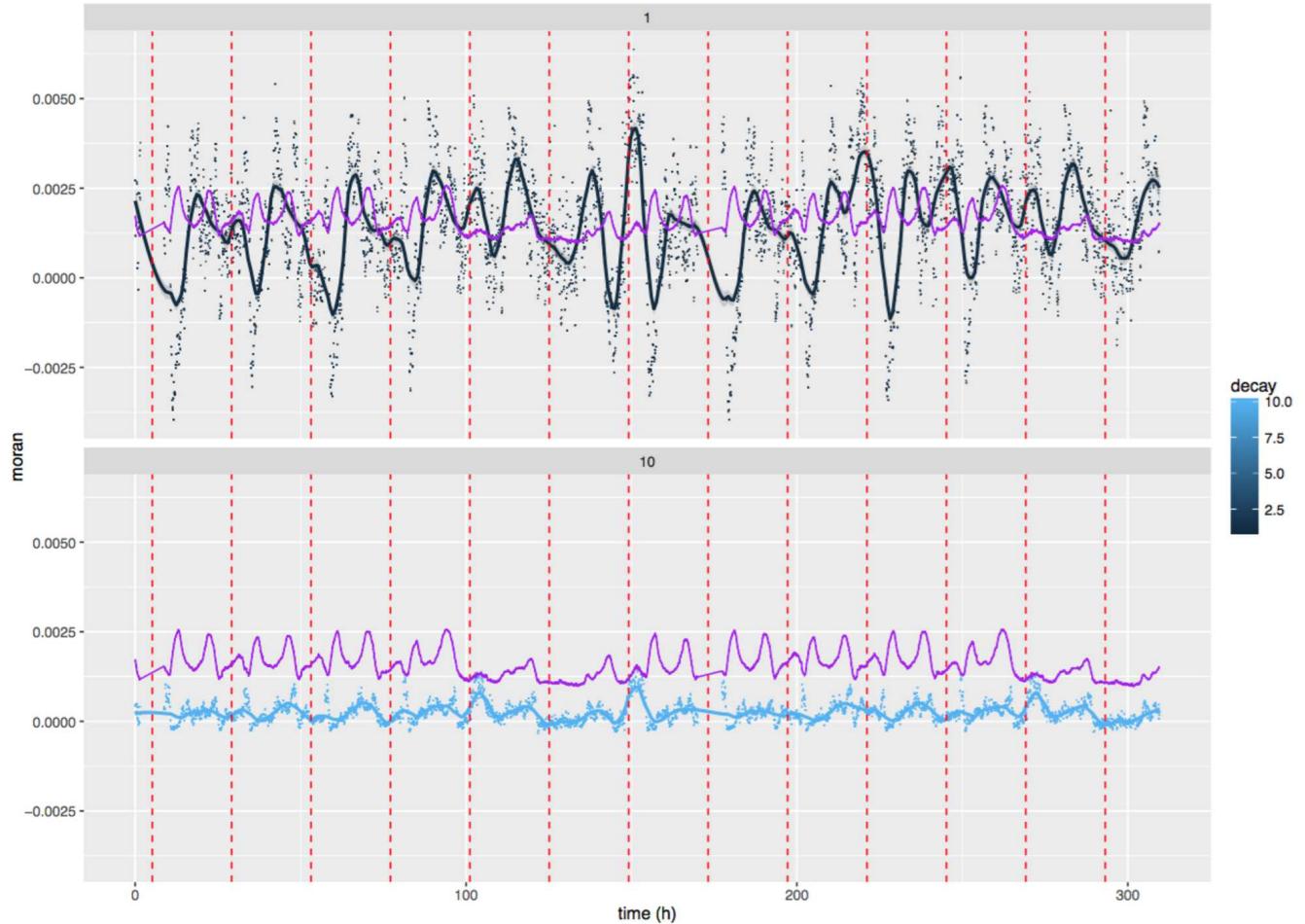


Figure 20: Spatial auto-correlations for relative travel speed on two weeks. We plot for varying value of decay parameter (1,10km) values of auto-correlation index in time. Intermediate values of decay parameter yield a rather continuous deformation between the two curves. Points are smoothed with a 2h span to ease reading. Vertical dotted lines correspond to midnight each day. Purple curve is relative speed fitted at scale to have a correspondence between auto-correlation variations and peak hours.

Perspective

Nous avons ainsi détaillé dans cette section certains enjeux liés à la reproductibilité et à la science ouverte, complétant notre positionnement spécifiques en termes de modélisation à un positionnement plus général correspondant à la pratique scientifique.

Nous allons finalement dans la dernière section qui suit encore monter en généralité et préciser nos positionnement épistémologiques, c'est-à-dire concernant les disciplines elles-mêmes et la production de connaissance. Cette étape sera cruciale, puisque notre positionnement au regard des systèmes sociaux et des systèmes biologiques permettra d'introduire les éléments fondamentaux pour une définition plus complète de la co-évolution.



3.3 EPISTEMOLOGICAL POSITIONING

The last section of this chapter aims at clarifying our epistemological positioning, since it has only been sketched at different points previously. Such a positioning is never harmless, since it strongly conditions the approaches, experiments and the interpretation of results: as [Morin, 1980] recalls, a positioning that pretends to be objective by rejecting any subjective component is much more biased than a conscious subjective approach.

The points we wish to develop can be put into both a vertical perspective in terms of levels of abstraction and in a perspective of scientific domains: linearly, we first give the general epistemological context (typical to history of science, at a medium abstraction level), then switch at a less generic level to conceptually precise our particular objects (epistemology of the living and of the social), and finally take a broader perspective at the level of knowledge production itself (epistemology of complexity).

3.3.1 Cognitive approach and perspectivism

Our epistemological positioning relies on a cognitive approach to science, given by Giere in [Giere, 2010b]. The approach focuses on the role of cognitive agents as carriers and producers of knowledge. It has been shown to be operational by [Giere, 2010a] that studies an agent-based model of science. These ideas converge with CHAVALARIAS' Nobel Game [Chavalarias, 2016] which tests through a stylized model the balance between exploration and falsification in the collective scientific enterprise.

This epistemological positioning has been presented by GIERE as *scientific perspectivism* [Giere, 2010c], which main feature is to consider any scientific enterprise as a *perspective* in which *agents* use *media* (models) to represent something with a certain purpose. To make it more concrete, we can position it within Hacking's "check-list" of constructivism [Hacking, 1999], a practical tool to position an epistemological position within a simplified three dimensional space which dimensions are different aspects on which realist approaches and constructivist approach generally diverge: first the contingency (path-dependency of the knowledge construction process) is necessary in the pluralist perspectivist approach which assumes parallel paths of knowledge construction. Secondly the "degree of constructivism" is quite high because agents produce knowledge. Finally, concerning the endogenous or exogenous explanation of the stability of theories, this stability depends on the complex interaction between the agents and their perspectives, and is thus strongly endogenous, close to the positioning of constructivism. It was presented for these reasons as an intermediate and alternative way between absolute realism and skep-

tical constructivism [Brown, 2009]. The concept of *perspective* will play thus a central role in the framework developed in 8.3.

Since this approach puts the emphasis on auto-organization, we consider it to be fully compatible with an anarchist view of science as advocated by [Feyerabend, 1993]. He formulates doubts on the relevance of political anarchism but introduces *scientific anarchism*, which must not be understood as a full refusal of any “objective” method, but of an artificial authority and legitimacy that some scientific methods or currents would like to impose. He demonstrates through a precise analysis of Galileo’s work that most of his results were based on beliefs and that most were not accessible with the current tools and methods at that time, and postulates that a similar logic should apply to contemporary works. There is thus no *perspective* that is objectively more legitimate than others as soon as they are evidence-based and peer review validated - and even in this case legitimacy should be questionable, since questioning is one foundation of knowledge. It corresponds exactly to the plurality of perspectives we defend.

Assuming an auto-organization and emergence of knowledge can be interpreted as a priority given to the *bottom-up* construction of paradigms, trying to take some distance with preconceptions or dogmas that impose a top-down view. In other words, it is similar to practicing the scientific anarchism proposed by FEVERABEND. Indeed, anarchist positioning have found a very relevant echo in the different currents of complexity, from cybernetics to self-organization during the 20th century [Duda, 2013]. Our knowledge framework developed in 8.3 illustrates this emergence of knowledge. Moreover, our will for reflexivity and to give to this work diverse reading paths beyond linearity (see Appendix F), shows the application of these principles. Methodological recommendations and positioning given previously in this chapter could sound as totalitarian if they were given roughly out of context, but these are indeed exactly the contrary since they sprout from a recent dynamic of open science which is well bottom-up founded, and in part a consequence of opening and plurality.

3.3.2 *From life to culture*

Biological systems and social systems

The parallel between social and biological systems is not rare, sometimes more from an analogy perspective as for example in WEST’s *Scaling* theory which applies similar growth equations starting from scaling laws, with however inverse conclusions concerning the relation between size and pace of life [Bettencourt et al., 2007]. Scaling relations do not hold when we try to apply them to a single ant, and they must be applied to the whole ant colony which is then the organism studied. When adding the property of cognition, we confirm that it is the relevant level, since the colony shows advanced cognitive

properties, such as the resolution of spatial optimization problems, or the quick answer to an external perturbation. Human social organizations, cities, could be seen as organisms ? [Banos, 2013] extends the metaphor of the *urban anthill* but recalls that the parallel stops quickly. We will however see to what extent some concepts from the epistemology of biology can be useful to understand social systems that we propose to study.

We start from the fundamental contribution of MONOD in [Monod, 1970], which aims at developing crucial epistemological principles for the study of life. Thus, living organisms answer to three essential properties that differentiate them from other systems: (i) the teleonomy , i.e. the property that these are “objects with a project”, project that is reflected in their structure and the structure of artifacts they produce²⁹; (ii) the importance of morphogenetic processes in their constitution (see 5.1); (iii) the property of the invariant reproduction of information defining their structure. MONOD furthermore sketches in conclusion some paths towards a theory of cultural evolution. Teleonomy is crucial in social structures, since any organization aims at satisfying a set of objectives, even if in general it will not succeed and the objectives will co-evolve with the organization. This notion of multi-objective optimization is typical of complex socio-technical systems, and will be more crucial than for biological systems.

Moreover, we postulate that the concept of morphogenesis is an essential tool to understand these systems, with a definition very similar to the one used in biology. A more thorough work to build this definition is done in 5.1, that we will sum up as the existence of relatively autonomous processes guiding the growth of the system and implying causal circular relations between form and function, that witness an emergent architecture. For social systems, isolating the system is more difficult and the notion of boundary will be less strict than for a biological system, but we will indeed find this link between form and function, such as for example the structure of an organization that impacts its functionalities.

Finally, the reproduction of information is at the core of cultural evolution, through the transmission of culture and *memetics*, the difference being that the ratio of scales between the frequency of transmission and mutation and cross-over processes or other non-memetic processes of cultural production is relatively low, whereas is many orders of magnitude in biology.

An example shows that the parallel is not always absurd : [Gabora and Steel, 2017] proposes an auto-catalytic network model for cognition, that would explain the apparition of cultural evolution through processes that are analogous to the ones that occurred at the apparition of life, i.e. a transition allowing the molecules to be self-sustained

²⁹ That must not be mistaken with teleology, typical of animist thoughts, that consists in giving a project or a meaning to the universe.

and to self-reproduce, mental representations being the analogous of molecules.

But even if processes are at the origin analogous, the nature of evolution is then quite different, as show [Leeuw, Lane, and Read, 2009], darwinian criteria for evolution being not sufficient to explain the evolution of our organized societies. This is a complexity of a different nature in which the role of information flows is crucial (see the role of informational complexity in the next subsection).

One point that also must retain our attention is the greater difficulty to define levels of emergence for social systems: [Roth, 2009] underlines the risk to fall into ontological dead-ends if levels were badly defined. He argues that more generally we must go past the single dichotomy micro-macro that is used as a caricature of the concepts of weak emergence, and that ontologies must often be multi-level and imply multiple intermediate levels.

This last question must also be put into perspective with the problem of the existence of strong emergence in social structures, that in sociological terms corresponds to the idea of the existence of “collective beings” [Angeletti and Berlan, 2015]. MORIN indeed distinguishes living systems of the second type (multi-cellular) and of the third type (social structures), but precises that the *subjects* of the latest are necessarily unachieved[Morin, 1980] (p. 852). Thus, emergences from the biological to the social are analogous by stay fundamentally different.

Co-evolution

This positioning on biological and social systems finds a direct echo for the concept of co-evolution. It indeed comes from biology, where it was developed following the concept of evolution, to be used more recently in social sciences and humanities. To what extent the concept was transferred ? Is there a parallel similar to the one between biological evolution and cultural evolution ? We propose, in order to answer these questions, to develop a brief multidisciplinary point of view on co-evolution³⁰. We will in the following review a broad spectrum of disciplines, starting from biology where the concept originated to progressively come to disciplines closer to territorial sciences.

³⁰ The approach here is slightly different from the one lead in 5.1 in the case of morphogenesis, that will be *interdisciplinary* in the sens that it aims at integrating approaches, whereas we stay here in an overview of concepts and thus more in a *multidisciplinary* approach. The concept of *co-evolution* being key for our empirical work in the following, we will therefore give an original characterization to it, and make the choice to not go into an integrative syncretism for this concept, but indeed to approach it from a *geographical point of view*, and even more precisely in the frame of territorial systems. We could postulate a congruence between the empirical and modeling specialization and the one for theory, reading our process of knowledge production in a particular profile of knowledge domains dynamics (see 8.3).

Biology

The concept of co-evolution in biology is an extension of the well-known concept of *evolution*, that can be tracked back to DARWIN. [Durham, 1991] (p. 22) recalls the components and systemic structures that are necessary to have evolution³¹.

1. Process of *transmission*, implying transmission units and transmission mechanisms.
2. Process of *transformation*, that necessitates sources of variation.
3. Isolation of sub-systems such that the effects of previous processes are observable in differentiations.

This way, a population submitted to constraints (often conceptually synthesized as a *fitness*) that condition the transmission of the genetic heritage of individuals (transmission), and to random genetic mutations (transformation), will indeed be in evolution in the spatial territories it populates (isolation), and by extension the species to which it can be associated.

Co-evolution is then defined as an evolutionary change in a characteristic of individuals of a population, in response to a change in a second population, which in turn responds by evolution to the change in the first, as synthesized by [Janzen, 1980]. This author furthermore highlights the subtlety of the concept and warns against its unjustified uses: the presence of a congruence between two characteristics that seem adapted one to the other does not necessarily imply a co-evolution, since one species could have adapted alone to one characteristic already present in the other.

This rough presentation partly hides the real complexity of ecosystems: populations are embedded in trophic networks and environments, and co-evolutionary interactions would imply communities of populations from diverse species, as presented by [Strauss, Sahli, and Conner, 2005] under the appellation of diffuse co-evolution. Similarly, spatio-temporal dynamics are crucial in the realization of these processes: [Dybdahl and Lively, 1996] study for example the influence of the spatial distribution on patterns of co-evolution for a snail and its parasite, and show that a higher speed of genetic diffusion in space for the parasite drive the co-evolutionary dynamics.

The essential concepts to retain from the biological point of view are thus: (i) existence of evolution processes, in particular transmission and transformation; (ii) in circular schemas between populations in the case of co-evolution; and (iii) in a complex territorial frame (spatio-temporal and environmental in the sense of the rest of the ecosystem).

³¹ And in that general context, evolution is not restricted to the biology of life and the presence of genes, but also to physical systems verifying these conditions. We will come back to that later.

Cultural evolution

This development on co-evolution was brought by the parallel between biological and social systems. The evolution of culture is theorized within a proper field, and witnesses many co-evolutive dynamics. [Mesoudi, 2017] recalls the state of knowledge on the subject and future issues, such as the relation with the cumulative nature of culture, the influence of demography in evolution processes, or the construction of phylogenetic methods allowing to reconstruct branches of past evolutionary trees.

To give an example, [Carrignon, Montanier, and Rubio-Campillo, 2015] introduces a conceptual frame for the co-evolution of culture and commerce in the case of ancient societies for which there are archeological data, and proposes its implementation with a multi-agent model which dynamics are partly validated by the study of stylized facts produced by the model. The co-evolution is here indeed taken in the sense of a mutual adaptation of socio-spatial structures, at comparable time scales, in this more general frame of cultural evolution.

Cultural evolution would even be indissociable from genetic evolution, since [Durham, 1991] postulates and illustrates a strong link between the two, that would themselves be in co-evolution. [Bull, Holland, and Blackmore, 2000] explores a stylized model including two types of replicant populations (genes and memes) and shows the existence of phase transitions for the results of the genetic evolution process when the interaction with the cultural replicant is strong.

Sociology

The concept was used in sociology and related disciplines such as organisation studies, following the parallel done before the same way as cultural evolution. In the field of the study of organisations, [Volberda and Lewin, 2003] develop a conceptual frame of inter-organisational co-evolution in relation with internal management processes, but deplore the absence of empirical studies aiming at quantifying this co-evolution. In the context of production systems management, [Tolio et al., 2010] conceptualize an intelligent production chain where product, process and the production system must be in co-evolution.

Economic geography

In economic geography, the concept of co-evolution has also largely been used. The idea of evolutionary entities in economy comes in opposition to the neo-classical current which remains a majority, but finds a more and more relevant echo [Nelson and Winter, 2009]. [Schamp, 2010] proceeds to an epistemological analysis of the use of co-evolution, and opposes the view of a neo-schumpeterian approach to economy

which considers the emergence of populations that evolve from micro-economic rules (what would correspond to a direct and relatively isolationist reading of biological evolution) to a systemic approach that would consider the economy as an evolutive system in a global perspective (what would correspond to diffuse co-evolution that we previously developed), to propose a precise characterization that would correspond to the first case, assuming co-evolving *institutions*. The most important for our purpose is that he underlines the crucial aspect of the choice of populations and of considered entities, of the geographical area, and highlights the importance of the existence of causal circular relations.

Diverse examples of application can be given. [Wal and Boschma, 2011] introduce a conceptual frame to allow to conciliate the evolutionary nature of companies, the theory of clusters and knowledge networks, in which the co-evolution between networks and companies is central, and which is defined as a circular causality between different characteristics of these subsystems. [Colletis, 2010] introduces a framework for the co-evolution of territories and technology (questioning for example the role of proximity on innovations), that reveals again the importance of the institutional aspect. The framework proposed by [Ter Wal and Boschma, 2011] couples the evolutionary approach to companies, the literature on industries and innovation in clusters, and the approach through complex networks of connexions between the latest in the territorial system.

In environmental economics, [Kallis, 2007] show that "broad" approaches (that can consider most of co-dynamics as co-evolutive) are opposed to stricter approaches (in the spirit of the definition given by [Schamp, 2010]), and that in any case a precise definition, not necessarily coming from biology, must be given, in particular for the search of an empirical characterization.

Geography

For geography, as we already presented in introduction, the works that are the closest to notions of co-evolution empirically and theoretically are closely linked to the evolutive urban theory. It is not easy to track back in the literature at what time the notion was clearly formalized, but it is clear that it was present since the foundations of the theory as recalls DENISE PUMAIN (see D.3): the complex adaptive system is composed of subsystems that are interdependent in a complex way, often with circular causalities. The first models indeed include this vision in an implicit way, but co-evolution is not explicitly highlighted or precisely defined, in terms that would be quantifiable or structurally identifiable. [Paulus, 2004] brings empirical proofs of mechanisms of co-evolution through the study of the evolution of economic profiles of French cities. The interpretation used by [Schmitt, 2014] is based on an entry by the evolutive urban theory, and funda-

mentally consists in a reading of systems of cities as highly interdependent entities.

Physical geography

In the study of landscapes, [Sheeren et al., 2015] evoke the co-evolution of landscape and agricultural activities, but in fact do not consider any circular effect of one on the other. Their result show a priori that the evolution of agricultural practices yield an evolution of the landscape, and it is not clear to what extent the conceptual frame of co-evolution, evoked without any more details, is used.

Physics

Finally, we can mention in an anecdotal way that the term of co-evolution has also been used by physics. Its use for physical systems may induce some debates, depending if we suppose or not that the transmission assumes a transmission of *information*³². In the case of a purely physical ontological transmission (*physical beings*), then a large part of physical systems are evolutive. [Hopkins et al., 2008] develop a cosmological frame for the co-evolution of cosmic heterogenous objects which presence and dynamics are difficultly explained by more classical theories (some types of galaxies, quasars, supermassive black holes). [Antonioni and Cardillo, 2017] study the co-evolution between synchronisation and cooperation properties within a Kuramoto oscillators network³³, showing on the one hand that the concept can be applied to abstract objects, and on the other hand that a complex network of relations between variables can be at the origin of dynamics witnessing circular causalities, i.e. a co-evolution in that sense.

Synthesis

Most of these approaches fit in the theory of complex adaptive systems developed by HOLLAND, in particular in [Holland, 2012]: it takes any system as an imbrication of systems of boundaries, that filter signals or objects. Within a given limit, the corresponding subsystem is relatively autonomous from the outside, and is called an *ecological niche*, in a direct correspondence with highly connected communities

³² Information is defined within the shanonian theory as an occurrence probability for a chain of characters. [Morin, 1976] shows that the concept of information is indeed far more complex, and that it must be thought conjointly to a given context of the generation of a self-organizing negentropic system, i.e. realizing local decreases in entropy in particular thanks to this information. This type of system is necessarily alive. We will follow here this complex approach to information.

³³ The Kuramoto model studies synchronization within complex systems, by studying the evolution of phases θ_i coupled by interaction equations $\dot{\theta}_i = \vec{\omega} + \vec{W}[\vec{\theta}] + \vec{B}$ where $\vec{\omega}$ are proper forcing phases and the coupling strength between i and j is given by $\vec{W}_{ij} = \sum_j w_{ij} \sin(\theta_i - \theta_j)$ and \vec{B} is noise.

within trophic or ecological networks. This way, interdependent entities within a niche are said to be co-evolving. We will come back on that approach in our theoretical construction in 8.2 when we will have developed other concepts that are necessary for it.

We retain from this multidisciplinary view of co-evolution the fundamental following points, that are precursors of a proper definition of co-evolution that will be given further, concluding the first part.

1. The presence of *evolution processes* is primary, and their definition is almost always based on the existence of transmission and transformation processes.
2. Co-evolution assumes entities or systems, belonging to distinct classes, which evolutive dynamics are coupled in a circular causal way. Approaches can differ depending on the assumptions of populations of these entities, singular objects, or components of a global system then in mutual interdependency without a direct circularity.
3. The delineation of systems and subsystems, both in the ontological space (definition of studied objects), but also in space and time, and their distribution in these spaces, is fundamental for the existence of co-evolutionary dynamics, and it seems in a large number of cases, of their empirical characterization.

3.3.3 *Nature of complexity and knowledge production*

The two previous epistemological points that we just developed were related respectively related first to the positioning in itself, i.e. the framework to read processes of production of scientific knowledge, and then to the nature of the concepts considered. We propose to again gain in generality compared to the first one and to introduce a development modestly contributing (i.e. in our context) to the *knowledge of knowledge*. The aim is to interrogate the links between complexity and processes of knowledge production.

One aspect of knowledge production on complex systems, that we encounter several times here (see chapter 8), and that seems to be recurrent and even inevitable, is a certain level of reflexivity (and that would be inherent to complex system in comparison to simple systems, as we will develop further). We mean by this term both a practical reflexivity, i.e. a necessity to increase the level of abstraction, such as the need to reconstruct in an endogenous way the disciplines in which a reflexion aims at positioning as proposed in 2.2, or to reflect on the epistemological nature of modeling when constructing a model such as in B.5, but also a theoretical reflexivity in the sense that theoretical apparels or produced concepts can apply recursively to themselves. This practical observation can be related to old epistemological debates questioning the possibility of an objective knowledge

of the universe that would be independent of our cognitive structure, somehow opposed to the necessity of an “evolutive rationality” implying that our cognitive system, product of the evolution, mirrors the complex processes that led to its emergence, and that any knowledge structure will be consequently reflexive³⁴. We do not pretend here to bring a response to such a broad and vague question as such, but we propose a potential link between this reflexivity and the nature of complexity.

Complexity and complexities

What is meant by complexity of a system often leads to misunderstandings since it can be qualified according to different dimensions and visions. We distinguish first the complexity in the sense of weak emergence and autonomy between the different levels of a system, and on which different positions can be developed as in [Deffuant et al., 2015]. We will not enter a finer granularity, the vision of social complexity giving even more nightmares to the Laplace daemon, and since it can be understood as a stronger emergence (in the sense of weak and strong emergence as developed before in 3.1). We thus simplify and assume that the nature of systems plays a secondary role in our reflexion, and therefore consider complexity in the sense of an emergence.

Moreover, we distinguish two other “types” of complexity, namely computational complexity and informational complexity, that can be seen as measures of complexity, but that are not directly equivalent to emergence, since there exists no systematic link between the three. We can for example consider the use of a simulation model, for which interactions between elementary agents translate as a coded message at the upper level: it is then possible by exploiting the degrees of freedom to minimize the quantity of information contained in the message. The different languages require different cognitive efforts and compress the information in a different way, having different levels of measurable complexity [Febres, Jaffé, and Gershenson, 2013]. In a similar way, architectural artefacts are the result of a process of natural and cultural evolution, and witness more or less this trajectory.

Numerous other conceptual or operational characterizations of complexity exist, and it is clear that the scientific community has not converged on a unique definition [Chu, 2008]³⁵. We propose to focus on these three concepts in particular, for which the relations are already not evident.

³⁴ We thank here D. PUMAIN to have formulated this alternative view on the problem that we will develop in the following.

³⁵ In an approach that is in a way reflexive, [Chu, 2008] proposes to continue exploring the different existing approaches, as proxies of complexity in the case of an essentialism, or as concepts in themselves. The complexity should emerge naturally from the interaction between these different approaches studying complexity, hence the reflexivity.

Indeed, links between these three types of complexity are not systematic, and depend on the type of system. Epistemological links can however be introduced. We will develop the links between emergence and the two other complexities, since the link between computational complexity and informational complexity is relatively well explored, and corresponds to issues in the compression of information and signal processing, or moreover in cryptography.

Computational complexity and emergence

Different clues suggest a certain necessity of computational complexity to have emergence in complex systems, whereas reciprocally a certain number of adaptive complex systems have high computational capabilities.

A first link where computational complexity implies emergence is suggested by an algorithmic study of fundamental problems in quantum physics. Indeed, [Bolotin, 2014] shows that the resolution of the Schrödinger equation with any Hamiltonian is a NP-hard and NP-complete problem, and thus that the acceptance of $P \neq NP$ implies a qualitative separation between the microscopic quantum level and the macroscopic level of the observation. Therefore, it is indeed the complexity (here in the sense of their computation) of interactions in a system and its environment that implies the apparent collapse of the wave function, what rejoins the approach of GELL-MANN by quantum decoherence [Gell-Mann and Hartle, 1996], which explains that probabilities can only be associated to decoherent histories (in which correlations have led the system to follow a trajectory at the macroscopic scale)³⁶. The paradox of the Schrödinger cat appears then as a fundamentally reductionist perspective, since it assumes that the superposition of states can propagate through the successive levels and that there would be no emergence, in the sense of the constitution of an autonomous upper level. In other terms, the work of [Bolotin, 2014] suggests that computational complexity is sufficient for the presence of emergence³⁷.

³⁶ The *Quantum Measurement Problem* arises when we consider a microscopic wave function giving the state of a system that can be the superposition of several states, and consists in a theoretical paradox, on the one hand the measures being always deterministic whereas the system has probabilities for states, and on the other hand the issue of the non-existence of superposed macroscopic states (collapse of the wave function). As reviewed by [Schlosshauer, 2005], different epistemological interpretations of quantum physics are linked to different explanations of this paradox, including the “classical” Copenhagen one which attributes to the act of observation the role of collapsing the wave function. GELL-MANN recalls that this interpretation is not absurd since it is indeed the correlations between the quantum object and the world that product the decoherent history, but that it is far more specific, and that the collapse happens in the emergence itself: the cat is either dead or living, but not both, before we open the box.

³⁷ This effective separation of scales does not a priori imply that the lower level does not play a crucial role, since [Vattay et al., 2015] proves that the properties of quan-

Reciprocally, the link between computational complexity and emergence is revealed by questions linked to the nature of computation [Moore and Mertens, 2011]. Cellular automata, that are moreover crucial for the understanding of several complex systems, have been shown as Turing-complete³⁸, such as the Game of Life [Beer, 2004]³⁹. Some organisms without a central nervous system are capable of solving difficult decisional problems [Reid et al., 2016]. An ant-based algorithm is shown by [Pintea, Pop, and Chira, 2017] as solving a Generalized Travelling Salesman Problem (GTSP), problem which is NP-difficult. This fundamental link had already been conceived by TURING, since beyond his fundamental contributions to contemporary computer science, he studied morphogenesis and tried to produce chemical models to explain it [Turing, 1952b] (that were far from actually explaining it - it is still not well understood today, see 5.1 - but which conceptual contributions were fundamental, in particular for the notion of reaction-diffusion). We moreover know that a minimum of complexity in terms of constituting interactions in a particular case of agent-based system (models of boolean networks), and thus in terms of possible emergences, implies a lower bound on computational complexity, which becomes significant as soon as interactions with the environment are added [Tošić and Ordóñez, 2017].

Informational complexity and emergence

Informational complexity, or the quantity of information contained in a system and the way it is stored, also bears some fundamental links with emergence. Information is equivalent to the entropy of a system and thus to its degree of organisation - this what allows to solve the apparent paradox of the Maxwell Daemon that would be able to diminish the entropy of an isolated system and thus contradict the second law of thermodynamics: it indeed uses the information on positions and velocities of molecules of the system, and its action balances to loss of entropy through its captation of information⁴⁰.

tum criticality are typical of molecules of the living, without a priori any specificity for life in this complex determination by lower scales: [Verlinde, 2017] has recently introduced a new approach linking quantum theories and general relativity in which it is shown that gravity is an emergent phenomenon and that path-dependency in the deformation of the original space introduces a supplementary term at the macroscopic level, that allows to explain deviations attributed up to now to *dark matter*.

³⁸ A system is said to be Turing-complete if it is able to compute the same functions than a Turing machine, commonly accepted as all what is "computable" (CHURCH's thesis). We recall that a Turing machine is a finite automaton with an infinite writing band [Moore and Mertens, 2011].

³⁹ There even exists a programming language allowing to code in the *Game of Life*, available at <https://github.com/QuestForTetris>. Its genesis finds its origin in a challenge posted on *codegolf* aiming at the conception of a Tetris, and ended in an extremely advanced collaborative project.

⁴⁰ The Maxwell Daemon is more than an intellectual construction: [Cottet et al., 2017] implements experimentally a daemon at the quantic level.

This notion of local increase in entropy has been largely studied by CHUA under the form of the *Local Activity Principle*, which is introduced as a third principle of thermodynamics, allowing to explain with mathematical arguments the self-organization for a certain class of complex systems that typically involve reaction-diffusion equations [Mainzer and Chua, 2013].

The way information is stored and compressed is essential for life, since the ADN is indeed an information storage system, whose role at different levels is far from being fully understood. Cultural complexity also witnesses of an information storage at different levels, for example within individuals but also within artefacts and institutions, and information flows that necessarily deal with the two other types of complexities. Information flows are essential for self-organization in a multi-agent system. Collective behaviors of fishes or birds are typical examples used to illustrate emergence and belong to the canonical examples of complex systems. We only begin to understand how these flows structure the system, and what are the spatial patterns of information transfer within a *flock* for example: [Crosato et al., 2017] introduce first empirical results with transfer entropy for fishes and lay the methodological basis of this kind of studies.

Knowledge production

We know have enough material to come to reflexivity. It is possible to position knowledge production at the intersection of interactions between types of complexity developed above. First of all, knowledge as we consider it can not be dissociated from a collective construction, and implies thus an encoding and a transmission of information: it is at another level all problematics linked to scientific communication. The production of knowledge thus necessitates this first interaction between computational complexity and informational complexity. The link between informational complexity and emergence is introduced if we consider the establishment of knowledge as a morphogenetic process. It is shown in 5.1 that the link between form and function is fundamental in psychology: we can interpret it as a link between information and meaning, since semantics of a cognitive object can not be considered without a function. HOFSTADER recalls in [Hofstadter, 1980] the importance of symbols at different levels for the emergence of a thought, that consist in signals at an intermediate level. Finally, the last relation between computational complexity and emergence is the one allowing us a positioning in particular on knowledge production on complex systems, the previous links being applicable to any type of knowledge.

Therefore, any *knowledge of the complex* embraces not only all complexities and their relations in its content, but also in its nature as we just showed. The structure of knowledge in terms of complexity is analog to the structure of systems its studies. We postulate that

this structural correspondence implies a certain recursivity, and thus a certain level of *reflexivity* (in the sens of knowledge of itself and its own conditions).

We can try to extend to reflexivity in terms of a reflexion on the disciplinary positioning: following [Pumain, 2005], the complexity of an approach is also linked to the diversity of viewpoints that are necessary to construct it. To reach this new type of complexity⁴¹, that would be a supplementary dimension linked to the knowledge of complex systems, reflexivity must be at the core of the approach. [Read, Lane, and Leeuw, 2009] recall that innovation has been made possible when societies reached the ability to produce and diffuse innovation on their own structure, i.e when they were able to reach a certain level of reflexivity. The *knowledge of the complex* would thus be the product and the support of its own evolution thanks to reflexivity which played a fundamental role in the evolution of the cognitive system: we could thus suggest to gather these considerations, as proposed by PUMAIN, as a new epistemological notion of *evolutive rationality*.

To conclude, we can remark that given the law of *requisite complexity*, proposed by [Gershenson, 2015] as an extension of *requisite variety* [Ashby, 1991]⁴², the *knowledge of the complex* will necessarily have to be a *complex knowledge*. This other point of view reinforces the necessity of reflexivity, since following MORIN (see for example [Morin, 1991] on the production of knowledge), the *knowledge of knowledge* is central in the construction of a complex thinking.

Practical implications

To conclude this epistemological section, we propose to synthesize all the ideas introduced as concrete manifestations that directly yield from them, and that strongly condition all the forms and semantics of knowledge introduced in the following. These directions (that we will not go up to name principles since they are only at the state of sketch) can be grouped into three large families: modeling practices, Open Science practices, and epistemology. On the domain of model-

⁴¹ For which links with the previous types naturally appear: for example, [Gell-Mann, 1995] considers the effective complexity as an *Algorithmic Information Content* (close to Kolmogorov complexity) of a Complex Adaptive System *which is observing an other* Complex Adaptive System, what gives their importance to informational and computational complexities and suggests the importance of the observational viewpoint, and by extension of their combination - what furthermore must be related to the perspectivist approach of complex sciences presented above.

⁴² One of the crucial principles of cybernetics, the *requisite variety*, postulates that to control a system having a certain number of states, the controller must have at least as much states. GERSHENSON proposes a conceptual extension of complexity, which can be justified for example by [Allen, Stacey, and Bar-Yam, 2017] which introduce the multi-scale *requisite variety*, showing the compatibility with a theory of complexity based on information theory.

ing practices, in each section emerge different axis that are more or less complementary:

- Modeling, which will be in most cases equivalent to simulation, must be understood as an indirect tool of knowledge on processes within a complex system or on its structure (according to the section on “why modeling”), and models will necessarily have to be complex (following the reflexion on the different types of complexity) in the sense that they capture a phenomenon of weak emergence, but still respecting constraints of parsimony.
- The exploration of models is fully contained in the modeling enterprise (see reproducibility), and intensive computation is a cornerstone to efficiently explore simulation models (see intensive computation). Sensitivity analysis methods must be questioned and extended if needed (as illustrates the example of the sensitivity to space).
- As suggested by the perspectivist positioning, the coupling of models will have to play a crucial role in the capture of complexity.

Concerning open science, we can extract the following points:

- The necessity of all measures linked to open science to allow the construction always more complex models, towards the co-construction of models by different disciplines.
- In this frame, the full opening of source code, together with its readability are crucial. The complete explication of the model in the scientific reporting, and a self-sustaining code documentation, are two aspect of it.
- The question of open data is not negotiable in that frame. The quasi-totality of our treatments is based on initially open data, and when it is not the case we work at an aggregated level for which data can be opened. Constructed open data are open.
- Concerning the methods of interactive exploration, which are an aspect of opening science, we develop some, but stay limited compared to the ideal requirement that these should be fully compatible with a reproducible approach.

Finally, from the epistemological point of view, we can also find “practical” implications that will naturally be more implicit in our approach, but not less structuring:

- Our inspiration will essentially be interdisciplinary and will aim at combining different points of view.

- Different knowledge domains (notion that we will precise in 8.3, but that we can understand for now in the sense of theoretical, modeling and empirical domains introduced by [Livet et al., 2010]) can not be dissociated for any approach of scientific production, and we will use them in a strongly dependent way.
- Our approach will have to imply a certain level of reflexivity.
- The construction of a complex knowledge ([Morin, 1991]) is neither inductive nor deductive, but constructive in the idea of a morphogenesis of knowledge: it can be for example difficult to clearly identify precise “scientific deadlocks” since this metaphor assumes that an already constructed problem has to be unlocked, and even to constrain notions, concepts, objects or models in strict analytical frameworks, by categorizing them following a fixed classification, whereas the issue is to understand if the construction of categories is relevant. Doing it a posteriori is similar to a negation of the circularity and recursivity of knowledge production. The elaboration of ways to report that translate the diachronic character and the evolutive properties of it is an open problem.

★ ★

★

CHAPTER CONCLUSION

La lecture d'un article ou d'un ouvrage est toujours bien plus éclairante lorsqu'on connaît personnellement l'auteur, d'une part car on peut profiter des *private joke* et extrapolier certains développements des narrations qui se doivent synthétiques (même si l'art de l'écriture est justement d'essayer de transmettre la majorité de ces éléments, l'ambiance en quelque sorte), et d'autre part car la personnalité a des implications complexes sur la manière d'appréhender la nature de la connaissance et une certaine structure a priori du monde. Pour cela, la connaissance scientifique serait très probablement moins riche si elle était produite par des machines aux capacités cognitives équivalentes, aux connaissances et expériences empiriques subjectives équivalentes et aussi diverses que celles humaines, mais qui auraient été programmées pour minimiser l'impact de leur personnalité et de leur convictions sur l'écriture et la communication (toujours en supposant qu'elles aient une certaine forme de données et fonctions plus ou moins équivalentes). Dans ces laboratoires de recherche dignes de *Blade Runner*, nous doutons que la production d'une connaissance du complexe serait effectivement possible, puisqu'il manquerait à ces machines justement la *rationalité évolutive* développée en 3.3, et nous doutons fortement que celle-ci puisse être produite du moins dans l'état des connaissances actuelles en intelligence artificielle.

Le but de ce chapitre était donc "de faire connaissance" sur les points de positionnements incontournables pour l'ensemble de notre réflexion. Ceux-ci en sont d'autant plus cruciaux car conditionnent très fortement certaines directions de recherche.

Notre positionnement sur la reproductibilité développé en 3.2 implique certains choix de modélisation, notamment l'utilisation unique de plateformes ouvertes, de workflow et d'implémentations ouverts ; il implique aussi un choix de données qui se doivent au maximum d'être accessibles ou rendues accessibles, et donc certains choix d'objets et d'ontologie, ou plutôt le non-choix de certains : nos problématiques pourraient être mobilisées sur des données d'entreprise fines tout en gardant une cohérence avec l'approche théorique et thématique (la théorie évolutive des villes a largement mobilisé ce type d'étude comme par exemple [Paulus, 2004]), mais la relative fermeture de ce type de données ne les rend pas utilisables dans notre démarche.

Ensuite, notre positionnement sur le rôle du calcul intensif et les besoins d'exploration des modèles 3.1 est source de l'ensemble des expériences numériques et des méthodologies utilisées ou développées.

Enfin, notre positionnement épistémologique 3.3 percole dans l'ensemble de notre travail, et permet de poser les premières briques pour des

formalisations théoriques plus systématiques qui seront développées en chapitre 8.

CONCLUSION OF PART I: A DEFINITION OF CO-EVOLUTION

This first part allows us to formulate much more precisely our research question. Indeed, the first chapter allowed us to draw a sketch of the diversity of processes involved and of the temporal and spatial scales concerned. The second chapter gave us a very general view of existing modeling approaches and of their precise scientific context. Finally, the third chapter positions the question in an epistemological way, shed some light on co-evolution through a multi-disciplinary perspective, and clarifies the complexity with which we are dealing. It allows us to open on the directions to take in order to lead successfully the project of modeling the co-evolution.

Defining co-evolution

After the literature review given in 2.1, that includes different degrees of coupling between components of networks and territories, we are first able to precise what will be meant by *modeling co-evolution*, by giving a definition of co-evolution in view of the multidisciplinary overview given in 3.3.

We propose the following entry for the specific case of transportation networks and territories, which echoes to the three main points (existence of evolutive processes, definition of entities or populations, isolation of subsystems in space and time) that we gave in 3.3. It verifies the three following specifications.

First of all, evolutive processes correspond to transformations of components of the territorial system at the different scales: transformation of cities on the long time, of their networks, transmission between cities of socio-economic characteristics carries by microscopic agents but also cultural transmission, reproduction and transformation of agents themselves (firms, households, operators)⁴³.

These evolutive processes may imply a co-evolution. Within a territorial system, can simultaneously co-evolve: (i) given entities (a given infrastructure and given characteristics of a given territory for example, i.e. individuals), when their mutual influence will be circularly causal (at the corresponding scale); (ii) populations of entities, what will be translated for example as such type of infrastructure and given territorial components co-evolve at a statistical level in a given geo-

⁴³ This list is based on assumptions of the evolutive urban theory that we already briefly introduced and that we will develop in itself in Chapter 4. It can not be exhaustive, since what would be the "ADN of a city" remains an open question as recalls DENISE PUMAIN in a dedicated interview D.3.

graphical region; (iii) all the components of a system at a small geographical scale when there exists strong global interdependencies. Our approach is thus fundamentally *multi-scale* and articulates different significations at different scales.

Finally, the constraint of an isolation implies, in relation with the previous point, that co-evolution and the articulation of significations will have a meaning if there exists spatio-temporal isolations of subsystems in which different co-evolutions operate, what is directly in accordance with a vision in *Multi-scalar systems of systems*.

This extended definition will constitute our reference in the following when we will evoke the co-evolution of transportation networks and territories.

We can then synthesize the fundamental results of this first part in the two following significant facts:

1. The hypothesis of co-evolution of transportation networks and territories is supported from a theoretical and thematic point of view, et we construct a precise definition for it.
2. Co-evolution remains relatively poorly explored in the literature of urban modeling, the characteristic of concerned disciplines and their interactions being a potential cause for it.

We develop now the perspective that open at this stage.

On the need of an empirical characterization

The broadest signification, i.e. generalized interdependency, is rapidly limited if its patterns are not finely characterized. It allows as an epistemological premise to consider certain ontologies and certain modeling approaches, but allows difficultly to finely understand the structure and processes of a system. The object will be then to decrease in generality and consider subsystems, in which we can consider the co-evolution of entities and of population. An understanding at this level necessitates a fine empirical characterization, without which our distinction would have no sense. A question that opens, and that we will tackle in the following, is then which are the possible empirical methods to characterize a co-evolution between entities or populations of entities.

Two complementary tracks

The state of the art done in 2.1 above witnesses a weakness in the literature in the domain of strong coupling between the evolution of territories and network growth, given the restricted range and the disparity of reviewed works. The gap to fill on this point would thus be linked to the introduction of models strongly coupled in time more

or less multi-processes and multi-scale, for which a part of the models described in 2.1 then in 2.3 are precursors.

The first exploratory research we will lead will have to answer to different conceptual tensions that result from the conclusions we just obtained:

- allowing both an empirical approach, and more particularly a characterization method, and a modeling approach;
- allowing to take into account different scales;
- allowing the inclusion of ontologies for territories and for networks that are not always directly compatible.

The scales will especially be a mesoscopic and a macroscopic scale since as we suggested in 3.2 with the study of traffic flows, and as shows [Yasmin, Morency, and Roorda, 2017] for the validation of an activity model, the microscopic scale witnesses complex trajectories that are difficult to reproduce.

We will choose to answer simultaneously to these different problematics with an original strategy of a double thematic entry.

Part II

BRIQUES ÉLÉMENTAIRES

This part provides building blocks for the final objective of constructing models of co-evolution. These contain both stylized facts from empirical analyses and toy and hybrid modeling. They correspond to three distinct components of our overall construction: first analyses at the micro-scale confirming the chaotic and non-stationary nature of interactions between networks and territories, secondly a morphogenetic vision of these that corresponds roughly to a meso-scale, and finally an application of the evolutive urban theory at the macro-scale.

INTRODUCTION DE LA PARTIE II

He finally would have realized his trip. No cities, or a very few. Which soul in these perpendicular streets and avenues, that we necessarily discover by car. Fill the tank again, maybe it's on purpose, just for the charm of the fuel smell. Anyway it would be funny to look at what these stations have to say, to keep in mind. A running return journey to Mount Elbert, then Longs Peak. Soon out of Colorado, goodbye my gummy bears. Damn it, so close to Denver, it could be worth it. Never mind, the mountains are calling and I must go, as someone used to say. What do we finally know of a territory as a consequences of our so selective discoveries ? A very narrow band on the spectrum of scales ? A tiny spatial extent: one supplementary dimension is not invented so easily. Maybe at least an awakening of conscience for antagonisms, of dualities. And the conscience to necessarily choose one of the aspects each time. To build bridges one must be prepared. To see the world with an eye catching multiple views, one must already have understood, i.e. subjectively integrated, the corresponding processes. Memory of one of the first serious routes: Meije ridge traverse, 23 hours without interruption to end with hallucinations on the path to follow whereas the sparkles of crampons on the scree were not anymore enough to light the fallen night. This concrete feeling of the void on each side which imposes to grope, anchors in the subconscious even before reaching the stage of hallucinations: we travel at each moment a fine line, which is as much that of the arbitrariness of road trips as that of the bridges which difficultly resist the flood. On this ridge, anchors that are of course strong but also heterogenous are a pledge of life: diversity overcomes adversity.

A paradox which is intrinsic to numerous knowledge production approaches is a need of an intrinsic consistence and of a reasonable reach of explication for concerned phenomena, which is opposed to an inevitable reduction of explored dimensions, but also to the fragility of bridges that it aims at creating towards other corpora of knowledge. The image we took above suggests that groping, i.e. a step-by-step progression without precipitation, and also the solidity of anchors, are solid assets to tackle this paradox.

This part directly opens thematic directions of answer for modeling co-evolution that we mentioned when concluding the first part, and thus builds these strong anchors. It however constructs basements without going into the heart of the problem in a spirit of robustness through progressive entries, and constructs therefore the *elementary bricks* of our approach. Two chapters deal thus successively with the following thematics:

1. A first chapter focuses on the evolutive urban theory, which is a privileged entry on urban systems from an evolutive point of view, and integrates in its core a multi-scalar approach to these systems. It unveils fundamental properties of territorial systems implied by the evolutive theory, by introducing a first empirical analysis of the spatial variability of interactions between urban form and network topology, then by developing a methodology to statistically characterize co-evolution (in the intermediate sense of population). It introduces then a first model of interaction between urban system and flows of the transportation network, with a static network.
2. A second chapter explores the concept of morphogenesis, which allows a conceptual entry to the characteristic of modularity necessary to have co-evolution. After having developed an interdisciplinary definition of morphogenesis, it introduces a model of urban morphogenesis based on aggregation-diffusion processes for population density, and is then sequentially coupled to a network generation model.

* * *

*

MATHEMATICAL PRELIMINARIES

In order to be readable by the largest audience possible, we propose to precise in the preliminary interlude the definitions of notions or key methods that will be regularly used in the following, often out of a mathematical framework. This choice allows to keep a rigorous frame without making indigestible the reading of this manuscript to a large part of its legitimate audience. Without noted otherwise, the specifications given here will be the reference during the use of the corresponding terms.

Statistics

We will denote by $\mathbb{P}[\cdot]$ a probability, $\mathbb{E}[\cdot]$ an expectation, $\hat{\mathbb{E}}[\cdot]$ an associated estimator, and $\text{Cov}[\cdot, \cdot]$ a covariance.

CORRELATION Unless otherwise stated, we will estimate the covariance between two processes with a Pearson estimator, i.e. if $(X_i, Y_i)_i$ is a set of observations of processes X, Y , the correlation is estimated by

$$\hat{\rho} = \frac{\hat{\text{Cov}}[X, Y]}{\sqrt{\hat{\text{Cov}}[X] \cdot \hat{\text{Cov}}[Y]}}$$

where the covariance is estimated with the unbiased estimator $\hat{\text{Cov}}$.

GRANGER CAUSALITY A multi-dimensional time-series $\vec{X}(t)$ exhibits a Granger causality if with

$$\vec{X}(t) = \mathbf{A} \cdot (\vec{X}(t-\tau))_{\tau>0} + \varepsilon$$

there exists τ, i such that $a_{i\tau} > 0$ significantly. We will use a weak version of Granger causality, i.e. a test on lagged correlations defined by

$$\rho_\tau [X_i, X_j] = \hat{\rho} [X_i(t-\tau), X_j(t)]$$

with τ lag or advance. This will allow us to quantify relations between random variables defined in space and time.

GEOGRAPHICALLY WEIGHTED REGRESSION The Geographically Weighted Regression is an estimation technique for statistical models

that allow to take into account the spatial non-stationarity of processes. If Y_i is an explicited variable and X_i a set of explicative variables, measures in the same points in space, we estimate a model $Y_i = f(X_i, \vec{x}_i)$ at each point \vec{x}_i , by taking into account the observations with spatial weighting around the point, where weights are fixed by a kernel that can take several forms, for example an exponential kernel is of the form

$$w_i(\vec{x}) = \exp(-\|\vec{x} - \vec{x}_i\|/d_0)$$

The stationarity scale assumed by the model is then of the same order as d_0 . It can be adjusted by cross-validation for example.

MACHINE LEARNING We will designate by *Supervised Learning* any method to estimate a relation between variables $Y = f(X)$ where the value of Y is known on a data sample. This will be a classification when the variable is discrete. Non-supervised classification consists in constructing Y when only X is given. In order to classify data, we will use a basic technique which gives good results on data without an exotic structure: the method of *k-means*, repeated a sufficient number of times to take into account its stochastic character. The complexity of *k-means* is in average polynomial, even if the exact solution of the partitioning problem is NP-hard.

OVERFITTING The issue of *overfitting* is particularly important during the estimation of models, since a too large number of parameters can lead to the capture of the realization noise as a structure. During the estimation of statistical models, information criteria can be used to quantify the gain in information produced by the addition of a parameter, and obtain a compromise between performance and parsimony.

The *Akaike Information Criteria* (AIC) allows to quantify the gain in information allowed by the addition of parameters in a model. For a statistical model which has a Likelihood function, the AIC is then defined by

$$AIC = 2k - 2 \ln \mathcal{L}$$

if k is the number of parameters in the model and \mathcal{L} the maximal value of the likelihood function. [Akaike, 1998] shows that this expression corresponds to an estimation of the gain in Kullback-Leibler information. A correction for small samples of size n is given by

$$AICc = 2 \cdot \left(k + \frac{k^2 + k}{n - k - 1} - \ln \mathcal{L} \right)$$

A similar criteria but derived within a bayesian framework is the *Bayesian Information Criterion* (BIC) [Burnham and Anderson, 2003],

which leads to a stronger penalty for the number of parameters: $BIC = \ln n \cdot k - 2 \ln \mathcal{L}$.

These criteria are applied to model selection by studying their differences between models (only differences have a meaning, since they are defined with an arbitrary constant): the “best” model is the one having the lowest criterion. In the case of models with comparable performances, it can be more relevant to combine models with the Akaike weights $w_i = \exp(-\Delta AIC/2)$.

This issue of overfitting is also implicit in the case of simulation models, but to the best of our knowledge there does not exist an established method allowing to tackle it.

Stochastic processes: stationarity

Stationarity properties inform on the variability of the distribution of a stochastic process. Let $(\vec{X}_i)_{i \in I}$ a multidimensional stochastic process. It will be said to be strongly stationary if its law does not depends on i , i.e. if $\mathbb{P}[\vec{X}_i] = \mathbb{P}[\vec{X}_{i+1}]$. Strong stationarity implies the equality of all moments for all i .

We will use a weaker notion of stationarity for stochastic processes, or *Weak Stationarity*, which uses the first two moments: $(\vec{X}_i)_{i \in I}$ is weakly stationary if

1. $\mathbb{E}[\vec{X}_i] = \mathbb{E}[\vec{X}_0]$ for all i
2. $\text{Cov}[\vec{X}_i, \vec{X}_j]$ depends only on $i - j$

We can have a weak stationarity at the first order if only the condition on the expectation is verified, and at the second order if there is also the condition on autocovariance [Zhang and Zhou, 2014].

Exploration of simulation models

We will designate by simulation model any algorithm that associates a realization $\mathcal{M}[\vec{x}, \vec{\alpha}]$ to data \vec{x} given parameters $\vec{\alpha}$. The question is then to understand the behavior of the model in an empirical way, by simulating it, possibly with several repetitions for the same parameters if it is stochastic. It is then for example possible to calibrate the model, i.e. find a set of parameters allowing to fulfill given objectives (that can be distances to observed data).

Experience plan by sampling

The dimensionality curse corresponds to the fact that the size of the parameter space is exponential in the number of parameters. When it increases but we want to keep an overview of the behavior of a model

on a large variety of input parameters, we can sample the space with a given number of points.

The Latin Hypercube Sampling (LHS) allows to ensure that for each dimension, the full range of values is covered when generated points are projected on the dimension. The Sobol sampling allows to generate point clouds with a weak discrepancy (see B.4 for a precise definition of discrepancy, that should be understood as a covering of space), and is particularly suited for the computation of integrals.

Sampling can become cumbersome if the model is very irregular, or for a precise calibration objective. Therefore, there exists specific algorithms for exploration and calibration, for which we can give some examples.

Genetic algorithm calibration

Genetic algorithms are an alternative largely used in optimization, and are more generally a case of evolutionary computation meta-heuristics [Rey-Coyrehourcq, 2015]. We will generally use for the calibration of models the standard algorithm implemented in OpenMole, described in details by [Pumain and Reuillon, 2017b]. It is a stochastic extension of the NSGA2 algorithm for multi-objective optimization. It has the following main characteristics:

- given a population of parameters that are candidates as solutions of the multi-objective problem, the Pareto front is determined as the non-dominated points;
- a set is constructed from this front by taking a constraint of diversity into account;
- an offspring is generated from this set by crossovers and mutations, and evaluated for its performance;
- the algorithm iterates on the new population.

[Pumain and Reuillon, 2017b] adds the objective of the number of replications to the objectives of the algorithm, in order to take into account stochasticity and find a compromise between optimality and robustness of solutions.

Specific algorithms

Based on genetic algorithms, various algorithms have been proposed to refine model exploration. We can mention two examples developed in the frame of OpenMole: the *Pattern Space Exploration* algorithm (PSE) [Chérel, Cottineau, and Reuillon, 2015] aims at discovering the set of outputs of a model, in the idea of a search for all feasible behaviors. The *Calibration Profile* algorithm [Reuillon et al., 2015] aims on the other hand at establishing the necessary character of a parameter to fulfill an objective, independently of other parameters.

★ ★

★

4

CO-ÉVOLUTION : UNE ENTRÉE PAR LA THÉORIE ÉVOLUTIVE URBAINE

The study of interactions between transportation networks and territories can be studied from the standpoints of urban systems. Did the opening of the first High Speed Line in France between Paris and Lyon have an impact on the concerned territorial dynamics ? [Bonnafous, 1987] shows that it could have had some at the regional scale, in particular areas, as for example tourism in Burgundy. Did it have effects on the long time, beyond the decade ? At which scales, following which processes ? We rejoin the question of *structuring effects*, that we evoked in chapter 1 through a multi-scalar entry (micro, meso and macro), and also through the progressive development of the idea of co-evolution. These characteristics are indeed at the core of the evolutive urban theory, of which we propose therefore here to detail implications for our problematic.

After having recalled in preliminary the essential characteristics of the evolutive urban theory, we study in a first section at the mesoscopic scale the interactions between territories and networks, that we capture in morphological indicators for each, and for which we study the spatial correlations.

We then introduce the dynamical aspect by studying the notion of spatio-temporal causality in section 4.2. The multiple configurations highlighted for a simple urban growth model that strongly couples network growth and density, that we will designate as *causality regimes*, witness of circular causalities which are indeed markers of a co-evolution. The application to the case of rail network growth and urban populations in South Africa shows that this method empirically allows to reveal different regimes. This method is crucial on the one hand from a methodological point of view through the introduction of an original method allowing in some cases to better understand the respective influences between territories and networks, but also from a thematic point of view concerning the empirical presence of a co-evolution.

We finally explore in a last section 4.3 the possibilities offered by interaction models coming from the evolutive urban theory, at a small spatial scale and a long time scale, what suggests the existence of network effects in an indirect way, without even introducing co-evolution aspects in a first time.

This way, we build the first building bricks for different aspects of interactions and of co-evolution between networks and territories, in particular in the empirical domain for the characterization of co-

evolution, and in the modeling domain by the introduction of a first model relating territories and networks.

★ ★

★

This chapter is composed by various works. The first section includes a part from [Raimbault, 2018b] for the morphological analysis, and the results presented by [Raimbault, 2016a] for the analysis of correlations; the second section corresponds to the majority of [Raimbault, 2017a] for the theoretical formulation and the illustration on synthetic data, and then presents results of [Raimbault and Baffi, 2017] for the application. Finally the last section corresponds entirely to [Raimbault, 2018].

EVOLUTIVE URBAN THEORY

We have already evoked various aspects of the evolutive urban theory, in relation to complexity in geography, and then to some models of urban systems it produced. A synthesis is here necessary to precisely draw the frame in which our developments will take place. This theory has initially been introduced in [Pumain, 1997] which argues for a dynamical vision of systems of cities, in which self-organisation is crucial.

The core of the evolutive urban theory is perfectly synthesized by DENISE PUMAIN herself (interview in D.3): it is “*a geographical theory with the ambition to gather most of stylized facts known on cities and their organisation within territories, in an out-of-equilibrium and non-static perspective, by following them on long time periods and putting an emphasis on structuring factors and bifurcations.*”

Cities are interdependent evolutive spatial entities whose interrelations lead to the emergence of the macroscopic behavior at the scale of the system of cities. The system of cities is also seen as a network of cities, in correspondance with an approach through complex systems. Each city is itself a complex system in the spirit of [Berry, 1964], the multi-scalar aspect, in the sense of autonomous scales but that each have a specific role in the dynamics of the system, being essential in this theory, since microscopic agents carry processes of evolution of the system through complex retroactions between scales. The positioning of this theory within complexity approaches has later been confirmed [Pumain, 2003].

It has been shown that the evolutive urban theory gives an explanation to scaling laws, which are pervasive in urban systems¹, which would be a consequence of the diffusion of innovation cycles between cities [Pumain et al., 2006]. These have furthermore been exhibited empirically for several urban systems [Pumain, Paulus, and Vacchiani-Marcuzzo, 2009]. The notion of resilience of a system of cities, inducted by the adaptive character of these complex systems, implies that cities are drivers and incubators of social change [Pumain, 2010]. Finally, the path-dependancy is a source of non-ergodicity within these systems, making the “universal” interpretations of scaling laws developed by physicist not compatible with the evolutive urban theory [Pumain, 2010].

The evolutive urban theory has been conjointly elaborated with models of urban systems. For example the first Simpop model, described by [Sanders et al., 1997], is a multi-agent model which works with the following rules: (i) settlements are initially villages with a uniquely agricultural production, and can in time transform into commercial cities, then administrative, then eventually industrial, the

¹ We recall that a scaling laws allows to link the size of cities in terms of population P_i and an aggregated quantity Z_i , under the form $Z_i = Z_0 \cdot (P_i/P_0)^\alpha$.

transition rules depending on threshold parameters in terms of population and neighborhood resources for the industrialization; (ii) settlements produce different types of goods depending on their functions and populations; (iii) these are exchanged through the intermediary of spatial interactions (depending on distance) in order to satisfy demands; (iv) populations evolve according to the size of the city and the level of demand satisfaction. This first model allows to simulate the evolution of an urban system in a stylized way.

The Simpop2 model introduced by [Bretagnolle, Daudé, and Pumain, 2006] extends this model, allowing to include for example innovation cycles and the role of administrative boundaries in exchanges. It is applied on long time scales to urban growth patterns for Europe and the United States [Bretagnolle and Pumain, 2010b].

The most recent accomplishments of evolutive urban theory rely on the production of the ERC project GeoDiversity, presented in [Pumain and Reuillon, 2017d], which include considerable progresses from the technical point of view (OpenMole software² [Reuillon, Leclaire, and Rey-Coyrehourcq, 2013]), from the thematic point of view (knowledge issued from the SimpopLocal model [Schmitt, 2014] and the Marius model [Cottineau, 2014]), and from the methodological point of view (incremental modeling [Cottineau, Chapron, and Reuillon, 2015]). For an epistemological analysis through mixed methods of the evolutive theory, which allows to reinforce this bibliographical picture by a study of its genesis, in a sense of its *form*, refer to 8.3 which uses it as a case study to build a knowledge framework. In particular, an analysis of interviews with DENISE PUMAIN and ROMAIN REUILLOON, reveals the cross-fertilisation between geographical knowledge and computer science knowledge, allowed by the interdisciplinary effort of model development and of their exploration methods.

Implications

We can therefore consider the complexity of systems of cities in the sense of the evolutive urban theory as a morinian macro-concept [Morin, 1976], i.e. the complex combination of multiple concepts each necessary to the construction. The following concepts are thus necessary:

- Out-of-equilibrium aspect of urban systems. The spatial character of systems often leads to complex spatio-temporal dynamics, and thus properties of non-stationarity for spatio-temporal associated processes.
- Systemic dynamics, i.e. existence of a strong interdependency between cities that can be interpreted as a co-evolution (in the last sense in the definition we gave).

² <http://openmole.org/>

- Central role of interactions between cities as drivers of growth processes, existence of structure effects on the long time.

These concepts will be thus explored following different perspectives in this chapter, in the following sections:

1. From an empirical point of view, we will first study an example of non-stationarity properties of characteristics for territories and networks, and also of their interactions.
2. We introduce then from a methodological point of view an approach allowing to better understand patterns of spatio-temporal interdependency, and thus *co-evolution* that we will link to its intermediate statistical sense we gave.
3. Finally, a modeling approach allows to explore interactions between cities on the long time, in particular in relation with the network in the context of our questionings.

* * *

*

4.1 CORRELATIONS BETWEEN FORM OF TERRITORIES AND NETWORK TOPOLOGY

Through relocation processes, sometimes induced by networks, we can expect the latter to influence the distribution of populations in space. Reciprocally, network characteristics can be influenced by this distribution. We propose here to study these potential links by the intermediate of characterizations given by synthetic indicators for these two subsystems, and by correlations between these indicators.

At the scale of the system of cities, the spatial nature of the urban system is captured by cities position, associated with aggregated city variables. We will work here at the mesoscopic scale, at which the precise spatial distribution of activities is necessary to understand the spatial structure of the territorial system. We will therefore use the term of morphological characteristics for population density and the road network.

The choice of “relevant” boundaries for the territory or the city is a relatively open problem which will often depend on the question we are trying to answer [Páez and Scott, 2005]. This way, [Guérois and Paulus, 2002] show that the entities obtained are different if we consider an entry by the continuity of the built environment (morphological), by urban functions (employment area for example) or by administrative boundaries. We choose here the mesoscopic scale of a metropolitan center, of an order of one hundred kilometers, first for the relevance of the spatial field computed, and secondly because smaller scales become less relevant for the notion of urban form, whereas larger scales induce a too large variability.

At this scale, we can assume that territorial characteristics, for population and network, are locally defined et vary in an approximately continuous way in space. Thus, the construction of fields of morphological indicators will allow to endogenously reconstruct territorial entities through the emergent spatial structure of indicators at larger scales. For examples, cities should be distinguishable within non-urban spaces. The aim of this section is thus to study properties of these indicators and their interactions, and thus indirectly interactions between the territory and the network.

4.1.1 *Morphological measures*

Urban morphology

The approaches to quantify and qualify *urban form* at the considered scale, and by extension to any population distribution in space what we can call *territorial form*, are numerous.

We need however quantities having a certain level of invariance to extract typical shapes. For example, two monocentric cities, i.e. concentrated around a given point, should be measured as morphologi-

cally close by a monocentricity indicator, whereas a direct comparison of population distributions can give a very high distance³ between configurations depending on the position of centers.

We choose here to refer to the literature in urban morphology which proposes various set of indicators to describe urban form [Tsai, 2005]. [Le Néchet, 2009] recalls the necessity of a multi-dimensional measure of the urban form. It is possible to obtain a robust description with a small number of independant indicators by a reduction of the dimension [Schwarz, 2010].

Other solutions exist to quantify urban form⁴. [Guérois and Pumain, 2008] study the form of European cities using a simple measure of density slopes from the center to the periphery. It is also possible to use indexes from fractal analysis, such as for example systematically applied by [Chen, 2016] to classify urban forms. The link between urban morphology and topology of the underlying relational network has been suggested in a theoretical approach by [Badariotti, Banos, and Moreno, 2007]. Other more original indexes can be proposed, such as by [Lee et al., 2017] which use the variations of trajectories for routes going through a city to establish a classification and show that it is strongly correlated with socio-economic variables.

Note that we consider here indicators on the spatial distribution of population density only, and that more elaborated considerations on urban form can include for example the distribution of economic opportunities and the combination of these two fields through accessibility measures. For the choice of indicators, we follow the analysis done in [Le Néchet, 2015] where a morphological typology of large European cities is obtained. Its consistence suggests the ability of the indicator set used to capture urban form at this scale. We work at a comparable scale and must capture diverse aspects such as hierarchy, concentration, level of acentrism of the population distribution, hence the use of similar indicators.

Indicators

We give now the formal definition of morphological indicators. We consider gridded population data $(P_i)_{1 \leq i \leq N^2}$, write $M = N^2$ the number of cells, d_{ij} the distance between cells i, j , and $P = \sum_{i=1}^M P_i$ total population. We measure urban form using:

1. Rank-size slope γ , expressing the degree of hierarchy in the distribution, computed by fitting with Ordinary Least Squares a

³ Spatial distributions can be compared by an euclidian distance between corresponding matrices, or by more elaborated distances such as the Monge distance which solves a minimal transport problem and gives the quantity of displacements necessary to go from one distribution to the other.

⁴ In operational urbanism, urban morphology is defined as “the characteristics of the material form of cities and fabrics” [Paquot, 2010]. We use this term here for fabrics at a mesoscopic scale, seen through the spatial distribution of populations.

power law distribution by $\ln(P_{\tilde{i}}/P_0) \sim k + \gamma \cdot \ln(\tilde{i}/i_0)$ where \tilde{i} are the indexes of the distribution sorted in decreasing order (the constant k of the adjustment does not play a role in hierarchy). It is always negative, and values close to zero mean a flat distribution.

2. Entropy of the distribution [Le Néchet, 2015], which expresses how uniform the distribution is, what is a way to capture a level of concentration:

$$\mathcal{E} = \sum_{i=1}^M \frac{P_i}{P} \cdot \ln \frac{P_i}{P} \quad (5)$$

$\mathcal{E} = 0$ means that all the population is in one cell whereas $\mathcal{E} = 1$ means that the population is uniformly distributed.

3. Spatial-autocorrelation given by Moran index [Tsai, 2005], with simple spatial weights given by $w_{ij} = 1/d_{ij}$

$$I = M \cdot \frac{\sum_{i \neq j} w_{ij} (P_i - \bar{P}) \cdot (P_j - \bar{P})}{\sum_{i \neq j} w_{ij} \sum_i (P_i - \bar{P})^2}$$

Its theoretical bounds are -1 and 1 , and positive values will imply aggregation spots ("density centers"), negative values strong local variations, whereas $I = 0$ corresponds to totally random population values.

4. Average distance between individuals [Le Néchet, 2009], which captures a spatial dispersion of population and quantifies a level of acentrism (distance to a monocentric model):

$$\bar{d} = \frac{1}{d_M} \cdot \sum_{i < j} \frac{P_i P_j}{P^2} \cdot d_{ij}$$

where d_M is a normalisation constant taken as the diagonal of the area on which the indicator is computed in our case.

The first two indexes are not spatial, and are completed by the last two that take space into account. Following [Schwarz, 2010], the effective dimension of the urban form justifies the use of all.

RESULTS We compute the morphological measures given above on real urban density data, using the population density grid of the European Union at 100m resolution provided openly by Eurostat [EUROSTAT, 2014]⁵. The choice of the resolution, the spatial range, and the shape of the window on which indicators are computed, is made

⁵ This database has some precision issues that have been recognized [Bretagnolle et al., 2016] but the aggregation at a larger resolution should allow to remove possible bias.

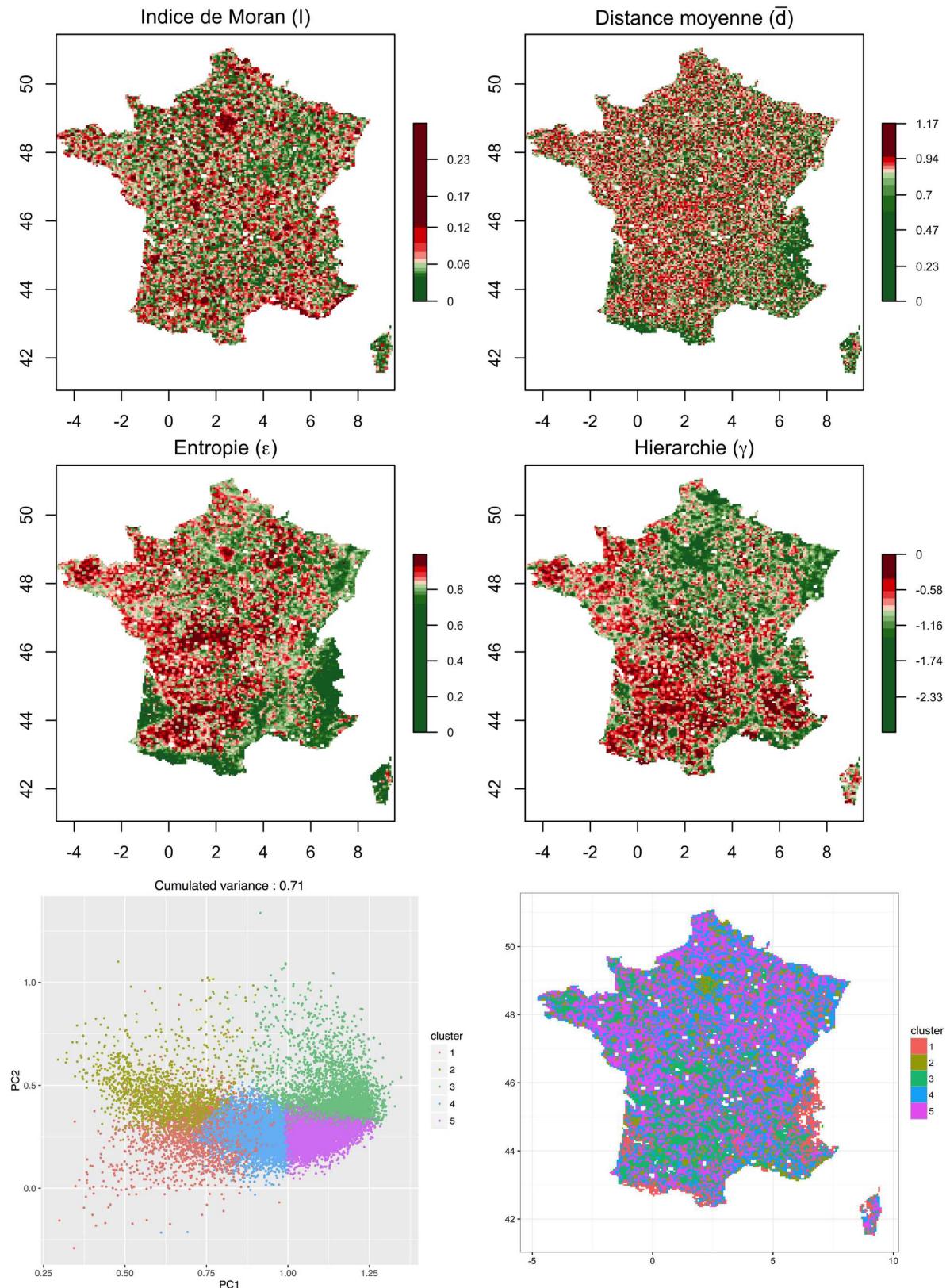


Figure 21: **Empirical values of morphological indicators.** (Top four maps) Spatial distribution of the morphological indicators for France. Scale color discretization is done using quantiles to ease map readability. (Bottom Left) Projection of morphological values on the two first components on a Principal Component analysis. Color gives cluster in an unsupervised classification (see text). (Bottom right) Spatial distribution of clusters. See text for details on the process to estimate spatial indicators and for the classification.

according to the thematic specifications given before. We consider 50km wide square windows. As it also does not make sense to have a too detailed resolution because of data quality⁶, we take $N = 100$ and aggregate the initial raster data at a 500m resolution to meet this size on real windows of size 50km. To have a rather continuous distribution of indicators in space, we overlap windows by setting an offset of 10km between each, what induces a smoothing of values and allows to limit bord effects due to the shape. We have furthermore tested the sensitivity to window size by computing samples with 30km and 100km window sizes and obtained rather similar spatial distributions, and also strong correlations between the fields and their smoothing at a finer resolution, as detailed in Appendix A.4.

The implementation of indicators must be done carefully, since computational complexities can reach $O(N^4)$ for the Moran index for example: we use convolution through Fast Fourier Transform, which is a technique allowing the computation of the Moran index with a complexity in $O(\log^2 N \cdot N^2)$ ⁷.

We show in Fig. 21 maps giving values of indicators, for France only to ease maps readability. The first striking feature is the diversity of morphological patterns across the full territory. The auto-correlation is naturally high in Metropolitan areas (Paris, Lyon, Marseille for example), with the Parisian surroundings clearly detached. When looking at other indicators, it is interesting, regarding possible areas in which a co-evolution could happen, to denote regional regimes: rural areas have much less hierarchy in the South than in the North, whereas the average distance is rather uniformly distributed except for mountain areas. Regions of very high entropy are observed in the Center and South-West.

To have a better insight into existing morphological classes, we use unsupervised classification⁸ with a simple k-means algorithm⁹. The number of clusters $k = 5$ witnesses a transition in inter-cluster variance, what means that a variation of structure occurs at this number, that we then choose as the number of clusters. The split between classes is plotted in Fig. 21, bottom-left panel, where we show measures projected on the two first components of a Principal Component

⁶ According to [Silva, Gallego, and Lavalle, 2013] which details the construction of the dataset, good results were obtained after validation for seven countries on samples with a grid of resolution 1km. We are thus closer of this resolution with a resolution of 500m.

⁷ I.e. having an execution time bounded by $\log^2 N \cdot N^2$ if N is the data size, what is a considerable gain compared to N^4 : to process a grid of width 100, the asymptotic gain factor will be approximatively 10000.

⁸ Which consists in partitioning the data space according to their endogenous structure.

⁹ Given the distribution of points which have a relatively homogenous density, alternative methods such as the DBScan algorithm are relatively equivalent. We take here a number of repetitions $b = 100$ of the algorithm to have a result robust to stochasticity.

Analysis (explaining 71% of variance, what is relatively large). The map of morphological classes confirms a North-South opposition in a background rural regime (clear green against blue), the existence of mountainous (red) and metropolitan (dark green) regimes. Such a variety of settlements forms will be the target for the model in 5.2. A similar computation of morphological indicators was done for China using the gridded population data from [Fu, Jiang, and Huang, 2014]. Maps are available in Appendix A.4.

4.1.2 Network Measures

We consider network aggregated indicators as a way to characterize transportation network properties on a given territory, the same way morphological indicators yielded information on urban structure. We propose to compute some simple indicators on same extents as for morphology, to be able to explore relations between these static measures.

Static network analysis has been extensively documented in the literature, such as for example [Louf and Barthelemy, 2014a] for a cross-sectional study of cities or [Lagesse, 2015] for the exploration of new measures for the road network. [Moosavi, 2017] uses techniques from deep learning to establish a typology of urban road networks for a large number of cities across the world.

The questions behind such approaches are multiple: they can aim at finding typologies or at characterizing spatial networks, at understanding underlying dynamical processes in order to model morphogenesis, or even at being applied in urban planning such as *Space Syntax* approaches [Hillier and Hanson, 1989]. We are positioned here more within the two first logics since we aim at characterizing the shape of networks in a first step, and then to include their dynamics in models in a second step. Our significant contribution is the characterization of the road network on large spatial extents, covering Europe and China.

Indicators

We introduce indicators to have a broad idea of the form of the network, using a certain number of indicators to capture the maximum of dimensions of properties of networks, more or less linked to their use. These indicators summarize the mesoscopic structure of the network and are computed on topological networks obtained through simplification steps that will be detailed later. If we denote the network with $N = (V, E)$, nodes have spatial positions $\bar{x}(V)$ and populations $p(v)$ obtained through an aggregation of population in the corresponding Voronoï polygon¹⁰, and edges E have *effective distances*

¹⁰ A Voronoï diagram is a partition of the plan, constructed from a point cloud. The cell associated to each point is composed by the set of points closer to it than other

$\mathbb{L}(E)$ taking into account impedances and real distances (to include the primary network hierarchy). We then use:

- Characteristics of the graph, obtained from graph theory, as defined by [Haggett and Chorley, 1970]: number of nodes $|V|$, number of links $|E|$, density d , average length of links \bar{d}_l , average clustering coefficient \bar{c} , number of components c_0 .
- Measures linked to shortest paths: diameter r , euclidian performance v_0 (defined by [Banos and Genre-Grandpierre, 2012]), average length of shortest paths \bar{l} .
- Centrality measures: these are aggregated at the level of the network by taking their average and their level of hierarchy, computed by an ordinary least squares of a rank-size law, for the following centrality measures:
 - Betweenness centrality [Crucitti, Latora, and Porta, 2006], average \bar{b}_w and hierarchy α_{bw} : given the distribution of centrality on all nodes, we take the slope of a rank-size adjustment and the average of the distribution.
 - Closeness centrality [Crucitti, Latora, and Porta, 2006], average \bar{c}_l and hierarchy α_{cl} .
 - Accessibility [Hansen, 1959], which is in our case computed as a closeness centrality weighted by populations: average \bar{a} and hierarchy α_a .

The concept of accessibility is measured here by a network indicator, since its computation implies to attribute weights to the nodes with a corresponding population, and can be interpreted than as a potential of access to the rest of the population (as we did in chapter 1). This indicator is interesting a priori since it lies at the interface between the urban form and network topology, since the distribution of population on nodes is taken into account.

Network performance is close to the rectilinearity measure (*straightness*) proposed by [Josselin, Labatut, and Mitsche, 2016], which show that it efficiently differentiate rectilinear networks and radio-concentric networks, that are both recurring urban networks.

Our indicators are conceived around network topology but not its use: developments with suited data could extend these analyses to the functional aspect of networks, such as for example performance measures computed by [Trépanier, Morency, and Agard, 2009] using massive data for a public transportation network.

points of the cloud. The graph of a Voronoï diagram is the dual of the associated Delaunay triangulation.

Data preprocessing

We work here with the road network, which structure is finely conditioned to territorial configuration of population densities. Furthermore, data for the current road network is openly available through the OpenStreetMap (OSM) project [OpenStreetMap, 2012]. Its quality was investigated for different countries such as England [Haklay, 2010] and France [Girres and Touya, 2010]. It was found to be of a quality equivalent to official surveys for the primary road network. Concerning China, although [Zheng and Zheng, 2014] underlined a quick acceleration of OSM road data completeness and accuracy, its use for computation of network indicators may be questioned at a very fine scale. [Zhang et al., 2015] highlights different regimes of data quality, partitioning China into regions among which qualitative behavior of OSM data varies. We will have to keep in mind this variability, and to ensure the robustness of results, we will simplify the network at a sufficient level of aggregation.

The network constituted by primary road segments is aggregated at the fixed granularity of the density grid to create a graph. It is then simplified to keep only the topological structure of the network, normalized indicators being relatively robust to this operation. This step is necessary for a simple computation of indicators and a thematic consistence with the density layer. We keep only the nodes with a degree strictly greater or smaller than two, and corresponding links, by taking care to aggregate the real geographical distance when constructing the corresponding topological link. Given the order of magnitude of data size (for Europe, the initial database has $\simeq 44.7 \cdot 10^6$ links, and the final simplified database $\simeq 20.4 \cdot 10^6$), a specific parallel algorithm is used, with a *split-merge* structure. It separates the space into areas that can be independently processed and then merged. It is detailed in Appendix A.4.

Results

Network indicators have been computed on the same areas than urban form indicators, in order to put them in direct correspondance and later compute the correlations. We show in Fig. 22 a sample for France.

The spatial behavior of indicators unveils local regimes as for the urban form (urban, rural, metropolitan), but also strong regional regimes. They can be due to the different agricultural practices depending on the region for the rural for example, implying a different partition of parcels and also a particular organization of their serving. For network size, Brittany is a clear outlier and rejoins urban regions, witnessing very fragmented parcels (and a fortiori also of a land property fragmentation in the simplifying assumption of corresponding parcels and properties). This is partly correlated to a low hierarchy of

accessibility. The South and the East of the extended *Bassin Parisien* are distinguishable by a strong average betweenness centrality, in accordance with a strong hierarchy of the network.

The same way as for urban form, this spatial variability suggests the search of variables regimes of interactions between indicators, as we will do for later through their correlations.

For China, for which a selection of indicators is also given in A.4, we observe even stronger local and regional variations. Highly populated urban areas detach themselves, corresponding to a particular regime.

The accessibility indicator is finally strongly correlated with the same unweighted indicator, i.e. closeness centrality: we obtain a correlation of $\rho = 0.86$ estimated on all measure points for China.

4.1.3 *Effective static correlations and non-stationarity*

Spatial correlations

Local spatial correlations are computed on windows gathering a certain number of observations, and thus of windows on which indicators have been computed. We denote by l_0 (which is equal to 10km in preceding results) the resolution of the distribution of indicators. The estimation of correlations is then done on squares of size $\delta \cdot l_0$ (with δ which can vary typically from 4 to 100). δ gives simultaneously the number of observations used for the local estimation of correlation, and the spatial range of the corresponding window. Its value thus directly influences the confidence of the estimation.

We show in Fig. 23 examples of correlations estimated with $\delta = 12$ in the case of France. With 20 indicators, the correlation matrix is significantly large in size, but the effective dimension (the number of components required to reach the majority of variance) is reduced: principal components analysis shows that 10 components already capture 62% of variance, and the first component already captures 17%, what is considerable in a space where the dimension is 190¹¹.

It is possible to examine the bloc for urban form, for the network, or for crossed correlations, which directly express a link between properties of the urban form and of the network. For example, the relation between average betweenness centrality and morphological hierarchy that we visualize allows to understand the process corresponding to the correspondance of hierarchies: a hierarchical population can induce a hierarchical network or the opposite direction, but it can also induce a distributed network or such a network create a population hierarchy - this must be well understood in terms of correspondence

¹¹ This corresponds to the dimension of the correlation matrix between 20 indicators, i.e. the number of elements of its half without the diagonal. If correlations were randomly distributed, the first component would capture $1/190 = 0.5\%$ only, and the 10 first 5%, since the variance is equally shared between independent dimensions.

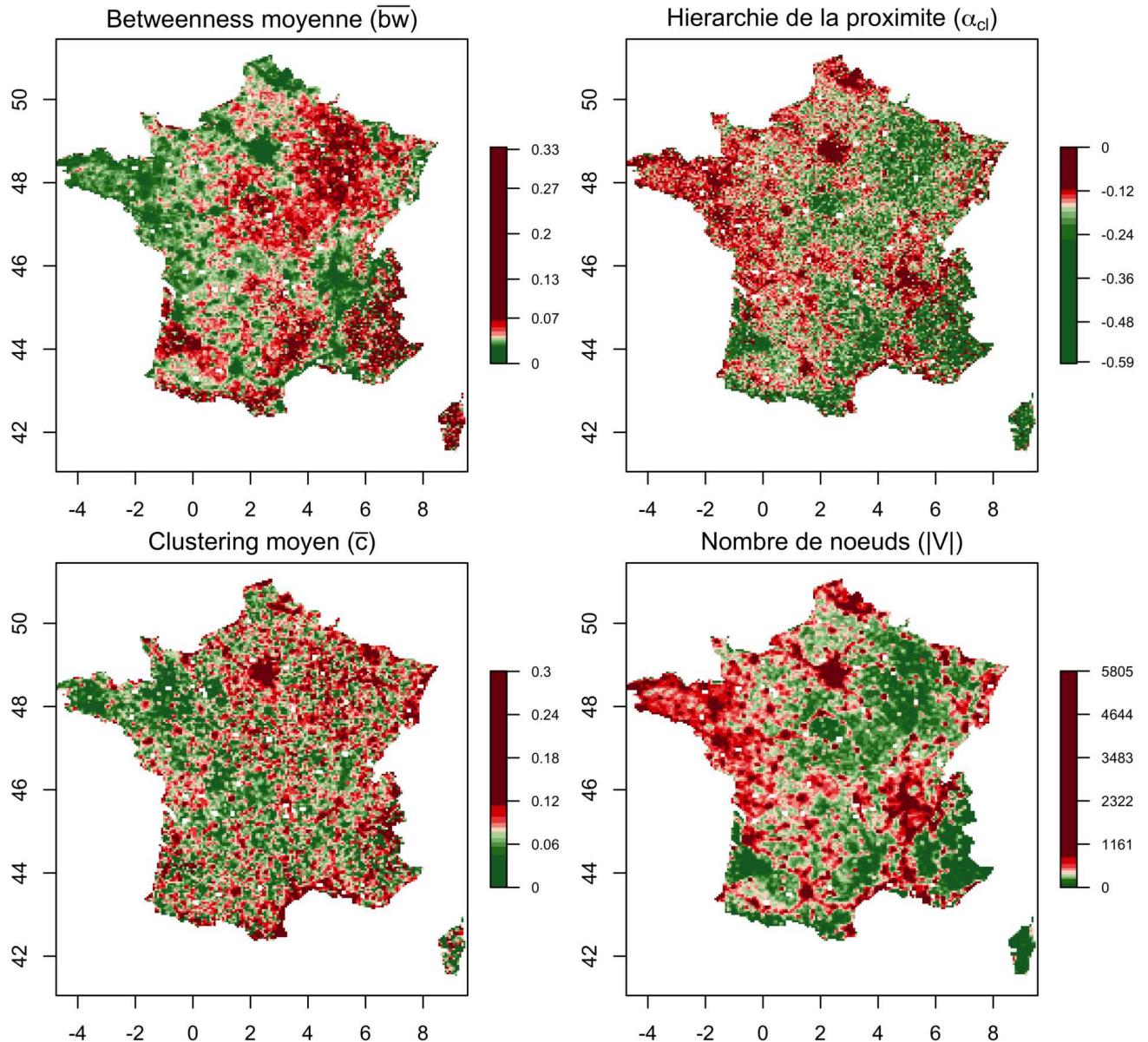


Figure 22: **Spatial distribution of network indicators.** We show indicators for France, in correspondance with morphological indicators described previously. We give here the average betweenness centrality \bar{bw} , the hierarchy of closeness centrality α_{cl} , the average clustering coefficient \bar{c} and the number of nodes $|V|$.

and not causality, but this correspondance informs on different urban regimes. Metropolitan areas seem to exhibit a positive correlation for these two indicators, as shows the Fig. 23, and rural spaces a negative correlation.

In order to give a picture of global relations between indicators, we can refer to the full correlation matrix in Fig. 78 (Appendix A.4): for example, a strong population hierarchy is linked to a high and hierarchical betweenness centrality, but is negatively correlated to the number of edges (a diffuse population requires a more spread network to serve all the population). However, it is not possible this way to systematically link indicators, since they especially strongly vary in space. We give also in Appendix A.4, Fig. 79, maps for different correlation coefficients for all Europe.

This suggests a very high variety of interaction regimes. The spatial variation of the first component of the reduced matrix confirms it, what clearly reveals the spatial non-stationarity of interaction processes between forms, since the first and second moments vary in space. The statistical significance of stationarity can be verified in different ways¹². We use here the method of [Leung, Mei, and Zhang, 2000] which consists in estimating through bootstrap the robustness of Geographically Weighted Regression models. These will be developed below, but we obtain for all tested models a significant non-stationarity without doubt ($p < 10^{-3}$).

Furthermore, the statistical distribution of correlations given in Fig. 80 in Appendix A.4 follows an asymmetric law for the morphology alone, and rather symmetric for the network and the cross-correlations, what would mean that some areas have rather strong morphological constraints whereas the shape of the network is rather free. Finally, we observe on the point clouds of the same figure, relating the values of correlations in the different blocs, that configurations for which cross-correlations are the strongest correspond to the ones for which morphological and network correlations are also strong, confirming the intrication of processes in that case.

Variations of the estimated correlations

We show in Fig. 24 the variation of the estimation of correlation as a function of window size. More precisely, we observe a strong variation of correlations as a function of δ , what is reflected in the average value of the matrix given here (which extends for example from $\rho(4) = 0.22$ to $\rho(80) = 0.12$ for average absolute cross-correlations). An increase of δ leads for all measures a shift towards positive values, but also a narrowing of the distribution, these two effects resulting in a decrease of average absolute correlations, which approximatively

¹² There does not exist to the best of our knowledge a generic test for spatial non-stationarity. [Zhang and Zhou, 2014] develops for example a test for rectangular regions of any dimension, but in the specific case of *point processes*.

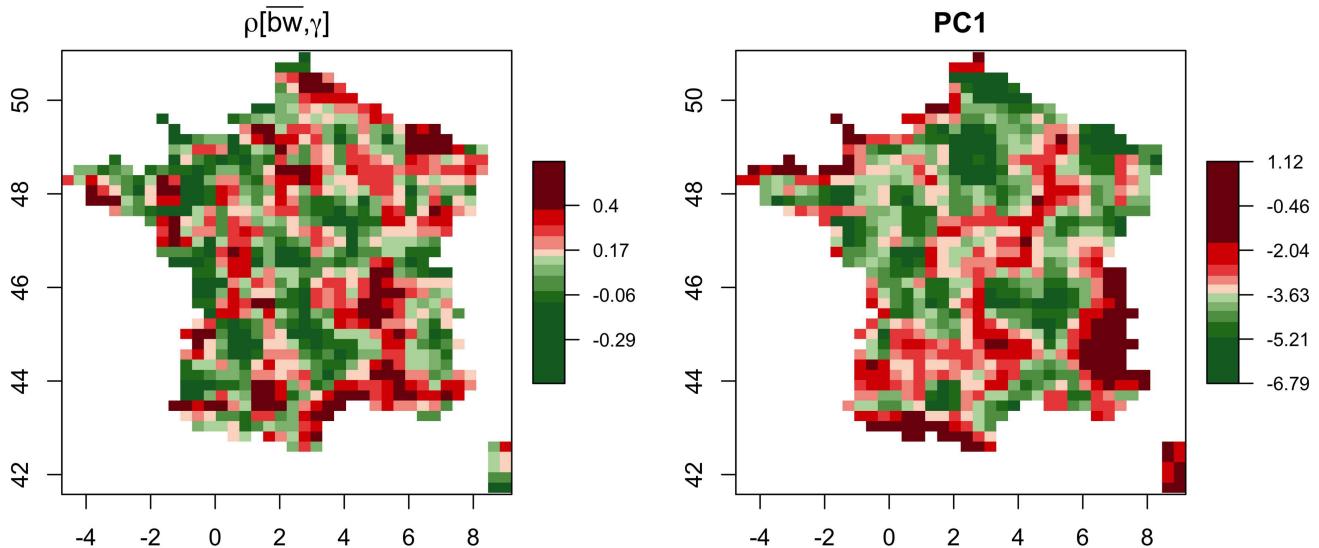


Figure 23: **Examples of spatial correlations.** For France, the maps give $\rho[\bar{bw}, \gamma]$, correlation between the average betweenness centrality and the hierarchy of population (Left) and the first component of the reduced matrix (Right).

stabilize for large values of δ . Such a variation could be a clue of a multi-scalar behavior: a change in window size should not influence the estimation if a single process would be implied, it should only change the robustness of the estimation. The development in Appendix A.4 illustrates this link in the case of processes superposed at two scales, and demonstrates that this structure of process implies a variation of the estimated correlation as a function of δ , at least in low values, which is what we observe here in Fig. 24.

Furthermore, the variation of the normalized size of the confidence interval for correlations, which in theory under an assumption of normality should lead $\delta \cdot |\rho_+ - \rho_-|$ to remain constant, since bounds vary asymptotically as $1/\sqrt{N} \sim 1/\sqrt{\delta^2}$ (the demonstration is given in Appendix A.4), follows the direction of this hypothesis of processes superposed at different scales as proposed previously.

Thus, processes are both non-stationary, and clues suggest that they result of the superposition of processes at different scales¹³.

Typical scales

We also propose to explore the possible property of multi-scalar processes by the extraction of endogenous scales which are present in the data. A Geographically Weighted Principal Component Analy-

¹³ The notion of multi-scalar process is otherwise very broad, and can manifest itself in scaling laws for example [West, 2017]. An approach closer to the one we took is given by [Chodrow, 2017] which measures intrinsic scales to segregation phenomena by using measures from Information Theory.

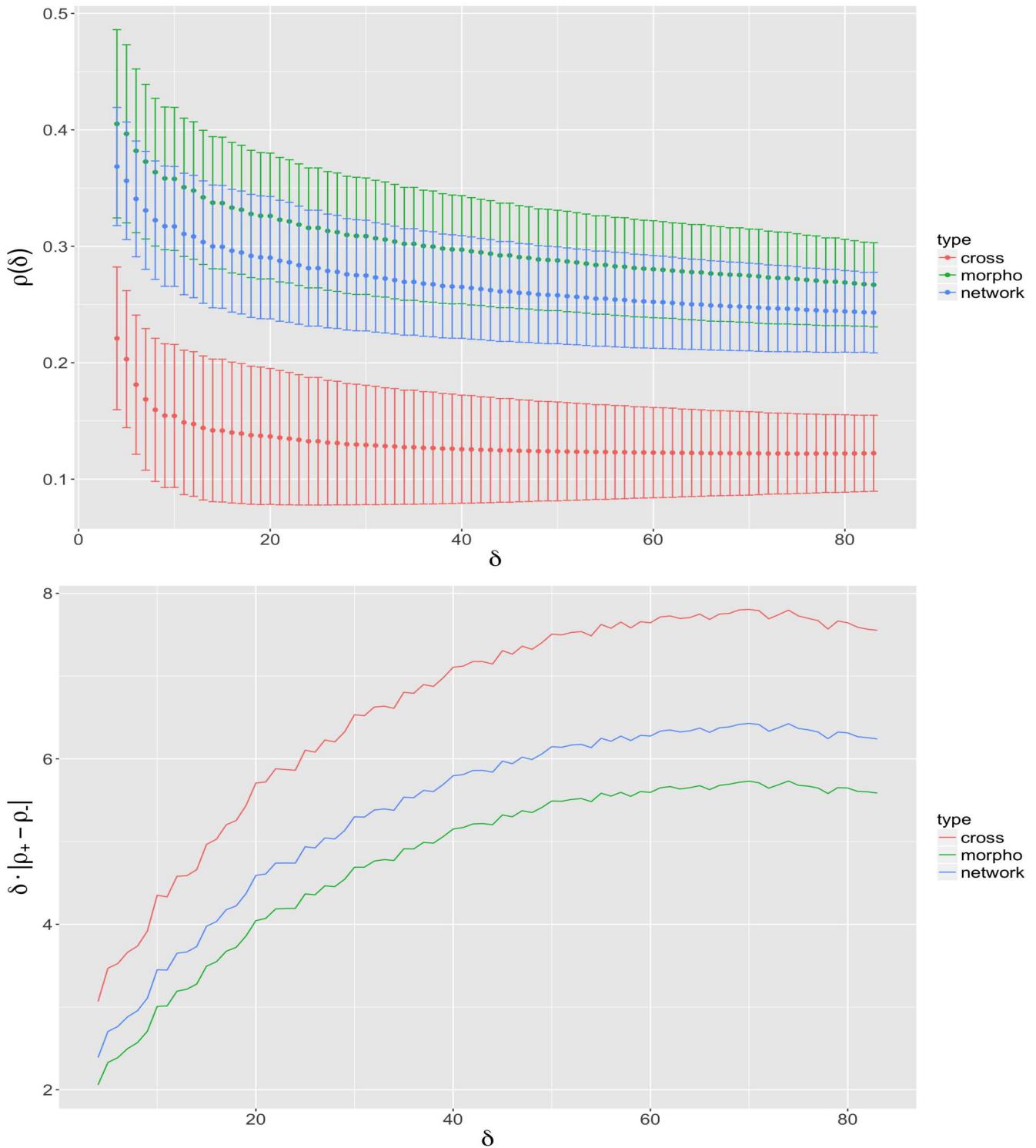


Figure 24: **Variation of correlations with scale, for correlations computed on Europe.** (Top) Average absolute correlations and their standard deviations, for the different blocs, as a function of δ ; (Bottom) Normalized size of the confidence interval $\delta \cdot |\rho_+ - \rho_-|$ (confidence interval $[\rho_-, \rho_+]$ estimated by the Fisher method) as a function of δ .

sis (GWRPCA) [Harris, Brunsdon, and Charlton, 2011] in exploration suggests weights and importances that vary in space, what is in consistence with the non-stationarity of correlation structures obtained above. There is no reason a priori that the scales of variation of the different indicators are strictly the same. We propose thus to extract typical scales for crossed relations between the urban form and network topology.

We implement therefore the following method: we consider a typical sample of indicators (four for each aspect, see the list in Table 11), and for each indicator we formulate all the possible linear models as a function of opposite indicators (network for a morphological indicator, morphological for a network indicator), aiming at directly capturing the interaction without controlling on the type of form or of network. These models are then adjusted by a Geographically Weighted Regression (GWR) with an optimal range determined by a corrected information criteria (AICc)¹⁴. For each indicator, we keep the model with the best value of the information criteria. We adjust the models on data for France, with a *bisquare* kernel and an adaptative bandwidth in number of neighbors.

Results are presented in Table 11. It is first interesting to note that all models have only one variable, suggesting relatively direct correspondances between topology and morphology. All morphological indicators are explained by network performance, i.e. the quantity of detours it includes. On the contrary, network topology is explained by Moran index for centralities, and by entropy for performance and the number of vertices. There is thus a dissymmetry in relations, the network being conditioned in a more complex way to the morphology than the morphology to the network. The adjustments are rather good ($R^2 > 0.5$) for most indicators, and *p-values* obtained for all models (for the constant and the coefficient) are lower than 10^{-3} . Concerning the scales corresponding to the optimal model, they are very localized, of the order of magnitude of ten kilometers, i.e a larger variation than the one obtained the correlations. This analysis confirms thus statistically on the one hand the non-stationarity, and on the other hand give a complementary point of view on the question of endogenous scales.

Developments

We have thus shown empirically the non-stationarity of interactions between the morphology of the distribution of populations and the topology of the road network. Various developments of this analysis are possible.

¹⁴ By using the R package GWModel [Gollini et al., 2013].

Table 11: Interrelations between network indicators and morphological indicators. Each relation is adjusted by a Geographically Weighted Regression, for the optimal range adjusted by AICc.

Indicator	Model	Range (km)	Adjustment (R^2)
Average distance \bar{d}	$\bar{d} \sim v_0$	11.6	0.31
Entropy \mathcal{E}	$\mathcal{E} \sim v_0$	8.8	0.75
Moran I	$I \sim v_0$	8.8	0.49
Hierarchy γ	$\gamma \sim v_0$	8.8	0.68
Average betweenness $b\bar{w}$	$b\bar{w} \sim I$	12.3	0.58
Average closeness $c\bar{l}$	$c\bar{l} \sim I$	13.9	0.26
Performance v_0	$v_0 \sim \mathcal{E}$	8.6	0.86
Number of nodes $ V $	$ V \sim \mathcal{E}$	8.6	0.88

Population density grids exist for all regions of the world, such as for example the ones provided by [Stevens et al., 2015]¹⁵. The analysis may be repeated with other regions of the world, to compare the correlation regimes and test if urban system properties stay the same, keeping in mind the difficulties linked to the differences in data quality.

The research of local scales, i.e. with an adaptative estimation window in terms of size and shape for correlations, would allow to better understand the way processes locally influence their neighborhood. The validation criteria for window size would still be to determine: it can be as above an optimal range for explicative models that are locally adjusted.

The question of ergodicity should also be explored from a dynamical point of view, by comparing time and spatial scales of the evolution of processes, or more precisely the correlations between variations in time and variations in space, but the issue of the existence of databases precise enough in time appears to be problematic. The study of a link between the derivative of the correlation as a function of window size and of the derivatives of the processes is also a direction to obtain indirect informations on dynamics from static data.

Finally, the search of classes of processes on which it is possible to directly establish the relation between spatial correlations and temporal correlations, is a possible research direction. It stays out of the scope of this present work, but would open relevant perspectives on co-evolution, since it implies evolution in time and an isolation in space, and therefore a complex relation between spatial and temporal covariances.

¹⁵ Available at <http://www.worldpop.org.uk/>. The potential variability of data quality depending on the areas should however lead to stay cautious on their use.

★ ★

★

This section allowed us thus to study non-stationarity properties of morphological characteristics of territories and networks, and of their interactions in terms of static correlations. The indicators we computed will also be useful in the following.

We propose in the next section to tackle a statistical approach to co-evolution, corresponding to the preliminary definition we gave. On the contrary to the previous approach, it will be based on dynamics.

★ ★

★

4.2 SPATIO-TEMPORAL CAUSALITIES

This section contributes to the understanding of strongly coupled spatio-temporal processes by describing a generic method based on Granger causality, which is a method introduced in economics to characterize possible causal relationships from correlation relations between variables lagged in time. We indeed introduce here a method allowing to characterize co-evolution at the statistical level.

The method is validated by the robust identification of causality regimes and of their phase diagram for an urban morphogenesis model that couples network growth with density. The application to the real case of South Africa unveils interactions that change in time, witnessing historical events between territorial demographic dynamics and network growth.

The exists in literature a small number of examples using statistical relationships on dynamical relations between network and territories, i.e. trying to establish a causal relationship between the two. For example, [Levinson, 2008] explains for the case of London population and connectivity to network variables by these same variables lagged in time, unveiling circular causal effects. [[doi:10.1068/b39089](https://doi.org/10.1068/b39089)] uses similar techniques for a region in Italy with historical data on long time, but stays moderate on possible conclusions of systematic effects by recalling the importance of historical events on the estimated relations. [Cuthbert, Anderson, and Hall, 2005] proceeds to econometric estimations of reciprocal influence, and concludes that in their Canadian case study at a sub-regional scale, the development of the network induces the development of land-use but not the contrary. Space and time scales influence thus significantly the results of such analysis. [Koning, Blanquart, and Delaplace, 2013] proposes an estimation of relations between the existence of a High Speed Rail connection and economic variables on French Urban Units, and shows a negative effect of the connection itself, after controlling on the endogenous nature of the connection by a selection model, and a significant effect of the characteristics of Urban Units: for example, for urban units benefiting from a TGV connection without LGV, the effect is of -1% on employments between 1982 and 2006. This study remains however limited as it takes neither a time lag larger than one time step nor spatial relations between entities. Finally, still in the same spirit but without explicit inclusion of space, [[MANCMANC1073](#)] shows on long time a causality link between infrastructure stock and economic growth on a global panel, but that these effects are moderated locally by under or over-investments: in that case, macro-economic effects are revealed.

4.2.1 Spatio-temporal causalities

The study of strongly coupled spatio-temporal processes implies to understand tangled intrications generally highly difficult to isolate. These interactions are the essence of complexity approaches, and are indeed at the origin of the emergent behavior of the system. They make sense as an object of study in itself and a separation of processes appears then contradictory with an integrated view of the system. In the case of territorial systems, the example of interactions between transportation networks and territories is a good illustration of this phenomenon, as shows the debate on structuring effects developed in chapter 1. We recall that we have suggested that the reality of territorial processes is in fact much more complicated than a simple causal relationship between the construction of an infrastructure and spillovers on local development, but indeed corresponds to a *co-evolution*.

At another scale, still for relations between networks and territories, we can point at the relations between mobility practices, urban sprawl et ressource localisation in a metropolitan framework that are as much complex: [Cerqueira, 2017] shows for example a strong correspondence between conditioning of mobility practices by the accessibility and socio-professional category.

This kind of issue is naturally present in other fields: in Economic Geography, the example of links between innovation, local spillovers of knowledge and aggregation of economic agents is a typical illustration of spatio-temporal economic processes exhibiting circular causalities difficult to disentangle [Audretsch and Feldman, 1996]. Specific methods are introduced, as the use of statistical instruments: [Aghion et al., 2015] shows that the geographical origin of US Congress members that attribute local subsidies is a powerful instrumental variable to link innovation and income inequalities for higher incomes, what confirms that the significant correlation between the two is indeed a causality of innovation on inequalities¹⁶.

Causality in geography

Strong coupling in space and time generally implies a notion of causality, that geography has always studied: [Loi, 1985] shows that fundamental issues tackled by contemporary theoretical geography (isolation of objects, link between space and causal structures, etc.) were already implicit in VIDAL's classical geography.

Beside, [Claval, 1985] criticizes the new determinisms having emerged, in particular the one advocated by some scholars of systemic analy-

¹⁶ This example is important from the methodological point of view, but not only since it implicitly links to the thematic of the diffusion of innovation which is crucial in the evolutive urban theory.

sis¹⁷: in its beginning, this approach inherited from cybernetics and thus of a reductionist vision implying a determinism even for a probabilistic formulation. CLAVAL observes that works contemporary to his writings could allow to capture the complexity that characterizes human decisions: the Prigogine School and the Theory of Catastrophes by René Thom.

This viewpoint has anticipated posterior developments, since as Pumain recalls in [Pumain, 2003], the shift from system analysis to self-organisation and complexity has been long and progressive, and these works have played a fundamental role for it. FRANÇOIS DURAND-DASTÈS sums up this picture more recently in [Durand-Dastès, 2003], by focusing on the importance of bifurcations and path-dependency in the initial moments of the constitution of a system that he defines as *systemogenesis*¹⁸. This type of complex dynamics generally implies a co-evolution of system components, that can be understood as circular causalities between processes: the issue of identifying them is thus crucial regarding the notion of causality for contemporary complex geography.

This view of a complex causality [Morin, 1976] can also be put into perspective with the concept of *cumulative causality* in economics [Skott and Auerbach, 1995], which insists on the role of path-dependency and the possibility for small perturbations to cause significant effects by negative feedback: it is then impossible to separate the effects from their causes in infinitesimal perturbations.

Identification of causalities

The regimes under which identification of causalities are relevant are not obviously known. These will depend of the definitions used, as well as available methods for which we give now a few examples. [Liu et al., 2011] proposes to detect spatio-temporal relations between perturbations of traffic flows, introducing a particular definition of causality based on correspondence of extreme points. Associated algorithms are however specific and difficult to apply to other kind of systems. The use of spatio-temporal correlations has been shown to have in some cases a strong predictive power for traffic flows [Min and Wynter, 2011]. Also in the field of transportation and land-use, [Xie and Levinson, 2009a] applies a Granger causality analysis, that can be interpreted as lagged correlation, to show for a case study that network growth induces urban development and is itself driven by externalities such as mobility habits.

Neuroscience has developed numerous methods answering similar issues. [Luo et al., 2013] defines a generalized Granger causality

¹⁷ See [Chamussy et al., 1984] for an example of model with a planning purpose positioned within that research stream.

¹⁸ This notion can be put closer to the one of *morphogenesis* that we study more deeply in chapter 5.

that takes into account non-stationarity and applies to abstracts regions produced by functional imaging. This kind of method is also developed in Computer Vision, as illustrated by [Ke, Sukthankar, and Hebert, 2007] that exploits spatio-temporal correlations of forms and flows between successive images to classify and recognize actions. Applications can be quite concrete such as compression of video files by extrapolation of motion vectors [Chalidabhongse and Kuo, 1997]. In all these cases, the study of spatio-temporal correlations meets the weak notions of causality described above.

This contribution aims to explore the possibility of a similar methods for spatio-temporal data exhibiting a priori complex circular causalities, and thus to realize the difficult exercise to couple a certain level of simplicity with a grasping of complexity. We introduce therefore a method to analyse spatio-temporal correlations, similar to a Granger causality estimated in space and time. The robustness of the method is demonstrated in a systematic way by the application to a complex model of simulation of urban morphogenesis, what leads to the unveiling of distinct causality regimes in the phase space of the model. We also include the application to an empirical case study, what positions this work at the interface between knowledge domains of methodology, modeling and empirical within the epistemological framework introduced by [2017arXiv170609244R].

The rest of this section is organized as follows: the generic framework of the method is described in the next section. We then apply it to a synthetic dataset to partially validate it and test its potentialities, what allows us to apply it then to the real case study of Grand Paris transportation network. We finally discuss to proximity with existing methods and possible developments.

Method

We formalize here the method in a generic way, based in a weak formulation of Granger causality, to try to identify causal relations in spatial systems. Let $X_j(\vec{x}, t)$ spatio-temporal unidimensional random processes, which realizations occur in space and time. We give a set of fundamental spatial units (u_i) that can be for example raster cells or any paving of the geographical space. We assume the existence of functions $\Phi_{i,j}$ allowing to make the correspondance between the realization of each components and spatial units, possibly through a first spatial aggregation or by a more elaborated process driven by a network for example. A realization of a system is given by a set of trajectories for each process $x_{i,j,t}$, and we write a set of realizations $x_{i,j,t}^{(k)}$ (accessible by stochastic repetitions in the case of a model of simulation for example, or by assumption of comparability of territorial sub-systems in real cases). We assume to have a correlation estimator $\hat{\rho}$ applying in time, space and repetitions, i.e. $\hat{\rho}[X, Y] = \hat{E}_{i,t,k}[XY] - \hat{E}_{i,t,k}[X]\hat{E}_{i,t,k}[Y]$.

$$\hat{\text{Cov}}[X, Y] = \hat{\mathbb{E}}_{i,t,k}[XY] - \hat{\mathbb{E}}_{i,t,k}[X]\hat{\mathbb{E}}_{i,t,k}[Y]$$

It is important to note here the hypothesis of spatial and temporal stationarity, that can however easily be relaxed in the case of local stationarity.

Furthermore, spatial auto-correlation is not explicitly included, but is taken into account either by the initial spatial aggregation if the characteristic scale of units is larger than the one of neighborhood effects, either by an adequate spatial estimator (weighted spatial statistics of type GWR [Brunsdon, Fotheringham, and Charlton, 1998] for example). It allows us to define the lagged correlation by

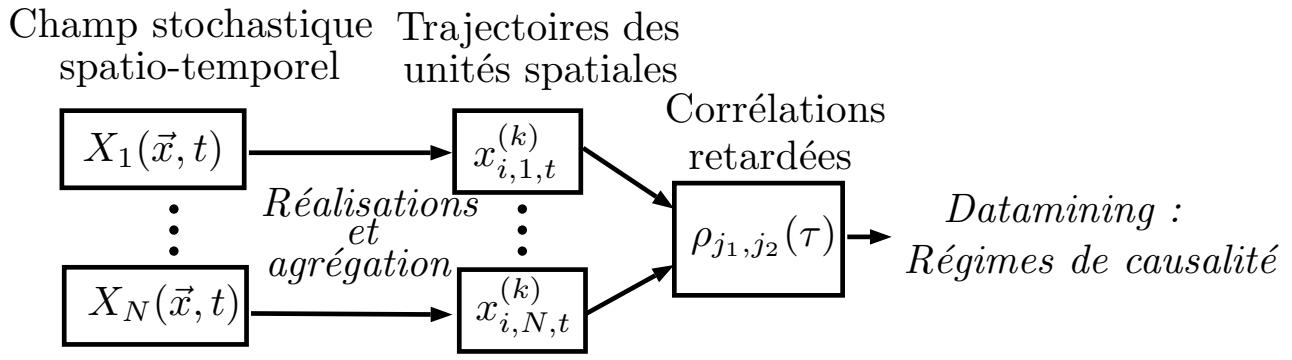
$$\rho_\tau[X_{j_1}, X_{j_2}] = \hat{\rho}\left[x_{i,j_1,t-\tau}^{(k)}, x_{i,j_2,t}^{(k)}\right] \quad (6)$$

The lagged correlation is not symmetric, but we have directly $\rho_\tau[X_{j_1}, X_{j_2}] = \rho_{-\tau}[X_{j_2}, X_{j_1}]$. This measure is applied in a simple way: if $\text{argmax}_\tau \rho_\tau[X_{j_1}, X_{j_2}]$ or $\text{argmin}_\tau \rho_\tau[X_{j_1}, X_{j_2}]$ are “clearly defined” (both could be simultaneously), their sign will give the sense of causality between components j_1 and j_2 and their absolute value the propagation lag.

Par exemple, X_{j_1} pourra être une propriété liée au réseau comme la centralité de proximité, et X_{j_2} une propriété liée aux territoires, comme la densité de population. Cette mesure permettra alors de déterminer un sens de causalité (éventuellement réciproque) entre ces propriétés. Le retard τ sera typiquement un certain nombre d’années, en association avec l’échelle spatiale des unités d’estimation qui pourra varier de l’échelle du quartier à celle des aires urbaines, comme nous le verrons dans les différents cas d’application par la suite.

The criteria for significance will depend on the case of application and of the estimator used, but can for example include the significance of the statistical test (Fisher test in the case of a Pearson estimator), the position of extremities of a confidence interval of a given level, or even an exogenous threshold θ on $|\rho_\tau|$ to ensure a certain level of correlation.

Avant de nous plonger dans l’exploration empirique de la méthode, donnons-en une vision intuitive pour mieux comprendre son lien avec la co-évolution. L’encadré 9 synthétise des situations stylisées pouvant se produire dans le cas de deux variables. De manière caricaturale, avec deux variables X, Y , le profil de $\rho_\tau[X, Y]$ est traduit selon les caractéristiques suivantes : existence d’un extrémum ou non pour $\tau < 0$ et existence d’un extrémum ou non pour $\tau > 0$, c’est-à-dire possibilités de causalité de X vers Y et/ou de causalité de Y vers X . Nous illustrons quatre exemples de profils et représentons les interactions entre variables sous forme graphique, dans le temps et de manière synthétique.



FRAME 8: Structure of the methodology.

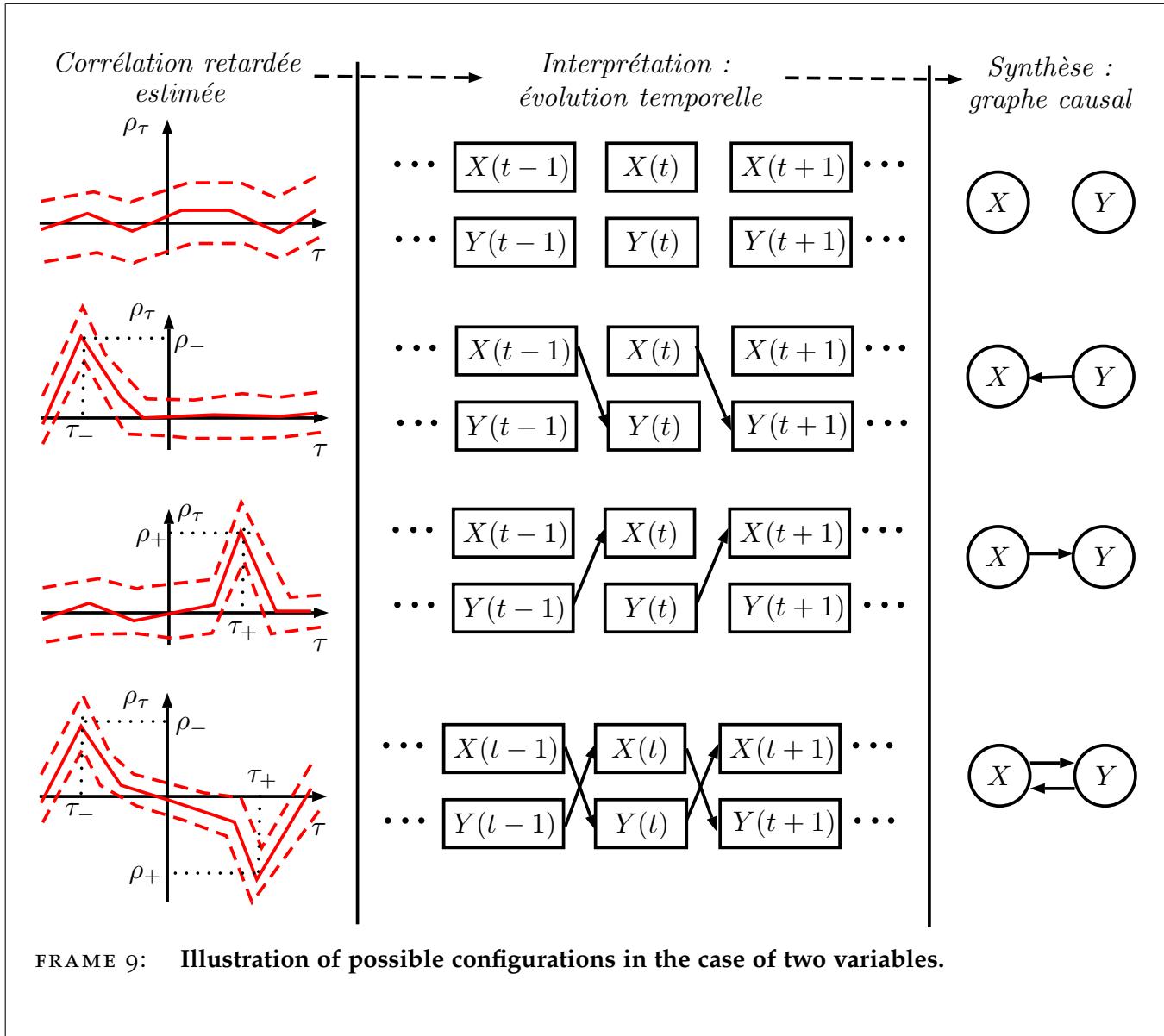
Emergence and a proxy to measure co-evolution ?

Prenons également un court instant pour clarifier le statut épistémologique et ontologique attendu par l'application de cette méthode, et dans quelle mesure on peut espérer l'utiliser comme mesure indirecte de la co-évolution. La causalité de Granger est estimée à la fois *dans le temps, dans l'espace et entre les répétitions*. Dans le cas où l'on observe un phénomène historique, on a une unique trajectoire et l'estimation est faite dans le temps et l'espace uniquement, mais dans tous les cas on passe de caractéristiques à l'échelle microscopique à une mesure macroscopique¹⁹. Ainsi, on peut avoir des interactions microscopiques circulaires, mais émergence d'un sens de la causalité au niveau macroscopique, ou l'inverse. Rejoignant la question des populations et individus pour la définition de la co-évolution en biologie (voir 3.3), pour laquelle les adaptations mutuelles émergent au niveau des espèces, nous postulons que la caractérisation des motifs de causalité est une manière de caractériser des dynamiques co-évolutives pour les systèmes territoriaux, correspondant alors à notre définition intermédiaire de la co-évolution au niveau d'une population.

Est-il alors possible de répondre de manière équivoque à la question "*y a-t-il co-évolution dans un cas particulier*"²⁰ ? Cela se saurait si nous pouvions réinventer l'eau chaude mais qui se chauffe elle-même. Nous voulons dire par là, et nous le verrons dans les multiples développements, que de nombreux problèmes fondamentaux intrinsèques à l'étude des systèmes géographiques (la question des échelles, de la définition du système, des variables prises en compte, le problème de l'observation de trajectoire uniques, de données bruitées et

¹⁹ Nous utilisons ici ces termes pour simplifier, il s'agit en fait d'un échelle donnée à une échelle supérieure qui dépend de l'étendue temporelle et spatiale totale.

²⁰ A laquelle nous ajoutons : pour ces composantes, sur cette portée spatiale et temporelle et sur ces échelles spatiale et temporelle.



FRAME 9: Illustration of possible configurations in the case of two variables.

éparses, le problème du MAUP, etc.) seront bien toujours présents, et que la question ci-dessus qui y est naturellement soumise s'avère naïve. Mais nous verrons qu'il sera bien possible d'isoler des signaux clairs, et mettrons en évidence des cas où il existe un sens causal et d'autres où il y a circularité au niveau macroscopique.

4.2.2 Synthetic data

Nous explorons et validons la méthode dans un premier temps sur données synthétiques, c'est-à-dire générées par l'intermédiaire d'un modèle avec un certain niveau de contrôle.

Auto-regressive time series

Illustrons les motifs qui peuvent être attendus, notamment ceux stylisés donnés précédemment en Encadré 9, sur des données synthétiques avec une structure simple. L'idée est de générer des séries temporelles sur lesquelles le retard et le niveau de corrélation sont contrôlés, ainsi que les résultats théoriques connus.

Soit $\vec{X}(t)$ un processus stochastique suivant l'équation d'auto-régression $\vec{X}(t) = \sum_{\tau>0} \mathbf{A}(\tau) \cdot \vec{X}(t-\tau) + \vec{\epsilon}(t)$. Dans le cas où $\mathbf{A}(\tau) = 0$ pour $\tau \neq \tau_0$ et $\mathbf{A}(\tau_0) = \begin{pmatrix} 0 & a \\ a & 0 \end{pmatrix}$ pour $-1 < a < 1$, le calcul des corrélations théoriques est possible (voir Annexe A.5), et on obtient, en notant $\mathbf{X} = (X, Y)$, pour $\tau > 0$

$$\rho[X(t), Y(t-\tau)] = \begin{cases} a^{2k+1} \text{ si } \tau = (2k+1)\tau_0 \text{ pour tout } k \in \mathbb{Z} \\ 0 \text{ sinon} \end{cases}$$

L'expression est la même pour $\tau < 0$ en échangeant X et Y . Ainsi, on contrôle la corrélation retardée au retard voulu et aux retards qui en sont multiples avec un facteur impair. En changeant l'un des coefficients en a ou en son opposé, on obtient pour les premiers maximums les trois profils stylisés donnés en Encadré 9.

Utilisons cet exemple pour explorer numériquement la possibilité de classifier les profils de corrélations retardées. Nous considérons le même processus pour $\tau_0 = 2$ et $\mathbf{A}(\tau_0) = \begin{pmatrix} 0 & a_1 \\ a_2 & 0 \end{pmatrix}$, avec $-1 < a_1, a_2 < 1$. Nous simulons avec ce modèle des séries temporelles de longueur $t_f = 10000$ en tirant $b = 10000$ valeurs aléatoires pour les paramètres (a_1, a_2) . Sur chaque série les corrélations retardées sont estimées, et nous procédons à une classification non-supervisée²¹ sur les séries temporelles $[\rho(\tau)]_{a_1, a_2}$. Nous montrons en Fig. 25 les profils typiques obtenus en correspondance avec leur position dans l'espace des paramètres (a_1, a_2) . Nous obtenons exactement les neuf profils stylisés possibles, en correspondance avec les valeurs relatives des paramètres comme attendu. À partir de profils très variés de corrélations retardées, nous sommes ainsi capable d'extraire des profils typiques d'interaction entre les variables. Cela nous renforce dans l'idée d'appliquer cette méthode sur des données plus complexes par la suite.

Urban growth model

This method must first be tested and partially validated, what we propose to do on synthetic data, what allows a more refined knowledge of the behavior of models [raimbaultshs01514415]. Echoing the

²¹ Par algorithme des *k-means* avec $k = 9$ et $b_c = 1000$ répétitions.

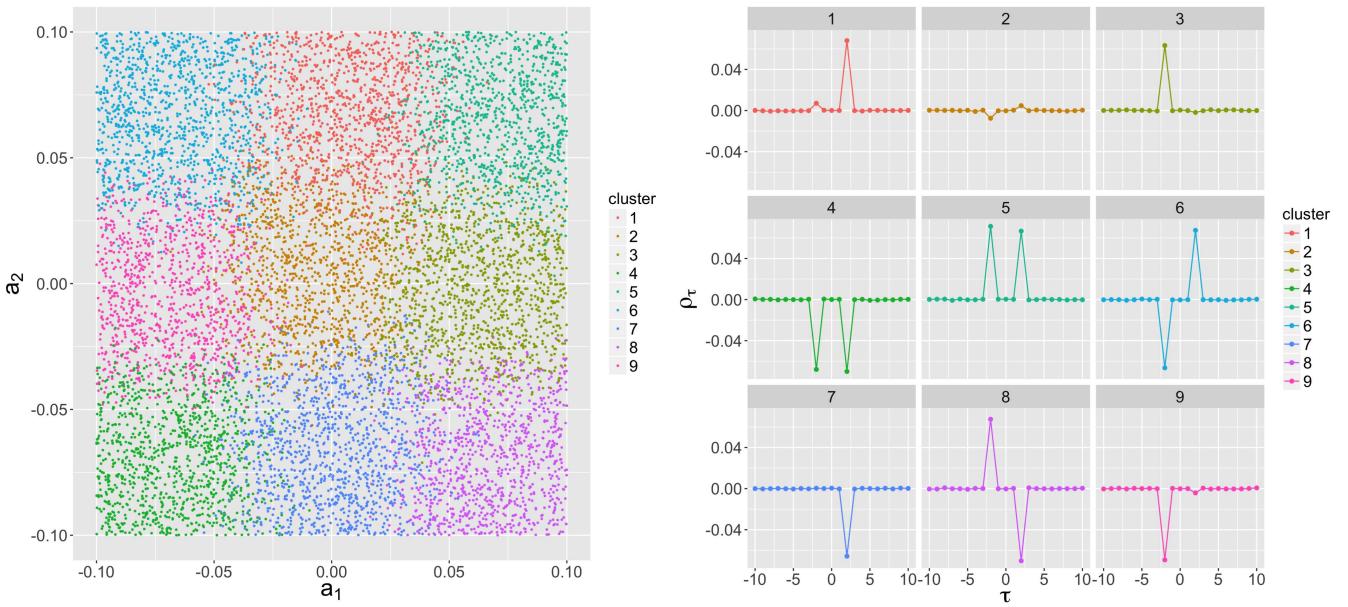


Figure 25: Estimation of correlation regimes of auto-regressive time series.

example of relations between transportation networks and territories that introduced the research question before, we generate stylized urban configurations in which network and density mutually interact, and for which causalities are not obvious *a priori* knowing the parameters of the generative model.

[Rimbault, Banos, and Doursat, 2014] describes and explores a simple model of urban morphogenesis (the RBD model) that fits perfectly these constraints. Indeed, explicative variables of urban growth, processes of network extension and the coupling between urban density and the network are relatively simple. However, except for extreme cases (for example when distance to the center solely determines land value, the network will depend on density in a causal way; when only the distance to the network counts, the causality will be inverted), mixed regimes do not exhibit obvious causalities. It is for this reason an ideal case to test if the method is able to detect some.

We use an applied implementation²² of the original model, allowing to capture the values of studied variables for each cell of the cellular automaton and for each time step, and to calculate the lagged correlations in the sense described before, between variables of the model. We explore a grid of the parameter space of the RBD model, making the weight parameters for density, distance to center and dis-

²² available on the open repository of the project at
<https://github.com/JusteRimbault/CityNetwork/tree/master/Models/Simple/ModelCA>

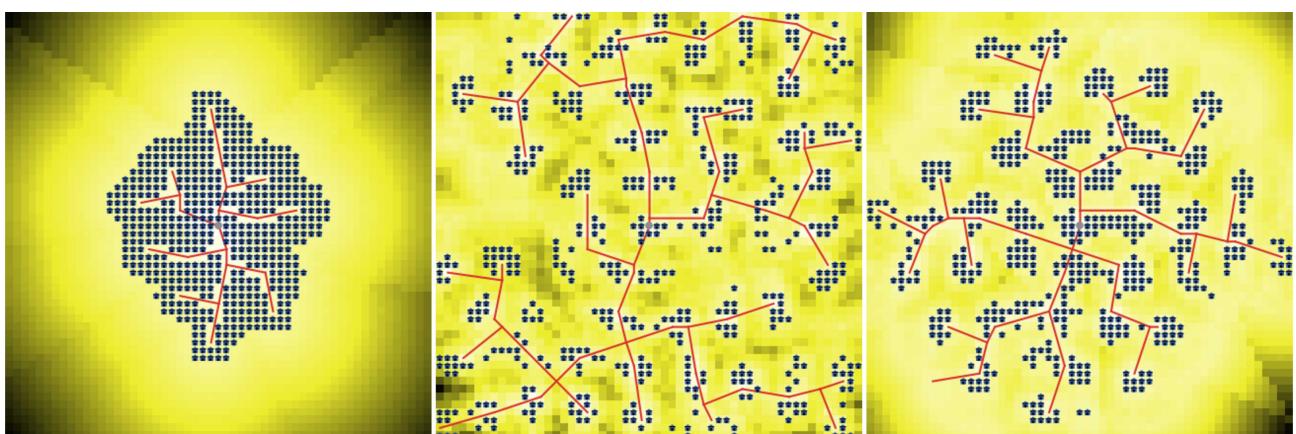
Le modèle RBD suppose une grille de côté N , dont les cellules ont un état binaire (occupée ou non). Dans la version utilisée, il existe un unique centre urbain (noeud particulier du réseau) et le réseau de transport est initialement nul. Chaque cellule i est caractérisée par les variables $x_d(i)$ (densité dans un rayon fixé $r = 5$), $x_r(i)$ (distance à la route la plus proche) et $x_c(i)$ (distance au centre via le réseau). Ces variables permettent de calculer une valeur de potentiel pour chaque cellule $U_i = \sum w_k \tilde{x}_k(i)$, où les w_k sont des paramètres du modèle permettant d'influencer les formes urbaines produites et $\tilde{x}_k(i)$ les variables normalisées sur l'ensemble des cellules par $\tilde{x}_k(i) = \frac{\max_i x_k(i) - x_k(i)}{\max_i x_k(i) - \min_i x_k(i)}$.

Le potentiel peut être interprété comme une utilité agrégeant les préférences des agents devant se localiser. Une répulsion à la densité donnera par exemple des formes urbaines très dispersées.

Le modèle évolue séquentiellement en peuplant progressivement la grille. À chaque pas de temps :

- les N_G cellules avec plus grande valeur U_i sont occupées de manière simultanée ;
- si une cellule nouvellement peuplée est à une distance au réseau supérieure à un seuil θ_d (que nous fixerons ici à $\theta_d = 5$), celle-ci est connectée au réseau par une nouvelle route prenant le chemin le plus court.

La croissance s'arrête à un temps final fixé t_f .



Exemples de configurations finales variées, obtenues avec les paramètres de poids (w_d, w_c, w_r) valant respectivement (0, 1, 1), (1, 0, 1), et (1, 1, 1).

FRAME 10: Description of the RBD model.

tance to the network vary²³, that we write respectively (w_d, w_c, w_r) , in $[0; 1]$ with a step of 0.1. Other parameters are fixed to their default values given by [Raimbault, Banos, and Doursat, 2014]. For each parameter value, we proceed to $N = 100$ repetitions, what is enough for a good convergence of indicators. Explorations are done with the OpenMole software [Reuillon, Leclaire, and Rey-Coyrehourcq, 2013], the large number of simulations (1,330,000) implying the use of a computation grid.

²³ The model works the following way: a value of cells is determined by the weighted average of these different explicative variables, value that determines the growth of new patches at the next time step.

We compute for all patches the lagged correlations with the unbiased Pearson estimator between the variations of the following variables²⁴: local density, distance to center and distance to network.

The figure ?? shows the behavior of ρ_τ for each couple of variable (undirected, τ taking negative and positive values), for the combination of extreme values of parameters. We can already see different regimes emerge: for example, $(1, 0, 1)$ leads to a causality of density on distance to center with a lag $\tau = 1$, and a negative causality of density on distance to network with the same lag, whereas distance to the center and to the network are correlated in a synchronous manner.

L'intérêt de la méthode se précise ici, puisqu'elle permet de dégager des motifs de causalité "macroscopiques" (c'est-à-dire effectivement mesurables à un niveau statistique), à partir de motifs "microscopiques" (par exemple la règle de connection de la route), et de manière non-linéaire. Des liens qu'on pourrait attendre intuitivement comme $D \rightarrow R$ sont dans certains cas inhibés. Cela confirme la pertinence de la distinction entre les deux premiers niveaux de coévolution, la co-évolution "processuelle" (au niveau des entités ou des processus) et la co-évolution statistique au niveau d'une population.

Causality regimes

To study these behaviors in a systematic way, we propose to identify regimes endogenously, by using non-supervised classification. We apply a *k-means* clustering, robust to stochasticity (5000 repetitions), with the following features: for each couple of variables, $\text{argmax}_\tau \rho_\tau$ and $\text{argmin}_\tau \rho_\tau$ if the corresponding value is such that $\frac{\rho_\tau - \bar{\rho}_\tau}{|\bar{\rho}_\tau|} > \theta$ with θ threshold parameter, 0 otherwise. The inclusion of supplementary features of values of ρ_τ does not significantly changes the results, these are therefore not taken into account to reduce the dimension. The choice of the number of clusters k is generally a difficult problem in this kind of approach [Hamerly and Elkan, 2003]. In our case the system exhibit an convenient structure: the curves of inter-cluster variance proportion and its derivative in figure ??, as a function of k for different values of θ , show a transition for $\theta = 2$, what gives for the corresponding curve a break around $k = 6$. A visual screening of clusters in a principal plan confirms the good quality of the classification for these values. A class corresponds then to a *causality regime*, for which we can represent the phase diagram as a function of model parameters, and also cluster centers profiles (computed as the barycenter in the full initial space) in figure ??.

²⁴ Computing the correlations directly on the variables makes no sense since their value has no absolute meaning.

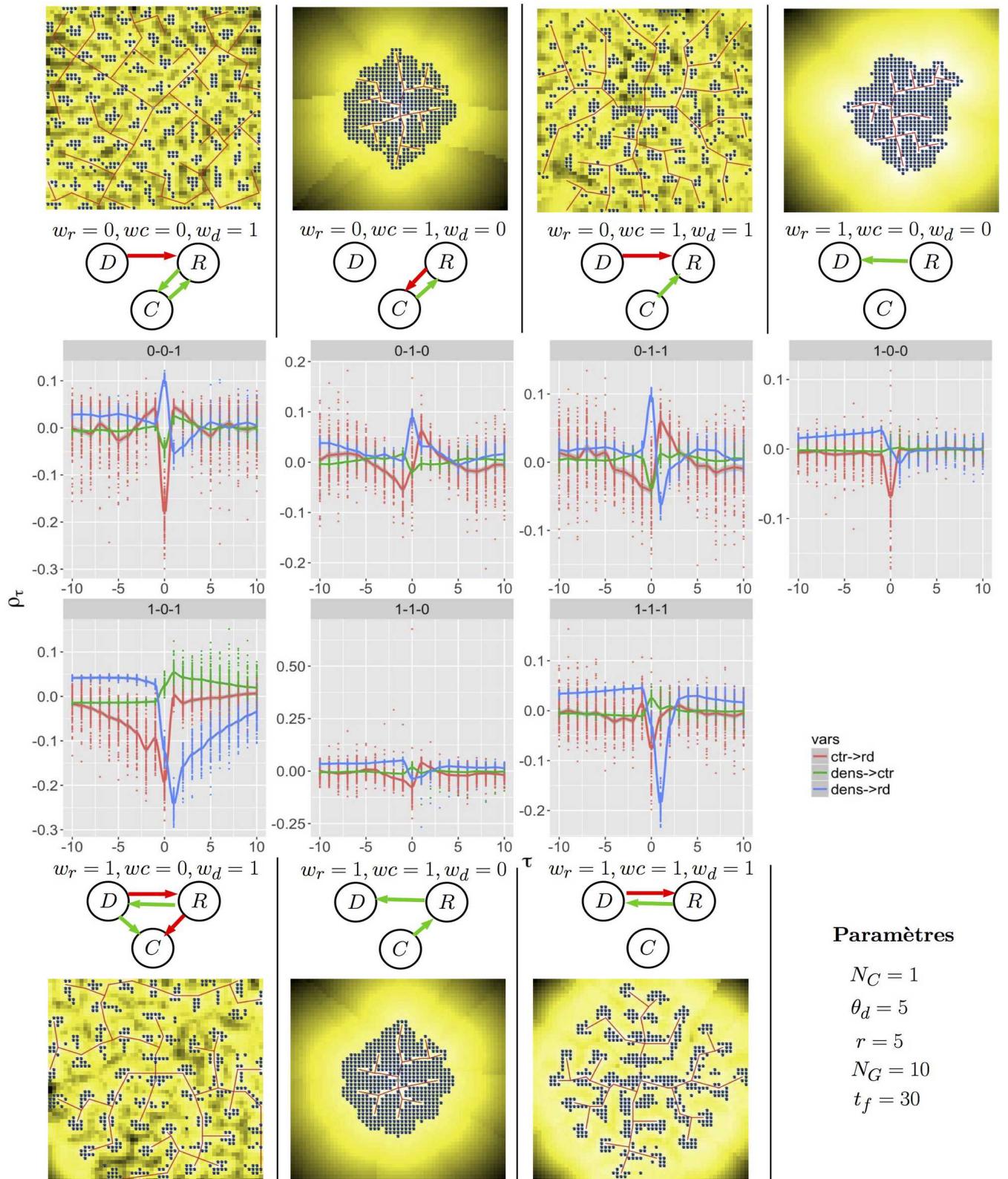


Figure 26: **Correlations in the RBD model.** (First row) Example of different final configurations, obtained with (w_d, w_c, w_r) being respectively $(0, 1, 1), (1, 0, 1)$, and $(1, 1, 1)$. (Second row) Lagged correlations, for each combination of parameters in $\{0, 1\}$, as a function of the lag τ . The different colors correspond to each couple of variables: distance to the center (ctr , C), density ($dens$, D) and distance to the network (rd , R). The dots show the extent on all the repetitions of the model (estimators on i and t only).

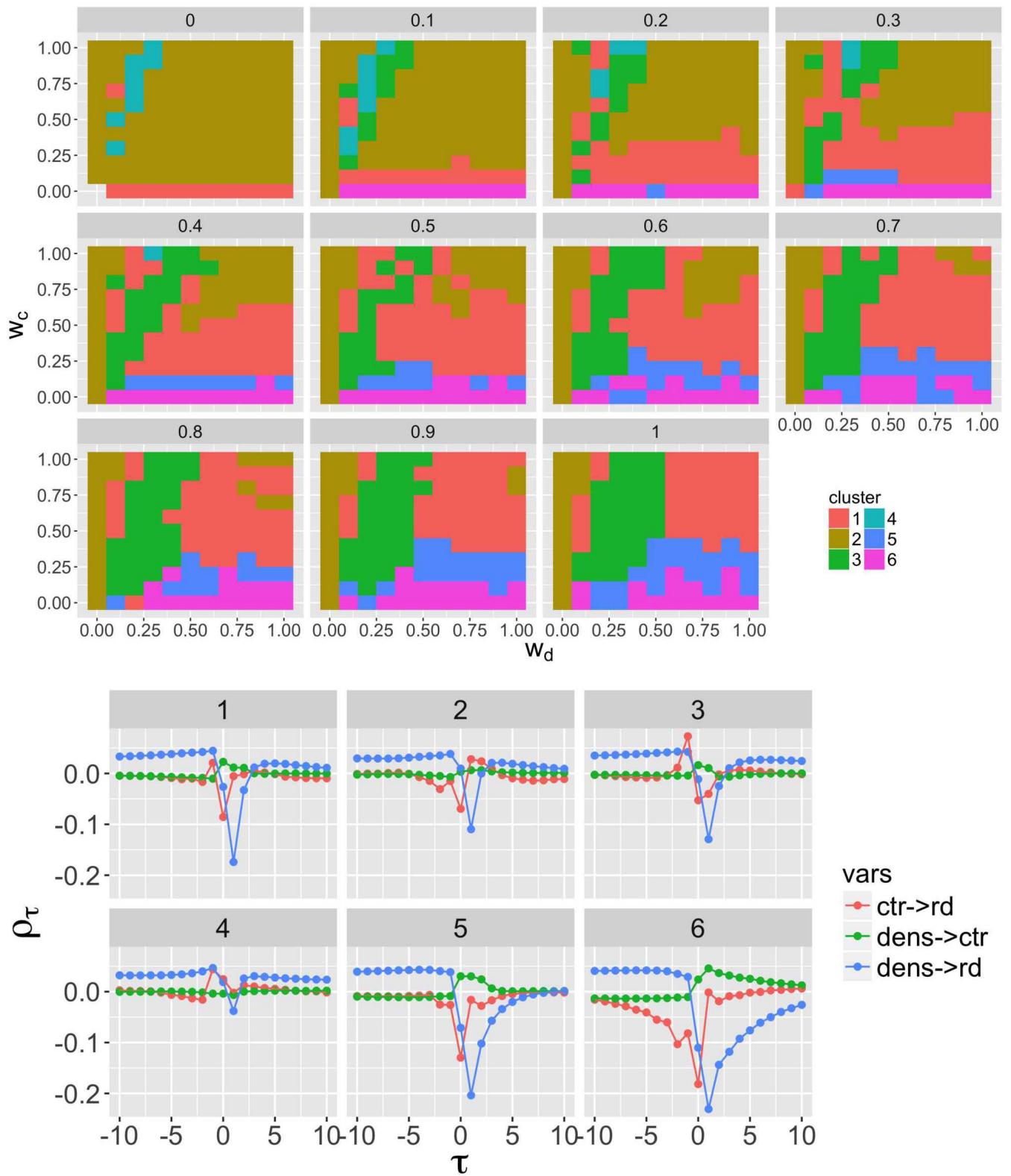


Figure 27: **Identification of regimes of interaction.** **(Top left)** Inter-cluster variance as a function of cluster number. **(Top middle)** Derivative of the inter-cluster variance. **(Top right)** Features in a principal plan (81% of variance explained by the two first components)**(Bottom left)** Phase diagram of regimes in the space (w_d, w_c, w_r) , w_r varying between the different sub-diagrams of (w_d, w_c) . **(Bottom right)** Corresponding profiles of centroids.

Interpretation

The behavior obtained is interesting, as regions in the diagram corresponding to the different regimes are clearly delimited and connected. For example, we observe the emergence of regime 6 in which distance to network causes strongly the density in a negative way, but distance to the center causes distance to the network. Its maximal extent on (w_d, w_r) is for an intermediate value $w_r = 0.7$. Thus, to maximize the impact of network on density, the corresponding weight must not be maximized, what can be counter-intuitive at first sight. It illustrates the utility of the method in the case of circular causal relations difficult to entangle a priori. The regime 5, in which distance to network influences the density the same way, but the relation between distance to center and to the network is inverted, is also interesting, and predominates for low w_r values. The regime 1 is an extreme one and corresponds to an isolated situation in which distance to the center has no role: this aspect dominates then totally the other interaction processes between density and network.

This application on synthetic data demonstrate on one hand the robustness of the method given the consistence of obtained regimes, and realizes this way a much more finer qualification of model behavior than the one done in the original paper. In this precise case, it can be taken as an instrument of knowledge for relations between networks and territories in itself, allowing the test of assumption or the comparison of processes in the stylized model.

4.2.3 Network-territory relations in South Africa

We assum that territorial dynamics and network dynamics responded differently to these. We expect to learn from these project informations on interactions at long time scale and large spatial scale, in a very particular context of constrained growth.

Context

Transportation Networks can be leveraged as a powerful socio-economic control tool, with even more significant outcomes when it percolates to their interaction with territories. The case of South Africa is an accurate illustration, as [Baffi, 2016] shows that during apartheid railway network planning was used as a racial segregation tool by shaping strongly constrained mobility and accessibility patterns. In particular, it is shown qualitatively that dynamics between territories and networks profoundly changed at the end of the apartheid, transforming a tool of planed segregation (network shaped was optimized to minimize unwanted accessibility) into an integration tool thanks to recent changes in network topology patterns. We propose to investigate the potential *structural* properties of this historical process, by

focusing on dynamical patterns of interactions between the railway network and city growth. More precisely, we try to establish if the segregative planning policies did actually modify the trajectory of the coupled system, what would correspond to deeper and wider impacts.

Data

We use a comprehensive database covering the full South African railway network from 1880 to 2000 with opening and closing dates for each station and link, together with a city database spanning from 1911 to 1991 for which consistent ontologies for urban areas have been ensured. These database are described by [Baffi, 2016], but they are not open so we make available only the aggregated data we used in the analysis.

Network Measures

First, a dynamical study of network measures seem to confirm the hypothesis: a trend rupture in closeness centrality (defined for a node as the average travel time to other nodes) at a roughly constant network size evolution, at a date corresponding to the beginning of official segregative policies, suggests that the planning process after this date had in the best case no global effect on network performance, and in the worst case had intended negative effects on accessibility with the aim to physically segregate more.

Causality patterns

We then turn to dynamical interactions between the railway network and city growth. For that, we study Granger causalities, in the large sense of correlations between lagged variables, estimated between cities growth rates and accessibility differentials due to network growth, for all cities and urban areas having a connection to the network. We test both travel-time and population weighted accessibilities, for varying values of distance decay parameter. Lagged correlations are fitted on varying length time windows, to test for potentially varying stationarity scales.

Results are shown in Figure 29. We find that results are significant with travel-time accessibility only, autocorrelation dominating with weighted accessibility. A time-window of 30 years appears to be a good compromise between the number of significant correlations ($p < 0.1$ for a Fisher test) and the absolute correlation level across all lags and distance decays, what should correspond roughly to the time-stationarity scale of the system. We observe furthermore a phase transition when distance decay increases, revealing the shift between the spatial scale of urban areas and the scale of the country, what gives local spatial stationarity scale.

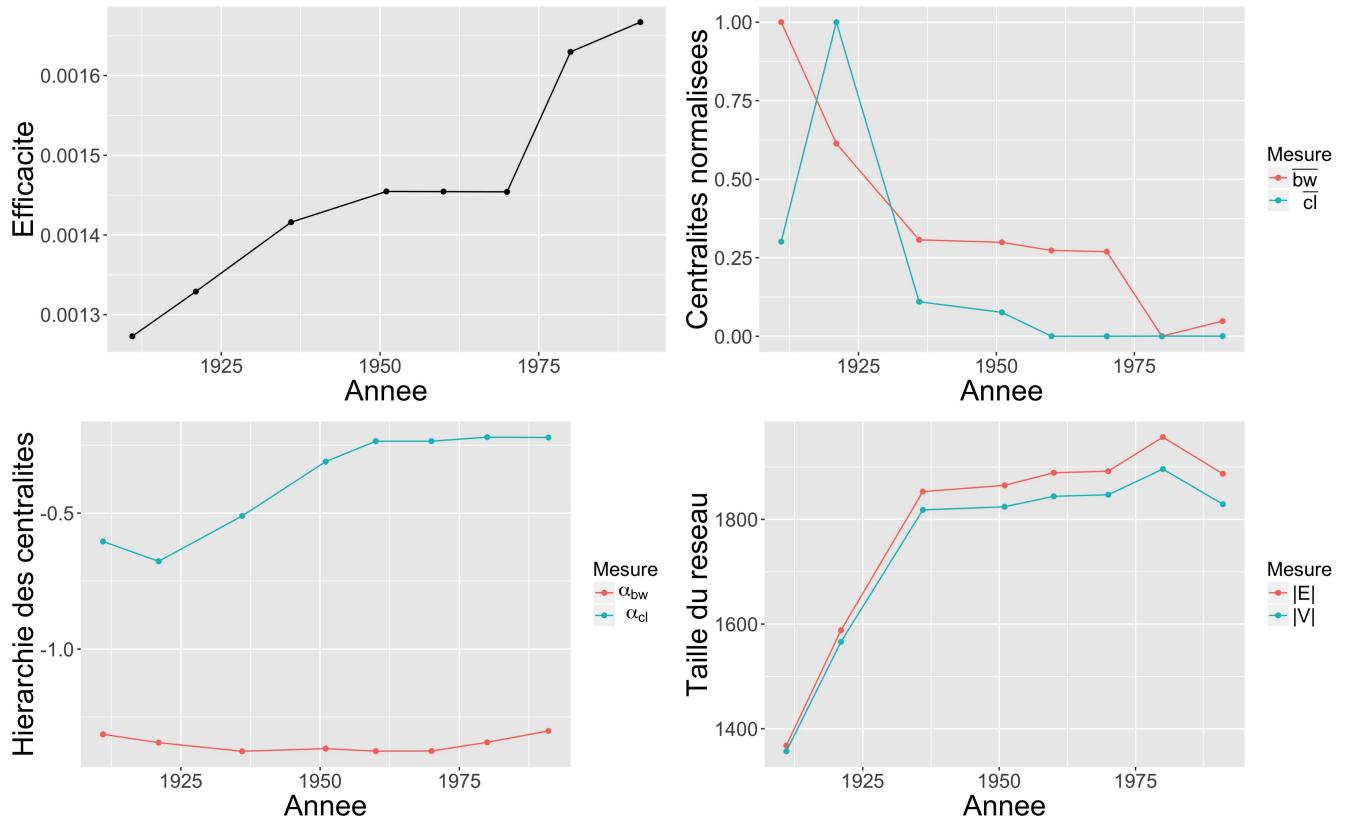


Figure 28:

We obtain therethrough clear causality patterns, namely an inversion of the Granger causality (lagged correlation up to 0.5 for several values of distance decay), from accessibility causing population growth with a lag of 10-20 years before the apartheid (1948), to the opposite after the apartheid (lag 20 years). We interpret these as *Structural segregation*, i.e. a significant impact of planning policies on dynamics of interactions between networks and territories. Indeed, the first regime corresponds to direct effect of transportation on migrations in a free context in opposition to the second one.

Possible developments

Further work should consist in similar study with more precise socio-economic variables, for example quantifying directly segregation patterns. The method of instruments in statistics [Angrist, Imbens, and Rubin, 1996] is used to identify causal relationships between variables, in a different way than Granger causality test for example. Trying to identify causalities between network dynamics and territorial dynamics is of crucial importance to test our theoretical assumption on the existence of co-evolution.

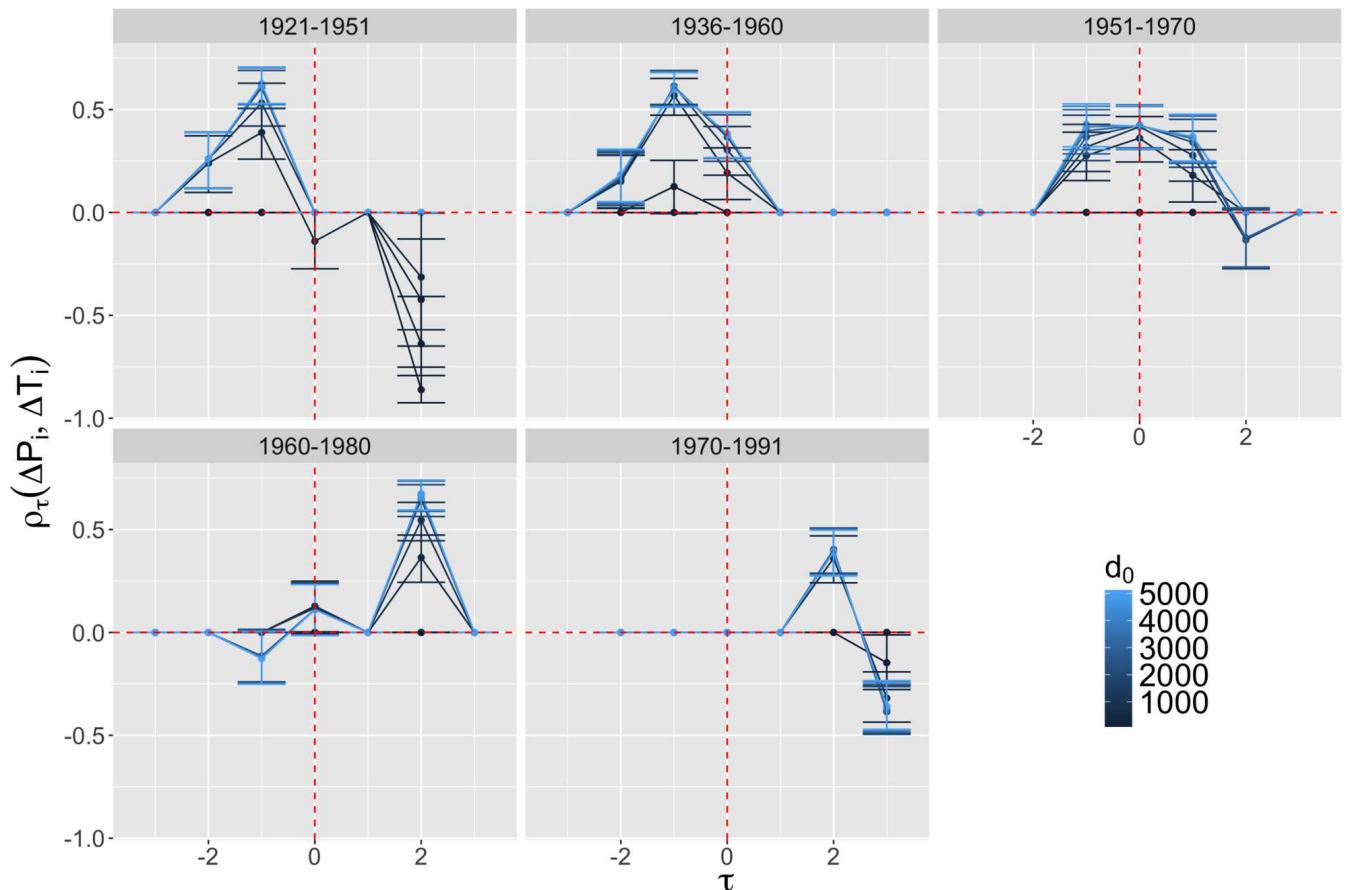


Figure 29:

* * *

*

Nous avons jusqu'ici dans ce chapitre exploré deux ingrédients de la théorie évolutive des villes, qui nous seront cruciaux pour comprendre la co-évolution entre réseaux de transport et territoires, à savoir les propriétés de non-stationnarité des corrélations, qui guideront la mise en place des modèles à une échelle similaire (chapitre 5 puis chapitre 7) et la possibilité de mise en évidence de régimes de causalité, qui nous servira d'outil de caractérisation de la co-évolution.

Nous proposons à présent d'introduire un dernier élément crucial de la théorie évolutive des villes, qui est l'appréhension des systèmes urbains par l'intermédiaire de modèles d'interaction entre villes. Ceux-ci ne seront pas co-évolutifs dans un premier temps mais leur ontologie visera à intégrer le rôle des réseaux dans le système urbain.

* * *

*

4.3 MACROSCOPIC GROWTH MODEL

Le dernier aspect de la Théorie Evolutive des Villes que nous proposons d'explorer se positionne sur le plan thématique et sur le plan de la modélisation : l'étude des villes elles-mêmes et de leur interactions, par l'intermédiaire de modèles de simulation. Comme nous allons le voir, la plupart des modèles de systèmes de villes issus de la théorie évolutive des villes se basent sur les interactions entre villes : cela nous permet une entrée directe dans notre problématique puisque celles-ci se font par l'intermédiaire des réseaux, qu'on pourra alors expliciter dans nos modèles.

We describe a simple spatial model of urban growth for systems of cities at the macroscopic scale, which combines direct interaction between cities and an indirect effect of physical network flows as population growth drivers. The model is parametrized on population data for the French system of cities between 1831 and 1999, which strong non-stationarity in correlation patterns suggest to apply the model on local time windows. The corresponding calibration of the model using genetic algorithms provide the evolution of interaction processes and network effects in time. Furthermore, the fit improvement when adding network module appears effective when controlling for additional parameters, what confirms the ability of the model to unveil network effects in the system of cities.

In this section we aim at exploring further the assumption, central to Pumain's Evolutive Urban Theory, that spatial interactions between cities are significant drivers of their growth. More precisely, we consider both abstract interactions and flow interactions mediated through the physical networks, mainly transportation network. We extend existing models accordingly.

Our contribution is twofold: (i) we show that very basic interaction models based on population only can be fitted to empirical data and that fitted parameter values are directly interpretable; and (ii) we introduce a novel methodology to quantify overfitting in models of simulation, as an extension of Information Criteria for statistical models, which applied to our calibrated models confirms that fit improvement is not only due to additional parameters, but that the extended model effectively capture more information on system processes. This will unveil network effects in an indirect way. We first review modeling approaches to urban growth based on spatial interactions.

4.3.1 *Spatial interaction models*

Models of Urban Growth

Cities are paradoxically both unsustainable and source of negative externalities, but also the best chance to reach sustainability and resilience to climate change (Glaeser, 2011). The dynamics of Urban

Systems at a macroscopic scale, and more precisely drivers of urban growth, are crucial to be understood to meet these potentialities. A better knowledge of how cities differentiate, interact and grow is thus a relevant topic both for policy application and from a theoretical perspective. [Pumain, Paulus, and Vacchiani-Marcuzzo, 2009] suggests that cities are incubators of social change, their fate being closely linked to the one of societies. Various disciplines have studied models of urban growth with different objectives and taking diverse aspects into account. For example, Economics are still reluctant to include spatial interactions in the models (Krugman, 1998) but are extremely detailed on market processes, even for models in Economic Geography, whereas Geography focuses more on territorial specificities and interactions in space but will produce general conclusion with more difficulty. The example of this two disciplines shows how it is difficult to make bridges, as it needed exceptional efforts to translate from one to the other (as P. Hall did for Von Thunen work (Taylor, 2016)), and therefore how it is far from evident to grasp the complexity of Urban Systems in an integrated way.

The simplest model to explain urban growth, the Gibrat model, that assumes random growth rates, has been shown by [Gabaix, 1999] to asymptotically produce the expected rank-size law (Zipf's law) for system of cities which is considered as one of the most regular stylized facts, at least in its generalized scaling law formulation (Nitsch, 2005). Explaining urban scaling laws is closely related to the understanding of urban growth, as [Bettencourt, Lobo, and West, 2008] suggests that these reflect underlying universal processes and that all cities are scaled version of each other. This approach however does not reflect the complex relation between economic agents for which [Storper and Scott, 2009] advocates.

Using a bottom reconstruction of urban areas using dynamical microscopic population data, [Rozenfeld et al., 2008] shows indeed that positive deviations to the rank-size law systematically exist, and that these must be an effect of spatial interaction between urban areas. Complexity approaches are good candidates to integrate these into models. [Andersson, Frenken, and Hellervik, 2006] introduce for example a model of urban economy as a growing complex network of relations. The Evolutive Urban Theory, introduced by [Pumain, 1997], focuses on cities as co-evolving entities and produces explanations for growth at the system of cities level. [Pumain et al., 2006] shows that scaling laws could be due to functional differentiation and diffusion of innovation between cities. The positioning regarding universality of laws is more moderate than Scaling theories, as [Pumain, 2012b] highlights that ergodicity can difficultly be assumed in the frame of complex territorial systems. One crucial feature of this paradigm is the importance of interactions between agents, generally the cities, to produce the emergent patterns at the scale of the system.

[Pumain and Sanders, 2013] has investigated the advantages of Agent-based models compared to more classical equation systems, and this methodological aspect is in accordance with the theoretical positioning, as it allows to take into account the heterogeneity of possible interactions, the geographical particularities, and to naturally translate emergence between levels and render multi-scale patterns.

Urban Growth and Spatial Interaction

First of all, we must precise that we consider only models at the macro-scale, ruling out the numerous and rich approaches at the mesoscopic scale, that include for example Cellular Automata models, models of Urban Morphogenesis or Land-use change models. We also naturally rule out economics models that do not include explicitly spatial interactions. Several models of Urban Growth at the macro scale have insisted on the role of space and spatial interactions. [Bretagnolle et al., 2000] proposed a spatial extension of the Gibrat model. The gravity-based interaction model that [Sanders, 1992] used to apply concept of Synergetics to cities is also close to this idea of interdependent urban growth, contained physically in the phenomenon of migration between cities. A more refined extension with economic cycles and innovation waves was developed by [Favaro and Pumain, 2011], yielding a system dynamics version of the core of Simpop models (Pumain, 2012a).

This family of models have started with a toy-model based on economic interactions between cities as agents, that yield hierarchical patterns at the scale of the system (Sanders et al., 1997). Later, the Simpop2 model, still based on distance interaction for commercial exchanges, including successive innovation waves, unveiled structural differences between the European and the US Urban Systems (Bretagnolle and Pumain, 2010a). The SimpopLocal model (Pumain and Reuillon, 2017c) is used to show the emergence of initial settlement patterns. The Marius model (Cottineau, 2014) couples population and economic growth with cities interaction, allowing to accurately reproduce real trajectories on the former Soviet Union after calibration with multi-modeling of processes.

Urban Growth and Transportation Networks

Nous situons ici l'aperçu que nous venons de donner en regard des modèles s'intéressant aux interactions entre territoires et réseaux que nous avons amplement revus en chapitre 2.

Under similar assumptions of previously reviewed models, the inclusion of transportation networks has been rarely pursued, contrary to the mesoscopic scale at which relations between networks and territories have been widely studied by Luti models for example (Chang,

2006). Network growth models (Xie and Levinson, 2009b), prolific in Economics and Physics, can not be utilized to explain urban growth.

[Bigotte et al., 2010] studies an optimization model for network design combining the effects of urban hierarchy and of transportation network hierarchy. [Baptiste, 1999] has modeled dynamical interplay between network links capacity and city growth on a subset of French city system. The SimpopNet model (Schmitt, 2014) goes a step further in modeling the co-evolution between cities and transportation networks, as it allows new network links to be created in time. These examples shows the difficulty of coupling these two aspects of urban systems in models of growth, and we will for this reason take into account network effects in a simplified way as detailed further.

The rest of this section is organized as follows : our model is introduced and formally described in next section; we then describe results obtained through exploration and calibration of the model on data for French cities, in particular the unveiling of network effects significantly influencing growth processes thanks to a novel methodology introduced. We finally discuss the implications of these results.

Le modèle de croissance à l'échelle macroscopique introduit et étudié en détails ici servira alors de brique élémentaire pour la construction des modèles de co-évolution que nous proposerons par la suite au chapitre 6.

4.3.2 Model and Results

Model Description

RATIONALE Some confusion may arise when surveying at stochastic and deterministic models of urban growth. To what extent is a proposed model “complex” and is the simulation of stochasticity necessary ? Concerning Gibrat model and most of its extensions, independence assumptions and linearity produce a totally predictable behavior and thus not complex in the sense of exhibiting emergence, in the sense of weak emergence (Bedau, 2002). In particular, the full distribution of random growth models can be analytically determined at any time (Gabaix, 1999), and in the case of studying only first moment, a simple recurrence relation avoids to proceed to any Monte-Carlo simulation. Under these assumptions, it is natural to work with a deterministic model, as it is done for example for the Marius model (Cottineau, 2014). We will work under that hypothesis, capturing complexity through non-linearity. We work on simple territorial systems assumed as regional city systems, in which cities are basic entities. The time scale corresponds to the characteristic scale associated to this spatial scale, i.e. around one or two centuries. Spatial interactions will be captured through gravity-type interactions, this formulation having the advantage of being simple and of capturing the first law of

Tobler, namely that interaction strength fades with distance. Other approaches introduced recently perform similarly at this scale (Masucci et al., 2013).

MODEL DESCRIPTION We consider on a deterministic extension of the Gibrat model, what is equivalent to consider only expectancies in time. Let $\vec{P}(t) = (P_i(t))_{1 \leq i \leq n}$ be the population of cities in time. Under Gibrat independence assumptions, we have $\text{Cov}[P_i(t), P_j(t)] = 0$. A linear extended version would write $\vec{P}(t+1) = \mathbf{R} \cdot \vec{P}(t)$ where \mathbf{R} is an independent random matrix of growth rates (a scalar times identity in the original case). It yields directly thanks to the independence assumption that $\mathbb{E}[\vec{P}(t+1)] = \mathbb{E}[\mathbf{R}] \cdot \mathbb{E}[\vec{P}](t)$.

Cela conduit directement grâce à l'hypothèse d'indépendance à $\mathbb{E}[\vec{P}(t+1)] = \mathbb{E}[\mathbf{R}] \cdot \mathbb{E}[\vec{P}(t)]$, ce qui revient à une formulation déterministe du modèle de Gibrat qui est équivalente à considérer seulement les espérances des populations dans le temps et ne plus simuler des trajectoires aléatoires.

We generalize this linear relation to a non-linear relation that allows to be more consistent with model interpretation and more flexible. Denoting $\vec{\mu}(t) = \mathbb{E}[\vec{P}(t)]$, we take $\vec{\mu}(t+1) = \Delta t \cdot f(\vec{\mu}(t))$. Note that in that case, stochastic and deterministic versions are not equivalent anymore, precisely because of the non-linearity, but we stick to a deterministic version for the sake of simplicity. The specification of the interdependent growth rate is given by

$$f(\vec{\mu}) = (1 + r_0) \cdot \mathbf{Id} \cdot \vec{\mu} + \mathbf{G}(\vec{\mu}) \cdot \vec{1} + \vec{N}(\vec{\mu}) \quad (7)$$

where $\vec{1}$ is the column vector full of ones, and $\mathbf{G} = G_{ij} = w_G \cdot \frac{V_{ij}}{\langle V_{ij} \rangle}$ such that the interaction potential V_{ij} follows a gravity-type expression given by, with d_{ij} distance between i and j (euclidian or network distance),

$$V_{ij} = \left(\frac{\mu_i \mu_j}{(\sum_k \mu_k)^2} \right)^{\gamma_G} \cdot \exp(-d_{ij}/d_G) \quad (8)$$

The network effect term \vec{N} is given by $N_i = w_N \cdot \frac{W_i}{\langle W_i \rangle}$ where the network flow potential W_i reads

$$W_i = \sum_{k < l} \left(\frac{\mu_k \mu_l}{(\sum_j \mu_j)^2} \right)^{\gamma_N} \cdot \exp(-d_{kl,i}/d_N) \quad (9)$$

where $d_{kl,i}$ is the distance of city i to the shortest path between k, l computed in the geographical space, which can be through a transportation network or in an impedance field of the euclidian space. All seven model parameters are detailed below.

The first component is the pure Gibrat model, that we obtain by setting the weights $w_G = w_N = 0$. The second component captures direct interdependencies between cities, under the form of a separable gravity potential such as the one used in [Sanders, 1992]. The rationale for the third term, aimed at capturing network effects by expressing a feedback of network flow between cities k, l on the city i . Intuitively, a demographic and economic flow physically transiting through a city or in its surroundings is expected to influence its development (through intermediate stops e.g.), this effect being of course dependent on the transportation mode since a high speed line with few stops will skip most of the traversed territories. Note that we don't use exactly gravity flows in the network term, since there is no decay of interactions generating flows with distance, but a decay of the effect of the flow as a distance to the network: it is equivalent to assuming long-range use of the network on average in time, and is this way complementary to the first gravity term.

MODEL PARAMETER SPACE We give in Table ?? the description of model parameters, detailing the associated processes and parameter ranges. Both direct interaction and second order network flows effect have the same structure, namely separability between effect of distance and population influence, an exponential decay parameter and a hierarchy parameter expressing the inequality of contribution depending on cities relative sizes: the highest the exponent, the more contribution of smaller cities will be negligible regarding larger cities. We propose to interpret the distance decay parameter the following way. Let fix an arbitrary fraction α and typical spatial ranges for a local urban system d_L and for a long range urban system d_R , consider a city i and two neighbors j, j' with same population $\mu_j = \mu_{j'}$, at distances d_L and d_R of i respectively. If we want to answer the question to what distance difference is equivalent an attenuation of α of the interaction potential with i , we obtain $d_L - d_R = -d_G \cdot \ln \alpha$. Therefore, d_G is exactly the proportionality coefficient answering this intuitive request. Finally, we will consider only positive weights w_G and w_N , to follow empirical observations as detailed below. Numerical values for the weights will be given normalized by number of cities implied in the process, i.e. $w'_G = w_G/n$ and $w'_N = w_N/(n(n-1)/2)$.

Data

Our model is assumed as hybrid as it relies on semi-parametrization on real data. It could be possible to study it as a full toy-model, initial configuration and physical environment being constructed as synthetic data. We however aim at unveiling stylized facts on real data rather than on model behavior in itself, and setup therefore the model from the data we now describe.

Table 12: **Interaction model parameters summary.** We give the parameters names, notations, associated processes, possible interpretations, typical variation ranges.

Paramètre	Notation	Processus	Interpretation	Domaine
Taux de Croissance	r_0	Croissance Endogène	Croissance Urbaine	$[0, 1]$
Poids gravitaire	w_G	Interaction directe	Croissance maximale	$[0, 1]$
Gamma gravitaire	γ_G	Interaction directe	Niveau de hiérarchie	$[0, +\infty]$
Décroissance gravitaire	d_G	Interaction directe	Portée d'interaction	$[0, +\infty]$
Poids de la rétroaction	w_N	Effet des flux	Croissance maximale	$[0, 1]$
Gamma de la rétroaction	γ_N	Effet des flux	Niveau de hiérarchie	$[0, +\infty]$
Décroissance de la rétroaction	d_N	Effet des flux	Portée de l'effet	$[0, +\infty]$

POPULATION DATA We work with the Pumain-INED historical database for French Cities (Pumain and Riandey, 1986), which give populations of *Aires Urbaines* (INSEE definition) at time intervals of 5 years, from 1831 to 1999 (31 observations in time). The latest version of the database integrates Urban Areas, allowing to follow them on long time-period, according to Bretagnolle's long time cities ontology (Bretagnolle, 2009), that constructs a functional definition of cities as entities with boundaries evolving in time. We work on the 50 bigger cities in 1999. We furthermore isolate periods of similar length excluding wars, obtaining 9 periods of 20 years on which semi-stationary in time fit of the model will be done.

PHYSICAL FLOWS As stated before, this modeling exercise focuses on exploring the role of physical flows, whatever the effective shape of the network. We choose for this reason not to use real network data which is furthermore not easily available at different time periods, and physical flows are assumed to take the geographical shortest path taking into account terrain slope. It avoids geographical absurdities such as cities with a difficult access having an overestimated growth rate. Using a 1km resolution Digital Elevation Model, we construct an impedance field of the form

$$Z = \left(1 + \frac{\alpha}{\alpha_0}\right)^{n_0}$$

where Z is the impedance of links of the 1km grid network in which each cell is connected to its eight neighbors. α is the terrain slope computed with elevation difference between the two cells. We take fixed parameter values $\alpha_0 = 3$ (corresponding to approximatively the real world value of a 5% slope) and $n_0 = 3$ which yielded more realistic paths than smaller values.

Performance indicators

We work on an explanatory rather than an exploratory model. Therefore, indicators to evaluate model outputs are not directly linked to intrinsic properties of trajectories or obtained final states, but rather to a distance to the phenomenon we want to explain, i.e. the data. Given real population $p_i(t)$ (historical realizations of $P_i(t)$) and simulated expected populations $\mu_i(t)$ obtained with $\bar{\mu}(t_0) = \bar{p}(t_0)$ on a period of length T , we can evaluate two complementary aspects of model performance:

- Overall model performance, given by logarithm of the mean-square error in space and time

$$\varepsilon_G = \ln \left(\frac{1}{T} \sum_t \frac{1}{n} \sum_i (p_i(t) - \mu_i(t))^2 \right)$$

- Average local model performance, given by the mean-square error on logarithms

$$\varepsilon_L = \frac{1}{T} \sum_t \frac{1}{n} \sum_i (\ln p_i(t) - \ln \mu_i(t))^2$$

Both are actually complementary, as using only ε_G as it is generally done will focus only on larger cities and give poor results on medium-sized and small cities (for France only Paris will have reasonable fit as it strongly dominates other urban areas and cities). ε_L allows therefore to take into account model performance in all cities simulated by the model.

Results

STYLIZED FACTS Basic stylized facts can be extracted from such a database, as it has already been widely explored in the literature [Guérin-Pace and Pumain, 1990]. We retrieve better fits of log-normal distributions of growth rates at all dates compared to normal fits, and also the fact that growth rates are mainly positive, on the cities we consider and when removing wars.

An interesting feature to look at in relation with our considerations on spatial interactions are correlations between time-series, and more particularly their variation as a function of distance. We consider 50 years overlapping time-windows to have enough temporal observation, finishing respectively in (1881, 1906, 1931, 1962, 1999), and estimate on each, for each couple of cities (i, j) , the correlation between log-returns $\hat{\rho}_{ij} = \rho [\Delta X_i, \Delta X_j]$ with a classical Pearson estimator, where $\Delta X_i = X_i(t) - X_i(t-1)$ and $X_i(t) = \ln \left(\frac{P_i(t)}{P_i(t_0)} \right)$. This method, used mainly in econophysics (Mantegna and Stanley, 1999), reveals dynamical interactions without being biased by sizes.

We show in Figure 30 the smoothed correlations curves as a function of distance, for each time period. First of all, the strong differences between each confirms the non-stationarity of growth rates over the whole time period, and justifies the use of local fit in time for the model. We can also interpret these patterns in terms of historical events for the system of city and the transportation network. System dynamic begins with a flat correlation in 1881, around 0.2, that could be spurious due to simultaneous similar growth for all cities. It then stays flat but goes to zero, witnessing strong differentiations in growth patterns between 1856 and 1906. After 1931, the effect of the distance is clear with decreasing curves, starting between 0.4 and 0.5. We postulate that this evolution must be partly linked to transportation network evolution: considering railway network for example (Thévenin, Schwartz, and Sapet, 2013), the initial overall development may have fostered long range interactions flattening thus the correlation curves, whereas its maturation over time has conducted to the return of more classical interactions decreasing quickly with distance.

MODEL EXPLORATION Data preprocessing, result processing and models profiling are implemented in R. For performances reasons and an easier integration into the OpenMole software for model exploration (Reuillon, Leclaire, and Rey-Coyrehourcq, 2013), a scala version was also developed. The question of trade-off between implementation performance and interoperability is a typical issue in this kind of model, as a fully blind exploration and calibration can be misleading for further research directions or thematic interpretations. A NetLogo implementation, allowing interactive exploration and dynamical visualization, was also developed for this reason. Source code for models, cleaned raw data, simulation data, and results used in this paper are available on the open repository of the project at <https://github.com/AnonymousAuthor1/InteractionGibrat.git>. We show in Figure 31 an example of model output. Cities color give city-level fit error and their size the population. Outliers can therefore easily be spotted (as Saint-Nazaire having the worst fit in the example shown) and possible regional effects identified. We illustrate in pink an example of geographical shortest path, from Rouen to Marseille, which reasonably corresponds to the actual current shortest path by highway. Top right plot shows trajectory in time for a given city, whereas the bottom right plot gives overall fit quality in time, by plotting simulated data against real data. The closest the curve is from the diagonal, the better the fit.

First model explorations, by simply sweeping fixed grids of the parameter space, already suggest the presence of network effects, in the sense that physical flow effectively have an influence on growth rates. We show in Figure 32 a configuration of such a case. At fixed gravity

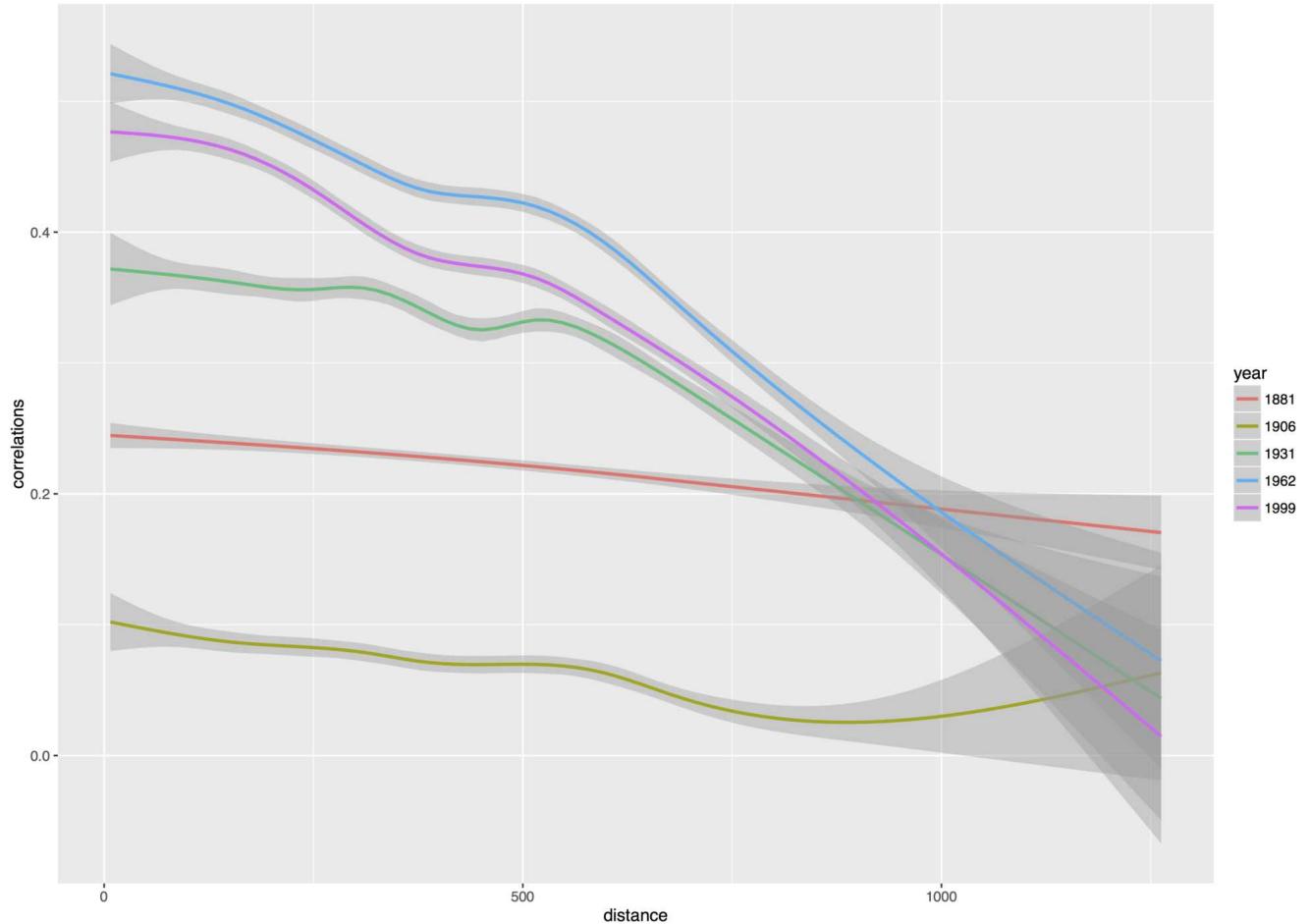


Figure 30: **Time-series correlations as a function of distance.** Solid line correspond to smoothed correlations, computed between each pairs of normalized log-returns of population time-series, on successive periods given by curve color.

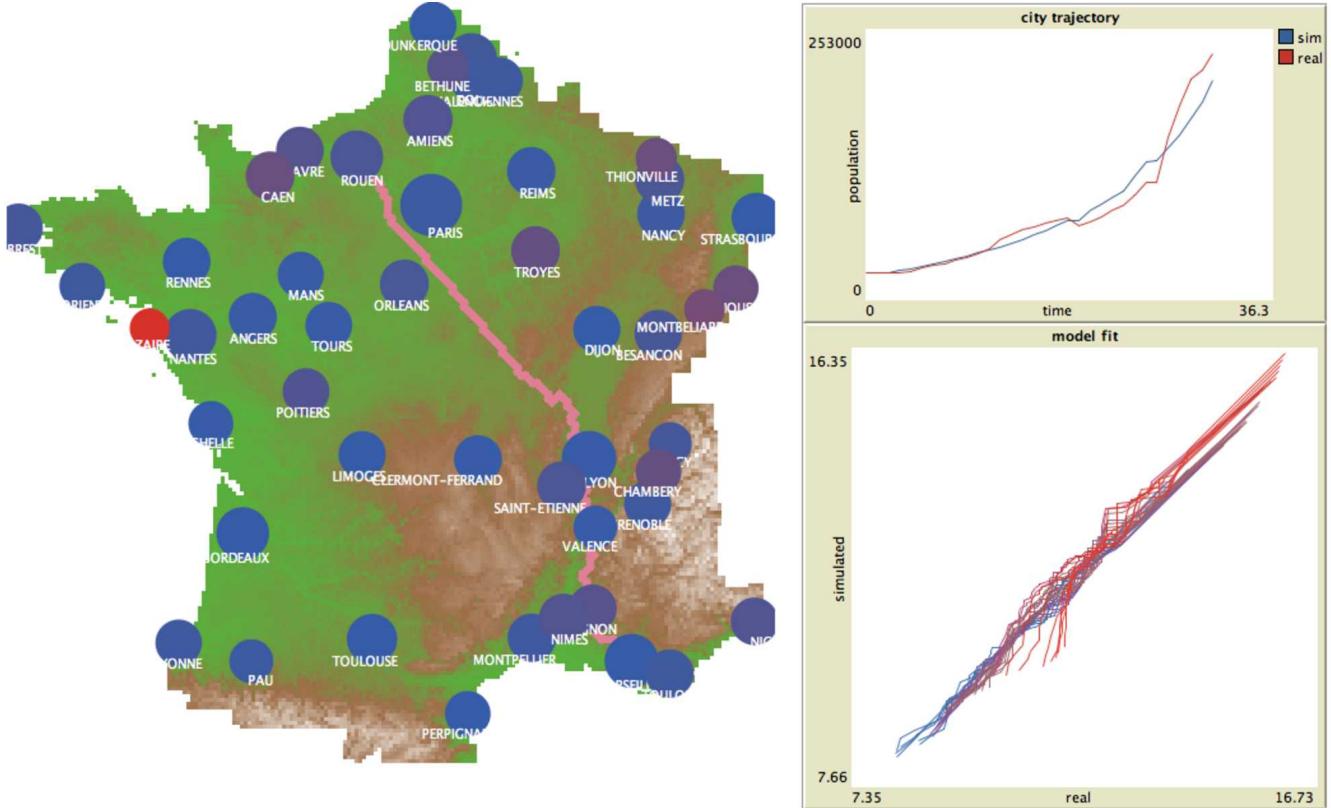


Figure 31: Example of output of the model. The graphical interface allows to explore interactively on which cities changes operate after a parameter change, what is necessary to interpret raw calibration results. The map gives adjustment errors by city (color) and their population (size). We illustrate in pink the geographical shortest path between Rouen and Marseille. The plot in top right panel gives in time the trajectory of a selected city, comparing simulated population with real population. The bottom right plot gives for each date the simulated data against real data: the closest the curve is to the diagonal, the better is the fit.

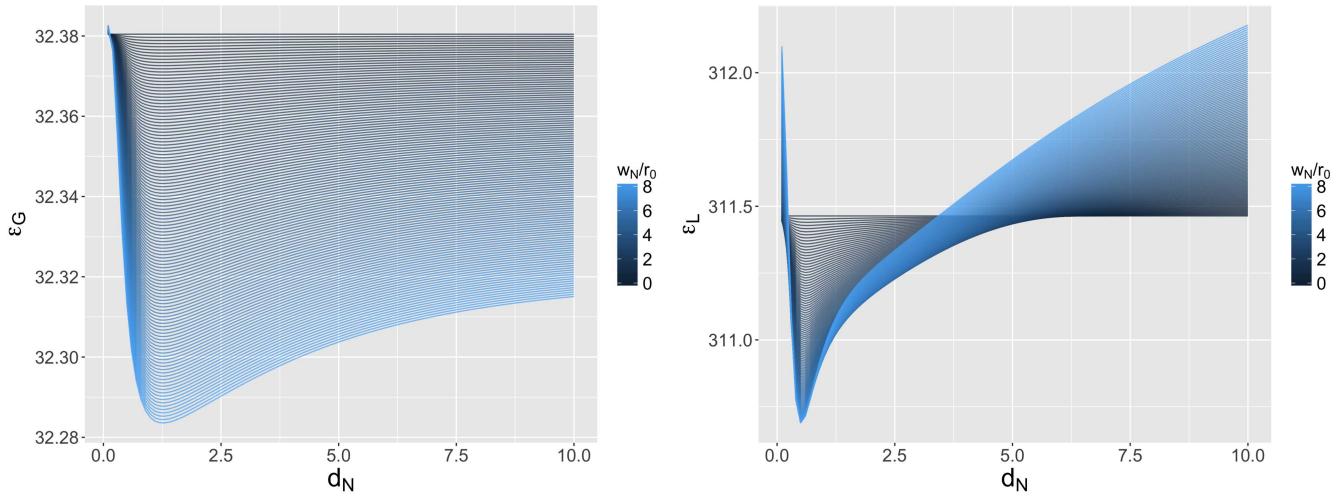


Figure 32: Evidence of network effects revealed by model exploration. Left plot gives ε_G as a function of d_N for varying r_0/w_N , at fixed gravity effect and $\gamma_N = 3$. Right plot is similar for ε_L .

parameters and growth rate, we study variations of the parameters w_N , d_N and γ_N and the corresponding response of ε_G and ε_L . At fixed values of γ_N , we observe similar behaviors of the indicators when w_N and d_N change. The existence of a minimum of both as a function of d_N , that becomes stronger when w_N increases, shows that introducing the network feedback terms improves local and global fits compared to the gravity model alone, i.e. that the associated process have potential explanatory power for growth patterns.

CALIBRATING THE GRAVITY MODEL We now use the model to indirectly extract information on processes in time. Indeed under assumption of non-stationarity, temporal evolution of locally fitted parameters show the evolution of the corresponding aspect of the processes. In a first experiment, we set $w_N = 0$ and calibrate the model with four parameters on the 9 successive 20 years time windows. The optimization problem associated to model calibration does not present features allowing an easy solving (closed-form of a likelihood function, convexity or sparsity of the optimization problem), we must rely on alternative techniques to solve it. Brute force grid search is rapidly limited by the dimensionality curse. Classical methods (Batty and Mackie, 1972) such as gradient descent fail because of the rather complicated shape of the optimisation landscape. Calibration using Genetic Algorithms (GA) are an efficient solution to find approximate solutions in a reasonable time. OpenMole embeds a collection of such meta-heuristics for different purposes: [Schmitt et al., 2015] demonstrates the capabilities of these methods to calibrate models of simulation. In our case, it furthermore allow to do a bi-objective calibration on $(\varepsilon_G, \varepsilon_L)$. We use the standard steady state GA

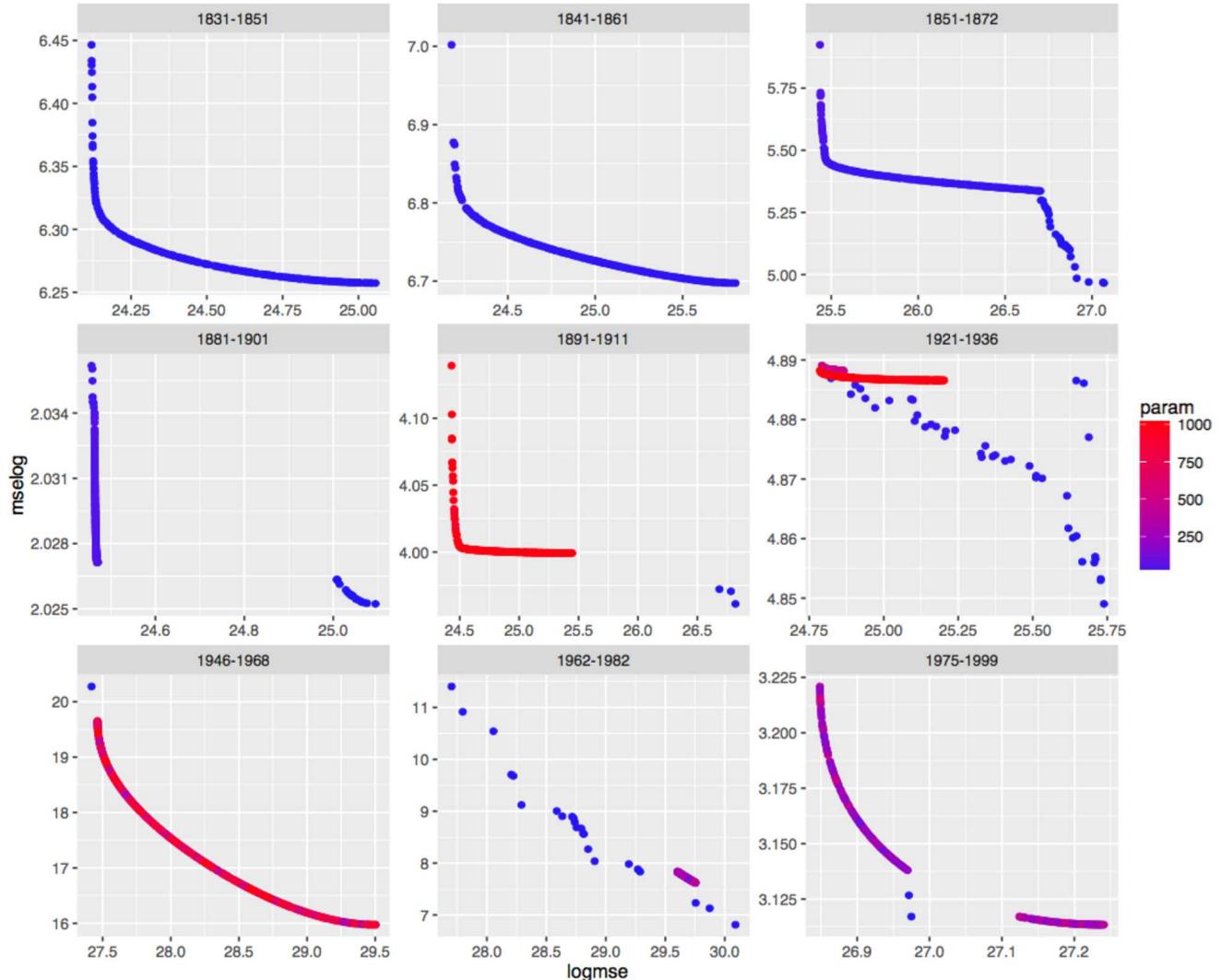


Figure 33: **Calibrating the Gravity Model.** Pareto-front on successive periods. Color level gives the value of distance decay parameter.

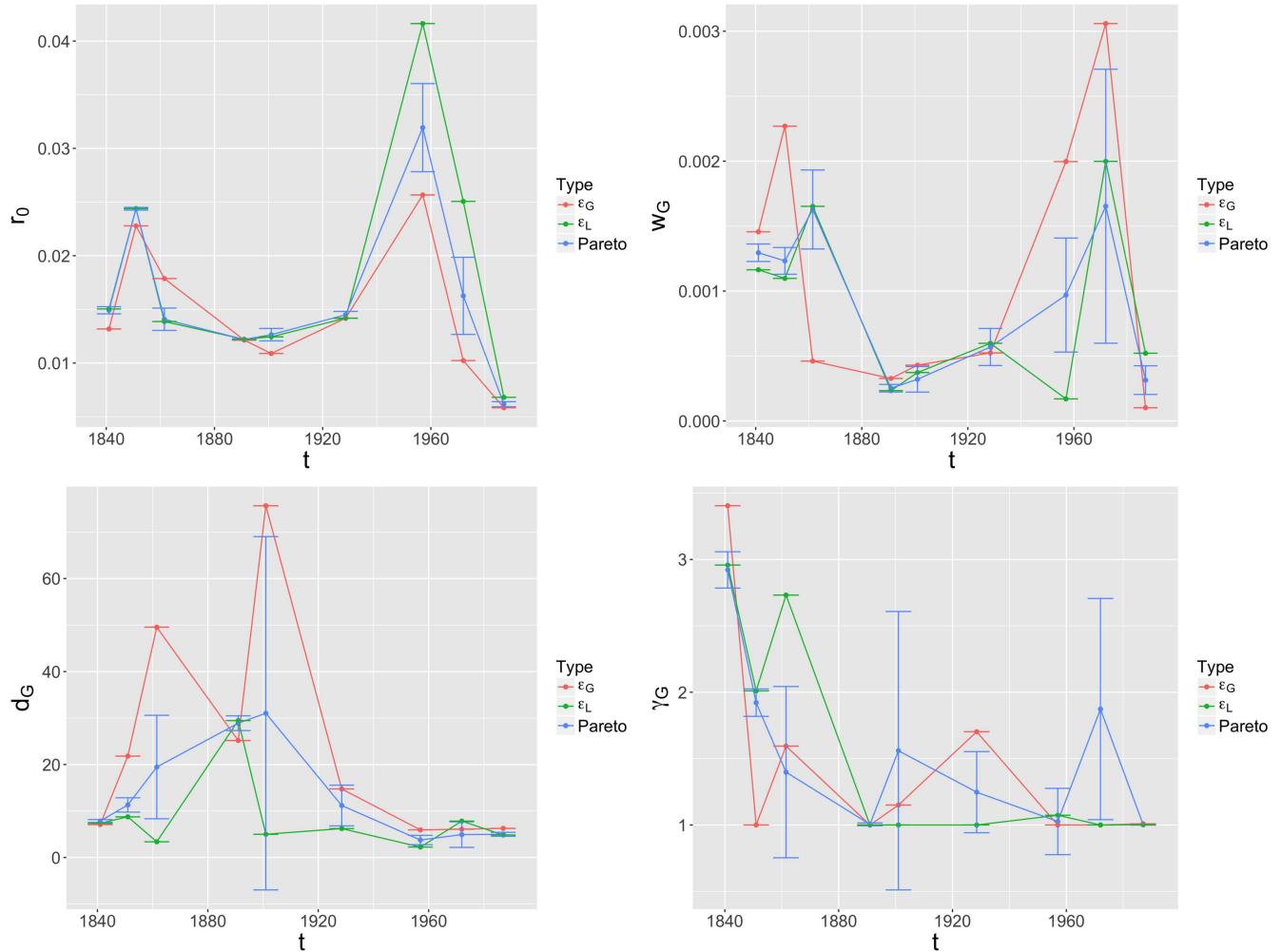


Figure 34: **Calibrated parameters for gravity model only.** Each plot gives fitted values in time for each parameter. Red and Green curves correspond to best points for ϵ_G (respectively ϵ_L), whereas the blue curves give the average value over the Pareto front with standard deviation.

provided by OpenMole, distributed on 25 islands, with population of 200 and 100 generations.

We show in Figure 33 the calibration results on successive periods, by plotting final population in the indicator space. As expected, Pareto fronts that corresponds to compromises between the two opposite objectives are the rule. It means that the model cannot be accurate both globally and locally, and an intermediate solution has to be found. This may due to the fact that interaction range changes with city size (i.e. that terms in the potential are no longer separable), that we keep as a possible model development. The shape of the Pareto front are revealing the chaotic optimisation landscape, as for some periods such as 1921-1936 or 1962-1982 fronts are not regular and sparse. The change in shapes also translates different dynamical regimes across the periods: for 1881-1901, the quasi-vertical shape followed by an isolated front at high ε_G values reveals a quasi-binary model behavior in the optimal regimes, in the sense that improving ε_L under the limit is only possible through a qualitative jump at a high price for ε_G .

The values taken by d_G for periods 1892-1911 and 1921-1936 show that larger cities have longer interaction range, as high value give better values of ε_G . We show in Figure 34 the values of fitted parameters in time, averaged over the Pareto front and for best single-objective solutions. First, the two peaks patterns for r_0 corresponds roughly to the patterns observed in average growth rates. The evolution of w_G has a similar shape but lagged by 20 years: it can be interpreted as a repercussion of endogenous growth on interaction patterns in the following years, which is consistent with an interpretation of the interaction process in terms of migration. The values of d_G , with an increase until 1900 followed by a progressive decrease, is consistent with the behavior of empirical correlations commented above: the first 50 years windows have higher interaction range what corresponds to flat correlation curves. Finally, the level of hierarchy γ_G has regularly decreased, corresponding to an attenuation of the power of large cities that can be understood in terms of progressive decentralization in France that has been fostered by the administration.

UNVEILING NETWORK EFFECTS We now turn to the calibration of the full model on successive periods, in order to interpret parameters linked to network flows and gain insight into network effects. The full calibration is done in a similar way with seven parameters being free. We plot in Figure 35 the fitted values in time for some of these parameters. The behavior of growth rate and of the gravity weight relative to growth rate, that is similar to the gravity model only, confirms that network effects are well at the second order and that endogenous growth and direct interactions are main driver. Net-

work effects are however not negligible, as they improve the fit as shown before in model exploration, capturing therein second order processes. The evolution of d_N , corresponding to the range on which network influences the territories it goes through, shows a minimum in 1921-1936 to stabilize again later, but at a value lower than past values. This could correspond to the “tunnel effect”, when high-speed transportation do not stop much. Indeed, the evolution of railway has witnessed a high decrease in local lines at a date similar to the minimum, and later the emergence of specific High Speed lines, explaining this lower final value. Hierarchy of flows have slightly decreased as for gravity, but are extremely high. This means that only flows between larger cities have a significant effect. This way, the model gives indirect information on the processes linked to network effects.

Nous retenons de la calibration du modèle complet les faits stylisés suivants.

- Des effets des réseaux sont capturés au second ordre par le modèle.
- Les variations de la portée de l'effet du réseau suggèrent l'émergence de l'effet tunnel.
- Les flux principaux dominent largement dans l'effet de réseau.

ESTIMATING THE COMPROMISE BETWEEN FITTING POWER AND NUMBER OF PARAMETERS We focus in this last experiment on quantifying the “performance” of the model, taking into account its predictive abilities, but also its structure. More precisely, we want to tackle the issue of overfitting, which has been for long recognized in Machine Learning for example [Dietterich, 1995], but for which there is a lack of methods for models of simulation. We need to introduce a tool to confirm that the improvement in model fit is not only artificially due to additional parameters.

The Akaike Information Criterion (AIC) provides for statistical models for which a likelihood function is available the gain in information between two models (Akaike, 1998), correcting fit improvement for number of parameters. Similar methods include the Bayesian Information Criterion (BIC), which relies on slightly different assumptions and corrects differently. [Biernacki, Celeux, and Govaert, 2000] proposes an integrated likelihood as a generalization of these criteria in unsupervised classification. [Pohle et al., 2017] shows that in the case of selecting the number of states in Hidden Markov Models, real cases induces too much pitfalls for standard methods to work robustly, and suggest pragmatic selection based on their results and expert judgement. In our case, the problem is that it is not even possible to define these.

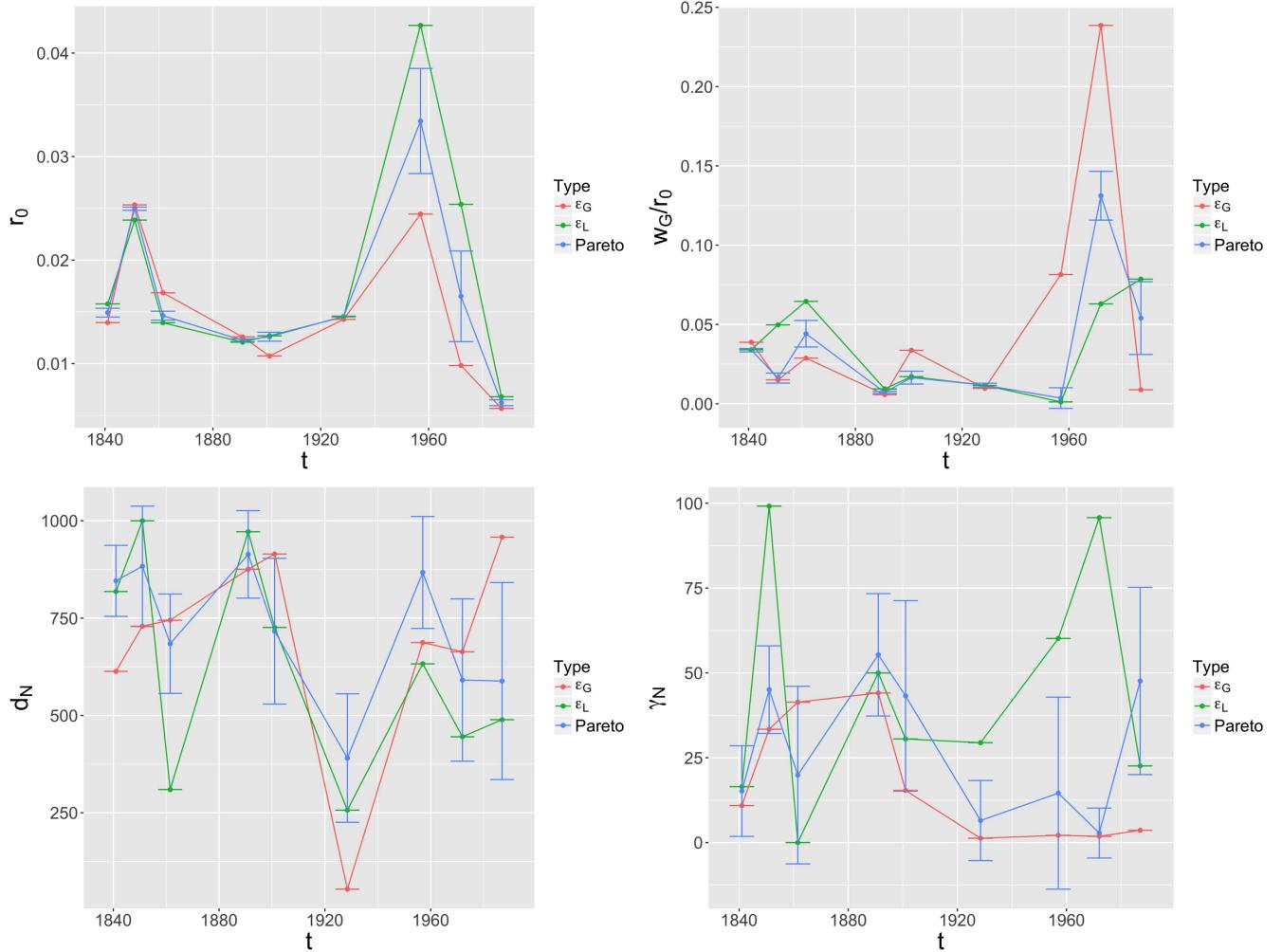


Figure 35: **Calibrated parameters for the full model.** We plot values of r_0 , w_G/r_0 , d_N and γ_N in time, for single-objective optimal points (Red and Green curves) and averaged over the Pareto front (Blue).

The method we propose is based on the intuitive idea of approaching models of simulation by statistical models and using the corresponding AIC under certain validity conditions. [Bastani, Kim, and Bastani, 2017] uses a similar trick of considering the models as black boxes and approaching them to gain insights, in their case to extract interpretable structure as decision trees.

Let (X, Y) be the data and observations. We consider computational models as functions $(X, \alpha_k) \mapsto M_{\alpha_k}^{(k)}(X)$ mapping data values to a random variable. What is seen as data and parameters is somehow arbitrary but is separated in the formulation as corresponding dimensions will have different roles. We assume that the models have been fitted to data in the sense that an heuristic has been used to find an approximate optimal solution $\alpha_k^* = \operatorname{argmin}_{\alpha_k} \|M_{\alpha_k}^{(k)}(X) - Y\|$, and we write $\varepsilon_k = \|M_{\alpha_k}^{(k)}(X) - Y\|^2$ the corresponding mean-square error.

For each optimized computational model, a statistical model $S^{(k)}$ with the same degree of freedom can be fitted on a set of realizations: $M_{\alpha_k^*}^{(k)}(X) = S^{(k)}(X)$, with an error $s_k = \|M_{\alpha_k^*}^{(k)}(X) - S^{(k)}(X)\|^2$. If statistical models are good approximations of models compared to models discrepancy to reality, namely $s_k \ll \varepsilon_k$, then the gain of information between the two should mostly capture the gain of information between simulation models.

We define therefore an *Empirical AIC* measure between two simulation models by

$$I(M^{(1)}, M^{(2)}) = \Delta AIC [S^{(1)}, S^{(2)}] \quad (10)$$

In practice we calibrate the gravity only model and the full model on the full time span, and choose two intermediate solutions giving $M^{(1)}$ at $r_0 = 0.0133, d_G = 4.02e12, w_G = 1.28e-4, \gamma_G = 3.82$ with $\varepsilon_G = 31.2375, \varepsilon_L = 302.89$ and the full model $M^{(2)}$ at $r_0 = 0.0128, d_G = 8.43e14, w_G = 1.230e-4, \gamma_G = 3.81, w_N = 0.60, d_N = 7.47e14, \gamma_N = 1.15$ with $\varepsilon_G = 31.2366, \varepsilon_L = 302.93$. It is not clear how the empirical method is sensitive to the type of statistical model used, we use therefore severals for robustness, each time with the corresponding number of parameters (4 for the first and 7 for the second model): a polynomial model of the form $a_0 + \sum_{i>0} a_i X^i$, a mixture of logarithm and polynomial as $a_0 + a_1 \ln X + \sum_{i>1} a_i X^i$ and a generalized polynomial with real power coefficients that are optimized for model fit using a genetic algorithm $a_0 + \sum_{i>0} a_i X^{\alpha_i}$. We fit the statistical models using successive years as different realizations. Results for each are shown in Table ???. We give the value of s_k/ε_k and the ΔAIC . We also provide the ΔBIC to check the robustness regarding the information criterion used. We find a positive value for 5 criteria out of 6, what means that information gain is indeed positive. The gain decreases when statistical model fit improves, and only the BIC

for the optimized model fails to show an improvement. The assumption of negligible errors is always verified as the rate is always around 1%. This approach is of course preliminary and further work should be done for a more systematic testing and more robust justification of the method. It suggests however that fit improvement in the model of simulation are effective, and that the model reveals therefore network effects.

Table 13: Empirical AIC results.

Modèle Statistique	Ajustement pour $M^{(1)}$	Ajustement pour $M^{(2)}$	ΔAIC	ΔBIC
Polynomial	0.01438	0.01415	19.59	3.65
Log-polynomial	0.01565	0.01435	125.37	109.43
Polynomial Généralisé	0.01415	0.01399	11.70	-4.23

4.3.3 *Towards co-evolutive models*

Our focus on network effects remains quite limited since (i) we do not consider an effective infrastructure but abstract flows only, and (ii) we do not take into account the possible network evolution, due to technical progresses (Bretagnolle et al., 2000) and infrastructure growth in time. An interesting development would be first the application of our model with real network data, using effective distance matrices in time, computed e.g. with the train network used by [Thévenin, Schwartz, and Sapet, 2013]. Then, allowing the network to dynamically evolve in time, as a function of flows, would yield a model of co-evolution between cities and transportation networks for a system of cities, which has been proven empirically by [Bretagnolle, 2009]. This kind of model is very rare, and [Schmitt, 2014] provides with Simpop-Net one of the few examples. It is shown by [Raimbault, 2016d] that disciplinary compartmentalization may be at the origin of the relative absence of such type of models in the literature. Indeed, it would imply to include heterogenous processes such as economic rules to drive network growth, that are quite far from the approach taken. It would however allow to investigate to what extent the refinement of network spatial structure and network dynamics can improve the explanation of urban system dynamics.

We have introduced a spatial model of growth for a system of cities at the macroscopic scale, including second order network effects among endogenous growth and direct interaction growth drivers. The model is parametrized on real data for the French city system between 1831 and 1999. The calibration of the model in time provides interpretations for the evolution of processes of interaction within the system of cities. We furthermore show that the model effectively unveils network effects by controlling for overfitting. This work paves

the way for more complicated models with dynamical networks, that would capture the co-evolution between transportation network and territories.

★ ★

★

CHAPTER CONCLUSION

La notion de co-évolution, qui était jusqu'ici dans notre travail relativement conceptuelle, apparaît sous de multiples angles nouveaux complémentaires. Ce chapitre permet d'éclairer son rôle au sein de la théorie évolutive des villes. Celle-ci sera également centrale pour la construction théorique que nous élaborerons en 8.2.

En effet, des interdépendances fortes peuvent se traduire par des corrélations locales variables, c'est-à-dire une non-stationnarité spatiale, induite d'une part par les motifs locaux correspondant à une régime d'interaction donné, dont nous avons pu capturer les manifestations statiques en section 4.1, d'autre part par le caractère multi scalaire des processus impliqués que nous avons également montré, et donc par les interactions à grande échelle et portée entre les différentes entités territoriales, que nous avons illustré sur un cas simple par le modèle d'interaction étudié en 4.3, qui a déjà permis de révéler indirectement des effets de réseaux dans les systèmes de villes.

On a également éclairé une approche dynamique de la co-évolution, en montrant la complexité potentielle de la structure des relations causales dans le cas d'un modèle de morphogenèse urbaine simple. La méthodologie développée s'est montrée également efficace sur les données réelles de l'Afrique du Sud sur le temps long, permettant de découvrir un effet des politiques de ségrégation au second ordre sur la co-évolution elle-même. Cette méthode nous servira de caractérisation empirique de la co-évolution par la suite.

* * *

*

MORPHOGENÈSE URBAINE

La géographie accorde une grande importance aux relations spatiales et à la mise en réseau, comme l'atteste par exemple la première loi de TOBLER combinée au fait que les réseaux sont vecteurs des interactions. Nous l'avons mis en évidence pour les relations entre réseaux et territoires par exemple en section 4.3. Toutefois, nos résultats sur la non-stationnarité, ainsi que la mise en valeur d'échelles locales endogènes, suggèrent une certaine pertinence de l'idée de sous-systèmes relativement indépendants. Il serait alors possible d'isoler certaines règles locales régissant un sous-système, un fois fixés certains paramètres exogènes capturant justement les relations avec d'autres sous-systèmes. Cette question porte à la fois sur l'échelle d'espace, de temps, mais aussi sur les éléments concernés.

Reprendons un exemple concret de terrain déjà évoqué au chapitre 1 : la laborieuse mise en place du tramway de Zhuhai. L'impact du retard de la mise en place et la remise en question de futures lignes (dus à un problème technique inattendu lié à une technologie de transfert de courant par troisième rail importée d'Europe qui n'avait jamais été testée dans les conditions climatiques locales assez exceptionnelles en termes d'humidité), aura une nature très différente selon l'échelle et les acteurs urbains considérés. Le manque de coordination générale entre transports et urbanisme laisse supposer que les dynamiques urbaines en termes de populations et d'emplois y sont relativement insensibles dans l'immédiat. Le Bureau des Transports de la Municipalité de Zhuhai ainsi que le bureau technique européen ayant conçu la technologie défectueuse ont pu subir des répercussions politiques et économiques bien plus conséquentes, tandis que par ailleurs, que ce soit à Zhongshan, Macao ou Hong-Kong, nous pouvons supposer que le problème a une répercussion quasi-nulle, le projet ayant un rôle uniquement local. Il existe ainsi des jeux complexes d'indépendances et d'interdépendances relatives dans les systèmes territoriaux.

Généralisant au système de transport local, celui-ci peut être relativement bien isolé des systèmes voisins, et donc ses relations avec la ville considérée dans un contexte local. Il est possible de supposer à la fois une certaine forme de stationnarité locale mais aussi une certaine indépendance avec l'extérieur. Le type de raisonnement que nous avons esquissé mobilise les éléments essentiels propres à l'idée de *morphogenèse urbaine*.

Nous allons dans ce chapitre clarifier sa définition et montrer les potentialités qu'elle donne pour éclairer les relations entre réseaux et territoires. Dans un premier temps, un effort d'épistémologie par

des points de vue complémentaires de plusieurs disciplines permet d'éclairer la nature de la morphogenèse dans la section 5.1. Cela permet de clarifier le concept en lui donnant une définition bien précise, distincte de celle de l'auto-organisation, qui appuie les relations causales circulaires entre forme et fonction.

Nous explorons ensuite un modèle simple de morphogenèse urbaine, basé sur la densité de population seule, à l'échelle mesoscopique, dans la section 5.2. La démonstration que les processus abstraits d'agrégation et de diffusion sont suffisants pour reproduire une grande diversité de formes d'établissements humains en Europe, en utilisant les résultats de la section 4.1, confirme la pertinence de l'idée de morphogenèse pour la modélisation à certaines échelles et pour les dimensions morphologiques.

Ce modèle est ensuite couplé de manière séquentielle à un modèle de morphogenèse de réseau dans la section 5.3, afin d'établir un espace possible des corrélations statiques entre indicateurs de forme urbaine et indicateurs de réseau, qui sont comme on l'a vu précédemment un témoin des relations locales entre réseaux et territoires.

Nous posons ainsi d'autres briques de modélisation de la co-évolution, à l'échelle mesoscopique par l'entrée de la morphogenèse urbaine.

* * *

*

Ce chapitre est composé de divers travaux. La première section est adaptée d'un travail en anglais en collaboration avec C. ANTELOPE (University of California), L. HUBATSCH (Francis Crick Institute) et J.M. SERNA (Université Paris VII) à la suite de l'école d'été 2016 du Santa Fe Institute [Antelope et al., 2016]; la deuxième section est traduite de [Rimbault, 2018b]; et enfin la troisième section a été écrite pour les Actes des Journées de Rochebrune 2016 [Rimbault, 2016b].

5.1 AN INTERDISCIPLINARY APPROACH TO MORPHOGENESIS

A first crucial step is a clarification of what is meant by morphogenesis.

The notion of morphogenesis seems to play an important role in the study of a broad range of complex systems. If the concept was introduced in embryology to design growth of organisms, it was rapidly used in various fields, e.g. urbanism, geomorphology and even psychology. However, the use of the concept seems generally fuzzy and to have a field-specific definition for each use. We propose in this section an epistemological study, starting with a broad interdisciplinary review and extracting essential notions linked to morphogenesis across fields. It allows to build a consistent general meta-framework for morphogenesis. Further work may include concrete application of the framework on particular cases to operate interdisciplinary transfers of concepts, and quantitative text analysis to strengthen qualitative results.

CONTEXT During every historical period, people use the main technological advance as a metaphor to explain other phenomena in nature. First, nature was mechanical, then electrical, and now computational. Here, we suggest that taking an alternative metaphor might allow us to better study some properties of a system, and study how the concept of morphogenesis that originated in the study of developmental biology, can be used across systems. Morphogenesis is a very powerful metaphor that is distinct from the previous three that have been very popular in history. Unlike the mechanical, electrical or computational explanations of nature, morphogenesis is not a human designed process. Morphogenesis emphasizes the role of change and growth, rather than a static state. As [Thompson, 1942] already pointed out, “natural history deals with ephemeral and accidental, not eternal nor universal things”. The goal of this paper is to study three questions:

1. How is morphogenesis defined in different fields?
2. Are there fields that use approaches and concepts that embody the notion of morphogenesis but do not use the word?
3. Can approaches to study morphogenesis be applied across different fields?

A similar effort is described in [Bourgine and Lesne, 2010], but it consists more of a collection of viewpoints from subjects that can be related to morphogenesis rather than an epistemological reconstruction of the notion as we propose to do. Furthermore, examples are far from exhausted and our review is thus complementary.

Dans le cadre de notre problématique globale, cet effort nous permettra d'une part de considérer des systèmes territoriaux cohérents

par l'hypothèse de l'existence de sous-système morphogénétiques, et d'autre part de lier territoires et réseaux par l'intermédiaire du lien crucial entre forme et fonction que nous allons développer ci-dessous.

The rest of this section is organized as follows : we provide first a compartmentalized review of the notion of morphogenesis across various fields, ranging from biology to social sciences, psychology and territorial sciences. A synthesis is then made and a framework as general as possible proposed. We finally discuss further developments and potential application of this epistemological analysis.

5.1.1 *Reviews*

Nous proposons un aperçu large de la manière dont est utilisée la notion de morphogenèse dans des domaines a priori très éloignés. Notre revue ne se prétend pas exhaustive et nous n'utilisons pas de méthode systématique, l'idée étant de mobiliser et de croiser différentes conceptions pertinentes de la notion.

Developmental Biology

In developmental biology, morphogenesis refers to the mechanisms of how an organism acquires its shape and different functional units, starting from only one cell. Generally, these mechanisms need to work reliably in order to guarantee similar outcomes for every individual. This often requires cells to know their position relative to some reference frame, in order to differentiate (a term used to describe how cells acquire a specific fate, becoming, say, skin cells, as opposed to neurons) or to decide whether or not to divide (which is often necessary for growth). The following section describes models that have been applied in developmental biology.

REACTION-DIFFUSION MECHANISMS Alan Turing used the term reaction-diffusion system in his seminal 1952 paper 'The Chemical Basis of Morphogenesis' to describe simple patterning in a (theoretical) ring of cells [Turing, 1952b]. Even though this work is now considered one of the most fundamental contributions to the field of pattern formation, it took many years until his work started getting recognition as an actual model for biological systems. Gierer&Meinhardt [Gierer and Meinhardt, 1972] then suggested using similar models also for intracellular polarity - a ubiquitous phenomenon in biology in which a cell establishes and maintains two different regions within itself - an important capability of most cell types. These reaction diffusion networks are one example of the emergence of patterns from a homogeneous state. Using this framework we can recapitulate many pattern formation mechanisms in development, such as coloration, segmentation as well as establishment and maintenance of cell polarity. These larger scale patterns are generated by the interaction of

a few species of chemicals. Every chemical species also undergoes diffusion, production and degradation. Thus it is possible to represent this model using a system of partial differential equations, and certain parameters will generate stable patterns from homogeneous initial condition, where random perturbations are amplified by the system. With only a few molecular species, very complex patterns can be formed [Kondo and Miura, 2010]. One of the most studied reaction-diffusion model capable of producing stable patterns comprises of two types of molecules, one activator and one repressor. The difference in diffusion rate between the two molecules is what amplifies random noise in the system [Gierer and Meinhardt, 1972; Turing, 1952a]. The most well-studied reaction-diffusion system explaining coloration is in zebrafish. Cells called melanophores and xanthophores produce black and yellow pigments respectively [Nakamasu et al., 2009]. The interaction of melanophores and xanthophores produces the striped pattern on zebrafish. Melanophore growth is promoted by long-distance interaction with xanthophores. Short distance interactions between the two cell types inhibit each other. Polarity formation in yeast division can also be explained by a reaction-diffusion mechanism involving the small Rho-GTPase Cdc42. Cdc42 has two forms, one active membrane bound and one inactive cytoplasmic [Goryachev and Pokhilko, 2008] Phenomena like body segmentation in *Drosophila melanogaster* usually involves a more complex system than the two previously discussed examples, because the pattern it generates needs to be robust to ensure the correct function, and thus cannot be sensitive to variation in initial conditions.

THE FRENCH FLAG MODEL Similar to the reaction-diffusion framework the French Flag Model was initially conceived to explain differentiation of cells in a regular fashion [Wolpert, 1969]. For example, stripes of cells within a tissue might need to assume different fates. Similar to a French flag, which has three differently colored stripes, cells within a tissue were thought to assume different fates if they are exposed to different concentrations of a certain protein - generically called a morphogen. This requires a graded concentration of the morphogen, which can be achieved if the morphogen is only produced locally and then diffuses across the tissue, away from the source. In order to achieve a stable gradient, on top of local protein production the system also needs either long range inhibition, a sink opposite to the source, or a degradation mechanism within the tissue (reviewed in [Rogers and Schier, 2011; Wolpert, 2011]). Once the gradient is set up, it can be interpreted linearly (for example by increasing the expression of a gene linearly with morphogen concentration) or switch-like by feedback mechanisms which are thought to further amplify the morphogen signal and is then translated to a specific cell fate at each position. This can be achieved via a variety of genetic circuits de-

pending on the tissue. To the author's knowledge there is no single well understood system, but there is evidence for such mechanisms at work at least on a coarse grain level [Wolpert, 2011]. For rigorous verification of these models, precise measurements of molecular mobilities as well as production and decay rates are necessary, alas very hard to obtain (fluorescent tags that are often used might change the molecules behavior, *in vivo* measurements are hard to perform, etc.). Also, experimenters are often confronted with biological redundancy, which can obscure effects of individual proteins.

FORCES AS DRIVERS OF CELL AND TISSUE SHAPE Epithelial rearrangements are often driven by intracellular forces that are generated by motor proteins acting on the cytoskeleton (e.g. kinesins walking along microtubules, actomyosin-mediated cortical tension) [Heisenberg and Bellaïche, 2013; Lecuit and Lenne, 2007]. These forces are then mediated between cells via cell-cell junctions forming an adhesive ring around a cell. These junctions are dynamic and can be remodelled, which can lead to seemingly fluid behavior when external stress is applied for a prolonged time. On short timescales, however, cells exhibit an elastic response, assuming their previous shape if an intermittent external force is no longer present. Tissues need to grow and often change shape during development. This can be driven by divisions, cell death, cell extrusion, or intercalation [Guillot and Lecuit, 2013]. An example of a well-studied tissue shape change can be found during mesoderm invagination in *Drosophila melanogaster*. In this case cells that initially form a flat layer become a long furrow by constricting their cell membrane area on one side [Lecuit and Lenne, 2007].

Ces considérations sont à la fois lointaines et proches de notre problématique générale : il existe par exemple des modèles bio-mimétiques appliqués aux systèmes urbains, comme pour la génération d'un réseau de transport [Tero et al., 2010]¹.

Artificial Life

As reviewed in [Crosato, 2014], the notion of *Programmable Self-Assembly* seems for students of Artificial Life to be very close to the biological concept of morphogenesis : "The greater example of Programmable Self-Assembly in nature is probably the cell organisation in multicellular organisms, which is encoded by the DNA." An important approach is Doursat's concept of Morphogenetic Engineering, that focuses on designing complex systems from the bottom-up. A review of the field is done in [Doursat, Sayama, and Michel, 2013].

¹ Voir aussi une initiative récente par des biologistes montant un projet interdisciplinaire visant à appliquer les principes complexes d'organisation spatiale complexe des protéines à l'espace urbain, qui s'est concrétisée dans la conférence "Gestion optimisée de l'Espace : des villes aux systèmes naturels" en décembre 2017 : <https://gopro2017.sciencesconf.org/>.

An essential distinction between self-organization and morphogenesis that it introduced is the presence of an architecture.

An example of a heterogeneous swarm of particles, yielding complex architectures is described in [Doursat, 2008]. The processes of local interactions (corresponding in biology to local physical forces) and positional information through gradient propagation are both integrated in the swarm model and allow bottom-up assembly of complex patterns. The combination of a chemical reaction layer with a hydrodynamic layer also provides an interesting model of morphogenesis in [Cussat-Blanc et al., 2012].

Territorial Sciences

The concept is used in various disciplines dealing with territories and the built environment: geography, urban planning and design, urbanism, architecture. There does not seem to be a unified view nor theory within these fields, not even within each field itself.

BUILT ENVIRONMENT Architecture and Urbanism are disciplines studying human settlements and the built environment at relatively small scales. OLSEN's theory of Urban Metabolism [Olsen, 1982] links city morphogenesis with urban metabolism and urban ecology. The city is seen as a living organism with different time scales of evolution (the life cycles). The study of Urban Morphology [Moudon, 1997], which focuses on morphogenetic processes, is presented as an emerging field in itself, across geography, architecture and urban planning: this view emphasizes the crucial role of the form in these kind of processes. [Burke, 1972] studies the growth of a particular city during a given period of time, and attributes the evolution of urban morphology to *morphogenetic agents*, i.e. people and developers.

At another scale, in architecture, a building can be seen as the results of micro-processes making sense and a particular architectural style interpreted through the use of generative shape grammars [Ceccarini, 2001]. This methodology is not far from the work of C. ALEXANDER, an architect who produced a theory of design process [Mehaffy, 2007], inspired from computer science and biology and linked in some aspects to complexity. The notion of morphogenesis is in that case however quite loose, as referring to the process of form generation in general, such as [Whitehand, Morton, and Carr, 1999] that studies concrete changes in house forms as witnesses of urban morphogenesis.

DOLLENS refers to autopoiesis [Dollens, 2014], implying a particular case of morphogenesis, to advocate Turing's influence on current design thinking, and to propose a more organic approach to architecture.

URBAN MODELING The Urban growth modeling literature often refers to the growth process as morphogenesis when the scale implied allows to exhibit shape patterns. An example of the emergence of qualitatively different urban functions, based on the Alonso-Muth model is proposed in [Bonin and Hubert, 2012].

[Makse et al., 1998] studies a model of urban growth involving the local urban form. In this case the local spatial correlations induce urban structure when the cities gain new inhabitants. More heterogeneous models imply a coupling between city components and transportation networks. [Achibet et al., 2014] describe a model of co-evolution between road network and urban blocks structure. At a larger scale and involving more abstract functions, [Raimbault, Banos, and Doursat, 2014] couples city growth with network growth, including local feedback of the form through a density constraint and global feedback of position through network centrality and accessibility to amenities. These two mechanisms are analogous to the local interaction and global information diffusion flow in biology.

ARCHEOLOGY The morphogenesis of past human settlements viewed from Thom's Catastrophe Theory point of view, is introduced by [Renfrew, 1978]. Sudden changes (qualitative changes, or regime shifts) have occurred at any time and can be viewed as bifurcations during the morphogenesis process. Another simplified way to see this is to interpret the transition as a change of meta-parameters of a stationary dynamic.

Social Science and Psychology

Morphogenesis has been occasionally used as a suitable metaphor to understand different processes in social science and various psychological fields. For example, in developmental psychology one can think of the relation to evolution of human cultural behavior and learning, epigenetic neural systems, and their influence on neural development and behavior throughout life [Hart, 2013]. In Clinical Psychology and Psychopathology, analogies are used for the emergence of psychical structures and the self-organization of forms of relation with the self and the Other. Additionally, "psychological morphogenesis" is akin to the outcome of the complexity of psychological dynamics undergoing creative emergence. Therefore, in "successful" psychotherapy this generation of novelty would be fostered [Piers, Muller, and Brent, 2007]. Moreover, in the field of neuroscience there are a plethora of morphogenetic phenomena related to the structure of the brain, dendritic morphogenesis and neural nets being some remarkable examples [INRA, 2013]. In social psychology we have noteworthy illustrations like the morphogenetic approach proposed by Margaret Archer as applied to the problem of structure and agency, that is, how we both shape society and are shaped by it in a dy-

namic interplay [Archer, 2010]. Nonetheless, more than a systematic and widespread unity throughout these different fields, we encounter multiple uses that are sometimes discontinuous, and one could argue that the utility of morphogenesis could be more tangible on an epistemological level. This would consist of a shared perception of morphogenesis's descriptive power to further understand the emergence and structure of various phenomena.

Epistemology

Morphogenesis is also used to study science itself: for example [Gilbert, 2003] studies the evolution of evolutionary developmental biology through the metaphor of morphogenesis. He sees scientific ideas as interacting agents from which emerge new phenotype through differentiation processes, what is designed as the morphogenesis of the field.

History of the notion

The study of morphogenesis started with embryology between just before 1930's. This is about the same time as Hodgkin and Lister, reported seeing red blood cells under a microscope, and less than 10 years before Dujardin's discovery of cellular movement in Amoeba. [Abercrombie, 1977] Using google book, the first use of the word morphogenesis in a book is in 1871, saw a large peak in usage between 1907-1909, and continued to increase in usage until the 1990's before slowing decreasing in usage.

Putting into perspective

Ces voyages par diverses disciplines nous ont permis déjà de dégager des idées clés et des concepts voisins à la morphogenèse. Nous concluons cette revue par une mise en perspective pour gagner en généralité.

A MATHEMATICAL APPROACH René Thom, in *Structural stability and Morphogenesis* [Thom, 1972] has developed a theory of system dynamics, the "catastrophe theory", that studies in deep the impact of topological structure of phase space manifolds on a system dynamics. Let M a differentiable manifold, in which system state (m, \dot{m}) is embedded. We assume the existence of a closed set K , called *Catastrophe set*. The topological type of K is indeed endogenously determined by system dynamics (in simple cases, it refers to the "classical" type of attractors/fixed points usually known: points, limit cycles). When m encounters K , the system follows a *qualitative* change in its form, what constitutes the basis of *morphogenesis*. This abstract theory of morphogenesis is independent of the nature of the system studied, its main contribution being to classify local catastrophes that occur

during morphogenesis. Differentiation and richness of patterns have thus a geometrical explanation through the topological types of catastrophes. Thom notes that at this time, the study of form has mainly been the focus of biology, but that many applications could be done in physics and geomorphology for example. He formulated the hypothesis that it is because it implies discontinuities and self-organisation, to which mathematicians were repulsive, that it was not applied easily to various fields. We can link this to the rise of complexity approaches, with complexity paradigms that slowly spreaded in various disciplines, and the study of morphogenesis seem to have followed.

AUTOPOIESIS It is interesting to note that Varela and Maturana's theory of autopoiesis in biology, from which they develop an observer-dependent interpretation of cognition, language, and consciousness, had a constructive epistemological impact on social science, philosophy and psychology, even if sometimes latent. For example, an application in sociology can be found in Niklas Luhmann's Systems Theory. His generalized view of autopoiesis conceptualizes systems as self-producing, not in terms of their physical components, but in terms of their organization, which can be measured in terms of information and complexity[Gershenson, 2015]. These views provide insight on the interpenetration between social and psychical systems. In Luhmann's theory, the 'human being' is not conceptualized as forming a systemic unity, but instead is understood as a conglomerate of organic and psychical systems, with language being the most important evolutionary achievement for the coupling of social and psychical systems. Language is thus a social phenomenon, yet thought processes are structured in a complementary way to language, as thoughts are broken down into chunks of sentences and words. [Seidl, 2004] We could further assess the epistemological significance of this if we consider the conception of the subject as dynamic and recursive, thus in a movement that can interact with its environment. This stance stems away from classically static conceptions of the human psyche, and echoes some contemporary clinical approaches in psychology and psychoanalysis. One concept that clearly illustrates this is Pichon Rivière's notion of ECRO (Schema Conceptual Referential and Operative), as the working processes which constitute the tools from which the subjects mental operations flow[Pichon Rivière, 2004]. Moreover, the interpenetration of the psychological and the social and the importance of language points us in the direction of psychoanalytical theory and clinical practice, with Jacques Lacan's views on linguistics and the big Other as well as Sigmund Freud's psychoanalytic anthropology that emphasizes the links between the neurotic patient's symptom and sociocultural phenomena [Freud, Strachey, and Freud, 1989].

The notion of *autopoiesis* expresses the ability for a system to reproduce itself. A basic characterization is a semi-permeable boundary produced within the system and the ability to reproduce its components. A more general definition is proposed by Bourgine and Stewart in [Bourgine and Stewart, 2004]: “*An autopoietic system is a network of processes that produces the components that reproduce the network, and that also regulates the boundary conditions necessary for its ongoing existence as a network*”. The notion of dynamical processes is key, and could be linked to Thom’s theory of morphogenesis. They furthermore introduce a definition of cognition (trigger actions as function of sensory inputs to ensure viability), and of living organism as autopoietic and cognitive, both notions being distinct [Bitbol and Luisi, 2004]. In that frame, for example, the arbotron [Jun and Hübner, 2005] is cognitive but not autopoietic. An example of link between autopoiesis and morphogenesis is shown in [Niizato, Shirakawa, and Gunji, 2010], where a type of Physarum organism has to play both on cell mobility and form evolution to be able to collect the food necessary for its survival. At this stage, we can postulate a strict inclusion from autopoietic systems, morphogenetic systems to self-organizing systems.

CO-EVOLUTION Since morphogenesis can be transposed to ecosystem or societies, and the components of the system are co-evolving in those cases, the existence of co-evolution may be linked with morphogenesis, as an other way of seeing the system. Symbiosis in biology can lead to very strong causalities in organism evolution (co-evolution) : this phenomenon has been designed as *symbiogenesis*. The symbiosis induce an change in morphogenetic patterns of symbiotic organisms as exemplified for different species in [Chapman and Margulis, 1998]. Thus the strong link between morphogenesis and co-evolution (here morphogenesis designing more evolutionary paths of morphogenetic patterns, i.e. at a different time scale).

SYSTEM DEFINITION AND BOUNDARIES La morphogenèse d’un système doit être considérée en même temps que la définition des limites d’un système, et la capacité des frontières à l’ouvrir et le fermer à la fois². La théorie des systèmes complexes adaptatifs de [Holland, 2012] se base sur une représentation de ceux-ci par des systèmes de frontières pouvant filtrer des signaux échangés entre systèmes. Cela rejoint la vision d’un système autopoïétique, et dans le cas morphogénétique, il est possible de supposer des limites floues (la difficulté dans la modélisation de tels systèmes étant alors la définition du système et de ses limites). Ces systèmes sont toutefois capables de maintenir une complexité par la combinaison complexe de

² Au chapitre 4, la définition de ces frontières avait été fixée de manière exogène. Leur rôle pour les systèmes morphogénétiques suggère la possibilité d’une approche endogène, comme finalement nous l’avons fait en révélant des régimes territoriaux endogènes en 4.1.

l'ouverture et de la fermeture [Morin, 1976]. Dans le cas des systèmes urbains, la morphogenèse est plus crédible qu'une autopoïèse stricte, puisque leur propriétés changent selon la définition des limites [Cotineau et al., 2015] (voir aussi l'Annexe B.2 qui montre de manière théorique la sensibilité des lois d'échelles à la définition du système).

5.1.2 *Synthesis*

Key notions

We list here important concepts that come out from this review, and from which a synthetic vision should emerge. Each may be domain-dependent, and underlying conceptions may vary from one field to the other.

- **Self-organisation** : Morphogenesis implies self-organisation but the contrary is not necessarily true, some aspects are specific of morphogenesis, such as the presence of functions resulting from the form.
- **Patterns and shape** : The “formation of shapes” seems to be common to all approaches to morphogenesis.
- **Embryogenesis / tissue modeling** In biology, typical processes of morphogenesis are generally observed at early stages of life, during embryogenesis, including the initial formation of tissues.
- **Apoptosis** Morphogenesis is often related to life (see section on autopoiesis), but also to death : the programmed death of cells, apoptosis, can in some cases be a part of morphogenetic processes.
- **Qualitative vs Quantitative** Qualitative bifurcations are a fundamental concept in morphogenesis : e.g. differentiation of organs in biology ; emergence of differentiated urban functions
- **Symmetry** Symmetry breaking occurs, mostly at early stages, but also at all stages of morphogenesis.
- **Unit and Scale** Are systems top-down or bottom-up designed, self-organized or exhibiting architecture ? Both are not necessarily incompatible, fundamental units and scales playing a crucial role in defining morphogenesis. Fractal-like systems, such as corals (collaborating tissues) or cities, but also the self and the society, can be studied from the point of view of morphogenetic processes at different levels.

- **Boundaries** Boundaries are a major aspect in Complex Adaptive systems (see e.g. Holland's approach as *Signals and Boundaries* [Holland, 2012]). Morphogenesis can imply clear boundaries (of an embryo e.g.) but not necessarily (social organisms, cities for which the definition of boundaries is still an open question [Cottineau et al., 2015]).
- **Relation between Form and Function** Causal relations between form and function are at the center of emerging architecture.

Common processes and differences

FROM LOCAL INTERACTIONS TO GLOBAL INFORMATION FLOW

The interplay between agent-to-agent interactions, either through neighborhood effects such as mechanistic interactions and diffusion, or through network interactions such as signaling, and the feedback of a global information flow (i.e. a downward causation of the upper level) appears to be common to most use of morphogenesis. It highlights the fundamental multi-level nature of morphogenetic processes and the central role of emergence.

FROM SELF-ORGANIZATION TO MORPHOGENESIS : THE NOTION OF ARCHITECTURE

Most system studied seem to have the particularity to exhibit an architecture, what would make the distinction between self-organization and morphogenesis. This idea comes from the field of morphogenetic engineering (which can be seen as a sub-field of artificial intelligence). This point may be a divergence point on some fields, as for example in physical science, where the "morphogenesis" of terrain patterns is a self-organization in our sense. The notion of architecture may be tricky to define. A way to do it is to consider the functions of macro-levels in the system : the emergence of function at an upper level implies an architecture, which is *the link between the form and the function*. Here this last concept takes all its sense and importance in regard to morphogenesis.

Proposition of a Meta-epistemological Framework

FRAMEWORK We propose a hierarchical organisation of concepts, that can be seen as a meta-epistemological framework, since definitions are built from synthesis of the many disciplines evoked here, and that their application in each particular discipline yields an epistemological frame. The concepts are organized the following way :

Self-organization \supseteq Morphogenesis \supseteq Autopoiesis \supseteq Life (11)

each having a generic definition, elaborated from the synthesis of disciplines.

Definition : Self-organization. A system is self-organized if it exhibits weak emergence [Bedau, 2002].

Definition : Morphogenesis. A self-organized system is the result of morphogenetic processes if it exhibits an emergent architecture, in the sense of causal relations between form and function at different levels.

Definition : Autopoiesis and Life. We take the definition of Bourgine [Bourgine and Stewart, 2004] for autopoiesis, that extends Bitbol's [Bitbol and Luisi, 2004], who also define life as autopoiesis with cognition.

The boundary between self-organization and morphogenesis is the existence of causal links between form and function, which can be defined as *architecture* [Doursat, Sayama, and Michel, 2013], generally emergent from the bottom-up. We observe that the complexity of systems increase with notion depth, what can be loosely translated in the fact that :

- Emergence strength [Bedau, 2002] diminishes with depth, in the sense that the number of autonomous scales increases.
- Number of bifurcations increases [Thom, 1972], i.e. path-dependency increases.

APPLICATION An ontological specification [Livet et al., 2010], i.e. the definition of entities to which the notion apply, yields an application to a particular field, each one developing its own properties and level of inclusion between concepts. There is a priori no reason for a direct correspondence or equivalence of projected concepts, thus transfer of knowledge between fields may be subject to caution.

Nous illustrons en Encadré 11 divers exemples de systèmes pouvant être qualifiés de morphogénétiques ou non, selon la perspective fonctionnelle que l'on en prend. Ceux-ci sont présentés par disciplines. Nous voyons ainsi le caractère générique du cadre ainsi que sa flexibilité.

5.1.3 Discussion

Avant de positionner l'utilité de la construction de ce concept par rapport à notre problématique générale, détaillons quelque développements potentiels propres à cet effort interdisciplinaire.

Towards a more systematic construction

Our work relies for now on a broad but not *systematic* review, in the sense of the methodology used for example in therapeutic evaluation, and where they play a role as important as primary studies, new knowledge being created through systematic comparison of results

	Physique	Biologique	Ingénieré
Non-Fonctionnel			
Fonctionnel			

Nous donnons des illustrations dans trois disciplines typiques, et proposons une interprétation des caractères fonctionnels ou non (rendant les systèmes de la première ligne non-morphogénétiques en notre sens). Dans l'ordre de gauche à droite et de haut en bas : modèle d'érosion (bibliothèque NetLogo) ; jeu de la vie (voir 3.3, bibliothèque NetLogo) ; *Swarm chemistry* (particules abstraites ayant des règles de mouvement et d'interaction), implémenté à partir de [Sayama, 2007] ; arbotron (billes de métal sous l'influence d'un potentiel électrique) [Jun and Hübler, 2005] ; modèle de fourmis (bibliothèque NetLogo) ; conception industrielle [Aage et al., 2017]. Le jeu de la vie, s'il est utilisé comme ordinateur, peut avoir des aspects fonctionnels. De même l'arbotron ou le nuage de particules de la *Swarm chemistry* peuvent être utilisés comme instruments. Tout est question de perspective dans laquelle le système est placé.

FRAME 11: Examples of morphogenetic systems.

and meta-analysis. It would imply in our case an iterative approach
:

- Blind systematic review, without any a priori on the fields concerned and on the way to express the notion.
- Extraction of main fields ; extraction of synonyms and close notions (such as we did here with autopoiesis and self-assembly for example ; if needed iteration of the first general review.

- Systematic reviews specific to each field, as each one has its own bibliographical databases, specific ways of communication, etc.
- Confrontation of each notion from one field to other fields.

L'objectif dans notre cas serait d'enrichir, comme nous l'avons déjà fait de manière préliminaire, mais systématiquement, le concept de morphogenèse urbaine.

Quantitative Epistemology

Our position may be also strengthen by quantitative approaches to literature analysis. With text-mining, keywords and concept extraction from abstracts (or even full texts) is possible, and would allow to confront our qualitative analysis to empirical data, by answering questions such as: is a concept indeed central, or what concept is used the same way in most disciplines. [Chavalarrias and Cointet, 2013] for example reconstructs scientific fields from the bottom-up through text-mining, and studies their lineage and dynamics in time. An other approach would be an iterative extraction of concept, by an algorithmic systematic review such as the one done in [Raimbault, 2017d].

Transfer of Knowledge between fields

Concrete applications of our framework include potential transfer of knowledge between fields. As biological systems inspire system architecture in morphogenetic engineering, or as the use of gravity models inspired from physics have flourishing applications in geography, we think that trying to decline the general framework in specific disciplines may bring analogies or new models that would have been difficult to formulate otherwise.

★ ★

★

The exploration of the concept of morphogenesis realized in the previous section allows to guide the conception of urban growth models. Models based on this concept will have to present the following properties:

1. Crucial role of the *form*, and thus inclusion both of a definition and a measure of the form, but also role of it in ontologies.
2. Strong coupling between form and function. In a first time, the function will not be explicit in the ontology but indeed present in abstract processes.
3. Autonomy of sub-systems, i.e. existence of a certain level of modularity in the global system. This property guides us both in the modeling scale, that we will take as “intermediate”, or mesoscopic, and also in the search for simple models, i.e. that are parsimonious in processes taken into account and in the number of parameters.

The strategy we follow to integrate these properties in morphogenesis models which will lead us to co-evolution models between transportation networks and territories, is progressive: progression in the span of ontologies (in terms of number of aspects taken into account) and progression in complexity, that we will interpret here as a coupling strength. The two following sections present thus first a morphogenesis model aimed at being minimalist for population density only, secondly the weak (sequential) coupling of it with a road network generation model. The strong coupling and the explicitation of functions through the network, producing the basis of a co-evolution model, will be the object of chapter 7.

★ ★

★

5.2 URBAN MORPHOGENESIS BY AGGREGATION-DIFFUSION

We study a stochastic model of urban growth generating spatial distributions of population densities at an intermediate mesoscopic scale. The model is based on the antagonist interplay between the two opposite abstract processes of aggregation (preferential attachment) and diffusion (urban sprawl). Introducing indicators to quantify precisely urban form, the model is first statistically validated and intensively explored to understand its complex behavior across the parameter space. We then compute real morphological measures on local areas of size 50km covering all European Union, and show that the model can reproduce most of existing urban morphologies in Europe. It implies that the morphological dimension of urban growth processes at this scale are sufficiently captured by the two abstract processes of aggregation and diffusion.

5.2.1 *Context*

Urban Growth

The study of urban growth, and more particularly its quantification, is more than ever a crucial issue in a context where most of the world population live in cities which expansion has significant environmental impacts [Seto, Güneralp, and Hutyra, 2012] and that have therefore to ensure an increased sustainability and resilience to climate change. The understanding of drivers for urban growth can lead to better integrated policies.

It is however a question far from being solved in the diverse related disciplines: Urban Systems are complex socio-technical systems that can be studied from a large variety of viewpoints. BATTY has advocated in that sense for the construction of a dedicated science defined by its objects of study more than the methods used [Batty, 2013b], what would allow easier coupling of approaches and therefore urban growth models taking into account heterogeneous processes. The processes that a model can grasp are also linked to the choice of the scale of study.

At a macroscopic scale, models of growth in system of cities are mainly the concern of economics and geography. We reviewed them in 4.3, and recall here that these can be more or less spatialized, and include interaction models to which belong for example Simpop models and their offspring.

Cellular automata

At larger scales, agents of models fundamentally differ. Space is generally taken into account in a finer way, through neighborhood effects for example. For example, [Andersson et al., 2002] propose a

micro-based model of urban growth, with the purpose to replace non-interpretable physical mechanisms with agent mechanisms, including interactions forces and mobility choices. Local correlations are used in [Makse et al., 1998], which develops the model introduced in [Makse, Havlin, and Stanley, 1995], to modulate growth patterns to resemble real configurations. The world of Cellular Automata (CA) models of Urban Growth [Batty and Xie, 1994] also offers numerous examples. [Xie, 1996] introduced a generic framework for CA with multiple land use, based on local evolution rules. A model with simpler states (occupied or not) but taking into account global constraints is studied by [Ward, Murray, and Phinn, 2000]. The Sleuth model, initially introduced by [Clarke and Gaydos, 1998] for the San Francisco Bay area, and for which an overview of diverse applications is given in [Clarke et al., 2007], was calibrated on areas all over the world, yielding comparative measures through the calibrated parameters.

Urban morphogenesis

Closely related to CA models but not exactly similar are Urban Morphogenesis models, which aim to simulate the growth of urban form from autonomous rules. We already described several in 5.1, and propose now to situate them regarding the models above. The link is clear, since for example [Frankhauser, 1998] suggests that the fractal nature of cities is closely linked to the emergence of the form from the microscopic socio-economic interactions, namely urban morphogenesis. [Courtat, Gloaguen, and Douady, 2011] develops a morphogenesis model for urban roads alone, with growth rules based on geometrical considerations. These are shown sufficient to produce a broad range of patterns analog to existing ones. Similarly, [Raimbault, Banos, and Doursat, 2014] couples a CA with an evolving network to reproduce stylized urban form, from concentrated monocentric cities to sprawled suburbs. The Diffusion-Limited-Aggregation model, originating from physics, and which was first studied for cities by [Batty, 1991], can also be seen as a morphogenesis model. These type of models, that sometimes can be classified as CA, have generally the particularity of being parsimonious in their structure. Similar models have also been studied in biology for the diffusion of population as for example [Bosch, Metz, and Diekmann, 1990].

The particularity of these models, compared to cellular automata, is the crucial role of the form in their evolution rules, and for some of the function, such as for [Bonin and Hubert, 2012]. We will follow here a similar logic of rules based on form (in a first time) and function (in chapter 7) to construct interaction models between territories and networks.

Objective

We study in this section a morphogenesis model, at the mesoscopic scale, aimed at being simplistic in its rules and variables, but trying to be accurate in the reproduction of existing patterns for the urban form (in the sense of 4.1). The underlying question is to explore the performance of simple mechanisms in reproducing complex urban patterns. We consider abstract processes, namely aggregation and diffusion, candidates as partially explanatory drivers of urban growth, based on population only, that will be detailed in model rationale below. An important aspect we introduce is the quantitative measure of urban form, based on a combination of morphological indicators, to quantify and compare model outputs and real urban patterns. Our contribution is significant on several points: (i) we compute local morphological characteristics on a large spatial extent (full European Union); (ii) we give significant insights into model behavior through extensive exploration of the parameter space; (iii) we show through calibration that the model is able to reproduce most of existing urban forms across Europe, and that these abstract processes are sufficient to explain urban form alone. The rest of this paper is organized as follows: we first describe formally the model and the morphological indicators. We then detail values of morphological measures on real data, study the behavior of the model by exploring its parameter space and through a semi-analytical approach to a simplified case, and we describe results of model calibration.

5.2.2 *Model and results*

Urban growth model

RATIONALE Our model is based on widely accepted ideas of diffusion-aggregation processes for Urban Processes. The combination of attraction forces with repulsion, due for example to congestion, already yield a complex outcome that has been shown under some simplifying assumptions to be representative of urban growth processes. A model capturing these processes was introduced by [Batty, 2006], as a cell-based variation of the *Diffusion-Limited-Aggregation* (DLA) model [Batty, 1991]. Indeed, the tension between antagonist aggregation and sprawl mechanisms may be an important process in urban morphogenesis. For example [Fujita and Thisse, 1996] opposes centrifugal forces with centripetal forces in the equilibrium view of urban spatial systems, what is easily transferable to non-equilibrium systems in the framework of self-organized complexity: a urban structure is a far-from-equilibrium system that has been driven to this point by these opposite forces. For example, concrete dispersion forces are congestion or the search for low density by residents, whereas aggregation forces can be the presence of amenities, of places

of interest, of increased possibilities of social interactions [Krugman, 1992].

The two contradictory processes of urban concentration and urban sprawl are captured by the model, what allows to reproduce with a good precision a large number of existing morphologies. We can expect aggregation mechanisms such as preferential attachment to be good candidates in urban growth explanation, as it was shown that the Simon model based on them generates power-laws typical of urban systems (scaling laws for example) [Dodds et al., 2017]. The question at which scale is it possible and relevant to define and try to simulate urban form is rather open, and will in fact depend on which issues are being tackled. Working in a typical setting of morphogenesis, the processes considered are local and our model must have a resolution at the micro-level. We however want to quantify urban form on consistent urban entities, and will work therefore on spatial extents of order 50~100km. We sum up these two aspects by stating that the model is at the *mesoscopic scale*.

FORMALIZATION We formalize now the model and its parameters. The world is a square grid of width N , in which each cell is characterized by its population $(P_i(t))_{1 \leq i \leq N^2}$. We consider the grid initially empty, i.e. $P_i(0) = 0$, but the model can be easily generalized to any initial population distribution. The population distribution is updated in an iterative way. At each time step,

1. Total population is increased by a fixed number N_G (growth rate). Each population unit is attributed independently to a cell following a preferential attachment such that

$$\mathbb{P}[P_i(t+1) = P_i(t) + 1 | P(t+1) = P(t) + 1] = \frac{(P_i(t)/P(t))^\alpha}{\sum(P_j(t)/P(t))^\alpha} \quad (12)$$

The attribution being uniformly drawn if all population are equal to 0.

2. A fraction β of population is diffused to cell neighborhood (8 closest neighbors receiving each the same fraction of the diffused population). This operation is repeated n_d times.

The model stops when total population reaches a fixed parameter P_m . To avoid bord effects such as reflecting diffusion waves, border cells diffuse their due proportion outside of the world, implying that the total population at time t is strictly smaller than $N_G \cdot t$.

We summarize model parameters in Table 14, giving the associated processes and values ranges we use in the simulations. The total population of the area P_m is exogenous, in the sense that it is supposed to depend on macro-scale growth patterns on long times. Growth rate

N_G captures both endogenous growth rate and migration balance within the area. The aggregation rate α sets the differences in attraction between cells, what can be understood as an abstract attraction coefficient following a scaling law of population. Finally, the two diffusion parameters are complementary since diffusing with strength $n_d \cdot \beta$ is different of diffusing n_d times with strength β , the later giving flatter configurations.

Table 14: **Summary of parameters of the morphogenesis model.** We give the corresponding processes and the typical variation range within the configuration we use.

Parameter	Notation	Processus	Range
Total population	P_m	Macroscopic growth	[1e4, 1e6]
Growth rate	N_G	Mesoscopic growth	[500, 30000]
Aggregation force	α	Aggregation	[0.1, 4]
Diffusion force	β	Diffusion	[0, 0.5]
Number of diffusions	n_d	Diffusion	{1, ..., 5}

MEASURING URBAN FORM As our model is only density-based, we propose to quantify its outputs through spatial morphology, i.e. properties of the spatial distribution of density. At the scale chosen, these will be expected to translate various functional properties of the urban landscape. The context and definition of indicators has already been given in section 4.1.

Real data

We work with values of indicators computed in section 4.1 for Europe, on windows of size 50km with a resolution of 100 cells. We set thus in the following $N = 100$ for model simulations.

Generation of urban patterns

IMPLEMENTATION The model is implemented both in NetLogo [Wilensky, 1999] for exploration and visualization purposes, and in Scala for performance reasons and easy integration into OpenMole [Reuillon, Leclaire, and Rey-Coyrehourcq, 2013], which allows a transparent access to High Performance Computing environments. Computation of indicator values on geographical data is done in R using the raster package [Hijmans, 2015]. Source code and results are available on the open repository of the project³. Raw datasets for real indicator values and simulation results are available on Dataverse⁴. We have

³ At <https://github.com/JusteRaimbault/Density>.

⁴ At <http://dx.doi.org/10.7910/DVN/WSUSBA>.

in the context of the `scala` implementation implemented the convolution of distributions in two dimensions through fast Fourier transform, allowing to capture a complexity of $O(N^4)$ in $O(N^2 \log^2 N)$ ⁵, and implemented indicators which have been integrated to a dedicated NetLogo extension (it is detailed in E.1.3).

GENERATED SHAPES The model has a relatively small number of parameters but is able to generate a large variety of shapes, extending beyond existing forms. We run the model for parameters varying in ranges given in Table 12, for a world size $N = 100$.

Fig. ?? shows examples of the variety of generated shapes for different parameter values, with corresponding interpretations. The four very different shapes can be obtained with variation of a single parameter sometimes: going from a peri-urban area from a rural area implies an increased aggregation at the same level of diffusion. Note that the model is density driven, and that the parameter P_m/N_G is what really influences the dynamics: the values of P_m are in some cases not directly corresponding to the interpretations we made (for the rural in particular) that are done on densities. A rescaling keeps the settlement form and solves this issue.

It appears that the dynamical nature of the model allows through the combination of parameters P_m and N_G to choose between configurations that can be non-stationary or semi-stationary, whereas the interaction between α and β modulates the sprawl and the compact character of forms.

These examples show the potentiality of the model to produce diverse forms. We have then to systematically study its stochasticity and explore its parameter space.

Model behavior

In the study of such a computational model of simulation, the lack of analytical tractability must be compensated by an extensive knowledge of model behavior in the parameter space [Banos, 2013]. This type of approach is typical of what Arthur calls the *Computational shift in modern science* [Arthur, 2015]: knowledge is less extracted through analytical exact resolution than through intensive computational experiments, even for “simple” models such as the one we study.

⁵ We recall that a measure of complexity of an algorithm corresponds to the evaluation of time necessary to solve a problem as a function of data size, denoted N . An asymptotic order of magnitude is written $O(f(N))$. Therefore, a switch from a fourth order of magnitude to an order very close to a square is significant for computation time, making quasi-instantaneous a computation that would take around 10 seconds for the grid size we have. The fast Fourier transform uses a sparse decomposition to compute the discrete Fourier transform in $O(N \log N)$ instead of $O(N^2)$. The morphism of the transform of product to convolution, i.e. $\mathcal{F}[f * g] = \mathcal{F}[f] \cdot \mathcal{F}[g]$, allows to transfer this gain to the computation of a convolution.

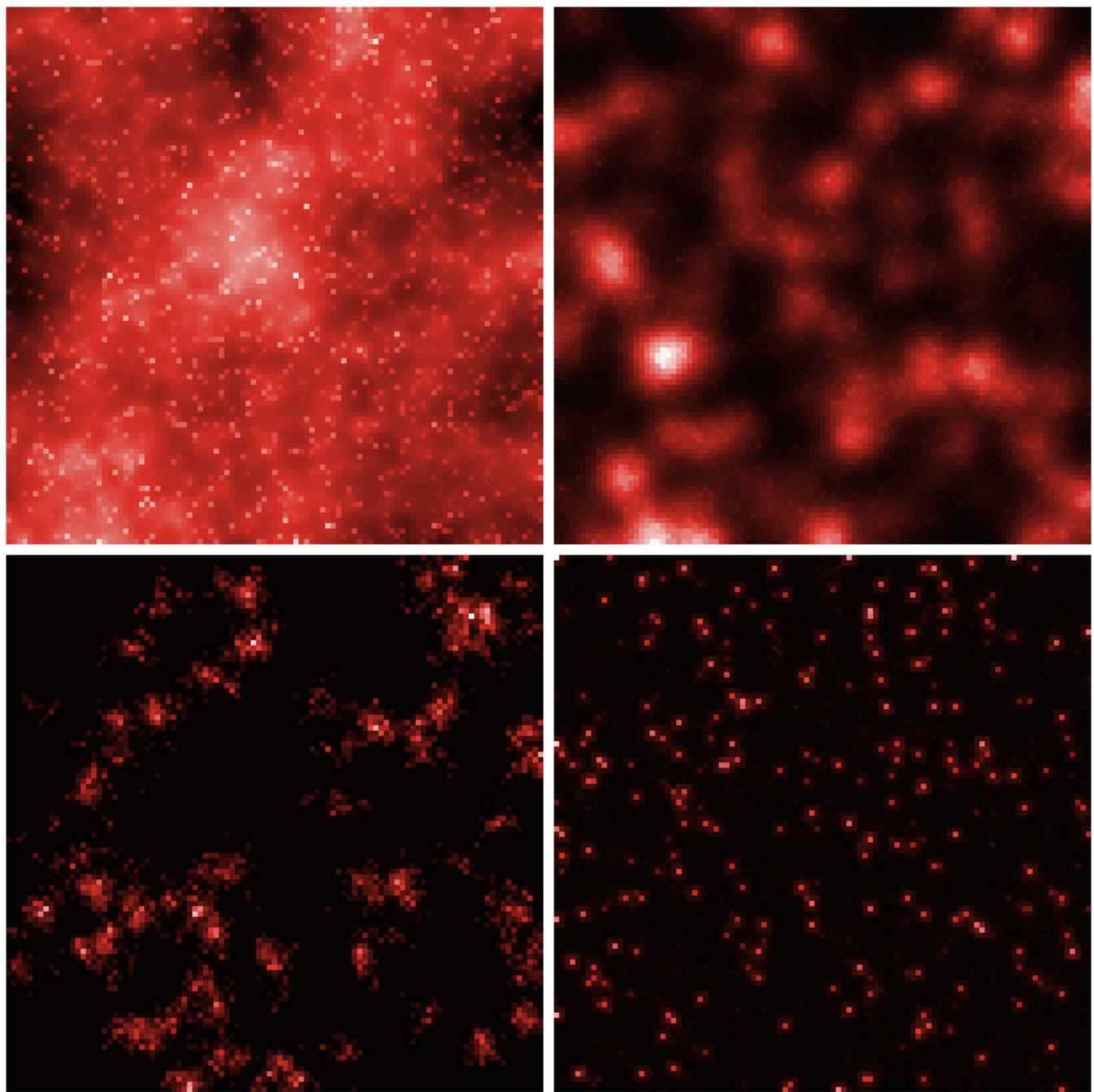


Figure 36: **Example of the variety of generated urban shapes.** (Top left) Very diffuse urban configuration, $\alpha = 0.4$, $\beta = 0.05$, $n_d = 2$, $N_G = 76$, $P_m = 75620$; (Top Right) Semi-stationary polycentric urban configuration, $\alpha = 1.4$, $\beta = 0.047$, $n_d = 2$, $N_G = 274$, $P_m = 53977$; (Bottom Left) Intermediate settlements (peri-urban or densely populated rural area), $\alpha = 0.4$, $\beta = 0.006$, $n_d = 1$, $N_G = 25$, $P_m = 4400$; (Bottom Right) Rural area, $\alpha = 1.6$, $\beta = 0.006$, $n_d = 1$, $N_G = 268$, $P_m = 76376$.

CONVERGENCE First of all we need to assess the convergence of the model and its behavior regarding stochasticity. We run for a sparse grid of the parameter space consisting of 81 points, with 100 repetitions for each point. Corresponding histograms are shown in ???. Indicators show good convergence properties: most of indicators are easily statistically discernable across parameter points: for example the Moran index, which is among the most dispersed, has a spread between 0 and 0.1 between parameters but a maximal variability of 0.01 between replications.

We use this experiment to find a reasonable number of repetitions needed in larger experiments. For each point, we estimate the Sharpe ratios for each indicators, i.e. mean normalized by standard deviation. The more variable indicator is Moran with a minimal Sharpe ratio of 0.93, but for which the first quartile is at 6.89. Other indicators have very high minimal values, all above 2. Its means than confidence intervals large as $1.5 \cdot \sigma$ are enough to differentiate between two different configurations. In the case of gaussian distribution, we know that the size of the 95% confidence around the average is given by $2 \cdot \sigma \cdot 1.96/\sqrt{n}$, what gives $1.26\bar{\sigma} \cdot \sigma$ for $n = 10$. We run therefore this number of repetitions for each parameter point in the following, what is highly enough to have statistically significant differences between average as shown above. In the following, when referring to indicator values for the simulated model, we consider the ensemble averages on these stochastic runs.

EXPLORATION OF PARAMETER SPACE We sample the Parameter space using a Latin Hypercube Sampling, with parameter as $\alpha \in [0.1, 4]$, $\beta \in [0, 0.5]$, $n_d \in \{1, \dots, 5\}$, $N_G \in [500, 30000]$, $P_m \in [1e4, 1e6]$. As we already explained, relative values of P_m and N_G have a stronger influence on the forms obtained than their values in absolute, and we thus set P_m to obtain territories containing at most 1 million inhabitants, what is a strong but not extreme density (for comparison, the Parisian region concentrates around 8 millions inhabitants on an area of a similar size). Values of N_G vary considerably to cover a large number of possible dynamical regimes. Values of α and β have been obtained through successive experimentations.

This type of cribbing is a good compromise to have a reasonable sampling without being subject to the dimensionality curse within normal computation capabilities. We sample around 80000 parameters points, with 10 repetitions each (as established in the previous experiment). We recall the protocol followed here to obtain the behavior of a simulation model, which can be put into perspective into the more general one presented in 3.1:

- sampling of parameter points;
- simulation of the models for each parameter point, repeated 10 times;

- computation for each model execution of urban form indicators;
- aggregation for each parameter point by computing averages on repetitions⁶.

Full plots of model behavior as a function of parameters are given in Appendix A.6. We show in 37 some particularly interesting behavior for slope γ and average distance \bar{d} . First of all, the overall qualitative behavior depending on aggregation strength, namely that lower alpha give less hierarchical and more spread configurations, confirms the expected intuitive behavior.

The effect of diffusion strength β is more difficult to grasp: the effect is inverted for slope between high and low growth rates but not for distance, that shows an inversion when α varies. In the low N_G case, low diffusion creates more sprawled configuration when aggregation is low, but less sprawled when aggregation is high. Furthermore, all indicators show a more or less smooth transition around $\alpha \simeq 1.5$. Slope stabilize over certain values, meaning that the hierarchy cannot be forced more and indeed depends of the diffusion value, at least for low N_G (right column). In general, higher valued for P_m/N_G increase the effect of diffusion what could have been expected.

The existence of a minimum for slope at $n_d = 1, P_m/N_G \in [13, 26]$ and lowest β is unexpected and witnesses a complex interplay between aggregation and diffusion. The emergence of this “optimal” regime is associated with shifts of the transition points in other cases: for example, lowest diffusion imply a transition beginning at lower values of α for average distance. This exploration confirms that complex behavior, in the sense of unpredictable emerging forms, occurs in the model: one cannot predict in advance the final form given some parameters, without referring to the full exploration of which we give an overview here.

Semi-analytical analysis

Our model can be understood as a type of reaction-diffusion model, that have been widely used in other fields such as biology as we synthesized in 5.1. An other way to formulate the model typical to these approaches is by using Partial Differential Equations (PDE). In the case of a firm growth model, which is a generalization of the Simon model with an arbitrary form of the attachment function, [Rushing Dewhurst, Danforth, and Sheridan Dodds, 2017] show that a PDE and its general solution can be derived. We propose to gain insights into long-time dynamics by studying them on a simplified case. We consider the system in one dimension, such that $x \in [0; 1]$ with $1/\delta x$ cells of size δx . A time step is given by δt . Each cell is characterized

⁶ Given the shape of distributions obtained for 100 repetitions, presented in A.6, the use of the average or the median given equivalent results.

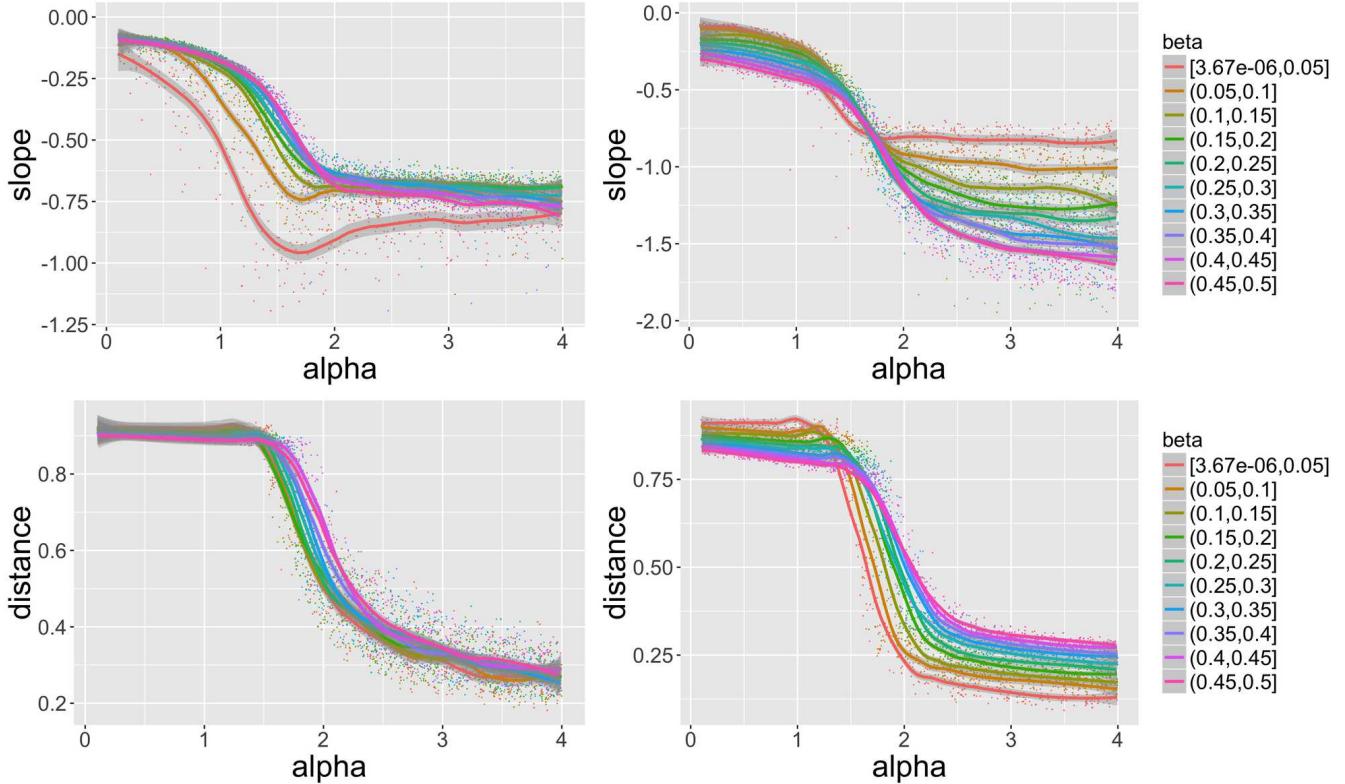


Figure 37: **Behavior of indicators.** Slope γ (top row) and average distance \bar{d} (bottom row) as a function of α , for different bins for β given by curve color, for particular values $n_d = 1, P_m/N_G \in [13, 26]$ (left column) and $n_d = 4, P_m/N_G \in [41, 78]$ (right column). We observe in each case a transition as a function of α , which properties are influenced by other parameters. For low values of P_m/N_G and of β emerges a counter-intuitive non-monotony.

by its population as a random variable $P(x, t)$. We work on their expected values $p(x, t) = \mathbb{E}[P(x, t)]$, and assume that $n_d = 1$. As developed in Appendix A.6, we show that this simplified process verifies the following PDE:

$$\delta t \cdot \frac{\partial p}{\partial t} = \frac{N_G \cdot p^\alpha}{P_\alpha(t)} + \frac{\alpha\beta(\alpha-1)\delta x^2}{2} \cdot \frac{N_G \cdot p^{\alpha-2}}{P_\alpha(t)} \cdot \left(\frac{\partial p}{\partial x} \right)^2 + \frac{\beta\delta x^2}{2} \cdot \frac{\partial^2 p}{\partial x^2} \cdot \left[1 + \alpha \frac{N_G p^{\alpha-1}}{P_\alpha(t)} \right] \quad (13)$$

where $P_\alpha(t) = \int_x p(x, t)^\alpha dx$. This non-linear equation can not be solved analytically, the presence of integral terms putting it out of standard methods, and numerical resolution must be used [Tadmor, 2012].

It is important to note that the simplified model can be expressed by a PDE analog to reaction-diffusion equations, as the one partially solved for a simpler model in [Bosch, Metz, and Diekmann, 1990]. We show in A.6 that because of the boundaries conditions, density (proportion of population) converges towards a stationary solution at long times, going through intermediate states in which the solution is partially stabilized, in the sense that its evolution speed becomes rather slow. These “semi-stationary” states are the ones used in two dimensions along with the dynamical ones. This study confirms that the variety of shapes obtained through the model is permitted both by the interplay of aggregation and diffusion as the equation couples them, but also by the values of P_m/N_G that allow to set the convergence level. Indeed, the sensitivity of the stationary solution to parameters is very low compared to the shape of the world, and using the model in stationary mode would make no sense in our case.

Finally, we use this toy case to demonstrate the importance of bifurcations in model dynamics. More precisely, we show that path-dependence is crucial for the final form. As illustrated in Fig. 38, using an initial condition making the choice ambiguous, corresponding to five equidistant equally populated cells, produces very different trajectories, as generally one of the spots will end dominating the others, but is totally random, witnessing dramatic bifurcations in the system at initial times. This aspect is typically expected in urban systems, since very precise characteristics will be included in the determinants of localization at the initial moments of system genesis: the existence of a very local resource, or the strategic advantage of the site (defensive or crossing site for example [Hypergeo 2017]), will determine on very long times the local territorial form. This aspect confirms the importance of robust indicators described before.

Model calibration

We finally turn to the calibration of the model, that is done on the morphological objectives. As a single calibration for each real cell is

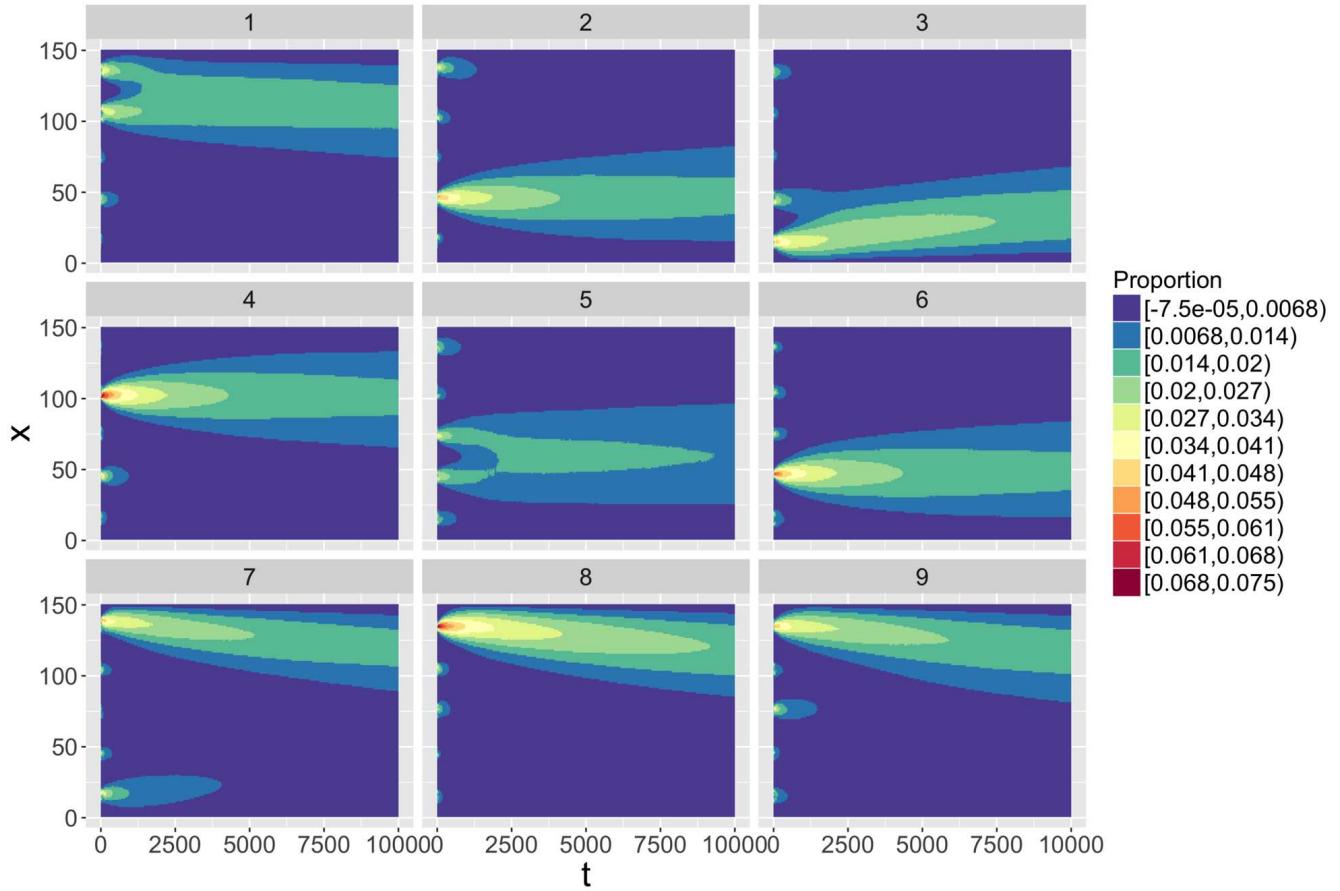


Figure 38: Randomness and frozen accidents. We show nine random realizations of the one dimensional system with similar initial conditions, namely five equidistant equally populated initial cells. Parameters are $\alpha = 1.4, \beta = 0.1, N_G = 10$. Each plot shows time against space, color level giving the proportion of population in each cell.

computationally out of reach, we use the previous model exploration and superpose the point clouds with real indicator values. Full scatterplots of all indicators against each other, for simulated and real configurations, are given in A.6. We find that the real point cloud is mostly contained within the simulated, that extend in significantly larger areas. It means that for a large majority of real configuration, there exist model parameters producing in average exactly the same morphological configuration. The highest discrepancy is for the distance indicator, the model failing to reproduce configuration with high distance, low Moran and intermediate hierarchy. These could for example correspond to polycentric configurations with many consequent centers.

We consider a more loose calibration constraint, by doing a Principal Component Analysis on synthetic and real morphological values, and consider the two first components only. These represent 85% of cumulated variance. The rotated point clouds along these

dimensions are shown in Fig. 39. Most of the real point cloud falls in the simulated one in this simplified setting. We illustrate particular points with real configurations and their simulated counterparts: for example Bucharest, Romania, corresponds to a monocentric semi-stationary configuration, with very high aggregation but also diffusion and a rather low growth rate. Other examples show less populated areas in Spain and Finland. From the plots giving parameter influence, we can show that most real situation fall in the region with intermediate α but quite varying β . It is consistent with real scaling urban exponents having a variation range rather small (between 0.8 and 1.3 generally [Pumain et al., 2006]) compared to the one we allowed in the simulations, whereas the diffusion processes may be much more diverse.

This way, we have shown that the model is able to reproduce most of existing urban density configuration in Europe, despite its rather simplicity. It confirms that in terms of urban form, most of drivers at this scale can be translated into these abstract processes of aggregation and diffusion. It also implies that urban functions, which can be quantified by similar indicators on their spatial distribution, play a reduced role in the location of populations, or that they must be quite correlated to the distribution of population (and are indeed taken into account in an abstract way in the aggregation function).

5.2.3 Discussion

CALIBRATION AND MODEL REFINEMENT Further work on this simple model may consist in extracting the exact parameter space covering all real situations and provide interpretation of its shape, in particular through correlations between parameters and expressions of boundaries functions. Its volume in different directions should furthermore give the relative importance of parameters.

Concerning the feasible space for the model of simulation itself, we tested a targeted exploration algorithm, giving promising results. More precisely, the Parameter Space Exploration (PSE) algorithm [Chérel, Cottineau, and Reuillon, 2015] which is implemented in OpenMole, is aimed at determining all the possible outputs of a simulation model, i.e. samples its output space rather than input space. We obtain promising results as shown in Fig. ??: we find that the lower bound in Moran-entropy plan, confirmed by the algorithm, unexpectedly exhibit a scaling relationship. It would mean that at a given level of auto-correlation, that one could want to attain for sustainability reasons for example (optimality through co-location), imposes a minimal disorder in the configuration of activities.

Other relations between indicators and as a function of parameters can be the object of similar future developments. The question of doing a dynamical calibration of the model, i.e. trying to reproduce

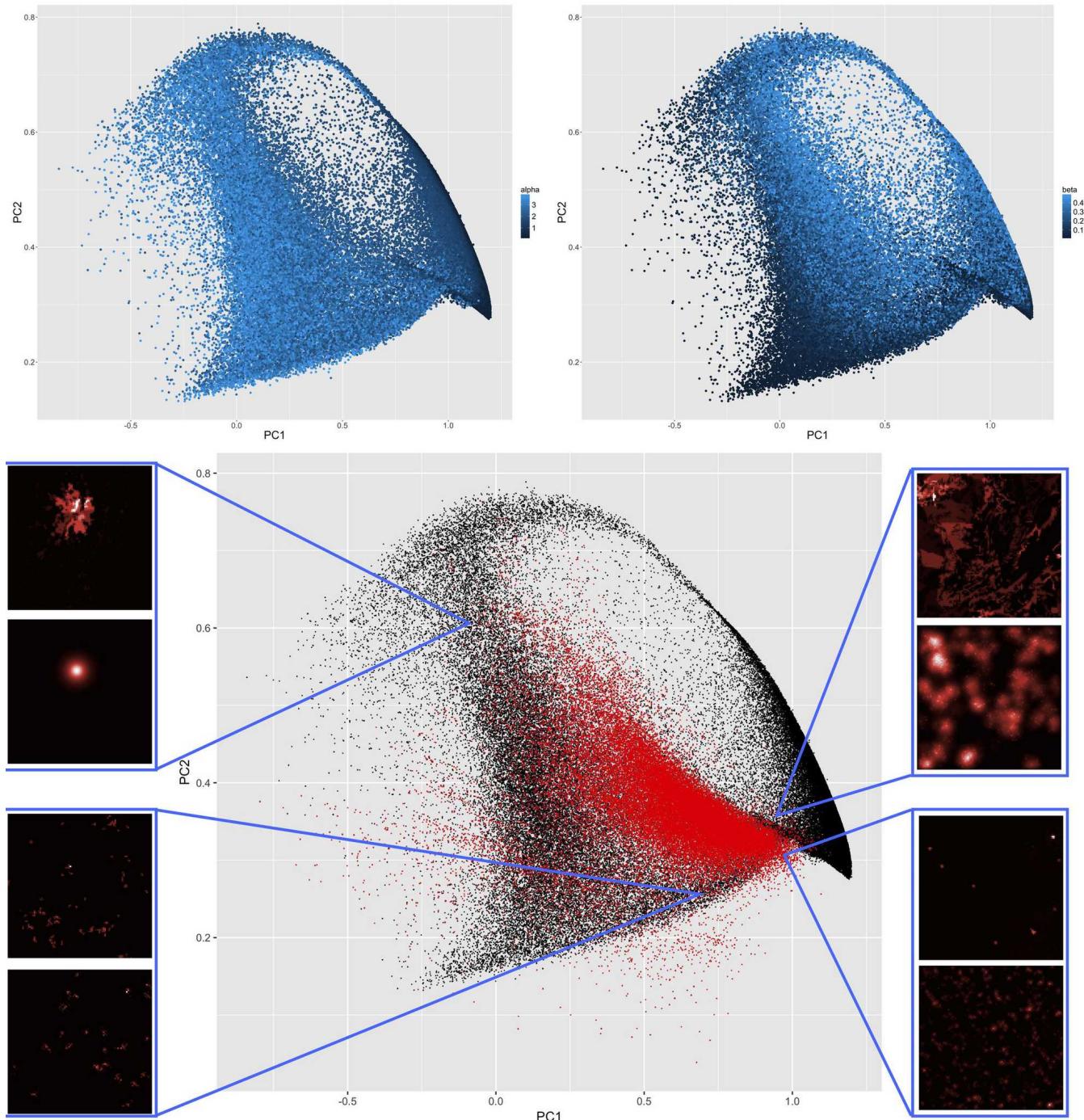


Figure 39: Model calibration. (Top) Simulated configurations in the two first principal components plan, color level giving the influence of α (left) and of β (right); (Bottom) Simulated points in the same space (in black) with real configurations (in red). We show around the plot typical examples of real configurations and their simulated counterparts in different regions of the space, the first being the real and the second the simulated in each case: Top left geographical coordinates 25.7361, 44.69989 - Romania, Bucharest - simulation parameters $\alpha = 3.87, \beta = 0.432, N_G = 1273, nd = 4, P_m = 63024$; Top right geographical coordinates -2.561874, 41.30203 - Spain, Castilla et Leon, Soria - simulation parameters $\alpha = 1, \beta = 0.166, N_G = 100, nd = 1, P_m = 10017$; Bottom left geographical coordinates 27.16068, 65.889 - Finland, Lapland - simulation parameters $\alpha = 0.4, \beta = 0.006, N_G = 25, nd = 1, P_m = 849$; Bottom right geographical coordinates -2.607152, 39.74274 - Spain, Castilla-La Mancha, Cuenca - simulation parameters $\alpha = 1.14, \beta = 0.108, N_G = 637, nd = 1, P_m = 13235$.

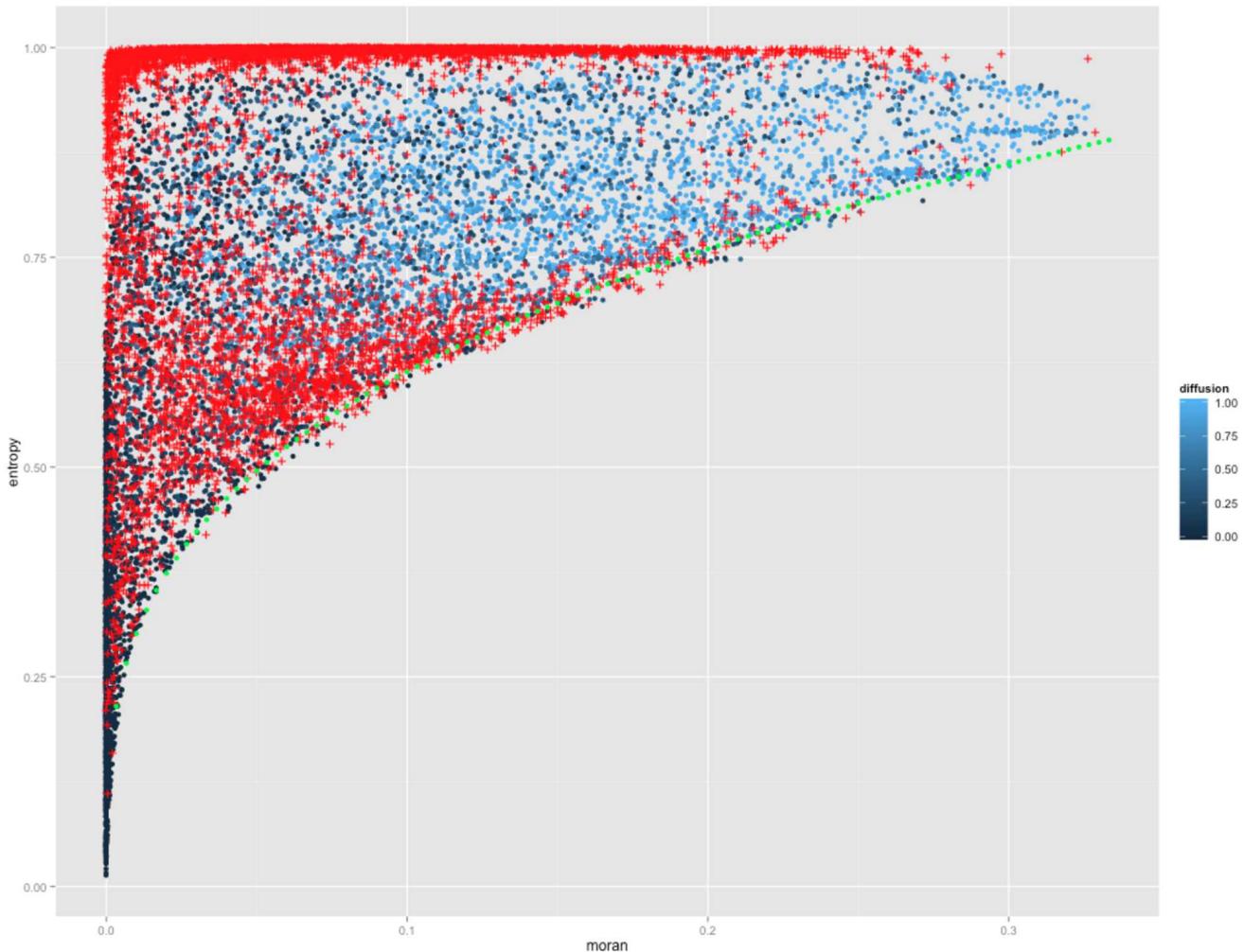


Figure 40: **PSE exploration.** Scatterplot of Moran against Entropy, with blue points obtained with LHS and red with PSE exploration. Green dashed line gives feasible lower bound.

configurations at successive times, is conditioned to the availability of population data at this resolution in time.

We aimed at using abstract processes rather than having a highly realistic model. Tuning some mechanisms is possible to have a model closer to reality in microscopic processes: for example thresholding the local population density, or stopping the diffusion at a given distance from the center if it is well defined. It is however far from clear if these would produce such a variety of forms and could be calibrated in a similar way, as being accurate locally does not mean being accurate at the mesoscopic level for morphological indicators. Allowing the parameters to locally vary, i.e. being non stationary in space, or adding randomness to the diffusion process, are also potential model refinements.

In conclusion, we have provided a calibrated spatial urban morphogenesis model at the mesoscopic scale that can reproduce a significant proportion of European urban pattern in terms of morphology. We demonstrate that the abstract processes of aggregation and diffusion are sufficient to capture urban growth processes at this scale. It is meaningful in terms of policies based on urban form such as energy efficiency, but also means that issues out of this scope must be tackled at other scales or through other dimensions of urban systems.

* * *

*

The first section of this chapter allowed to deepen the definition of morphogenesis, whereas the second lead to the construction of a simple model of urban morphogenesis, allowing to generate population distributions, that can be understood as territorial configurations.

We will now use this framework to introduce the morphogenesis of networks conditioned to a territorial morphogenesis, to progressively shift towards co-evolution models.

* * *

*

5.3 GENERATION OF CORRELATED TERRITORIAL CONFIGURATIONS

This section aims to explore the sequential coupling (or simple coupling) between previous model of density generation and an heuristic of network growth. We explore therein the feasible space of correlations between network measures and morphological measures. We first recall the issues linked to the notion of synthetic data and the role of correlation structures in these.

5.3.1 Correlated geographical data of density and network

One of the inspirations and applications of the current approach is the generation of synthetic data, for example to feed sensitivity analysis to the spatial configuration presented in section 3.1. The use of synthetic data in geography is generally directed towards the generation of synthetic populations within agent-based models (mobility, *LUTI* models) [Pritchard and Miller, 2009]. We can make a link with some spatial analysis techniques. The extrapolation of a continuous spatial field from a discrete spatial sample through a kernel density estimation for example can be understood as the creation of a synthetic dataset (even if it is not generally the initial view, as in Geographically Weighted Regression [Brunsdon, Fotheringham, and Charlton, 1998] in which variable size kernels do not interpolate data *stricto sensu* but extrapolate abstract variables representing interaction between explicit variables).

In the field of modeling in quantitative geography, *toy-models* or hybrid models require a consistent initial spatial configuration. A set of possible initial configurations becomes a synthetic dataset on which the model is tested. The first Simpop model [Sanders et al., 1997], precursor of a large family of models later parametrized with real data, could enter that frame but was studied on an unique synthetic spatialization. Similarly underlined was the difficulty to generate an initial transportation infrastructure in the case of the Simpop-Net model [Schmitt, 2014] although it was admitted as a cornerstone of knowledge on the behavior of the model.

A systematic control of spatial configuration effects on the behavior of simulation models was only recently proposed [Cottineau et al., 2015], and as we developed in 3.1, this approach that can be interpreted as a statistical control on spatial data. The aim is to be able to distinguish proper effects due to intrinsic model dynamics from particular effects due to the geographical structure of the case study. Such results are essential for the validation of conclusions obtained with modeling and simulation practices in quantitative geography. Indeed, as we reviewed in 3.1, most modeling experiments systematically explore the influence of parameters, but not of spatial initial

configurations, although they may have a stronger influence than parameters.

5.3.2 Model and results

Formalization

We propose in our case to generate territorial systems summarized in a simplified way as a spatial population density $d(\vec{x})$ and a transportation network $n(\vec{x})$. Correlations we aim to control are correlations between urban morphological measures and network measures. The question of interactions between territories and networks is already well-studied [Offner and Pumain, 1996] but stays highly complex and difficult to quantify [Offner, 1993]. A dynamical modeling of implied processes should shed light on these interactions [Bretagnolle, 2009] (p. 162). We develop in that frame a *simple* coupling (i.e. without any feedback loop) between a density distribution model and a network morphogenesis model.

DENSITY MODEL The density model is the model described and explored in the previous section 5.2. We use it for the conditional generation of network.

NETWORK MODEL On the other hand, we are able to generate a planar transportation network by a model N , at a similar scale and given a density distribution. Because of the conditional nature to the density of the generation process, we will first have conditional estimators for network indicators, and secondly natural correlations between network and urban shapes should appear as processes are not independent. The nature and modularity of these correlations as a function of model parameters are still to determine by exploration of the coupled model.

Concerning the choice of the heuristic to generate an infrastructure network, we have reviewed in 2.1 several models allowing it. Furthermore, we will compare different models in an operational manner in 7.1. The aim here being to demonstrate the feasibility of the coupling within a morphogenesis model and also to explore the feasible space of correlation, we propose a unique heuristic, which is inspired by the model of [Schmitt, 2014], and simplifies it by removing the stochastic character. This heuristic is detailed below.

The heuristic network generation procedure is the following :

1. A fixed number N_c of centers that will be first nodes of the network si distributed given density distribution, following a similar law to the aggregation process, i.e. the probability to be distributed in a given patch is $\frac{(P_i/P)^\alpha}{\sum(P_i/P)^\alpha}$. Population is then at-

tributed according to Voronoi areas of centers, such that a center cumulates population of patches within its extent.

2. Centers are connected deterministically by percolation [Callaway et al., 2000] between closest clusters : as soon as network is not connected, two closest connected components in the sense of minimal distance between each vertices are connected by the link realizing this distance. It yields a tree-shaped network.
3. Network is modulated by potential breaking in order to be closer from real network shapes. More precisely, a generalized gravity potential between two centers i and j is defined by

$$V_{ij}(d) = \left[(1 - k_h) + k_h \cdot \left(\frac{P_i P_j}{P^2} \right)^{\gamma_c} \right] \cdot \exp \left(-\frac{d}{r_g(1 + d/d_0)} \right)$$

where d can be euclidian distance $d_{ij} = d(i, j)$ or network distance $d_N(i, j)$, $k_h \in [0, 1]$ a weight to modulate role of populations in the potential, γ giving shape of the hierarchy across population values, r_g characteristic interaction distance and d_0 distance shape parameter (allowing to flatten the distribution in low values). This form of potential assumes on the one hand that the attenuation of interaction to distance is independent from the strength of interaction due to weights (standard assumption of gravity models); on the other hand that a constant term due to distance can weight more or less (weighting by k_h); and finally that the distance function take as parameter a characteristic distance, but also a shape parameter, allowing for example to control the decrease on low distances.

4. A fixed number $K \cdot N_L$ of potential new links is taken among couples having greatest euclidian distance potential ($K = 5$ is fixed, this value producing experimentally reasonable length links): this stage allows to eliminate very short links with a small population and very long links.
5. Among potential links, N_L are effectively realized, that are the one with smallest rate $\tilde{V}_{ij} = V_{ij}(d_N)/V_{ij}(d_{ij})$. At this stage only the gap between euclidian and network distance is taken into account : \tilde{V}_{ij} does indeed not depend on populations and is increasing with d_N at constant d_{ij} .
6. Planarity of the network is forced by creation of nodes at possible intersections created by new links (with the former network or between new links)⁷.

We insist on the fact that the network generation procedure is entirely heuristic and result of thematic assumptions (connected initial

⁷ Our model is different from [Schmitt, 2014] on that point, as we simplify and do not assume levels of hierarchy between links.

network, gravity-based link creation) combined with trial-and-error during first explorations. Other model types could be used as well, such biological self-generated networks [Tero et al., 2010], local network growth based on geometrical constraints optimization [Barthelemy and Flammini, 2008], or a more complex percolation model than the initial one that would allow the creation of loops for example. We could thus in the frame of a modular architecture, in which the choice between different implementations of a functional brick can be seen as a meta-parameter [Cottineau, Chapron, and Reuillon, 2015], choose network generation function adapted to a specific need (as e.g. proximity to real data, constraints on output indicators, variety if generated forms).

PARAMETER SPACE Parameter space for the coupled model⁸ is constituted by density generation parameters $\vec{\alpha}_D = (P_m/N_G, \alpha, \beta, n_d)$ (see section 5.2; we study for the sake of simplicity the rate between population and growth rate instead of both varying, i.e. the number of steps needed to generate the distribution) and network generation parameters $\vec{\alpha}_N = (N_C, k_h, \gamma, r_g, d_0)$. We denote $\vec{\alpha} = (\vec{\alpha}_D, \vec{\alpha}_N)$.

INDICATORS Urban form and network structure are quantified by numerical indicators in order to modulate correlations between these. Morphology is defined as a vector $\vec{M} = (r, \bar{d}, \varepsilon, a)$ giving spatial auto-correlation (Moran index), mean distance, entropy and hierarchy (see 4.1 for a precise definition of these indicators). Network measures $\vec{G} = (\bar{c}, \bar{l}, \bar{s}, \delta)$ are with network denoted (V, E)

- Mean centrality \bar{c} defined as average *betweenness-centrality* (normalized in $[0, 1]$) on all links.
- Mean path length \bar{l} given by

$$\frac{1}{d_m} \frac{2}{|V| \cdot (|V| - 1)} \sum_{i < j} d_N(i, j)$$

- with d_m normalization distance taken here as world diagonal $d_m = \sqrt{2}N$.
- Mean network speed [Banos and Genre-Grandpierre, 2012] which corresponds to network performance compared to direct travel, defined as $\bar{s} = \frac{2}{|V| \cdot (|V| - 1)} \sum_{i < j} \frac{d_{ij}}{d_N(i, j)}$.
- Network diameter $\delta = \max_{i,j} d_N(i, j)$.

⁸ Weak coupling allows to limit the total number of parameters as a strong coupling would involve retroaction loops and consequently associated parameters to determine their structure and intensity. In order to diminish it, an integrated model would be preferable to a strong coupling, what is slightly different in the sense where it is not possible in the integrated model to freeze one of the subsystems to obtain a model of the other subsystem that would correspond to the non-coupled model.

We do not have at this stage any “performance” indicator for the network generation process, i.e. aiming at reproducing typical patterns or optimizing some criteria. These will come later in 7.1 when we will calibrate similar models on real data. We consider the examples shown in 42 as elements of the feasible space, the question being if network shapes corresponding to realities or given stylized facts will be also the object of this calibration.

COVARIANCE AND CORRELATION We study the cross-correlation matrix $\text{Cov}[\vec{M}, \vec{G}]$ between morphology and network. We estimate it on a set of n realizations at fixed parameter values $(\vec{M}[D(\vec{\alpha})], \vec{G}[N(\vec{\alpha})])_{1 \leq i \leq n}$ with standard unbiased estimator. We will study the Pearson correlation associated to it, estimated in the same way.

Implementation

Coupling of generative models is done both at formal and operational levels. We interface therefore independent implementations. The OpenMole software [Reuillon, Leclaire, and Rey-Coyrehourcq, 2013] for intensive model exploration offers for that the ideal frame thanks to its modular language allowing to construct *workflows* by task composition and interfacing with diverse experience plans and outputs. For operational reasons, density model is implemented in `scala` language as an OpenMole plugin, whereas network generation is implemented in agent-oriented language NetLogo [Wilensky, 1999] because of its possibilities for interactive exploration and heuristic model construction. Source code is available for reproducibility on project repository⁹.

Results

The study of density model alone is developed in the previous section. We recall that it is calibrated on European density grid data, on 50km width square areas with 500m resolution for which real indicator values have been computed on whole Europe. Furthermore, a grid exploration of model behavior yields feasible output space in reasonable parameters bounds (roughly $\alpha \in [0.5, 2]$, $N_G \in [500, 3000]$, $P_m \in [10^4, 10^5]$, $\beta \in [0, 0.2]$, $n_d \in \{1, \dots, 4\}$). The reduction of indicators space to a two dimensional plan through a Principal Component Analysis (variance explained with two components $\simeq 80\%$) allows to isolate a set of output points that covers reasonably precisely real point cloud. It confirms the ability of the model to reproduce morphologically the set of real configurations.

Given the large relative dimension of parameter space, an exhaustive grid exploration is not possible. We use a Latin Hypercube sampling procedure with bounds given above for $\vec{\alpha}_D$ and for $\vec{\alpha}_N$, we take

⁹ at <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Synthetic>

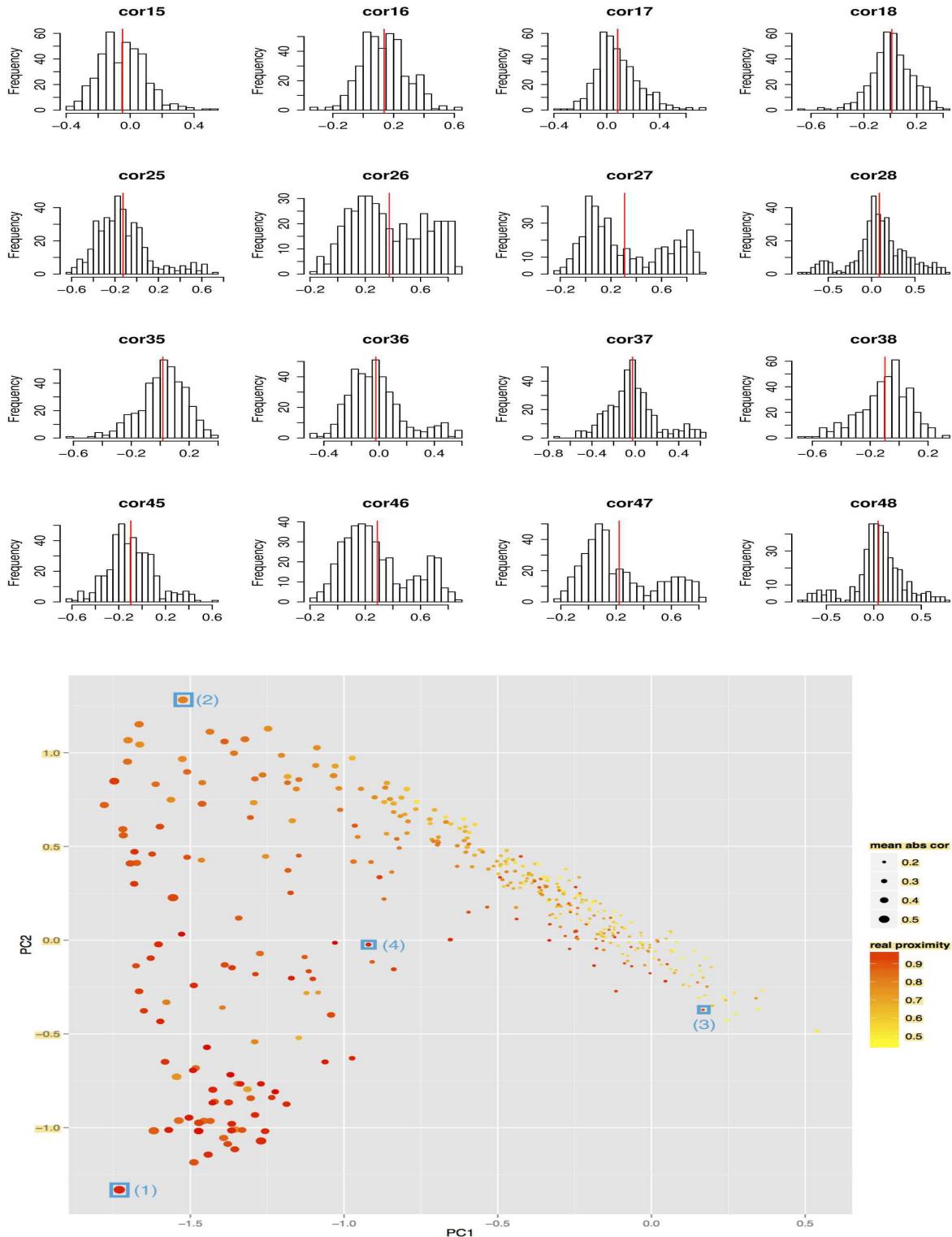


Figure 41: Exploration of feasible space for correlations between urban morphology and network structure.
 (a) Distribution of crossed-correlations between vectors \vec{M} of morphological indicators (in numbering order Moran index, mean distance, entropy, hierarchy) and \vec{N} of network measures (centrality, mean path length, speed, diameter). (d) Representation in the principal plan, scale color giving proximity to real data defined as $1 - \min_r \|\vec{M} - \vec{M}_r\|$ where \vec{M}_r is the set of real morphological measures, point size giving mean absolute correlation.

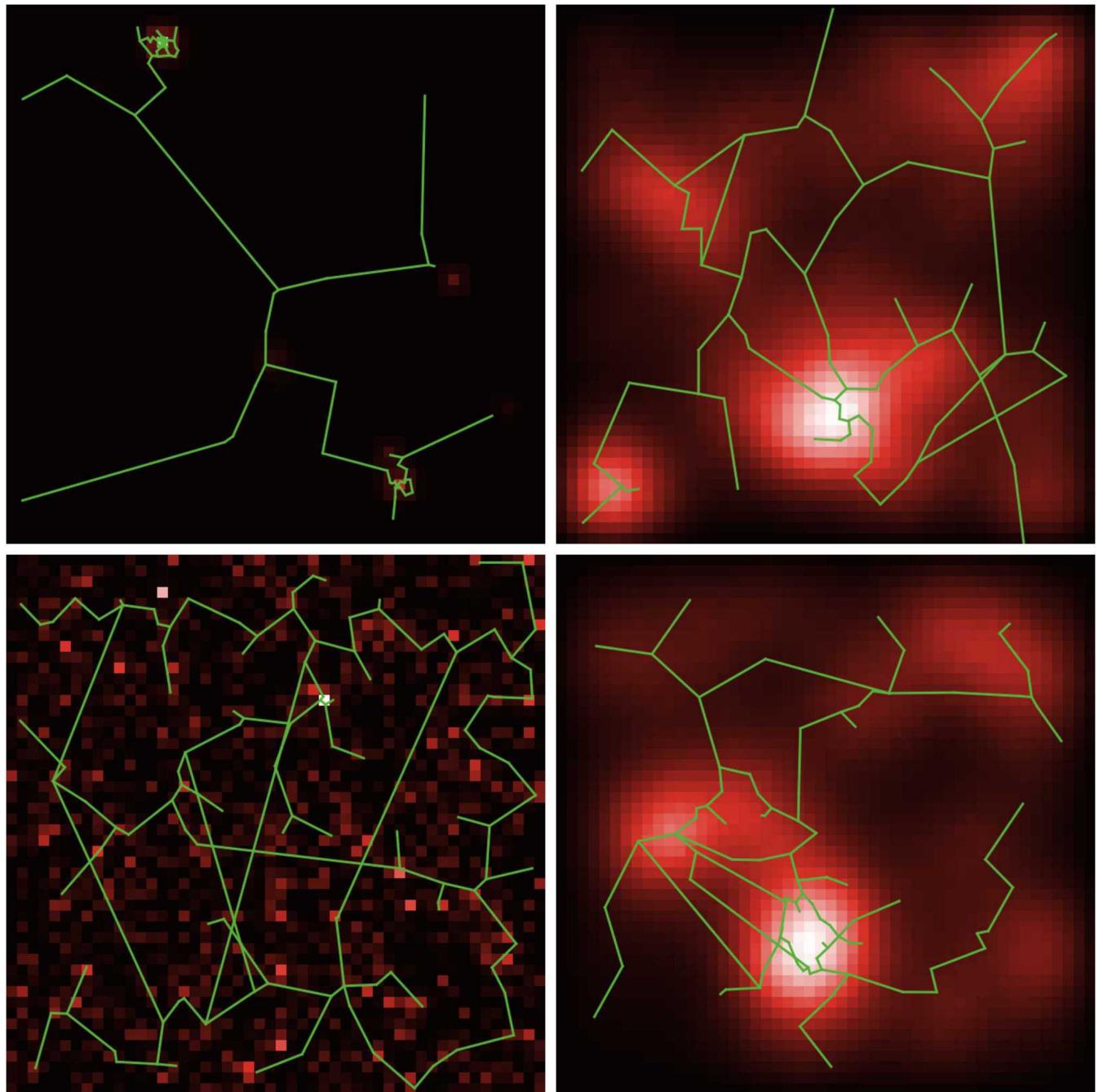


Figure 42: Configurations obtained for parameters giving the four emphasized points in (d), in order from left to right and top to bottom. We recognize polycentric city configurations (2 and 4), diffuse rural settlements (3) and aggregated weak density area (1). See appendix for exhaustive parameter values, indicators and corresponding correlations. For example \bar{d} is highly correlated with \bar{l}, \bar{s} (≈ 0.8) in (1) but not for (3) although both correspond to rural environments; in the urban case we observe also a broad variability: $\rho[\bar{d}, \bar{c}] \approx 0.34$ for (4) but ≈ -0.41 for (2), what is explained by a stronger role of gravitation hierarchy in (2) $\gamma = 3.9, k_h = 0.7$ (for (4), $\gamma = 1.07, k_h = 0.25$), whereas density parameters are similar.

$N_C \in [50, 120]$, $r_g \in [1, 100]$, $d_0 \in [0.1, 10]$, $k_h \in [0, 1]$, $\gamma \in [0.1, 4]$, $N_L \in [4, 20]$. For number of model replications for each parameter point, less than 50 are enough to obtain confidence intervals at 95% on indicators of width less than standard deviations. For correlations a hundred give confidence intervals (obtained with Fisher method) of size around 0.4, we take thus $n = 80$ for experiments.

Figure 41 gives details of experiment results. Regarding the subject of correlated synthetic data generation, we can sum up the main lines as following :

- Empirical distributions of correlation coefficients between morphology and network indicators are not simple and some are bimodal (for example $\rho[I, \bar{I}]$ between Moran index and mean path length which corresponds to cor46 on Fig. 41).
- it is possible to modulate up to a relatively high level of correlation for all indicators, maximal absolute correlation varying between 0.6 and 0.9. Amplitude of correlations varies between 0.9 and 1.6, allowing a broad spectrum of values. Point cloud in principal plan has a large extent but is not uniform : it is not possible to modulate at will any coefficient as they stay themselves correlated because of underlying generation processes. A more refined study at higher orders (correlation of correlations) would be necessary to precisely understand degrees of freedom in correlation generation.
- Most correlated points are also the closest to real data, what confirms the intuition and stylized fact of a strong interdependence in reality.
- Concrete examples taken on particular points in the principal plan show that similar density profiles can yield very different correlation profiles.

5.3.3 Discussion

Developments

This case study could be refined by extending correlation control method. A precise knowledge of N behavior (statistical distributions on an exhaustive grid of parameter space) conditional to D would allow to determine $N^{<-1>}|D$ and have more latitude in correlation generation. We could also apply specific exploration algorithms to reach exceptional configurations realizing an expected correlation level, or at least to obtain a better knowledge of the feasible space of correlations [Chérel, Cottineau, and Reuillon, 2015].

Direct applications

Starting from the second example which was limited to data generation, we propose examples of direct applications that should give an overview of the range of possibilities.

Calibration of network generation component at given density, on real data for transportation network¹⁰ should theoretically allow to unveil parameter sets reproducing accurately existing configurations both for urban morphology and network shape. It could be then possible to derive a “theoretical correlation” for these, as an empirical correlation is according to some theories of urban systems not computable as a unique realization of stochastic processes is observed. Because of non-ergodicity of urban systems [Pumain, 2012b], there are strong chances that involved processes are different across different geographical areas (or from an other point of view that they are in an other state of meta-parameters, i.e. in an other regime) and that their interpretation as different realizations of the same stochastic process makes no sense, the impossibility of covariation estimation following, except under simplified assumptions as we did in 4.1. By attributing a synthetic dataset similar to a given real configuration, we would be able to compute a sort of *intrinsic correlation* proper to this configuration. As territorial configurations emerge from spatio-temporal interdependences between components of territorial systems, this intrinsic correlation emerges the same way, and its knowledge gives information on these interdependences and thus on relations between territories and networks.

As already mentioned, most of models of simulation need an initial state generated artificially as soon as model parametrization is not done completely on real data. An advanced model sensitivity analysis implies a control on parameters for synthetic dataset generation, seen as model meta-parameters [Cottineau et al., 2015]. In the case of a statistical analysis of model outputs it provides a way to operate a second order statistical control.

We study stochastic processes for the study of synthetic data in C.3, in the sense of random time-series, whereas time did not have a role in the second case. We can suggest a strong coupling between the two model components (or the construction of an integrated model) and to observe indicators and correlations at different time steps during the generation. In a dynamical spatial models we have because of feedbacks necessarily propagation effects and therefore the existence of lagged interdependences in space and time [Pigozzi, 1980].

¹⁰ Typically road network given the shape of generated networks ; it should be straightforward to use OpenStreetMap open data that have a reasonable quality for Europe, at least for France [Girres and Touya, 2010], with however adjustments on generation procedure in order to avoid edge effects due its restrictive frame, for example by generating on an extended surface to keep only a central area on which calibration would be done.

It would drive our field of study towards a better understanding of dynamical correlations. A co-evolution model in that spirit will be proposed in chapter 7.

Generalization

We were limited to the control of first and second moments of generated data, but we could imagine a theoretical generalization allowing the control of moments at any order. However, as shown by the geographical example, the difficulty of generation in a concrete complex case questions the possibility of higher orders control when keeping a consistent structure model and a reasonable number of parameters. The study of non-linear dependence structures as proposed in [Chicheportiche and Bouchaud, 2013] is in an other perspective an interesting possible development.

We described a model allowing to generate synthetic datasets in which correlation structure is controlled, for which a generic formulation is given in Appendix B.3. Its partial implementation in two very different domains, in Appendix C.3 and here, shows its flexibility and the broad range of possible applications. More generally, it is crucial to favorise such practices of systematic validation of computational models by statistical analysis, in particular for agent-based models for which the question of validation stays an open issue.

Regarding our general problematic, we have introduced a first coupling between transportation networks and territories at the mesoscopic scale, through a sequential coupling of morphogenesis models. The development of this model into a co-evolution model will be the object of chapter 7.

★ ★

★

CHAPTER CONCLUSION

A general question relatively open regarding urban systems is the one of the *link between form and function*. Even if it is in some cases and at certain scales easily extricable, there does not seem to exist any general rule nor theory answering this fundamental problem. Will future *smart cities* be able to totally disconnect the form from the function as hypothesizes [Batty, 2017] ?

By situating oneself at the scale of a system of cities or a mega-urban region, for which the form will manifest in relative positions both geographically, but also according to multi-layer networks, of cities according to their specializations, or in the fine localization of the different types of activities within the region and the links formed by the transportation network, we can assume on the contrary that the new urban forms will be linked in ever more intricate and complex ways with their functions, at different scales and according to different dimensions.

The notion of morphogenesis, that we defined and partly explored, seems to be a good candidate to link form and function as we showed in 5.1. A simple model such as the one studied in 5.2 integrates this paradigm without providing any possible interpretation since functions are implicit in the processes considered. By coupling the model to the transportation network as done in 5.3, we explicitly introduce notions of functions since for example accessibility has now a role, but also because the network is a function in itself.

These paradigms will be used in the following to model co-evolution within a corresponding perspective in 7.2, i.e. at the mesoscopic scale with the same assumptions of autonomous processes and well defined sub-systems. We will deepen the reflexion on the role of functions with a multi-dimensional urban form in the study of the Lutecia model in 7.3, which will integrate the governance of the transportation system and relations between actives and employments within a metropolitan region.

* * *

*

CONCLUSION DE LA PARTIE II : CO-ÉVOLUTION, UN CONCEPT COMPLEXE AUX VISAGES MULTIPLES

Cette partie nous a permis d'apporter divers premiers éléments de réponse à notre problématique de modélisation de la co-évolution, en construisant à la fois des outils et en ouvrant des perspectives particulières.

Le premier chapitre, à la composition hétérogène, creuse des concepts fondamentaux issus de la théorie évolutive des villes, qui s'affirme ainsi comme partie intégrante de notre squelette conceptuel. L'étude des corrélations statiques confirme la non-stationnarité et suggère la multi-scalarité des interactions entre réseaux et territoires, et nous permet d'une part de confirmer la pertinence de notre approche à deux échelles distinctes, et d'autre part fournit une analyse empiriques construisant des données observées qui permettront de calibrer les modèles. Ensuite, la construction d'une caractérisation opérationnelle de la co-évolution, en termes de régimes de causalité, est essentielle à la fois du point de vue empirique et pour la caractérisation du comportement des modèles à venir. Enfin, nous explorons les potentialités des modèles d'interaction dans les systèmes de villes, ce qui nous permet de confirmer l'existence d'effet de réseau.

Le second chapitre explore le concept de morphogenèse, en commençant par en construire une définition interdisciplinaire qui suggère les paradigmes de modélisation par la forme et la fonction et introduit un lien implicite avec la co-évolution. Nous développons alors un modèle simple se basant uniquement sur la forme, par des principes d'agrégation-diffusion, et montrons que celui-ci reproduit une large gamme de formes territoriales existantes en Europe. Nous posons alors la première brique d'un couplage avec un modèle de croissance de réseau et explorons l'espace des corrélations statiques potentielles.

Nous pouvons faire à ce stade un bilan conceptuel de notre construction progressive.

Conceptual definition

Rappelons la définition conceptuelle de la co-évolution construite en particulier par transfert multidisciplinaire en première partie : des systèmes territoriaux évolutifs pourront présenter des propriétés de co-évolution à trois niveaux distincts : (i) entités locales en interactions réciproques ; (ii) population régionale d'entités présentant des

causalités circulaires d'un point de vue statistique ; (iii) interdépendances systémiques globales.

An operational approach

Cette partie aura également été cruciale puisqu'elle aura permis d'introduire une mesure opérationnelle de relations causales complexes, que nous proposons de considérer comme une méthode de caractérisation de la co-évolution, c'est-à-dire un proxy de celle-ci. Cette caractérisation, introduite et explorée en 4.2, se base sur l'idée de *régimes de causalité*, qui correspondent à des motifs de causalité au sens de Granger entre un ensemble de variables. Dans le cas de causalités réciproques entre deux populations d'entités, nous aurons bien *co-évolution* au deuxième sens donné ci-dessus. Nous avons donc ainsi une caractérisation empirique et opérationnelle de la co-évolution.

Morphogenesis

La morphogenèse appuie la question de l'autonomie et de l'interdépendance, des limites et de l'environnement, la question des échelles. Précisons dans quelle mesure celle-ci renforce la construction du concept de co-évolution. L'idée de sous-système indépendant rejoint celle de niche écologique équivalente à un système de frontières dans la théorie de HOLLAND [Holland, 2012]. Or celui-ci suppose les entités d'une même niche en co-évolution : on voit ainsi en filigrane que ce concept nous permet d'une part une entrée pertinente pour des modèles à l'échelle macroscopique, mais qu'il tisse d'autre part indubitablement bien que subrepticement des liens profonds avec la sphère conceptuelle que nous mettons progressivement en place.

Towards a modeling approach

En se raccrochant au triptyque des domaines de connaissance concepts-empirique-modèles [Livet et al., 2010], nous pouvons considérer être armé pour la composante encore manquante et qui est notre objectif final : celle des modèles, puisque nous avons amplement traité la co-évolution du point de vue conceptuel et du point de vue empirique.

L'enjeu de la partie suivante va donc être de produire une synthèse des briques que nous avons introduites, et construire progressivement des modèles de co-évolution aux deux échelles (macroscopique et mesoscopique), principalement en étendant les modèles déjà étudiés.

Part III

SYNTHÈSE

The buildings bricks, methods and tools are mainly set up for the culminating part of our work, which consists in the construction of models of co-evolution at different scales.

INTRODUCTION DE LA PARTIE III

Les contradictions ressenties au sein d'un contexte académique contraignant peuvent rapidement limiter les possibilités à la fois d'approfondissement mais aussi de synthèse. Le niveau de saturation est facilement atteint et la résignation à enterrer des illusions idéalistes passées est rapidement de mise. Mais la réceptivité du public permet d'échapper invisiblement à ces contraintes, et certains media y jouent un rôle déterminant. La plus marquante aura été celle par modélisation expérimentale. Le modèle comme outil de communication. Le modèle comme un jeu pour l'enseignement. Le modèle comme prétexte au développement d'une réflexion personnelle. Le modèle comme approfondissement de notions effleurées. Le modèle comme croisement des points de vue et des sensibilités. Le modèle à la convergence des concepts compris. Le modèle comme synthèse complexe. Le pessimisme ne doit finalement pas être de mise, les moyens les plus originaux de s'évader seraient aussi les plus efficaces.

Au coeur de notre sujet, nous devons à la fois faire la synthèse des entrées conceptuelle et empirique sur la co-évolution, et l'approfondissement des entrées thématiques. Les modèles vont être à la fois produits et producteurs de cette synthèse et de cet approfondissement, et permettre de nous extraire du cadre disciplinaire restrictif mis en valeur précédemment, en explorant des frontières floues des domaines et de la connaissance, à l'image de l'expérience de modélisation collective en enseignement imagée ci-dessus¹¹ dans laquelle le modèle a à la fois permis de s'extraire du cadre et d'opérer une synthèse et un approfondissement.

Cette partie vise ainsi à formuler et explorer des modèles de co-évolution, répondant à notre deuxième axe de la problématique, c'est-à-dire comment intégrer les processus de co-évolution dans des modèles. La question des échelles a été traitée de manière sous-jacente par les entrées thématiques complémentaires de la partie précédente : à une échelle mesoscopique, il sera plus pertinent de s'intéresser à la forme précise, tandis qu'à une échelle macroscopique les interactions entre agents sont fondamentales. Cette complémentarité des échelles fait par ailleurs écho à deux modèles séminaux de croissance urbaine, le modèle de Gibrat et le modèle de Simon. Nous démontrons en Annexe B.1 que ceux-ci sont deux spécifications d'un cadre plus global de modèles stochastique de croissance urbaine, ce qui suggère que nos deux approches sont non seulement complémentaires mais synthétisables.

¹¹ Qui a conduit à une réalisation concrète, voir <https://github.com/JusteRaimbault/ExperimentalModeling>.

Nous construisons ainsi les modèles dans deux chapitres, dont l'ordre a été fixé pour avoir un degré progressif de complexité des modèles. Le chapitre 6 développe les modèles à l'échelle macroscopique. Nous introduisons d'abord les indicateurs nécessaire pour qualifier le comportement de ce type de modèle, qui sont testés par application à un modèle de la littérature. Une extension directe du modèle d'interaction de 4.3 est ensuite proposée comme modèle de coévolution à l'échelle macroscopique.

Le chapitre 7 développe pour commencer un couplage fort du modèle de morphogenèse de 5.2 et de modèles de croissance de réseau, dans une démarche de multi-modélisation. Celui-ci est calibré sur données statiques calculées en 4.1. Nous introduisons ensuite un modèle à l'échelle métropolitaine prenant en compte des processus de gouvernance pour l'extension du réseau de transport.

* * *

*

6

CO-EVOLUTION AT THE MACROSCOPIC SCALE

Coupled dynamics between territories and networks can be grasped at the macroscopic scale through an approach by interactions, as we showed in chapter 4. The explicative power is then different to the one of classical economic models and concerns other types of processes, based on interactions at smaller spatial scales and longer time scales. In this frame, transportation networks and systems of cities co-evolve on long time.

To what extent the construction of the railway link through the Channel tunnel could have consolidated the economic power of London or reinforce its interactions with its close European neighbours, and to what extent the recent political events could lead to a modification of economic trajectories and then as a consequence to a modification of transportation patterns through a feedback of demand ? In a similar way, to what extent the projects of high speed lines on the East coast of the United States and in the California corridor are co-ordinated with regional dynamics, and if they are effectively realized, to what extent can they influence trajectories of the system of cities ?

We have already studied similar issues in the case of South Africa and with an empirical approach in 4.2, and we propose in this chapter to reflect it from the point of view of modeling, by introducing co-evolution processes in interaction models already developed.

To give an idea of the nature of conclusions we can expect to draw from such an approach, we begin in 6.1 by a systematic exploration of the SimpopNet model, approach which is the most advanced in terms of modeling the co-evolution of cities and transportation networks at this scale, as established in chapter 2. It also allows us to introduce the suited indicators for the evaluation of trajectories of systems of cities.

We then describe in 6.2 the generic model of co-evolution, which is tested on synthetic data at two levels of detail for network representation, and then on the French system of cities.

* * *

*

This chapter will be published as a book chapter [Raimbault, 2018b] for its first section. The second section describes the results of [Raimbault, 2017b]

for synthetic data, and will be published also as a book chapter [Raimbault, 2018a].

6.1 EXPLORING MACROSCOPIC MODELS OF CO-EVOLUTION

We first propose to introduce co-evolution models at the macroscopic scale by exploring the results produced by an existing model, what will also allow to introduce methods and indicators that are necessary to the exploration, and to grasp the typical questions linked to this type of models. In particular, we proceed to a systematic exploration of the SimpopNet model [Schmitt, 2014], which is to the best of our knowledge one of the rare initiatives to model co-evolution within a system of cities.

6.1.1 Context

A considerable gain in knowledge can be observed, from the conceptual or thematic description of a model, to its mathematical formalisation, its implementation, its systematic exploration, up to its exploration in depth with the help of specific meta-heuristics. Our postulate, that is a consequence of both our positioning (see Chapter 3 on simulation) and experiments of which previously developed models are part, is that it is important, but furthermore of a *qualitative* nature, in the sense that the nature of knowledge follows abrupt transitions during the advance of the investigation in this continuum.

The SimpopNet model introduced by [Schmitt, 2014], which is to our knowledge the only co-evolution model in the perspective of the evolutive urban theory. Its behavior was however not systematically explored, what makes it a good candidate for our approach.

Studied model

We briefly reformulate the model, following the notations for the formalization of the interaction model in 4.3, since a certain number of parameters and processes are similar. Cities grow following a specification that rejoins equation 7, i.e. in this specific case

$$\mu_i(t+1) - \mu_i(t) = \mu_i(t) \cdot \frac{\lambda^\beta}{N} \sum_j \frac{V_{ij}}{ < V_{ij} > }$$

where the potential is of the form $V_{ij} = \mu_j/d_{ij}^\beta$ and $V_{ii} = 0$, and β is a parameter for the distance decay and λ shape parameter for the decay function. We thus find our formulation, with $r_0 = 0$ and $w_G = \lambda^\beta \cdot N$. Since λ gives the typical distance of interaction, it will be noted d_G in the following, and β will be noted γ_G (it is indeed a level of hierarchy as a function of distance).

The network growths at each time step through a process that can be seen as a potential breakdown (as described in chapter 1): a couple of cities is chosen, the first according to populations with a hierarchy γ_N (i.e. with a probability proportional to $\mu_i^{\gamma_N}$) and the second following interaction forces $\mu_i\mu_j/d_{ij}^\beta$ with the same hierarchy

γ_N . A link is then created if the network is not efficient enough, i.e. if $d_{ij}/d_{ij}^{(N)} > \theta_N$. The links created at a date t have a speed $v(t)$, which will depend on current transportation technologies. The creation of new intersections to yield a planar graph is only done for links with a similar speed.

In order to study a stylized version of the model, we consider a configuration such that $v(t > 0) = v_0$ and $v(0) = 1$ (the initial model considers three values for speed that correspond to the reality of transportation technologies between 1830 and 2000).

Perspectives

We can put the structure of this model into perspective. Some modeling choices are not in direct consistency with the application it is used for: for example, such a precision in the parametrization of dates and speeds (historical dates from 1800 to 2000 and speed that approximately corresponds to transportation technologies) makes it a hybrid model, and should correspond to an application on a real spatial configuration. In a synthetic configuration as used in the model, these parameters have a sense only if we know the behavior of simulated dynamics, and in particular the role of the spatial configuration, i.e. if we are able to differentiate effects linked to the dynamics from effects linked to the initial spatial configuration.

Furthermore, the use of the interaction model without the endogenous Gibrat term would be difficultly adaptable to an application of the model on real data since the values we obtained in the precedent studies of interaction models, but stays relevant in a stylized model, in order to understand the interaction processes in an isolated way, as we will do later (keeping in mind that this knowledge does not necessarily describes the coupled behavior, since the interaction between processes can lead to the emergence of new behaviors).

The formulation of the potential, given above, as $(\lambda/d_{ij})^\beta$, implies that λ captures both the weight of the potential and the shape of the decreasing function, but imposes a dependence between these two effects, on the contrary to the specification we use previously. It furthermore does not allow an interpretation in terms of limit flows¹.

Finally, rules allowing variable values for $v(t)$ and the non-planarity mechanism², allows the introduction of a tunnel effect, which is as we recall is the absence of interaction of an infrastructure traversing a territory with it. The effect is however exogenous since explicitly specified in model rules, on the contrary to the interaction model with feedback of flows, in which the variations of w_N and d_N should capture an endogenous tunnel effect. The introduction of specific in-

¹ The weight parameter in our model in 4.3 gives indeed the value of the flow when the distance attenuation goes to infinity and for all the population.

² When a new link is constructed, it does create intersections only with links of similar speed.

dicitors to measure it would be an interesting development direction, but we stay here at considering the hierarchy of centralities which is already a good indicator for it³.

6.1.2 Methodology

Spatial configuration

An important aspect for understanding co-evolution processes implied in this model is the role of the initial spatial configuration in emerging patterns observed. We therefore apply the methodology developed in 3.1, which allows to extend the sensitivity analysis of a model to spatial meta-parameters⁴.

GENERATION OF SYNTHETIC CONFIGURATIONS A synthetic system of cities is constructed the following way (see Appendix B.3 for the notion of synthetic data, calibrated at the first and the second order). A fixed number N of cities is uniformly distributed in space, under the constraint of a minimal distance between each, and their population is attributed following a rank-size law which parameters P_m and α can be adjusted (the distribution of city sizes in the initial model corresponds to $\alpha \simeq 0.68$ with $R^2 = 0.98$).

A skeleton of network is created by progressive connection: the algorithm connects cities two by two by closest neighbour in terms of euclidian distance, and then iteratively selects randomly a cluster and connects it perpendicularly to the closest link outside the cluster. The network is then extended by the creation of local shortcuts, through a repetition n_s times of the random selection of a city according to populations, and its connection to a neighbour in a radius r_s under conditions of a maximal degree d_s . The final network is then made planar.

This process creates networks that visually correspond (in terms of the order of magnitude of the number of loops, and their spatial range) to the initialization of the model, knowing that a single instance of the network does not allow to determine distributions of topological parameters for which a more precise calibration could be done.

³ Indeed, a highly hierarchical distribution of accessibilities means that there exists a small number of cities very accessible and a large number with a low accessibility. If main cities reasonably cover the space, then their links necessarily ignore the overflowed cities with low accessibility, otherwise the distribution would be less hierarchical.

⁴ We recall that in our case a meta-parameter is a parameter allowing to generate an initial configuration upstream of the model.

Indicators

A crucial aspect of the study of simulation models is the definition of relevant indicators, particularly in the case of synthetic models where it is not possible to produce outputs that are directly linked to data for example. Very general stylized facts, as aiming at producing an urban hierarchy or a network hierarchy, are relatively limited. Moreover, the hierarchy is mechanically produced by most models including aggregation processes. We therefore need more elaborated indicators to understand the dynamics of the system. These indicators must in particular give elements of answer to the following questions:

- types of systems of cities produced by the model;
- change in time of the organization of the system of cities;
- typical profiles of trajectories;
- ability to “produce some co-evolution”.

In order to concentrate on the ability of the model to produce trajectories that are both diverse and complex, and for example its ability to produce bifurcations that would manifest as inversions in ranks, and also its ability to capture different aspects of co-evolutive dynamics, we propose a set of indicators, including for example lagged correlation measures in the spirit of causality regimes exhibited in 4.2, or a correlation measure as a function of distance, to understand the role of spatial interactions in the coupling of trajectories. Given a variable $X_i(t)$ defined for each city and in time (that will be the population or centrality measures for example), we define the following indicators.

- Indicators characterizing the distribution of X_i in time: hierarchy (slope of the least squares adjustment of X_i as a function of rank) $\alpha(t)$, entropy of the distribution $\varepsilon(t)$, descriptive statistics (average $\hat{E}[X](t)$ and standard deviation $\hat{\sigma}(t)$).
- Rank correlation between the initial time and the final time, which translates the quantity of change in the hierarchy during the evolution of the system, and is defined by $\rho_r = \hat{\rho}[\text{rg}(X_i(t=0)), \text{rg}(X_i(t=t_f))]$, where $\text{rg}(X_i)$ is the rank of X_i among all values.
- Diversity of trajectories $\mathcal{D}[X_i]$, which captures a diversity of time series profiles for the considered variable. With $\tilde{X}_i(t) \in [0; 1]$ the trajectories that have been individually rescaled, it is defined by

$$\mathcal{D}[X_i] = \frac{2}{N \cdot (N-1)} \sum_{i < j} \left(\frac{1}{T} \int_t (\tilde{X}_i(t) - \tilde{X}_j(t))^2 \right)^{\frac{1}{2}}$$

- Changes in direction of trajectories $C[X_i]$, that we take as the number of inflection points. In the context of such a type of model, which mainly produces monotonous trajectories, this indicator witnesses in a certain way of a “complexity” of trajectories.
- Correlations as a function of distance, to understand the way the effect of distance is translated at the macroscopic scale. The profile of this function, regarding interaction distance parameters included in the model, will translate the tendency of the model to lead to the emergence of one level of interaction or the other. It is computed as

$$\rho_d = \hat{\rho}[(X(\vec{x}_k), Y(\vec{x}_{k'}))]$$

where X_i, Y_i are the two variables considered and (k, k') the set of couples such that $\|\vec{x}_k - \vec{x}_{k'}\| - d \leq \varepsilon$ with ε a tolerance threshold (in practice taken to regroup couples by distance deciles).

- Lagged correlations between the variations of variables, to identify causality patterns between variables X and Y . The patterns $\hat{\rho}_\tau$ for all variables, and for τ lag or anticipation, must be understood in the sense of potential regimes, explored in 4.2.

$$\rho_\tau = \hat{\rho}[\Delta X(t - \tau), \Delta Y(t)]$$

These indicators are used on the following variables:

- populations $\mu_i(t)$,
- closeness centralities

$$c_i(t) = \frac{1}{N-1} \sum_{i \neq j} \frac{1}{d_{ij}(t)}$$

which capture the position within the urban system,

- accessibilities

$$x_i = \frac{1}{\sum_k \mu_k} \sum_{i \neq j} p_j \exp(-d_{ij}(t)/d_G)$$

which capture the insertion within the urban system.

We furthermore introduce diverse indicators for network topology, to understand the final forms produced by the heuristic: diameter, average path length, average betweenness centrality and its level of hierarchy, average performance, total length, as they have been defined in 4.1.

6.1.3 Results

Experience plan

Given an initial spatial configuration (i.e. a value of meta-parameters), we establish the behavior of indicators by exploring a grid of the parameter space. The number of parameters being low and the objective being a first grasp of the model behavior, in particular if it is able to produce co-evolution dynamics, we do not use more elaborated exploration methods. The parameters are $(d_G, \gamma_G, \gamma_N, \theta_N, v_0)$ and meta-parameters $(N_S, \alpha_S, d_S, n_S)$. We take also the meta-parameters into account in order to understand the sensitivity of the model to space.

We explore a grid of 16 configurations of meta-parameters, 324 configurations of parameters, and 30 random replications, what corresponds to 155520 simulations. They are executed on a computation grid with the intermediary of OpenMole⁵.

Convergence

Since the model is stochastic, it is important to control the convergence of indicators, that will be more or less easy depending on their variability. To quantify the variability of an indicator X regarding stochasticity, we use a measure similar to the one used in 5.2, given by $v[X] = \hat{E}[X] / \hat{\sigma}[X]$ with basic estimators for the expectance and the standard deviation. On the full set of replications, we obtain for all indicators given previously, a median for the ratio $v[X]$ estimated within replications, estimated on all parameter values, which takes a minimal value of 3.94, for the average accessibility at final time, what witnesses a low stochastic variability. We can furthermore use this value to estimate the level of convergence: it corresponds to a 95% confidence interval around the mean of relative size 0.18 (under the assumption of a normal distribution of the average), i.e. a good convergence. This aspect is crucial for the robustness of results.

Sensitivity to space

The Table 15 give values of \tilde{d} for 16 configurations of meta-parameters⁶, in comparison to an arbitrary reference configuration (first column). The hierarchy within the initial system of cities appears as the stronger determinant of variability, since all configurations with $\alpha_S = 1.5$ give values larger than 1.7, what witnesses a very strong sensitivity relative to this hierarchy.

Then, the number of cities plays a non negligible secondary role, giving the stronger effects of space. Thus, it is crucial to keep in mind this role of the initial configuration during the analysis of phase

⁵ Simulation results are available at <http://dx.doi.org/10.7910/DVN/RW8S36>.

⁶ The definition of the relative measure of sensitivity, given in 3.1, is for two phase diagrams f_1, f_2 and d euclidian distance, $\tilde{d} = 2d(f_1, f_2) / (\text{Var}[f_1] + \text{Var}[f_2])$.

diagrams. To stay within the same spirit than the model that was initially proposed, we will however comment a phase diagram for a given spatial configuration. The study of the extended model with integration of meta-parameters to which it is sensitive at their full extent is beyond the reach of this auxiliary analysis.

Table 15: Sensitivity to space of the SimpopNet model. Each column corresponds to an instance of the phase diagram, for which meta-parameters are given, with the relative distance to an arbitrary reference diagram. As inputs we have the meta-parameters N_S, α_S, d_S, n_S and as outputs of simulations the distance \tilde{d} .

N_S	40	40	40	40	40	40	40	40	80	80	80	80	80	80	80
α_S	0.5	0.5	0.5	0.5	1.5	1.5	1.5	1.5	0.5	0.5	0.5	1.5	1.5	1.5	1.5
d_S	5	5	10	10	5	5	10	10	5	5	10	5	5	10	10
n_S	10	30	10	30	10	30	10	30	10	30	10	30	10	30	10
\tilde{d}	0	0.05	0.26	0.21	1.79	1.80	1.79	1.72	0.44	0.36	0.42	0.42	2.25	2.23	2.24

Model behavior

The Fig. 43 reports the behavior of the model according to a selection among the diverse indicators given above. We comment a particular spatial configuration which corresponds to a low hierarchical system with a network having only local shortcuts, given by meta-parameters $N_S = 80, \alpha_S = 0.5, d_S = 10, n_S = 30$, which are the values giving configurations that are the most similar to the one of the initial model. Complete plots are available in Appendix A.8.

The values taken by the entropy for centralities (first panel of Fig. 43), as a function of time, for $\gamma_N = 2.5$ and $v_0 = 110$, exhibit different regimes depending on d_G and γ_G . A low hierarchy leads to an entropy stabilizing in time, what corresponds to a certain uniformization of distances. On the contrary, a strong hierarchy produces a regime with a minimum, and then an increase of disparities in time.

This variety of behaviors can be found again with the rank correlation ρ_R , that we show here for the population variable, as a function of d_G . It has a low sensitivity to θ_N and γ_N (see Appendix A.8), but strongly varies as a function of d_G and γ_G : interactions at a higher distance induce systematically a larger number of changes in the hierarchy of populations. These can occur when the hierarchy of distance is low. To summarize, the increase of the range of interactions will diminish the inertia of trajectories of the system of cities, whereas the increase of their hierarchy will increase it. This is relatively credible from a thematic point of view: longer and uniform interactions have more chances to make individual trajectories change.

The behavior of correlation indicators is shown in Fig. 44. Concerning the effect of distance on correlations between variables, i.e. the evolution of ρ_d , it is interesting to note that an increase of d_G sys-

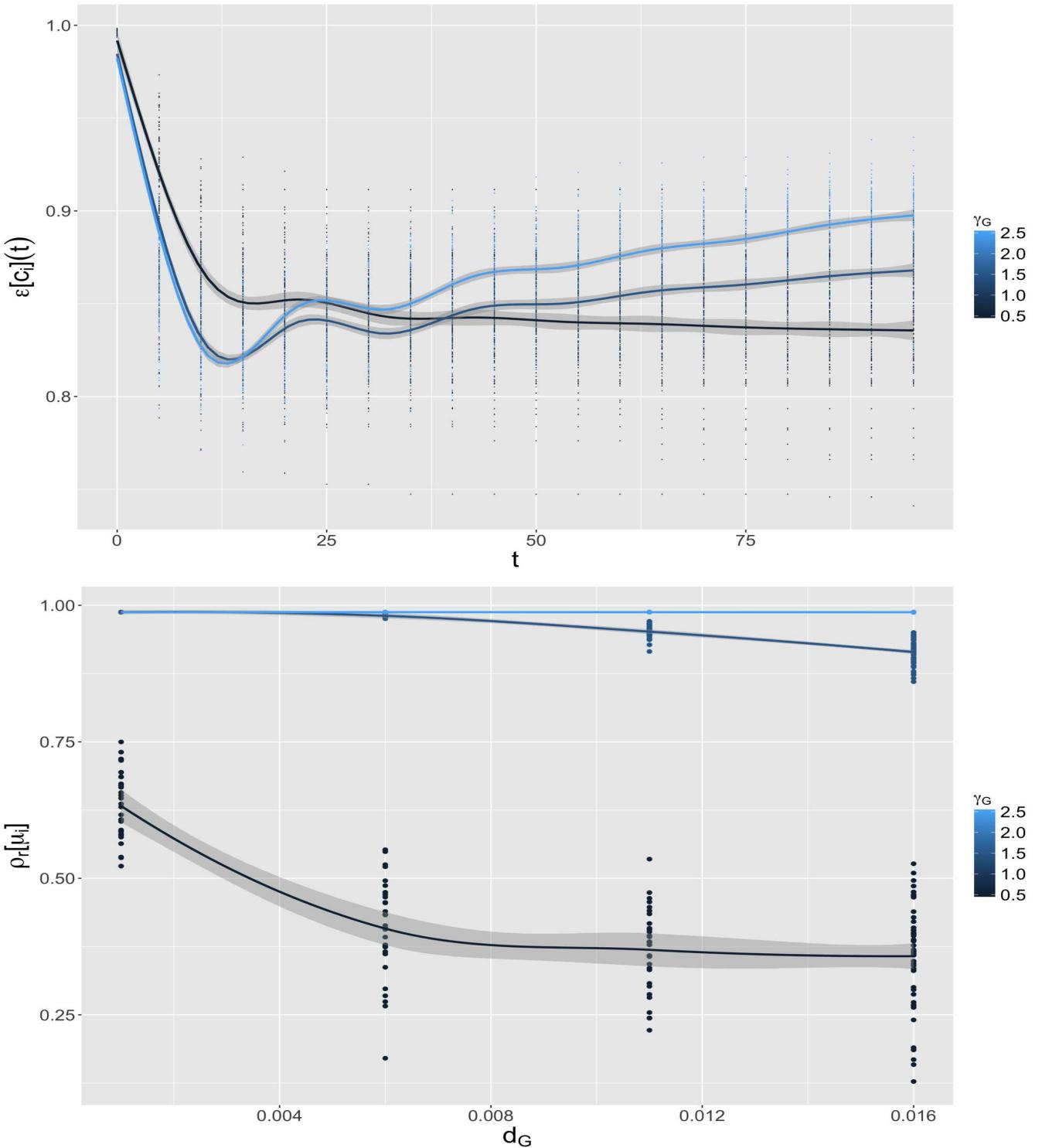


Figure 43: **Model behavior for the spatial configuration** $N_S = 80, \alpha_S = 0.5, d_S = 10, n_S = 30$. (Top) Temporal trajectories of the entropy for closeness centralities, for $\gamma_N = 2.5, v_0 = 110, d_G = 0.016, \theta_N = 11$, as a function of γ_G (color); (Bottom) Rank correlation for population, as a function of d_G and of γ_G (color), for $\theta_N = 11, \gamma_N = 2.5$.

tematically diminishes the levels of correlation, what corresponds to the complexification that we previously showed. As expected, $\rho_d[d]$ decreases as a function of distance, and exhibits non zero values for the correlation between population and centrality for a high hierarchy γ_G , what shows that simultaneous adaptation regimes are rare in this model.

Causality regimes

Finally, by studying ρ_τ (Fig. 44, bottom panel), we observe that causality regimes in the sense of 4.2 are not very varied (as the Fig. 95 in Appendix A.8 confirms it for a broader range of parameters). The population is systematically caused by the centrality, but there exists no regime in which we observe the contrary. This is a logic of an effect of reinforcement of hierarchy by centrality, but not a configuration with circular causalities, and thus not a co-evolution properly speaking as we defined in the statistical sense.

This brief exploration allows us to say that this model captures urban trajectories of a certain complexity, but that it does apparently not reproduces co-evolution regimes.

★ ★

★

We have thus in this section introduced the tools to understand trajectories produced by a co-evolution model, and tested these on the SimpopNet model.

In the following, we will explore in the same spirit a co-evolutive extension of the interaction model developed in 4.3, and will aim at establishing to what extent it is able to capture co-evolutive dynamics.

★ ★

★

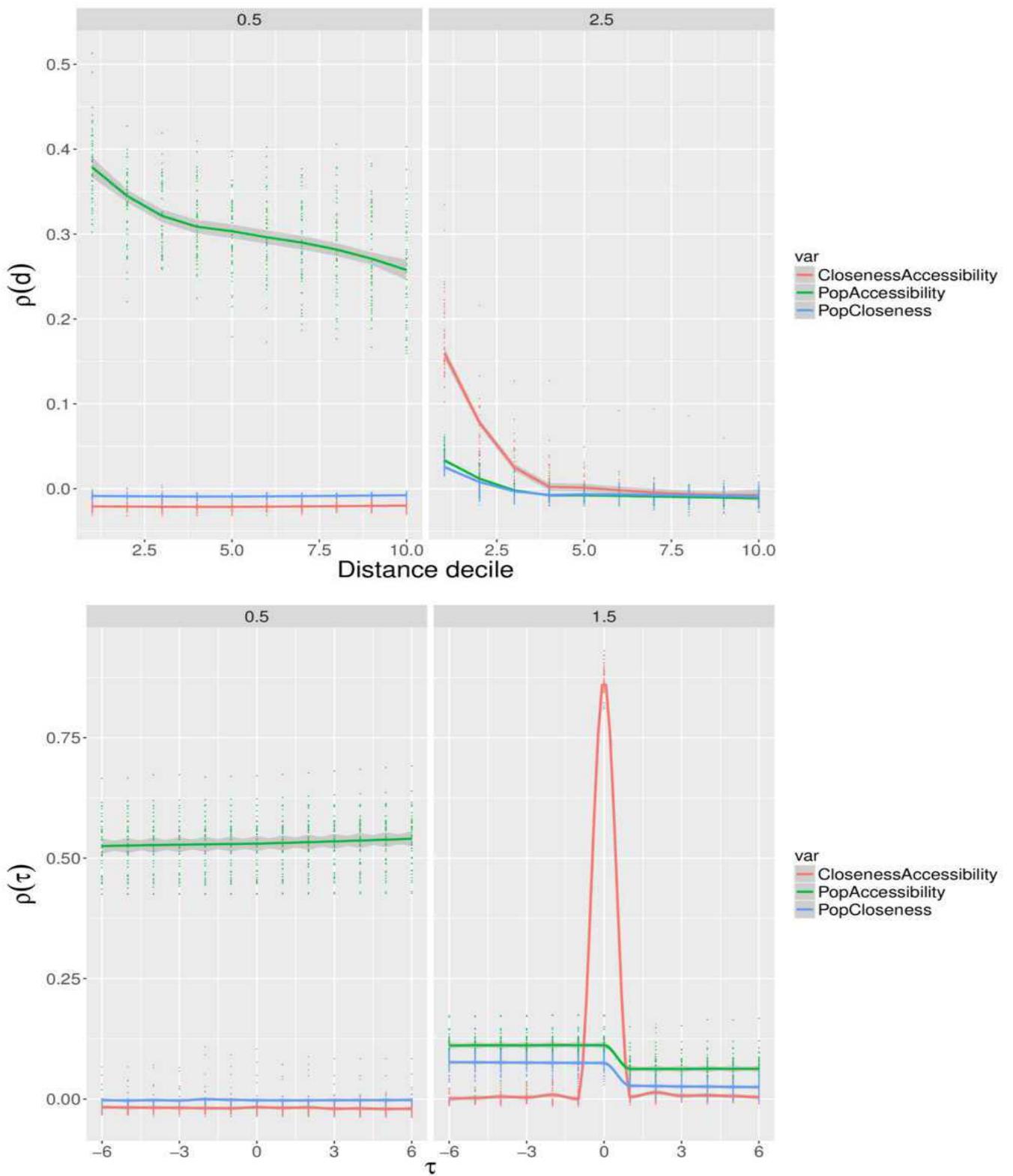


Figure 44: **Correlations in the model for the spatial configuration** $N_S = 80, \alpha_S = 0.5, d_S = 10, n_S = 30$. (Top) Correlations as a function of distance, for couples of variables (color), for $\gamma_N = 2.5, \theta_N = 21, v_0 = 10$, and for d_G (columns) and γ_G (rows) variables; (Bottom) Lagged correlations for the same parameters.

6.2 DYNAMICAL EXTENSION OF THE INTERACTION MODEL

This section extends the logic of integrating a system of cities with a transportation network, which has been pursued in a static way for network behavior in the interaction model developed and explored in section 4.3, to propose a *macroscopic model of co-evolution for systems of cities*.

6.2.1 Macroscopic Model of Co-evolution

Rationale

This first approach relies in a direct extension of the interaction model within a system of cities described in chapter 4, at a macroscopic scale with an ontology typical to systems of cities. For the sake of simplicity, we still stick to an unidimensional description of cities by their population.

Concerning network growth, we propose also to stay at a relatively aggregated and simplified level, allowing to test growth heuristics at different levels of abstraction. In order to be flexible on model mechanisms, diverse processes can be taken into account, such as direct interactions between cities, intermediate interactions through the network, the feedback of network flows and a demand-induced network growth.

Empirical characteristics emphasized by [Thévenin, Schwartz, and Sapet, 2013] for the French railway network suggest the existence of feedbacks of network use, or of flows traversing it, on its persistence and its development, whose properties have evolved in time: a first phase of strong development would correspond to an answer to a high need of coverage, followed by a reinforcement of main link and the disappearance of weakest links.

The coupling between cities and the network will be achieved by the intermediate of flows between cities in the network: these capture the interactions between cities and have simultaneously an influence on the network in which they flow.

General Formulation

The urban system is characterized by populations $\mu_i(t)$ and the network $G(t)$, to which can be associated a distance matrix $d_{ij}^G(t)$. Flows between cities ϕ_{ij} follow the expression given in 4.3 with network distance. The same way, the evolution of populations follows the specifications of the base model. The Fig. 45 shows the structure of the model.

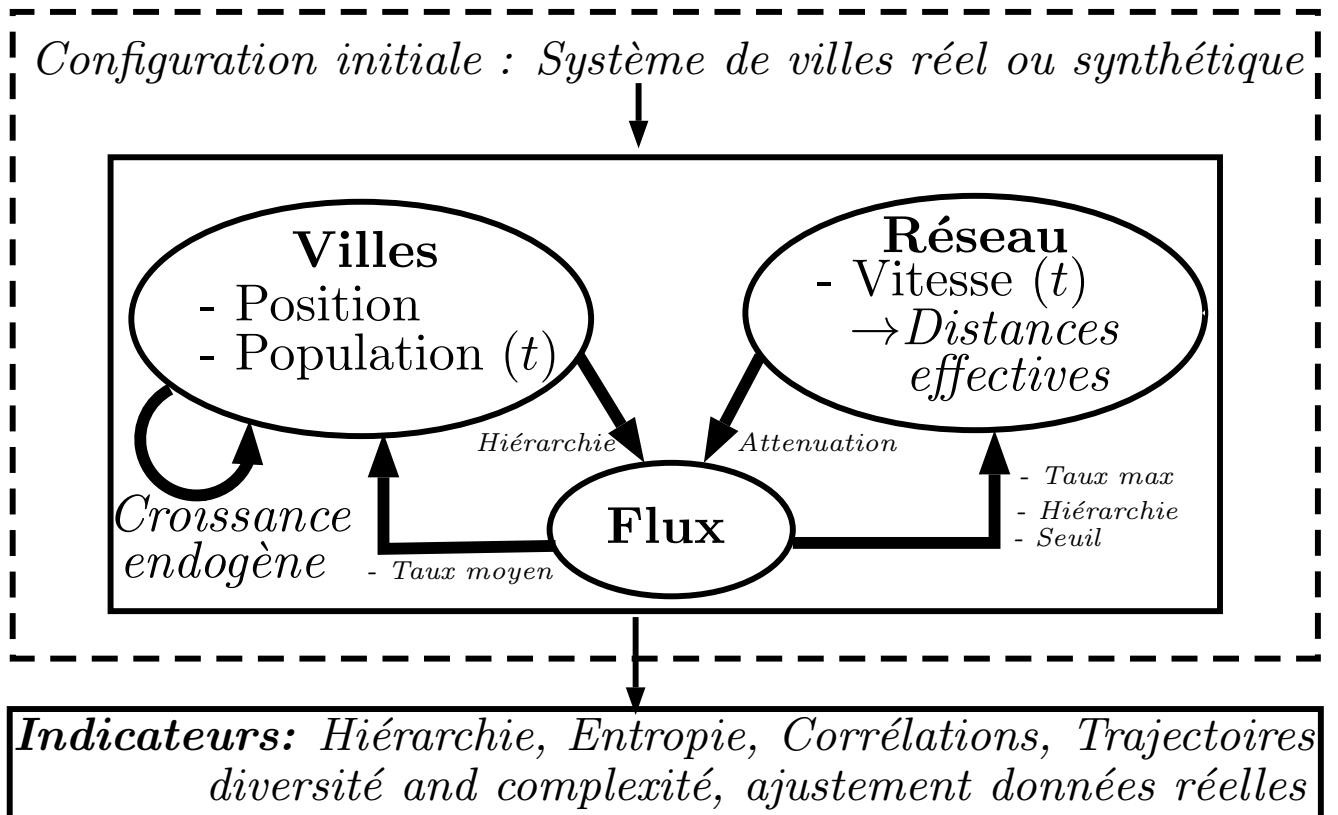


Figure 45: **Abstract representation of the model.** Ellipses correspond to main ontological elements (cities, network, flows), whereas arrows translate processes for which associated parameters are given. The model is described in its broader ecosystem of initialisation and output indicators.

NETWORK GROWTH Concerning the network, we assume that it evolves following the equation

$$\mathbf{G}(t+1) = F(\mathbf{G}(t), \phi_{ij}(t)) \quad (14)$$

such that the assignment of flows within the network and a local variation of its elements is possible. We propose in a first time to consider patterns linked to distance only, and to specify a relation on an abstract network as

$$d_{ij}^G(t+1) = F(d_{ij}^G(t), \phi_{ij}(t)) \quad (15)$$

i.e. an evolution of the distance matrix only. In this spirit, we keep an interaction model strictly at a macroscopic scale, since a precise spatialization of the network would imply to take into account a finer scale that includes the local shape of the network which determines shortest paths.

Following a thresholded feedback heuristic, given a flow ϕ in a link, we assume its effective distance to be updated by:

$$d(t+1) = d(t) \cdot \left(1 + g_{\max} \cdot \left[\frac{1 - \left(\frac{\phi}{\phi_0} \right)^{\gamma_s}}{1 + \left(\frac{\phi}{\phi_0} \right)^{\gamma_s}} \right] \right) \quad (16)$$

with γ_s a hierarchy parameter, ϕ_0 the threshold parameter and g_{\max} the maximal growth rate at each step. This auto-reinforcement function can be interpreted the following way: above a limit flow ϕ_0 , the travel conditions improve, whereas they deteriorate below. The hierarchy of gain is given by γ_s , and since $\frac{1 - \left(\frac{\phi}{\phi_0} \right)^{\gamma_s}}{1 + \left(\frac{\phi}{\phi_0} \right)^{\gamma_s}} \rightarrow_{\phi \rightarrow \infty} -1$, g_{\max} is the maximal distance gain. This function is similar to the one used by [Tero, Kobayashi, and Nakagaki, 2007]⁷.

Implementation

The coupling of the interaction model to a finer representation of the network (for example an encoding of the whole network structure) makes the full integration into an OpenMole plugin more difficult, as it was done for the model studied in 4.3. We need here an *ad hoc* implementation. The use of a workflow as a mediator for coupling is an interesting solution but which is realistic only for a weak coupling as in 5.3. One of the issues that the meta-modeling library for OpenMole that is currently being developed around OpenMole will have to tackle is the possibility to allow strong coupling (for example

⁷ Which uses $\Delta d = \Delta t \left[\frac{\phi^\gamma}{1 + \phi^\gamma} - d \right]$. This function yield similarly a threshold effect, since the derivative vanishes at $\phi^* = \left(\frac{d}{1-d} \right)^{1/\gamma}$, but it can not be adjusted.

in the sense of a dynamical coupling during the evolution of the simulation) of heterogeneous components in a transparent way, in order to benefit from the advantages of different languages or of already existing implementations.

We choose here a full implementation with NetLogo, for the simplicity of coupling between components. A particular care is taken for the duality of network representation, both as a distance matrix and as a physical network, in order to facilitate the extension to physical network heuristics.

6.2.2 Application to Synthetic Data

The model is first tested and explored on synthetic city systems, in order to understand some of its intrinsic properties. In this case, we consider the model with an abstract network as specified above, i.e. without spatial description of the network and with evolution rules acting directly on d_{ij}^G given the previous specifications.

Synthetic data

A synthetic city system is generated following the heuristic used in the previous section: (i) N_S cities are randomly distributed in the euclidian plan; (ii) populations are attributed to cities following an inverse power law, with a hierarchy parameter α_S and such that the largest city has a population equal to P_{\max} , i.e. following $P_i = P_{\max} \cdot i^{-\alpha_S}$.

To simplify, several meta-parameters are fixed: the number of cities is fixed at $N_S = 30$, the maximal population at $P_{\max} = 100000$ and the maximal network growth to $g_{\max} = 0.005$. Final time is fixed at $t_f = 30$, what corresponds to distances divided approximatively by 5⁸, in order to comply to an empirical constraint: this corresponds to the evolution of the travel time between Paris and Lyon from around ten hours at the beginning of the century to two hours today, showed for example by [Thévenin, Schwartz, and Sapet, 2013]. We also neglect network effects at the second order by taking $w_N = 0$.

We explore a grid in the parameter space $\alpha_S, \phi_0, \gamma_s, w_G, d_G, \gamma_G$. We use the indicators introduced in 6.1 to quantify model behavior in the parameter space. We describe the results for $\alpha_S = 1$, what is the closest to existing city systems (in comparison to 0.5 and 1.5, see the systematic review of the rank-size law estimations done by [Cotineau, 2017]).

⁸ Indeed, we can compute that the minimal multiplicative factor for distance is $(1 - g_{\max})^{t_f}$, what gives for these values $(1 - 0.05)^{30} \simeq 0.214$, i.e. a division by 5 of the travel time.

Trajectories

The evolution of the average closeness centrality in time is shown in Fig. 96 (top) for $w_G = 0.001$, and with variables (γ_G, ϕ_0) . The behavior is not sensitive to d_G (see the complete plots in A.9). This evolution witnesses a transition as a function of the level of hierarchy: when it decreases, we observe the emergence of trajectories for which the average centrality increases in time, what corresponds to configurations in which all cities profit in average from accessibility gains.

Concerning the entropy of populations, for which the temporal trajectory is shown in Fig. 96 (bottom), all parameters give a decreasing entropy, i.e. a behavior of convergence of cities trajectories in time⁹.

Looking at the complexity of accessibility trajectories, we observe for values of $\phi_0 > 1.5$ a maximum of complexity as a function of interaction distance d_G , stable when w_G and γ_G vary (see also the exhaustive plots in Fig. 97, Appendix A.9). This intermediate scale can be interpreted as producing regional subsystems, large enough for each to develop a certain level of complexity, et isolated enough to avoid the convergence of trajectories over the whole system. We reconstruct therein a spatial non-stationarity, typically observed in 4.1, and rejoin the concept of the ecological niche¹⁰ localized in space: the emergent subsystems that are relatively independent, are good candidates to contain processes of co-evolution. The emergence of this intermediate scale can be compared to the modularity of the French urban system showed by [Berroir et al., 2017].

Finally, the behavior of rank correlations for accessibility reveals that the interaction distance systematically increases the number of hierarchy inversions, what corresponds in a sense to an increase in overall system complexity. The hierarchy parameter diminishes this correlation, what means that a more hierarchical organization will impact a larger number of cities in the qualitative aspects of their trajectories. This effect is similar to the “first mover advantage” showed by [Levinson and Xie, 2011], which unveils a path dependency and an advantage to be rapidly connected to the network: in our case, the modifications in the hierarchy correspond to cities that benefit from their positioning in the network.

Correlations

We can in a first time focus on the variations of correlations between variables as a function of distance. Profiles of ρ_d for the three couples

⁹ Indeed, the entropy for the population variable gives the dispersion of the distribution of populations, and thus its decrease translate a trend to concentrate in time.

¹⁰ As it was already described in 5.1, an ecological niche in the sense of [Holland, 2012] corresponds to the relatively independent ecosystem in which there is co-evolution between the species.

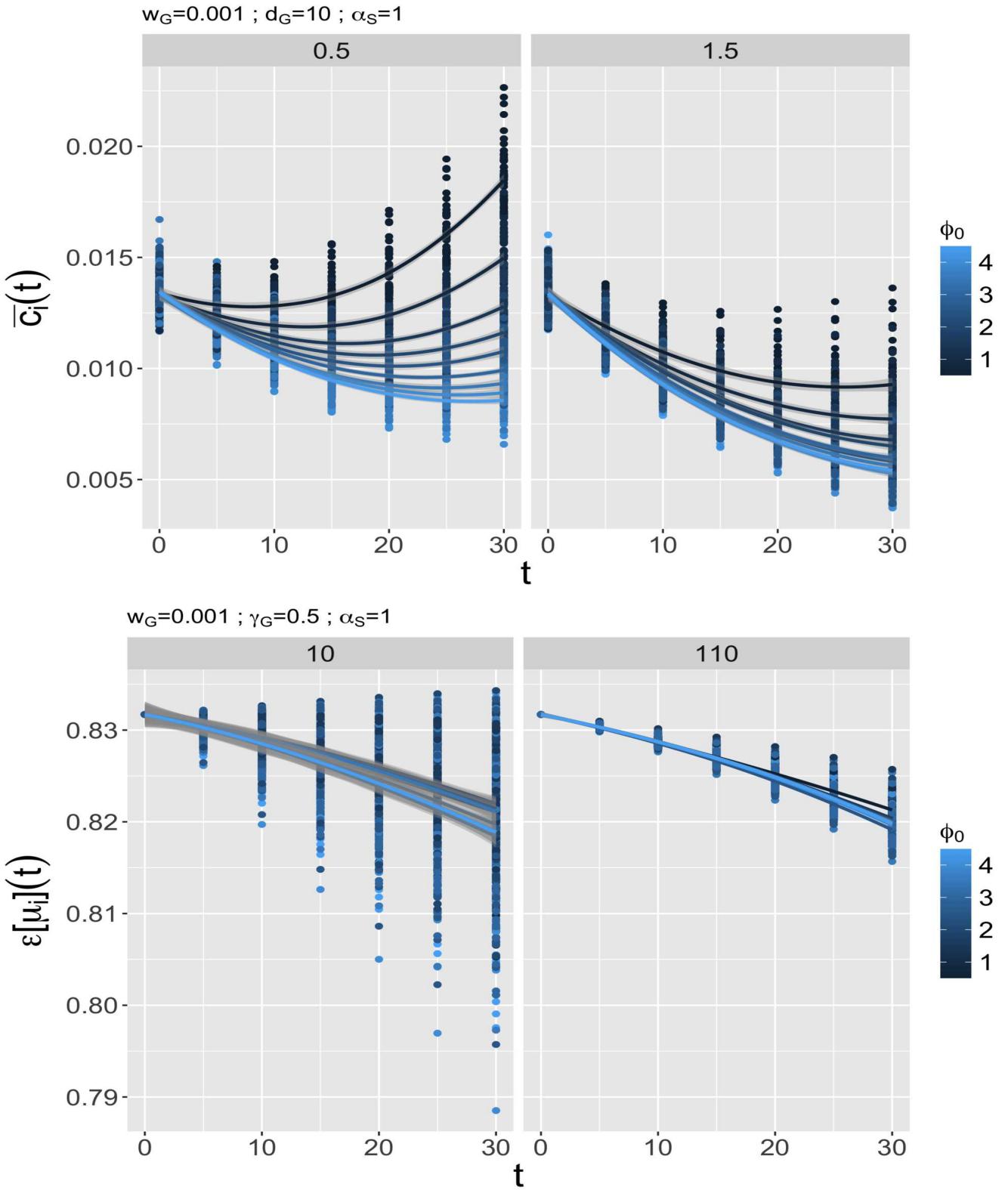


Figure 46: Temporal behavior of the co-evolution model with abstract network on a synthetic system of cities.
 (Top) Average closeness centralities, as a function of time, for γ_G (rows) and ϕ_0 (color) variable, at fixed $w_G = 0.001$ and $d_G = 10$; (Bottom) Entropy of populations, as a function of time, for d_G (columns) and ϕ_0 (color) variable, at fixed $w_G = 0.001$ and $\gamma_G = 0.5$. See main text for interpretation. Trajectories on the explored subspace of the parameter space are given in Fig. ??, Appendix A.9.

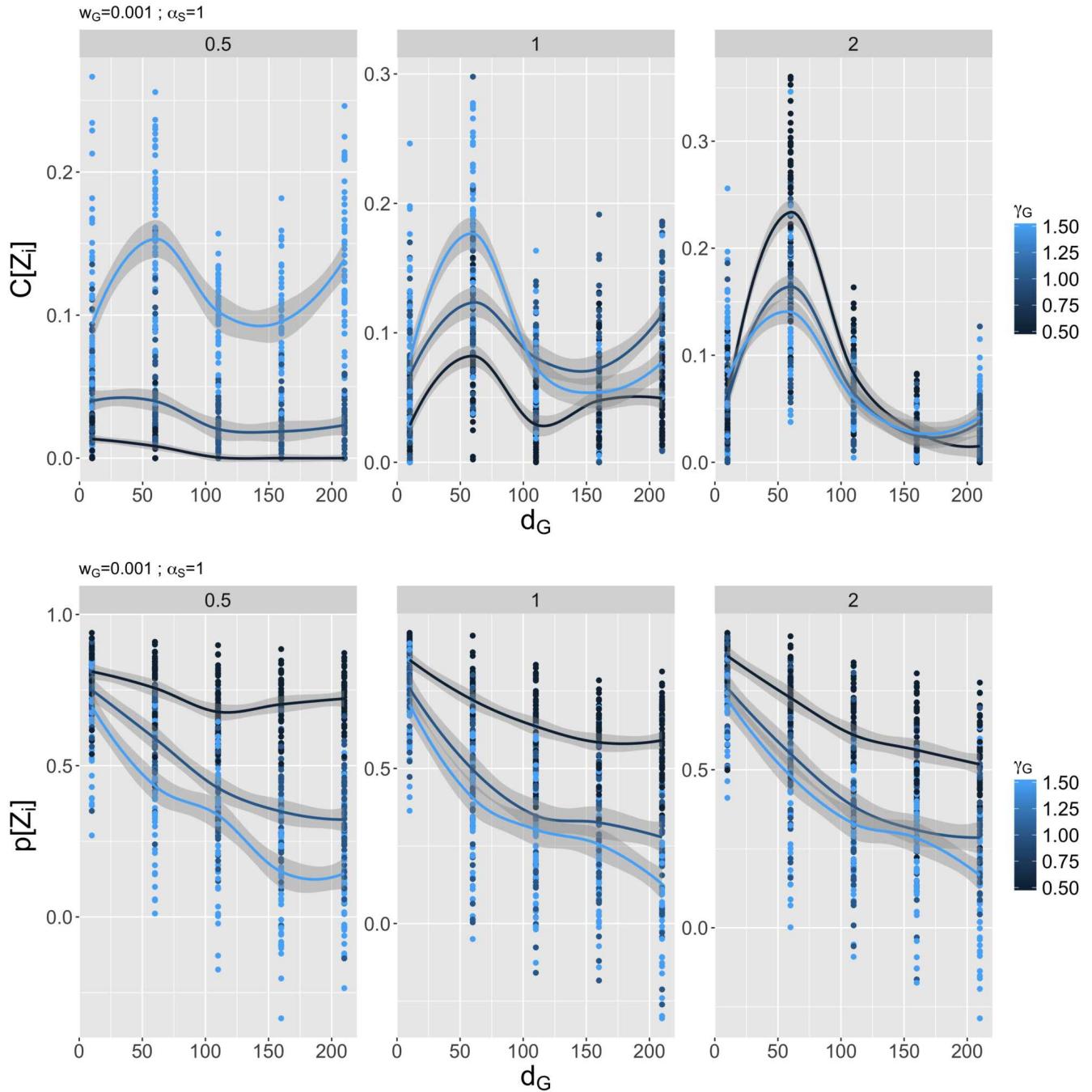


Figure 47: **Aggregated behavior of the co-evolution model.** (Top) Complexity of accessibilities, as a function of d_G , for ϕ_0 (columns) and γ_G (color) variable, at fixed $w_G = 0.001$; (Bottom) Rank correlations of accessibilities as a function of d_G , for the same parameters. The behavior on the explored subspace of the parameter space are given in Fig. 97, Appendix A.9.

of variables show that intermediate and large values of the interaction distance ($d_G > 50$) induce populations totally uncorrelated with centralities and accessibilities (Fig. 99, Appendix A.9). For small values of d_G , a decreasing then vanishing profile confirms the existence of strong local effects, where very close cities will have a strong reciprocal influence. The behavior of the correlation between accessibility and centrality is more difficult to interpret, and may be due to auto-correlation phenomena¹¹. Its level does not depend on distance but on d_G , and decreases to end at a negative correlation.

Causality regimes

We can now study lagged correlation patterns produced by the model, i.e. its ability to effectively produce co-evolution in the sense we defined.

The exploration of profiles for ρ_τ for varying parameter values is illustrated in Appendix A.9, and suggests the existence of multiple causality regimes. The Fig. 48 give examples of such profiles. We however observe (i) the systematic existence of a constant correlation at $\tau = 0$ and (ii) the small variations of correlations that impose the need for a statistical test to ensure that we isolate a significant effect.

We add here for this reason an additional criteria based on a statistical test: for $\tau_+ = \text{argmax}_{\tau>0} |\rho_\tau - \rho_0|$ and $\tau_- = \text{argmax}_{\tau<0} |\rho_\tau - \rho_0|$, a Kolmogorov-Smirnov test is used to compare the distributions of $\rho_{\tau\pm}$ and of ρ_0 . If they are declared different with a p-value smaller than 0.01, and if $|\rho_{\tau\pm}| > |\rho_0|$, we accept the causality link between variables in the corresponding direction.

A configuration is then coded by a representation of its graph between variables, given by the six discrete variables equal to 0 if there is no link between the variables (within all directed couples between population, accessibility and centrality) and 1 or -1 depending on the sign of the correlation if there exists a statistically significant link (in practice we observe only positive correlations).

We obtain overall 33 different configurations of links between variables, out of the 64 possible configurations (2^6 possible choices for positive correlations only). In comparison, the application of this method on the results of 6.1 give only 8 distinct configurations¹².

¹¹ These can not be computed, as it implies to decompose $\rho \left[\sum_{i \neq j} \frac{1}{d_{ij}}, \sum_{i \neq j} P_j \exp(-d_{ij}/d_G) \right]$. It is for example possible to approximate $\rho[X+Y; Z]$ under the condition that $\varepsilon = \sigma_Y/\sigma_X \ll 1$ at the first order by $\rho[X+Y; Z] \simeq (\rho[X; Z] + \varepsilon \rho[Y; Z]) \cdot \left(1 - \frac{1}{2} \rho[X; Y] \varepsilon - \frac{\varepsilon^2}{2}\right)$, by this assumption is too restrictive to be used for all terms in the sum.

¹² In which two configurations correspond to a negative circular causality between accessibility and centrality, what suggests that the SimpopNet model can produce a co-evolution between variables, but in a restricted number compared to the configurations obtained here and only between two network variables.

The type of relations we obtain are particularly interesting regarding co-evolution. We indeed observe:

- a configuration without any link between variables;
- 13 configurations of type “structuring effect”, i.e. for which the graph does not have any loop;
- a configuration of type “indirect co-evolution”, for which the graph has a loop of length three ($c_i \rightarrow X_i \rightarrow \mu_i \rightarrow c_i$) ;
- 18 configurations of type “co-evolution”, in which there exists at least a loop of length two (direct circular relation between two variables).

Among all these regimes, 8 correspond to a graph with at least 4 links (which are then necessarily co-evolutive): we show these profiles in Fig. 48. Two regimes witness a positive deviation of the correlation between population and accessibility for positive delays, increasing up to the maximal delay, what could be a clue of a reinforcement of population dynamics through centrality, stylized fact shown for the French system of cities by [Bretagnolle, 2009].

The regimes in which the centrality is co-evolving with population correspond to the ones where the co-evolution between the network and the territory is the strongest (since the accessibility depends on both), and are observed for large values of d_G (average $d_G = 183$ on 62 parameter points). This way, this co-evolution is favored by long interaction ranges.

Finally, the regime with the largest number of links¹³, is obtained for a long interaction range $d_G = 160$, a strong interaction hierarchy $\gamma_G = 1.5$, but a low hierarchy of the initial system of cities α_S : far-reaching but hierarchical interactions in an uniform system of cities lead to a maximum of entanglement between variables.

We finally confirm these results of variety in causality regimes produced by the model by applying the *Pattern Space Exploration* algorithm [Chérel, Cottineau, and Reuillon, 2015] to the model, with objectives the six correlations studied above (evaluated as zero in the case of a non-significance). A graphical presentation of results is given in Appendix A.9. We mainly obtain a number of regimes produced by the model larger than the ones obtained before (with negative correlations, 260 realized regimes out of $3^6 = 729$ possible). This short complementary study confirms the ability of the model to produce a large number of co-evolution regimes.

¹³ That corresponds to the regime coded by “10/11/11”, with co-evolution of population and centrality and of population and accessibility, and a causality of centrality on accessibility.

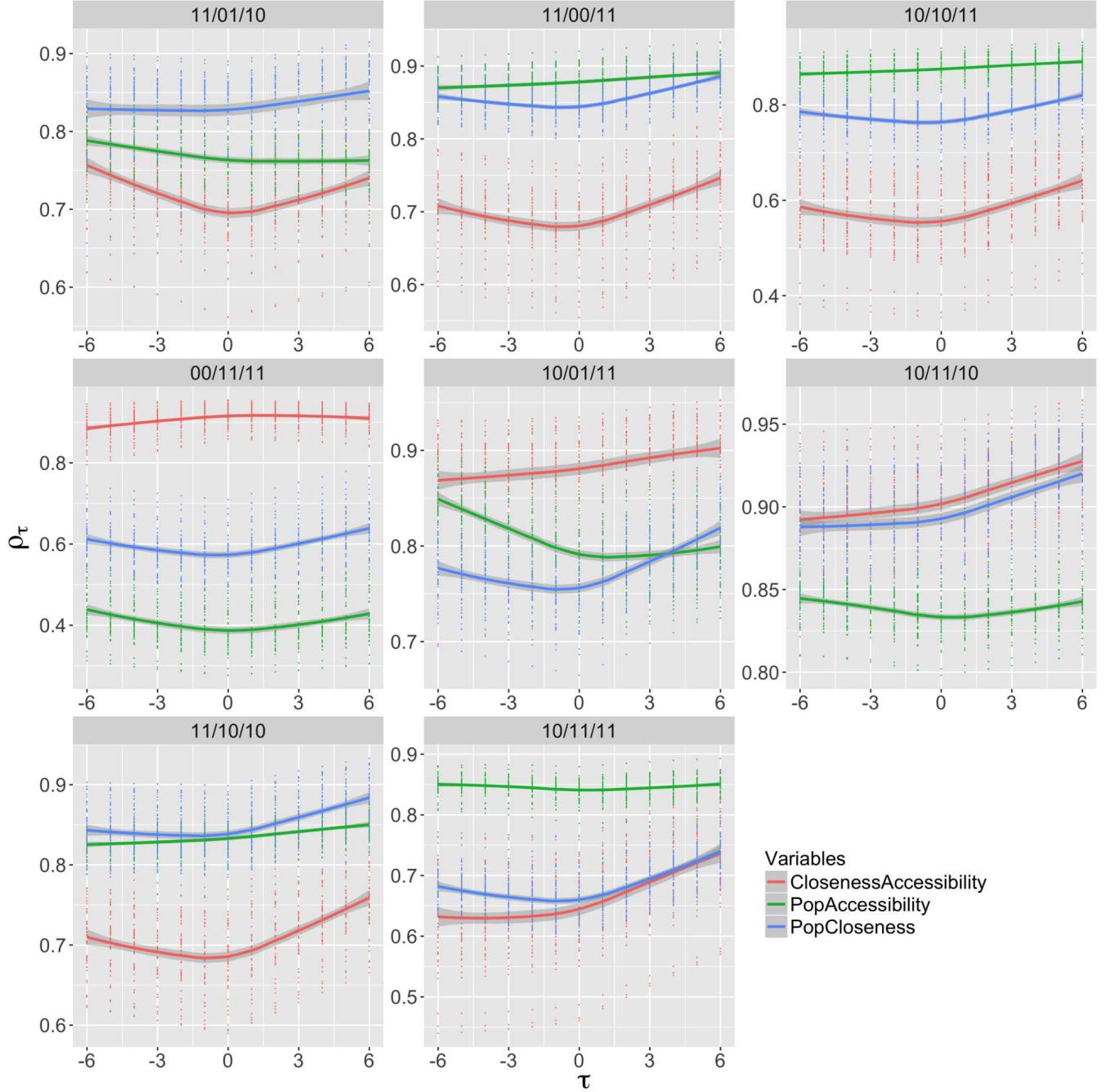


Figure 48: **Lagged correlations.** We give here for the 8 configurations showing at least 4 links between variables (coded in the order of couples, by the existence or not of a link for τ_+ and for τ_-), the lagged correlation profiles ρ_τ as a function of τ , for all couples of variables (color).

Synthesis

The important stylized facts that can be drawn from the exploration of the model on synthetic data are the following.

1. We observe the existence of an intermediate spatial scale allowing the evolution of relatively independent niches, corresponding to a maximal level of complexity for cities trajectories.
2. Lagged correlations unveil at least three different types of interaction regimes, that we interpret as an adaptation regime, a direct co-evolution regime, and an indirect co-evolution regime.

6.2.3 Applications to French City System

The model is then applied to the French system of cities on long time dynamical data: the Pumain-INED database for populations, spanning from 1831 to 1999 [Pumain and Riandey, 1986], with the evolving railway network from 1840 to 2000 [Thévenin, Schwartz, and Sapet, 2013]. Such a time span can be associated with structural effect on long time, as developed in 1.1. This application aims on the one hand at testing the ability of the model to reproduce a real dynamic of co-evolution, and on the other hand at extracting thematic information on processes through calibrated parameter values.

Network Data

We work on railway network data constructed by [Thévenin, Schwartz, and Sapet, 2013]. The French railway network is particularly interesting jointly with population data already presented, since the covered time span is relatively close, and as [Thévenin, Schwartz, and Sapet, 2013] recalls, this transportation mode has at any times materialized the implication of public and private actors. It corresponds to different processes depending on the period, from a more decentralized management to a more centralized recently, and different technological materializations with for example the recent emergence of high speed trains [Zembri, 1997]. For each date in the population database, we extract the simplified abstract network in which all stations and intersections with a degree larger than two are linked with abstract links which speed and length attributes correspond to real values, at a granularity of 1km¹⁴. This yields the time-distance matrices between the cities included in the model.

Stylized facts

Before calibrating the model, we can observe the lagged correlation patterns in the dataset, by applying the causality regimes method.

¹⁴ This processing is achieved thanks to the R package for transportation network analysis specifically developed for this thesis, see E.1.

This empirical study should on the one hand allow us to verify well known stylized facts, and on the other hand to produce a preliminary knowledge of empirical system behavior. We compute as detailed above the closeness centrality through the network, given by $T_i = \sum_j \exp -d_{ij}/d_0$, and we study the lagged correlation between its derivative ΔT_i and the derivative of the population ΔP_i , given by $\hat{\rho}_\tau = \hat{\rho} [\Delta P_i(t), \Delta T_i(t - \tau)]$ estimated on a moving window containing T_w successive dates. We show in Fig. 49 the results obtained.

These results are important for at least two reasons. First, the behavior of the number of significant correlations as a function of T_w and d_0 allows us to find stationarity scales in the system. We observe on the one hand a specific spatial scale that gives a maximum for all temporal windows, at $d_0 = 100\text{km}$, what suggests the existence of consistent regional subsystems, which existence is stable in time: indeed, this value corresponds to the interaction distance. It remarkably coincides with the intermediate scale isolated in the synthetic model. On the other hand, long spatial ranges induce an optimal temporal scale, for $T_w = 4$ what corresponds to around twenty years: we identify it as the overall temporal stationarity scale of the system and study the lagged correlations for this value.

Secondly, the behavior of lagged correlations does not seem to comply to the existing literature. At the intermediate spatial scale, the values of ρ_+, ρ_- exhibit no regularity. On the whole system, there is until 1946 close to no significant effect, then no causality between 1946 and 1975 (maximum at $\tau = 0$, non-significant minimum), and a 5 years shift of accessibility causing population after 1968 (the effect staying however doubtful). We do not reproduce the correlation effect between network centrality and place in the urban hierarchy advocated by [Bretagnolle, 2003]¹⁵, what lead us to question the existence of the “structural co-evolution” on long time described by BRETAGNOLLE in [Offner et al., 2014]. What [Bretagnolle, 2003] obtains is a simultaneous correspondence between growth rate and level of connectivity to the network (and not with network dynamic), but not in our sense a co-evolution, since no statistical relation is furthermore exhibited.

We rejoin the recent results of [Mimeur et al., 2017] that show the statistical non-significance of the correlation between growth rate and evolution of network coverage and accessibility, at a zero delay. Our results are less precise on the class of cities studied (they differentiate

¹⁵ As [Lemoy and Caruso, 2017] is not able to reproduce, for density profiles as a function of the distance to the center of European metropolis, the transition that allows [Guérois and Pumain, 2008] to define the peri-urban. These more or less recent works are not reproducible, producing neither code nor data, and giving only a superficial description of the methods, and it is thus impossible to know the origin of the qualitative divergence obtained. A good reproducibility together with the construction of systematic comparisons (*benchmarks*) of models, empirical analysis, that are recent but also to validate old studies, seems to be a reasonable solution to this kind of issue.

large and small cities, and work on a larger panel), but more general as they study variable delays and accessibility ranges, and are thus complementary.

Calibration of the abstract model

Expected results of the calibration on real data concern both the more or less accurate reproduction of real city population growth dynamics, i.e. to what extent the inclusion of a dynamical network can increase the explanatory power for trajectories, and also how realistic the evolution of network distance is. We still work with the abstract model.

MODEL EVALUATION We can add to the indicators used before a calibration indicator for distance. The particular property of adjustment for populations, that resides in the existence of a power law for the sizes of cities that made negligible the performance on medium and small cities in the case of a cumulated error, and suggested the addition of the indicator on the error on logarithms, is not present for distances that follow a distribution concentrated on a single order of magnitude. We use therefore a standard measure of fit, given by

$$\varepsilon_D = \log \left[\sum_t \sum_{i,j} (d_{ij}(t) - \tilde{d}_{ij}(t))^2 \right]$$

where $d_{ij}(t)$ are observed distances and $\tilde{d}_{ij}(t)$ the simulated distances. It is simply a cumulated squared-error, as used for the comparison of origin-destination matrices in a similar case of simulation of a transportation network in [Jacobs-Crisioni and Koopmans, 2016].

RESULTS We proceed to a non-stationary calibration, on the $(\varepsilon_P, \varepsilon_D)$ objectives, i.e. the squared-error on populations and on distances. The estimation is done with a moving window with the periods already used in 4.3. In order to have a limited dimension to explore, we take a fixed $w_N = 0$ to study the interactions only at the first order, knowing that the abstract network parameters $(g_{max}, \gamma_S, \varphi_0)$ are taken into account in the calibration. The calibration is done with a genetic algorithm in a way similar as in 4.3. The Fig. 101 shows the obtained Pareto fronts, and the Fig. 51 the evolution in time of parameter values for the optimal solutions.

We observe a large variability of the shape of Pareto fronts for the bi-objective calibration on population and distance, what witnesses more or less difficulty to simultaneously adjust population and distance. Some periods, such as 1891-1911 and 1921-1936, are close to have a simultaneous objective point for the two objectives, what would correspond to a good correspondence of the model to

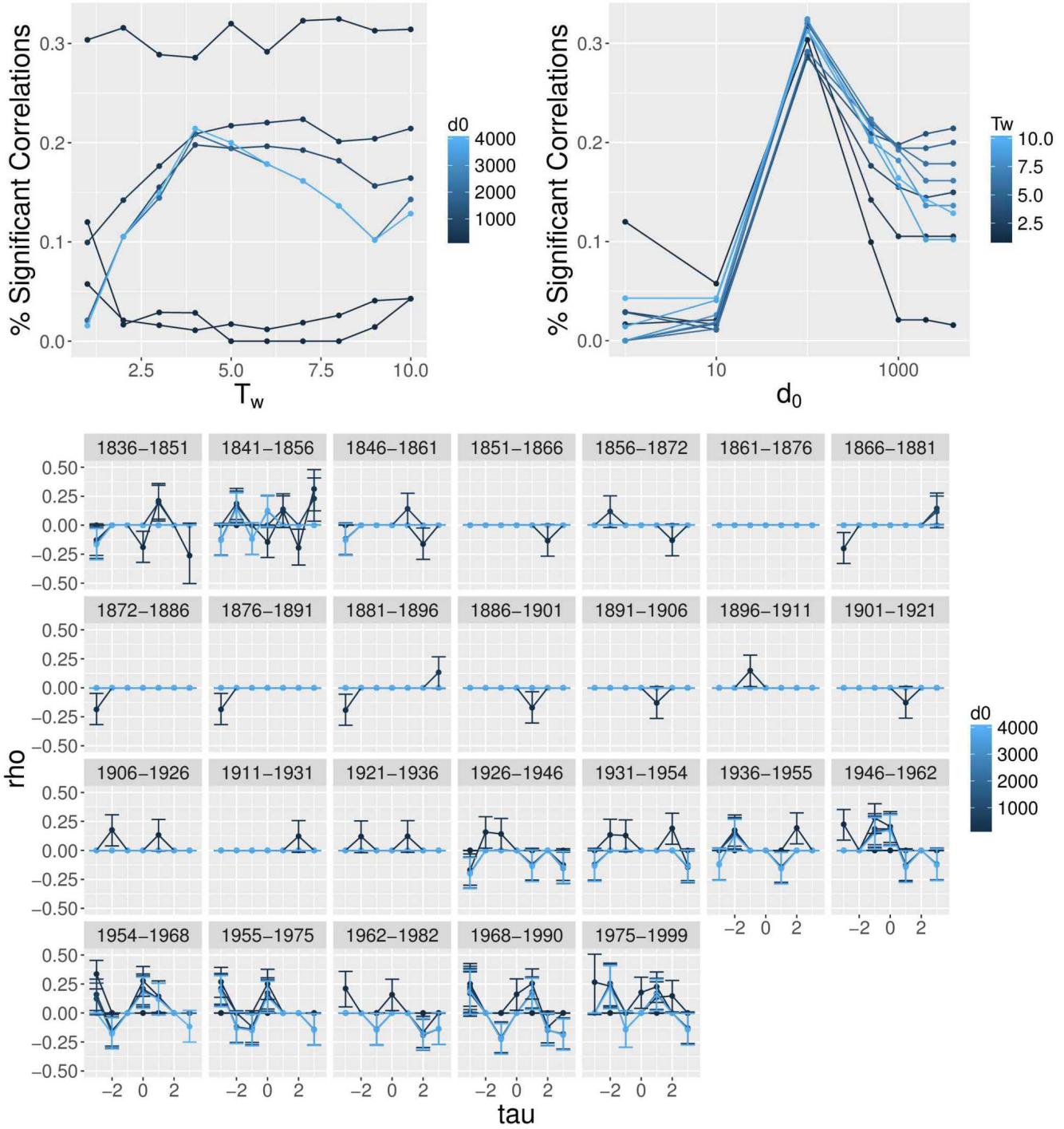


Figure 49: **Empirical lagged correlations for the French system of cities.** Correlations are estimated on a window of duration $5 \cdot T_w$, between population growth rates and the variations of closeness centrality with a decay parameter d_0 (see text). (*Top left*) Number of significant correlations (taken such that $p < 0.1$ at 95%) as a function of T_w for d_0 variable; (*Top right*) Number of significant correlations as a function of d_0 for T_w variable; (*Bottom*) For the “optimal” window $T_w = 4$, value of ρ_τ as a function of τ , for all successive periods.

both trajectories of cities and trajectory of the network on these periods.

In comparison with calibration results of the model with static network of 4.3, when comparing the performances for the objective ε_G , we find periods where the static is clearly better (1831 and 1841 for example) and others where the co-evolutive model is better (1946 and 1962): thus, taking into account the co-evolution helps in some cases to have a better reproduction of population trajectories.

The values of optimal parameters in time, shown in Fig. 51, seem to contain some signal. The evolution of w_G and γ_G are coherent with the evolutions observed for the static model. For d_G , the model principally saturates on the maximal distance and the evolution is difficult to interpret.

However, the evolution of ϕ_0 could be a sign of a “TGV effect” in recent periods, through the secondary peak for population after 1960. Indeed, the construction of high speed lines has shortened distances between cities on top of the hierarchy, and an increase of the threshold ϕ_0 corresponds to an increase of the selectivity for a potential diminution of distances.

The calibrated g_{\max} can finally be interpreted according to the history of the railway network (at least of all points in the Pareto front): a significant secondary peak in the first years, a minimum in the years corresponding to the stabilization of the network (1900), and an increase until today linked to the increase of train speeds and the opening of high speed lines.

We have this way in a certain extent indirectly quantify interaction processes through the network and the processes of network adaptation to flows, in the case of a real system.

Model with a physical network

We now sketch the outline of a specification of the model with a physical network, what would in a sense correspond to an hybrid model combining different scales. The objective of such a specification would be on the one hand to study the difference in trajectories compared to the abstract network, i.e. to quantify the importance of economies of scale (due to common links), of congestion and also the possible compromises to take in order to spatialize the network. On the other hand, it would help to understand to what extent it is possible to produce realistic networks in comparison to autonomous network growth models for example. These issues are tackled at an other scale and for other ontological specifications in chapter 7.

Such a specification follows the frame of [Li, Lu, and Tian, 2014], which model the co-evolution between transportation corridors and the growth of main poles at a regional scale.

The physical network we implement aims at satisfying a greedy criteria of local time gain. More precisely, we assume a self-reinforcement

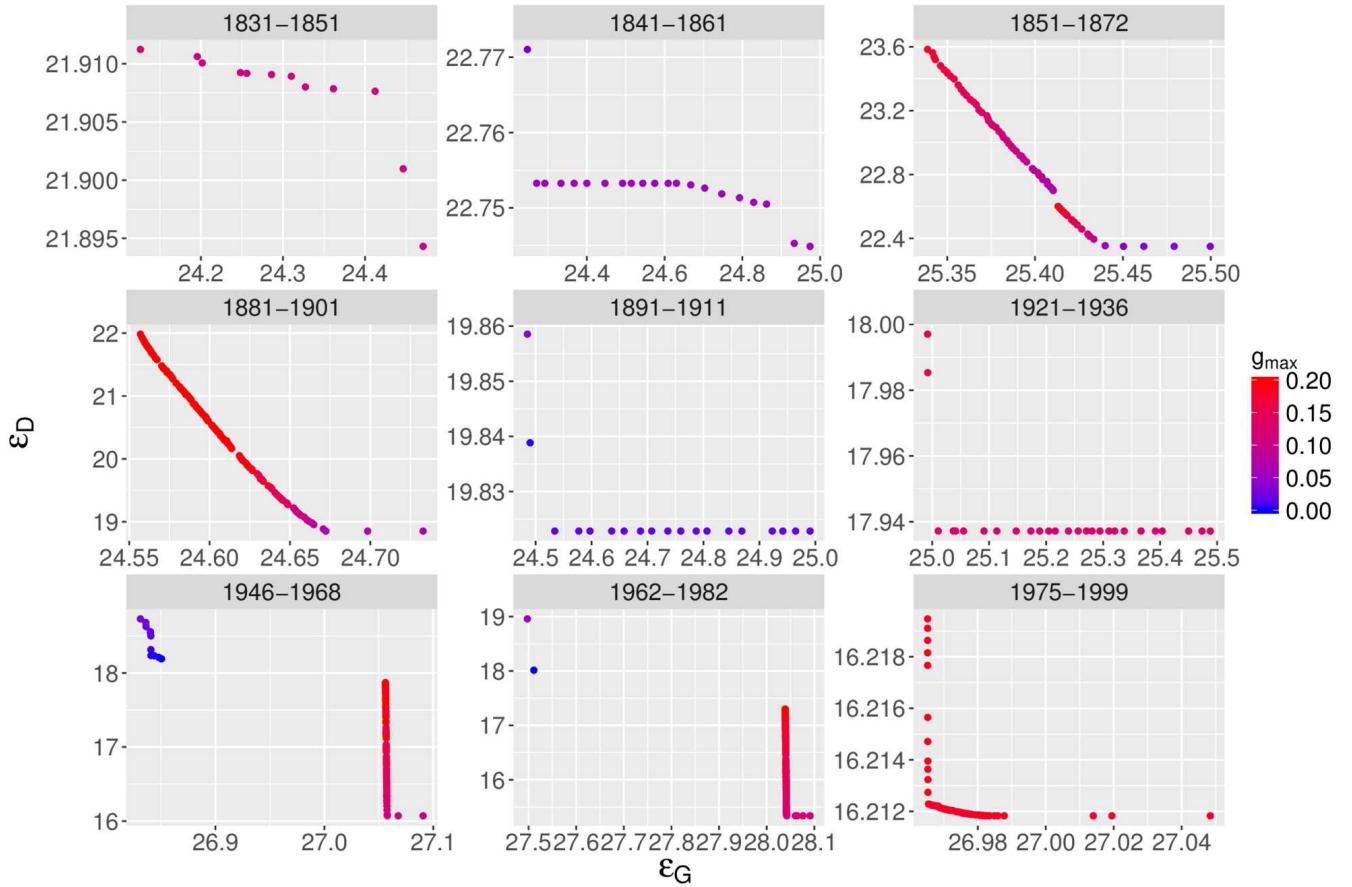


Figure 50: **Pareto fronts for the bi-objective calibration between population and distance.** Fronts are given for each calibration period and are colored according to g_{\max} .

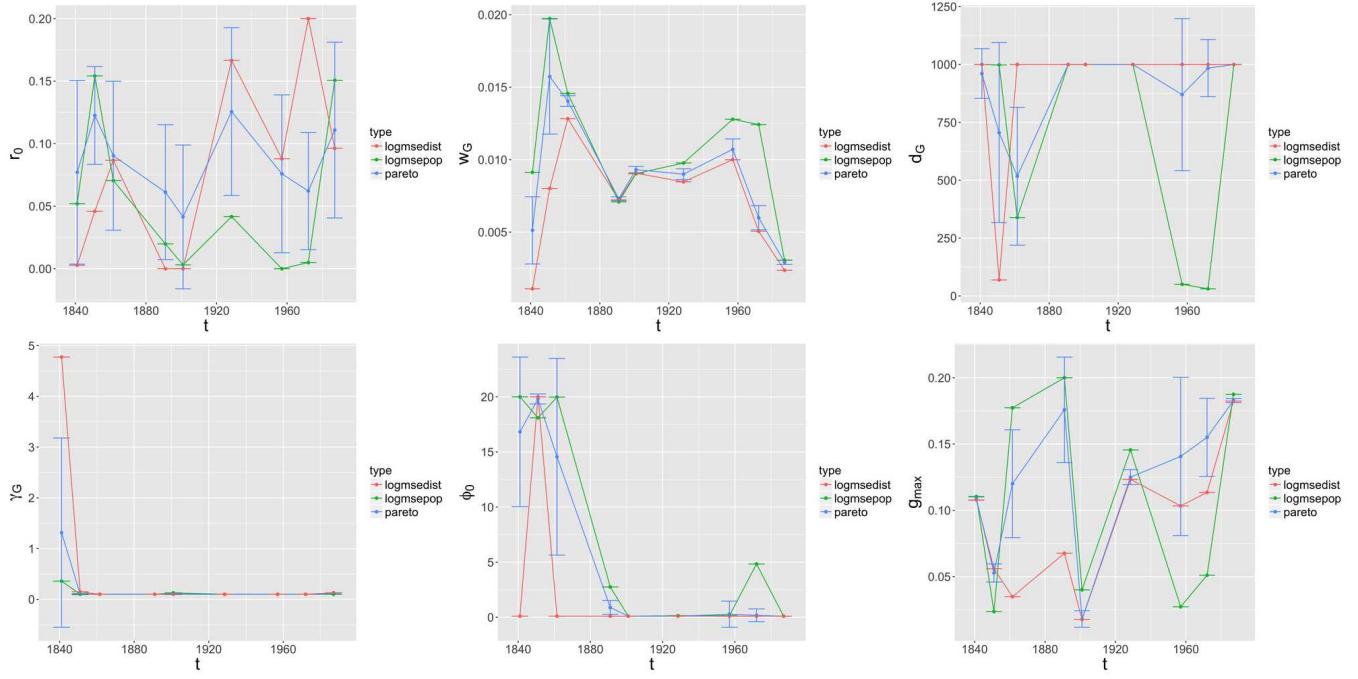


Figure 51: **Temporal evolution of optimal parameters.** From left to right and top to bottom, values of parameters ($r_0, w_G, d_G, \gamma_G, \phi_0, g_{\max}$), respectively for the full Pareto front (blue), for the optimal point in the sense of the distance (red) and the optimal point in the sense of the population (green).

similar to [Tero et al., 2010] A specification analog to the one used before assumes a growth for each link, given also in a logic of self-reinforcement by:

$$d(t+1) = d(t) \cdot \left(1 + g_{\max} \cdot \left[\frac{\phi}{\max \phi}\right]^{\gamma_s}\right)$$

if ϕ is the flow in the link and $d(t)$ its effective distance. The threshold specification used before does indeed not allow a good convergence in time, in particular with the emergence of local oscillation phenomena.

We generate a random initial network, by perturbing the position of vertices of a grid for which a fixed proportion of links has been removed (40%) and by linking cities to the network through the shortest path. Links have all the same impedance, which then evolves according to the equation above. An example of a configuration obtained with this specification is given in Fig. 52. The good convergence properties (visual stabilization of network structure during restricted experiments) suggest the potentialities offered by this specification, which systematic exploration is out of the scope of this work.

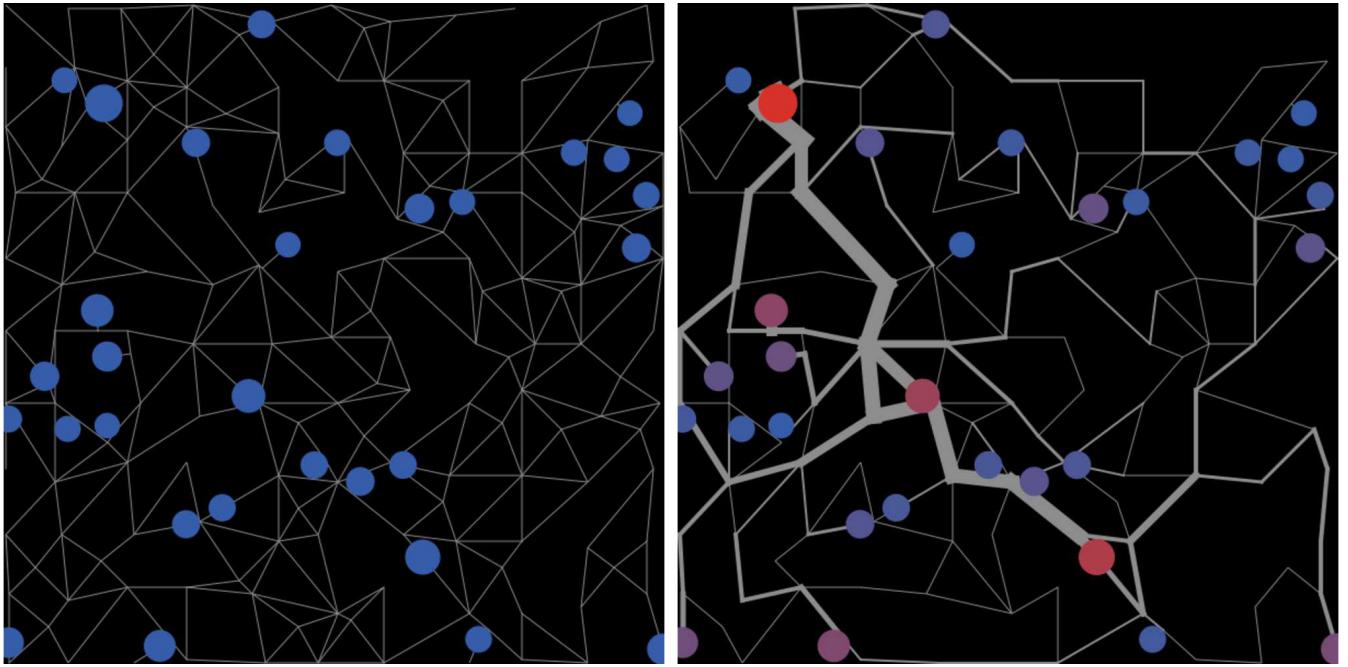


Figure 52: **Example of configuration obtained with a self-reinforcing network.** (Left) Initial random configuration, with uniform impedances; (Right) Final configuration obtained after 100 iterations.

Perspectives

PARTICULAR TRAJECTORIES The study of particular trajectories within a system of cities can allow to answer to specific thematic questions: for example, the influence of medium-sized cities on the global trajectory of the system, or the drivers of a more or less “successful” trajectory for this type of profile. In the case of the application to a real system, the mapping of deviation to the model in time can suggest regional particularities.

COMPARISON OF URBAN SYSTEMS We also finally expect to be able through the model to compare urban systems in different geographical and political contexts, and at different scales. This should foster the understanding the implications of planning actions on the interactions between networks and territories. For example, French railway network has emerged through multiple operators, on the contrary to the Chinese high speed railway network, for which a more precise development could be considered.

* * *

*

CHAPTER CONCLUSION

This macroscopic entry into co-evolution processes aimed at understanding them (i) within a system of cities, i.e. in an aggregated way and at an abstract level; and (ii) on a long time scale, of the order of a century. The processes we considered are: growth of city as a consequence of interactions which depend on the network; effect of flows at the second order on these growths (that we did not explore here); effect of feedback of flows on distances in the network in a thresholded way (the latest being refined with an effect of network topology in the case of SimpopNet).

We first show, through a systematic exploration of the SimpopNet model, that it is highly sensitive to the spatial configuration, suggesting that potential conclusions on processes will always have to be contextualized. We also show that it difficultly produces a co-evolution in the sense of circular causalities between network and cities, and that the dominating process is more an adaptation of cities to the network.

Our model we then explore allows on the other hand, at the price of an abstraction of the network, to reveal in a synthetic way first an intermediate scale of maximal complexity suggesting the emergence of regional subsystems, allowed by intermediate values of the interaction distance and high values of the feedback threshold for the network; secondly the existence of at least three regimes of causality, among which at least two can be qualified as co-evolutive. The study of real data for the French system of cities indeed confirms the existence of the regional scale, and also a short stationarity time scale of around twenty years, but very few significant interactions at this scale, in contradiction with the existing literature. The calibration of the model on real data reproduces well the known patterns of railway network growth, and suggest more recently a “TGV effect”.

We introduce a development with physical network, which allows to make the link with ontologies we will explore in the following in chapter 7: the co-evolution at a mesoscopic scale, by insisting on the role of form and function, and thus of precise mechanisms of network development.

★ ★

★

7

CO-EVOLUTION AT THE MESOSCOPIC SCALE

Processes underlying co-evolution are not exactly similar when switching from the macroscopic scale to the mesoscopic scale, as suggest our different empirical analysis: for example, causality regimes obtained at a small scale for South Africa in 4.2 are clearer than the ones for real estate transactions and the Grand Paris in 1.2. At the metropolitan scale, relocation processes are crucial to explain the evolution of the urban form, and these can partly be attributed to accessibility differentials, knowing that the evolution of networks answers on the other hand to complex logics conditioned by territorial distributions. Centrality, density, accessibility, as much properties potentially implied in co-evolutive processes, and that are proper to the concept of urban form.

We make the choice to insist on the role of the urban form at the mesoscopic scale, and use urban morphogenesis as a modeling paradigm for co-evolution: the strong coupling of the urban form with the network through co-evolution allows to consider urban functions more explicitly. This chapter follows the chapter 5, and extends the model that have been developed in it.

Different network generation heuristics are compared in a first section 7.1, still in a weak coupling paradigm, in order to establish the topologies produced by different rules.

This step allows to introduce a co-evolution model through morphogenesis in 7.2, which is calibrated on coupled objectives of urban morphology and network topology.

Finally, we describe in 7.3 a model allowing the exploration of complex processes for network growth, in particular endogenous governance processes implying deciding agents at the metropolitan scale.

★ ★

★

The results of the two first sections of this chapter have been presented at CCS 2017 as [Raimbault, 2017c], and will be published in a synthetic way as a book chapter [Raimbault, 2018a]; the structure of the model and preliminary results for the third section have been presented at ECTQG 2015 as [Le Néchet and Raimbault, 2015].

7.1 NETWORK GROWTH MODELS

We propose first to study with more details processes of network growth for the mesoscopic scale. The idea is to understand intrinsic properties of different network growth heuristics. This exercise is interesting in itself since there is to the best of our knowledge no systematic comparison of spatial networks morphogenesis models: [Xie and Levinson, 2009b] propose for example a review from the point of view of network economics, it does not include on the one hand some disciplines (see chapter 2), and on the other hand does not compare performances of models on dedicated comparable implementations.

7.1.1 Benchmarking network growth heuristics

Considering network growth in itself, several heuristics exist in order to generate a network under some constraints. As already developed especially in 2.1, from economic network growth approaches to local optimization heuristics, geographical mechanisms or biological network growth, each has its own advantages and particularities. We already tested in 5.3 an heuristic based on interaction potential breakdown. In order to be able to compare different network growth heuristics “everything else being equal”, it is necessary to explore them at fixed density, although the thematic meaning of results will not have any value, neither on long times nor for co-evolution.

The importance of heuristics capturing a topological structure allowing a certain compromise between performance, congestion and cost, is shown by empirical analyses such as [Whitney, 2012] for metropolitan networks, which shows that patterns of evolution for correlations between degrees witness an evolution of networks towards such a topology.

We precise in the following the core of the network growth model together with several heuristics from diverse origins, compared in similar conditions through their integration within the common basis.

Core of the network growth model

A common process to the different heuristics constitutes the core of the network growth model, and bridges population density distribution with the network. In concrete terms, the aim is to attribute new centers according to this density, and we make the choice of specifying this process exogenously to network growth itself¹.

¹ This intermediate stage is close in our case to the idea of procedural modeling, since the implemented rule aims at reproducing a shape without needing the actual processes. This raises the issue of equifinality and of the potential existence of equivalent models for this submodel or for the full model capturing a real process corresponding to it. The use of multi-modeling also at this stage could be a solution, but frame-

We recall the context used in 5.3, i.e. a grid of cells characterized by their population P_i , on which a network composed of nodes and links develops. The population distribution will here be fixed in time $P_i(t) = P_i(0)$, and the network evolves sequentially starting from an initial network.

A step of network growth is realized at fixed time intervals t_N (parameter which allows to adjust the respective evolution speeds for population and for the network). It corresponds to the following stages, of which the first two refine the logic of [Raimbault, Banos, and Doursat, 2014] (which stipulates that population centers must be connected to the existing network in a basic way).

1. A fixed number n_N of new nodes is added. Sequentially, the probability to receive a new node is given by

$$p_i = \frac{P_i}{P_{\max}} \cdot \frac{\delta_M - \delta_i}{\delta_M}$$

what means that an elementary node corresponds to the conjunction of events: (i) high density P_i of population in cell compared to the maximal population for each cell P_{\max} , (ii) density of nodes δ_i in a radius r_n low compared to a maximal density δ_M . Population of nodes is reattributed at each stage through triangulation the same way as in 5.3.

2. New nodes are then connected by a new link, following the shortest path to the network (perpendicular connexion or towards the closest node).
3. New links are added, until they reach a maximal number of added links l_m , following an heuristic that varies among: no heuristic (no supplementary links added), random, deterministic potential breakdown (see 5.3), random potential breakdown [Schmitt, 2014], cost-benefits [Louf, Jensen, and Barthelemy, 2013], biological network generation (heuristic based on [Tero et al., 2010]).

We fix to simplify the parameters $r_n = 5$, $\delta_M = 10$ and $n_N = 20$, and the parameters t_N and l_m will be variable.

Baseline heuristics

We consider two baseline heuristics to better situate the ones we will explore in the following: the one composed uniquely by the base described previously, which produces tree networks; and random network generation, which consists in creating a fixed number l_m of new

works allowing to tackle an arbitrary number of stationarity levels or even allowing the model to be autonomous on these choices do not exist yet.

links between randomly chosen nodes, and to make the final network planar².

Euclidian heuristic

This heuristic, which rationale relies on ideas of gravity potential breakdown, corresponds to the method developed in 5.3. It is a method close to the one introduced by [Schmitt, 2014], without the stochastic aspect and prone to miss path-dependency phenomena, but more refined in the mechanisms of gravity potentials.

Random potential breakdown

Random potential breakdown is the heuristic used in the SimpopNet model [Schmitt, 2014], which is inspired by the model introduced by [Blumenfeld-Lieberthal and Portugali, 2010]. At each step, two cities are randomly drawn, the first following a probability proportional to $P_i^{Y_R}$ and the second following $V_{i_0j}^{Y_R}$ such that i_0 is the first city drawn and V_{ij} are euclidian gravity potentials. If

$$d_N(i_0, j_0)/d(i_0, j_0) > \theta_R$$

i.e if the relative detour through the network is larger than a threshold parameter, a link is created between the two cities³. At each time step, l_m new links are created following this process. The final network is made planar.

Biological heuristic

[Raimbault and Gonzalez, May 2015] explores applications of biological network growth models (*slime mould*), in particular their ability to produce from the bottom-up optimal solutions in the Pareto sense for contradictory objectives, such as cost and robustness. The considered model comes from [Tero et al., 2010].

The advantage of such an heuristic is confirmed in some cases by the reality of multi-objective optimizations: [Padeiro, 2009] (p. 72) illustrates in particular the extension of the Parisian metro in Bobigny in the seventies, and the consideration of indicators for cost, served population, expected rush hour traffic, and average travel time.

The *slime mould* model works the following way. Given an initial network with links of uniform capacities, a fluid is distributed in the network from a source to a sink, establishing a flow in each link. An equilibrium of fluid pressures at network nodes can be found, which corresponds to the stationary state for flows⁴. Given an equilibrium

² The algorithm to obtain a planar network consists in the creation of nodes at the possible intersections of new links ("flattening" of the network).

³ To remain comparable to the other heuristics that do not include speeds in links, newly created links are of speed 1 and not v_0 as in the implementation of 6.1.

⁴ More precisely, the problem is equivalent to an electrostatic linear equations system that we just have to solve.

for pressures, capacities of links evolve according to the traversing flow. An iteration of equilibria and of tubes evolution allows then a convergence towards a stable hierarchical distribution of capacities. The detail of the procedure is described in Appendix A.10, following the mathematical details developed by [Tero, Kobayashi, and Nakagaki, 2007].

Our logic is to use this mechanism to determine at a given time a given number of realized links. Advantages of the heuristic we are going to detail are especially that (i) it can be used in an iterative way to represent a sequential topological evolution of the network, in comparison to most investment models that evolve only capacities in time; and (ii) it translates processes of network self-organization, and moreover produces optimal networks in the Pareto sense for cost and robustness.

The application of the slime-mould model to network generation is done according to the following steps, within the global frame described previously.

1. Starting from the existing network to which we add a grid network (with diameters two times smaller to take into account the preponderance of the existing network) with diagonal connexions, and in which 20% of links are randomly deleted to simulate perturbations linked to topology, we constitute the initial support in which slime-mould flows will be simulated.
2. We proceed by iteration of successive generations, which consist in the following steps, for an increasing value of k ($k \in \{1, 2, 4\}$ in practice):
 - given the distribution of population, the slime-mould model is iterated $k \cdot n_b$ times to obtain the emergent network through convergence of capacities;
 - links with a capacity inferior to a threshold parameter θ_b are removed;
 - the largest connected component is kept.
3. The final network is simplified⁵ and made planar.

We illustrate in Fig. 53 two stages of this generation process, showing the basis structure on which the self-reinforcement model is launched, and the convergence of link capacities after a certain number of steps.

⁵ The simplification algorithm consists in the replacement of link sequences which extremities have all a degree of two, excepted the start and end nodes, by a unique link.

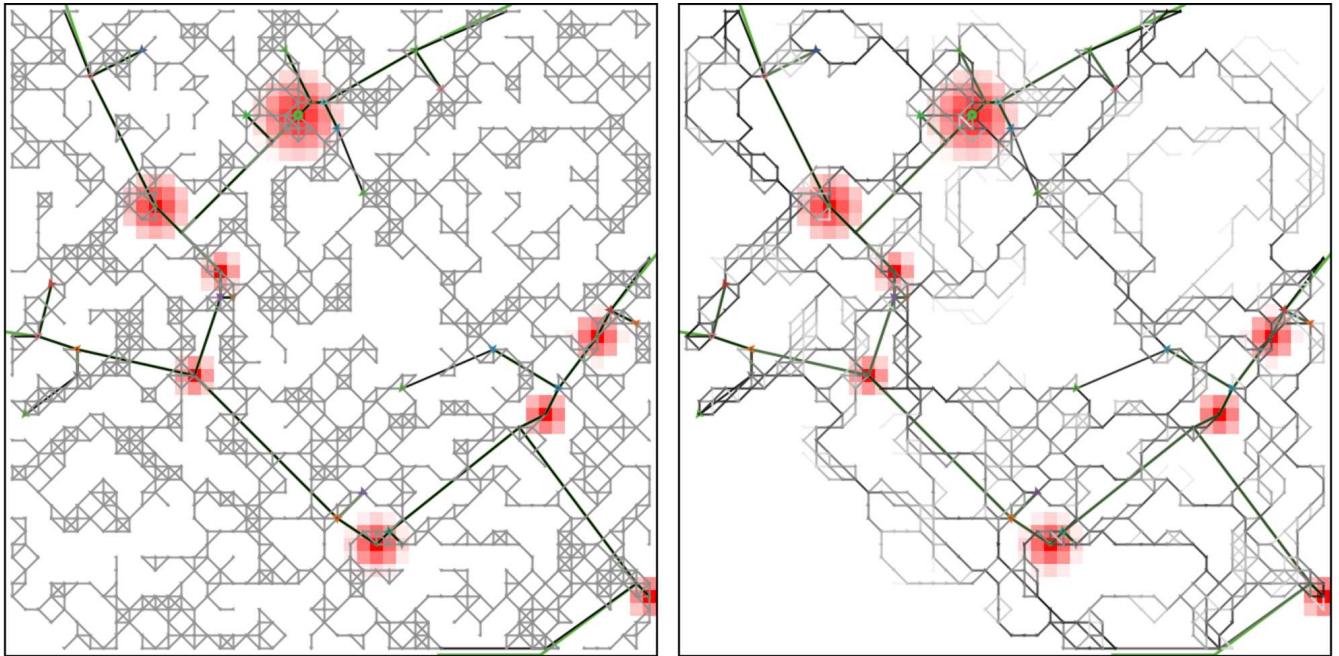


Figure 53: Biological heuristic for network generation. This visualization example illustrates the intermediate stages for the addition of links. (*Left*) The initial semi-random network in which the slime-mould is launched; (*Right*) same network after 80 iterations of the slime-mould, the thickness of links giving the capacity.

Cost-benefits evaluation

The notion of cost is not explicitly included in all the growth heuristics presented up to here - it is implicitly in gravity potentials through the distance decay parameter, and also in the slime-mould since it generates networks exhibiting a compromise between cost and robustness. We therefore add a simple heuristic which is focused on the cost of network links during their extension. It is the heuristic studied by [Louf, Jensen, and Barthélémy, 2013], which relies on a rationale in transportation economics. Following a logic of cost-benefits analysis by network developments actors, links are sequentially realized for the couple of non-connected cities with a minimal cost, with a cost of the form $d_{ij} - \lambda/V_{ij}$, where the parameter λ is the compromise between construction cost and gain in connected potential.

Parameters

We summarize the parameters that will vary in the following in Table 16. An additional “parameter”, or more precisely a meta-parameter, is the choice of the heuristic to add links.

Table 16: **Summary of network growth parameters for all heuristics.** We also give the corresponding processes, typical variation ranges and their default values.

Heuristic	Parameter	Name	Process	Domain	Default
Base	l_m	added links	growth	[0; 100]	10
	d_G	gravity distance	potential]0; 5000]	500
	d_0	gravity shape	potential]0; 10]	2
	k_h	gravity weight	potential	[0; 1]	0.5
	γ_G	gravity hierarchy	potential	[0.1; 4]	1.5
Random breakdown	γ_R	random selection hierarchy	hierarchy	[0.1; 4]	1.5
	θ_R	random threshold	breakdown	[1; 5]	2
Cost-benefits	λ	compromise	compromise	[0; 0.1]	0.05
Biological	n_b	iterations	convergence	[40; 100]	50
	θ_b	biological threshold	threshold	[0.1; 1.0]	0.5

7.1.2 Results

Model setup

The model is initialized on synthetic or semi-synthetic configurations, with a grid of size $N = 50$, with the following steps.

1. Population density is initialized either with an exponential mixture, which centers (network nodes) follow the configuration of a synthetic city system as done in 6.1; or from a real configuration extracted from the density raster for France. We will use the second option here in systematic explorations.
2. In the second case, a fixed number of network nodes are generated and located following a preferential attachment to density (see 5.3)⁶. We do not initialize on real networks, since these will be the calibration target, but impose an initial synthetic skeleton that can be interpreted as an archaic network.
3. An initial network is generated by connecting the nodes as detailed in 5.3.

Generated networks

A visual illustration of the different generated topologies is given in Fig. 54 for a synthetic density configuration. This allows us to compare the particularities of each heuristic. For example, links formed through random breakdown compared to deterministic breakdown

⁶ To avoid bord effects of a network with no connection to the exterior, we add a fixed number n_e of nodes (that we take as $n_e = 6$) at random locations on the border of the world.

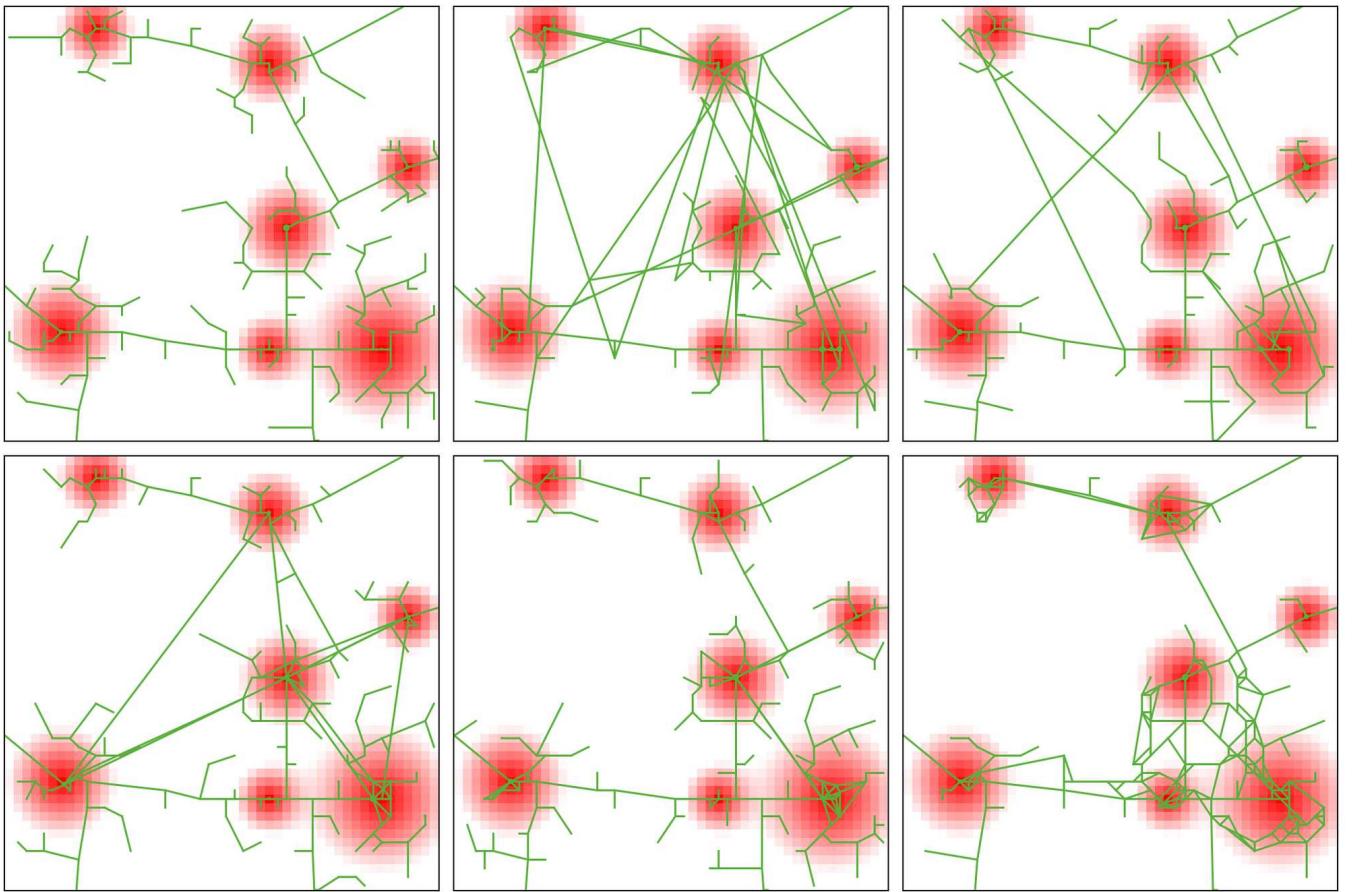


Figure 54: Examples of networks obtained with the different heuristics. Networks are obtained for the same density configuration composed of 7 centers, and for the same initial network connecting them. We take $l_m = 10$ and fix the final size to 200 nodes. Gravity parameters are $d_G = 2000$, $d_0 = 3$, $\gamma_G = 0.3$, $k_h = 0.6$. In the order from left to right and top to bottom: network with connexion only; random network; random potential breakdown with $\gamma_R = 2$ and $\theta_R = 1.6$; deterministic potential breakdown; cost-benefits with $\lambda = 0.009$; biological with $n_b = 50$ and $\theta_b = 0.6$.

witness the path-dependency and produce a less redundant network, whereas deterministic breakdown reinforces the strongest link between the two large cities that are close. The cost-based heuristic gives network that are dense in a very localized way, but avoids too long links. Finally, the biological heuristic produces a dense mesh in the sub-region where interactions are the strongest.

Experience plan

We detail now an experience plan to explore the space of networks generated by the different heuristics. Network generation is done with constant population densities, on real configurations that have been morphologically classified in 4.1. We consider 50 real density grids, corresponding to areas in France, classified into 5 morphological classes. Their description is given in Appendix A.10, and show

that they cover a set of morphologies spanning from very localized and sparse settlements to polycentric structures, and intermediate configurations.

Given the parameter ranges previously given for each heuristic, we compare the feasible space for a basic exploration with a Latin Hypercube Sampling of parameter space, for all density grids, with 5 repetitions for each parameter point⁷.

Obtained topologies

Networks are characterized here with the following indicators: average betweenness centrality \bar{bw} and average closeness centrality \bar{cl} , diameter r , average path length \bar{l} , relative speed v_0 . To visualize feasible spaces and then compare them to real networks, we reduce the space in a principal hyperplane, from points obtained in simulations. The first two components can be interpreted the following way⁸: the first will characterize networks in which paths are shorter, whereas the second corresponds to networks with a higher average distance, thus more spread in space, but more efficient.

The point cloud of the topological feasible space, obtained with the experience plan described above, is given in Fig. 55. The coverage is allowed by the complementarity of different clouds for each heuristic. For example, the random heuristic is at the total opposite of the reference heuristic along the first component: the reference tree network logically induces a larger number of detours, and thus longer paths. Random breakdown allows to cover a large span of PC1 and corresponds more to low values of PC2.

To better understand the complementarity of approaches, we can quantify the intersection of point clouds in Fig. 55 with a simple method: by dividing the plane into a grid (that we take of size 20x20), the proportions p_{ij} of points for each heuristic j for each cell i can be aggregated into a concentration index $h_i = \sum_j p_{ij}^2$ (Herfindhal index) which distribution describes the balance between heuristics in the different regions of space. We obtain for cells a first quartile at 0.54, a median at 0.76 and a third quartile at 1. For comparison, in the case of two types of points only, a repartition 65-35% gives an index of 0.55 and a repartition 85-15% an index of 0.75, what means that at least half of cells have more than three quarters of points in a unique category. This confirms the conclusion of a strong complementarity of heuristics.

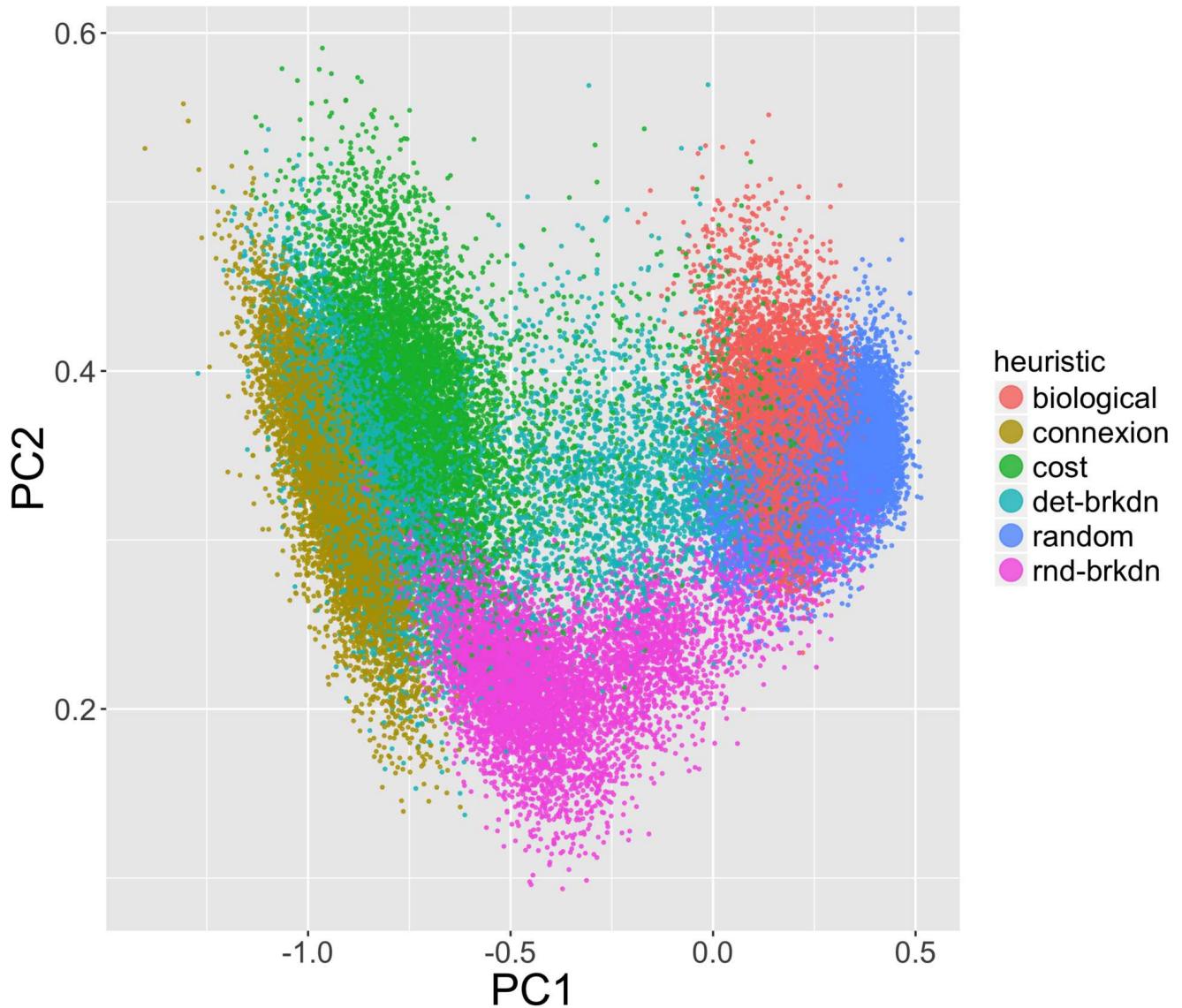


Figure 55: **Feasible topological space for the different generation heuristics.** Point clouds cover complementary regions of the topological space, the color giving the heuristic: biological (biological), reference (connexion), cost-benefits (cost), deterministic breakdown (det-brkdn), random (random) and random breakdown (rnd-brkdn). The same figure conditioned to the morphological class for density is given in Appendix A.10.

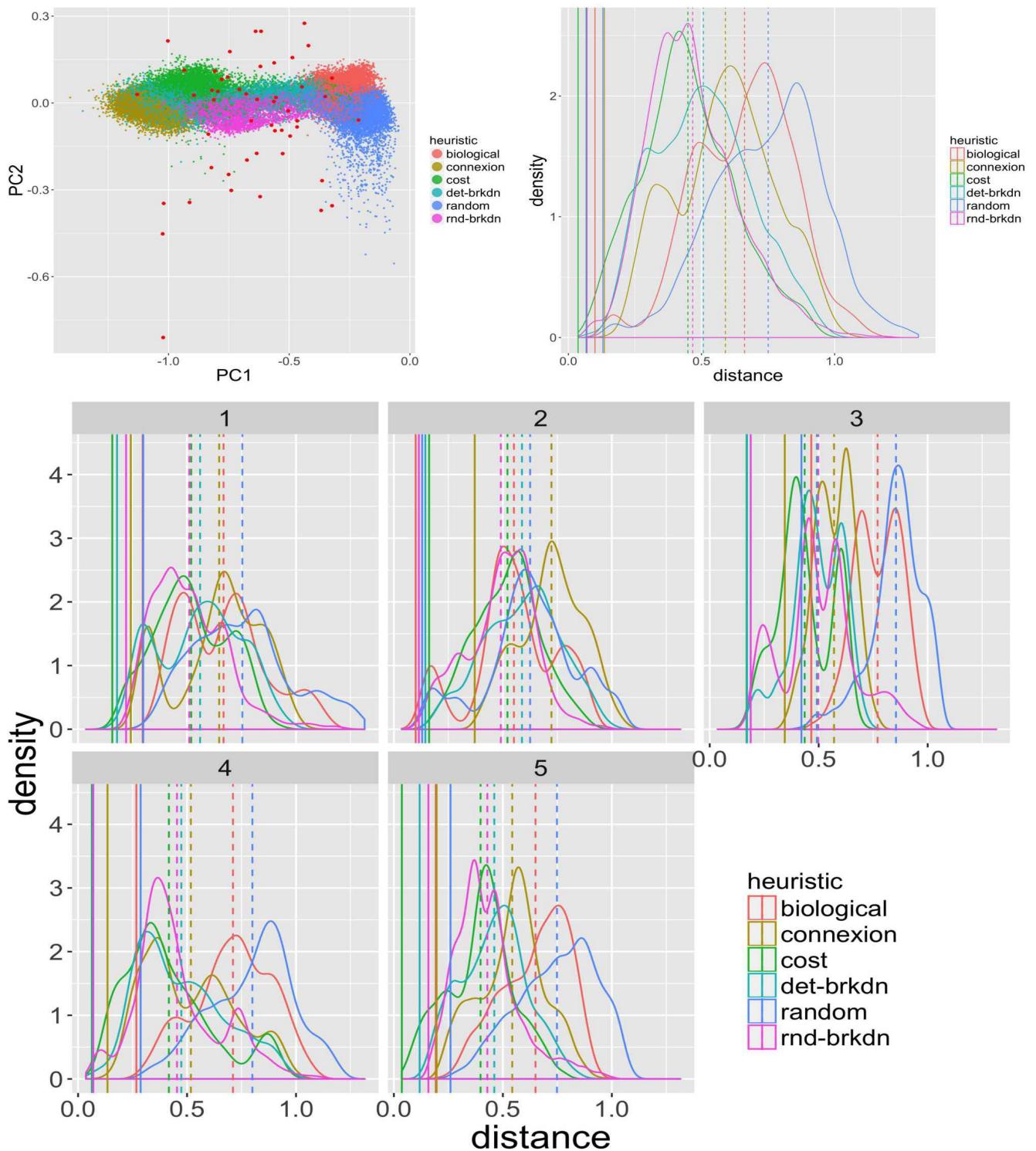


Figure 56: **Comparison to real networks.** (Top Left) Point clouds for simulated configurations (color in the legend) and for real configurations (in red), in a principal plan such that $PC1 = 0.12bw - 0.09cl + 0.98l$ and $PC2 = -0.20bw - 0.97cl - 0.06l$. (Top right) Distribution of distances d_{\min} for all simulated points, for each heuristic (color). Dashed vertical lines give the average and solid lines the minimum for each distribution. (Bottom) Same histograms, conditioned by morphological class for density distribution.

Comparison to real networks

We use the measures on real road networks obtained in 4.1 to compute a distance of generated configurations to observed configurations, by considering real networks corresponding to density configurations used for initialization. We take for a given parameter point the minimum of the euclidian distance on vectors of indicators for all real points⁹. This comparison is made possible since indicators are normalized, and indicators on real networks are comparable to indicators on synthetic networks.

Comparison results to real points are given in Fig. 56. We give a representation as a point cloud and histograms for distributions of distances, on all grids and by morphological class. We observe that around ten real configuration (one fifth) fall far outside the point cloud. Once again, heuristics are complementary to approach a larger number of points. Concerning distances, the random heuristic is the worse in terms of mode and average, followed by the biological, the reference (connexion only), the deterministic breakdown and finally the random-breakdown and the cost which are approximatively equivalent. All realize very low minimal distances.

When conditioning by morphological classes, we see that classes 3, 4 and 5 give the most difficulties for all heuristics in terms of minima - they are indeed the configurations with very localized settlements or a diffuse population (see A.10): it is therefore easier to reproduce real network configurations in the case of polycentric structures. In all cases, the biological heuristic is not very efficient, but it is not directly possible to know if this is a consequence of its under-exploitation and its fixed parameters, or of its intrinsic dynamics.

7.1.3 Discussion

If the slime-mould model is able to generate robust networks in a simplified way, its use for planning has been questioned, in particular because it does not take into account external factors and the urban environment [Adamatzky and Jones, 2010]. Our results seem to confirm these analyses, since this heuristic is the least performing in terms of distance to real networks.

We have thus explored and compared different network generation heuristics, at a fixed density. We note the following points.

⁷ What corresponds to around 240000 repetitions of the model. The simulation data is available at <http://dx.doi.org/10.7910/DVN/0BQ4CS>.

⁸ Their composition is given by: $PC1 = -0.51\bar{bw} - 0.45\bar{l} + 0.57v_0 - 0.43r + 0.05\bar{cl}$ and $PC2 = -0.45\bar{bw} + 0.17\bar{l} + 0.33v_0 + 0.8r + 0.1\bar{cl}$.

⁹ What means that if $d(1,2) = \sqrt{(\bar{bw}_1 - \bar{bw}_2)^2 + (\bar{cl}_1 - \bar{cl}_2)^2 + (\bar{l}_1 - \bar{l}_2)^2}$, we consider $d_{min} = \min_j d(S, R_j)$ if S is the simulated point and R_j the set of real points. We keep here only the indicators \bar{bw} , \bar{cl} and \bar{l} , for normalization reasons.

- Different models produce networks that appear as complementary in an indicator space.
- Similarly, they are complementary to resemble configurations of real networks, while showing different performances. Very localized or diffuse density configurations correspond to networks that are more difficult to reproduce, in comparison to polycentric structures.

★ ★

★

Armed with these network growth models, we will be able to couple them to a density model, in order to develop a co-evolution model at the mesoscopic scale, which will be the subject of the following section.

★ ★

★

7.2 CO-EVOLUTION AT THE MESOSCOPIC SCALE

Urban settlements and transportation networks have been shown to be co-evolving, in the different thematic, empirical and modeling studies of territorial systems developed up to here. As we saw, modeling approaches of such dynamical interactions between networks and territories are poorly developed. We propose in this section to realize a first entry at an intermediate scale, focusing on morphological and functional properties of the territorial system in a stylized way. We introduce a stochastic dynamical model of urban morphogenesis which couples the evolution of population density within grid cells with a growing road network.

7.2.1 *Model description*

General structure

The general principles of the model are the following. With an overall fixed growth rate, new population aggregate preferentially to a local potential, for which parameters control the dependance to various explicative variables. These are in particular local density, distance to the network, centrality measures within the network and generalized accessibility. [Rui and Ban, 2014] shows in the case of Stockholm the very strong correlation between centrality measures in the network and the type of land-use, what confirms the importance to consider centralities as explicative variables for the model at this scale. We generalize thus the morphogenesis model studied in 5.2, with aggregation mechanisms similar to the ones used by [Raimbault, Banos, and Doursat, 2014]. A continuous diffusion of population completes the aggregation to translate repulsion processes generally due to congestion. Because of the different time scales of evolution for the urban environment and for networks, the network grows at fixed time steps, following the submodel developed in 7.1: a first fixed rule ensures connectivity of newly populated patches to the existing network. The different network generation heuristics are then included in the model. We expect the different heuristics to be complementary since for example the gravity model would be more typical of planned top-down network evolution, whereas the biological model will translate bottom-up processes of network growth. The Fig. 57 summarizes the general structure of the morphogenesis model.

Formalization

The model is based on a squared population grid of size N , which cells are defined by populations (P_i). A road network is included in a way similar as in 7.1. We assume at the initial state a given population distribution and a network.

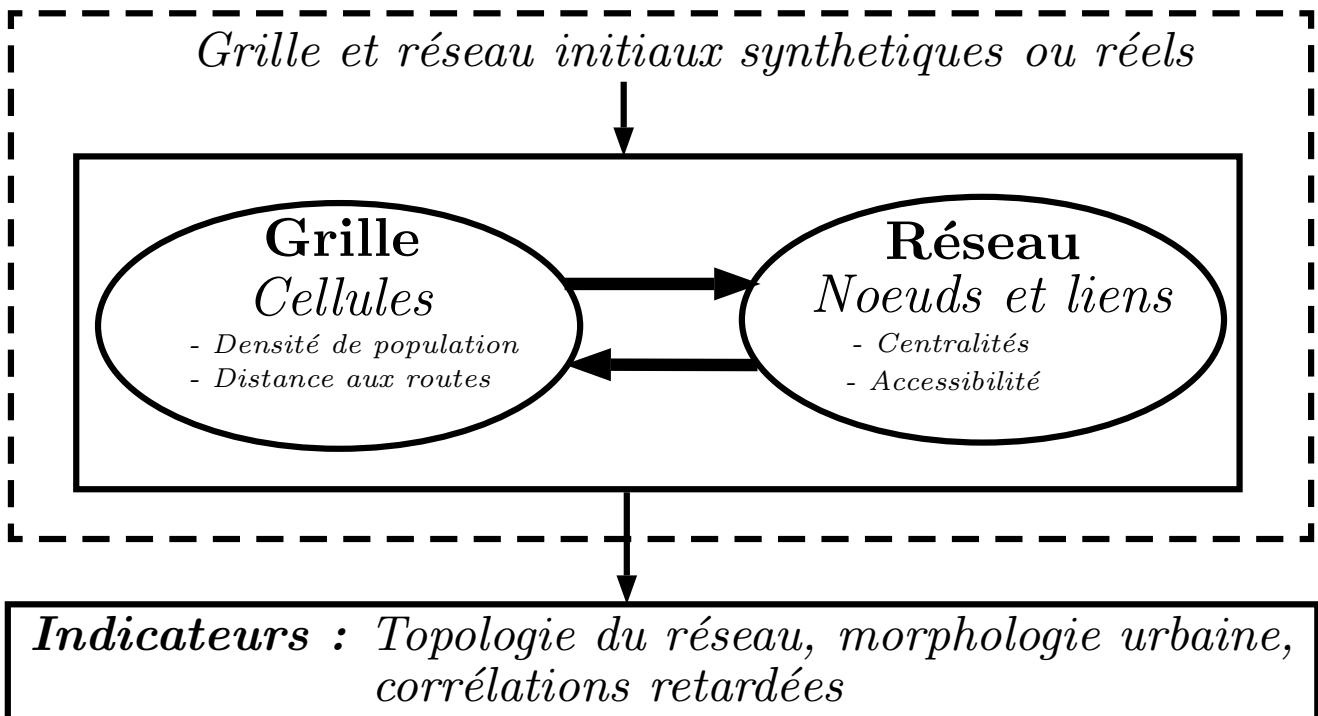


Figure 57: Structure of the co-evolution model at the mesoscopic scale.

The evolution of densities is based on a utility function, influenced by local characteristics of the urban form and function, that we call *explicative variables*. Let $x_k(i)$ a local explicative variable for cell i , which will be among the following variables:

- population P_i ;
- proximity to roads¹⁰;
- betweenness centrality;
- closeness centrality;
- accessibility.

For the last three, they are defined as previously for network nodes, and then associated to cells by taking the value of the closest node, weighted by a decreasing function of the distance to it¹¹. We consider then normalized explicative variables defined by $\tilde{x}_k(i) = x_k(i) - \min_j x_k(j) / (\max_j x_k(j) - \min_j x_k(j))$.

¹⁰ Taken as $\exp(-d/d_n)$ where d is the distance by projection on the closest road, and $d_n = 10$ is fixed.

¹¹ I.e. of the form $x_k = x_k^{(n)}(\operatorname{argmin}_j d(i,j)) \cdot \exp(-\min_j d(i,j)/d_0)$, with $x_k^{(n)}$ the corresponding variable for nodes, the index j being taken on all nodes, and the decay parameter d_0 is in our case fixed at $d_0 = 1$ to keep the property that network variables are essentially significant at close distances from the network.

The utility of a cell is then given by a linear aggregation¹²

$$U_i = \sum_k w_k \cdot \tilde{x}_k(i) \quad (17)$$

where \tilde{x}_k are the normalized local explicative variables, and w_k are weight parameters, which allow to weight between the different influences.

A time step of model evolution includes then the following stages.

1. Evolution of the population following rules similar to the morphogenesis model developed in 5.2. Given an exogenous growth rate N_G , individuals are added independently following an aggregation done with a probability $U_i^\alpha / \sum_k U_k^\alpha$, followed by a diffusion of strength β to neighbor cells, done n_d times.
2. Network growth following the rules described in 7.1, knowing that this takes place is the time step is a multiple of a parameter t_N , which allows to integrate a differential between temporal scales for population growth and for network growth.

The aggregation following a power of the utility yields a flexibility in the underlying optimization problem, since as [Josselin and Ciligot-Travain, 2013] recall, the use of different norms in spatial optimal location problems corresponds to different logics of optimization.

The parameters of the model that we will make vary are then:

- aggregation-diffusion parameters α, β, N_g, n_d , summarized in Table 14;
- the four weight parameters w_k for the explicative variables, which vary in $[0; 1]$;
- network growth parameters for the different heuristics, summarized in Table 16.

Output model indicators are the urban morphology indicators, topological network indicators, and lagged correlations between the different explicative variables.

7.2.2 Results

Implementation

The model is implemented in NetLogo, given the heterogeneity of aspects that have to be taken into account, and this language being particularly suitable to couple a grid of cells with a network. Urban morphology indicators are computed thanks to a NetLogo extension specially developed (see Appendix E).

¹² An alternative could be for example a Cobb-Douglas function, which is equivalent to a linear aggregation on the logarithms of variables.

Experience plan

We propose to focus on the ability of the model to capture relations between networks and territories, and more particularly the co-evolution. Therefore, we will try to establish if (i) the model is able to reproduce, beyond the form indicators, the static correlation matrices computed in 4.1; and (ii) the model produces a variety of dynamical relations in the sense of causality regimes developed in 4.2.

The model is initialized on fully synthetic configurations, with a grid of size 50. Configurations are generated through an exponential mixture in a way similar to [Anas, Arnott, and Small, 1998]: $N_c = 8$ centers are randomly located, to which a population is attributed following a scaling law $P_i = P_0 \cdot (i+1)^{-\alpha_s}$ with $\alpha_s = 0.8$ and $P_0 = 200$. The population of each center is distributed to all cells with an exponential kernel of shape $d(r) = P_{\max} \exp(-r/r_0)$ where the parameter r_0 is determined to fix the population at P_i , with $P_{\max} = 20$ (density at the center)¹³. The initial network skeleton is generated as detailed in 7.1.

We explore a Latin Hypercube Sampling of the parameter space, with 10 repetitions for around 7000 parameter points, corresponding to a total of around 70000 model repetitions¹⁴, realized on a computation grid by using OpenMole.

Static and dynamical calibration

The model is calibrated at the first order, on indicators for the urban form and network measures, and at the second order on correlations between these. Real data used are still the same as introduced in 4.1, which as we recall it are based on Eurostat population grid and the road network from OpenStreetMap. We use here the full set of points from Europe.

We introduce an *ad hoc* calibration procedure in order to take into account the first two moments, that we detail below. More elaborated procedures are used for example in economics, such as [Watson, 1993] which uses the noise of the difference between two variables to obtain the same covariance structure for the two corresponding models, or in finance, such as [Frey, McNeil, and Nyfeler, 2001] which define a notion on equivalence between latent variables models which incorporates the equality of the interdependence structure between variables. We avoid here to add supplementary models, and consider simply a distance on correlation matrices. The procedure is the following.

- Simulated points are the ones obtained through the sampling, with average values on repetitions.

¹³ We have indeed $P_i = \iint d(r) = \int_{\theta=0}^{2\pi} \int_{r=0}^{\infty} d(r) r dr d\theta = 2\pi P_{\max} \int_r r \cdot \exp(-r/r_0) = 2\pi P_{\max} r_0^2$, and therefore $r_0 = \sqrt{\frac{P_i}{2\pi P_{\max}}}$.

¹⁴ For which simulation results are also available at <http://dx.doi.org/10.7910/DVN/0BQ4CS>.

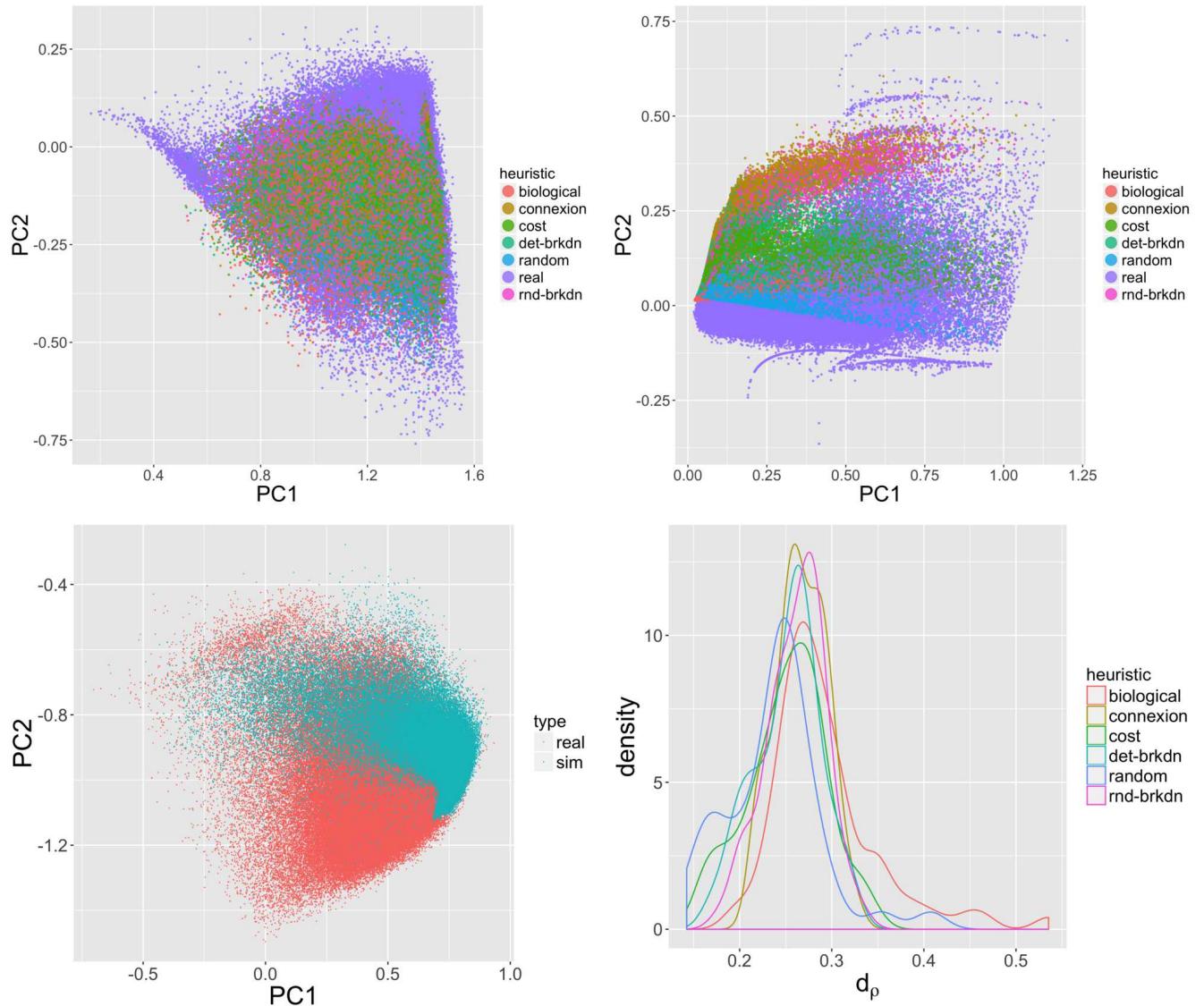


Figure 58: **Calibration of the morphogenesis model at the first and second order.** (Top Left) Simulated and observed point clouds in a principal plan for urban morphology indicators. (Top Right) Simulated and observed could points in a principal plan for network indicators. (Bottom Left) Simulated and observed point clouds in a principal plan for all indicators. (Bottom Right) Distributions of distances on correlations d_ρ , for the different heuristics.

- In order to be able to estimate correlation matrices between indicators for simulated data, we make the assumption that second moments are continuous as a function of model parameters, and split for each heuristic the parameter space into areas to group parameter points¹⁵, what allows to estimate for each group indicators and the correlation matrix.
- For each estimation done this way, that we write \bar{S} (indicators) and $\rho[S]$ (correlations), we can then compute the distance to real points on indicators $d_I(R_j) = d(\bar{S}, R_j)$ and on correlation matrices $d_\rho(R_j) = d(\rho[S], \rho[R_j])$ where R_j are the real points with their corresponding correlations¹⁶, and d an euclidian distance normalized by the number of components.
- We consider then the aggregated distance defined as $d_A^2(R_j) = d_I^2(R_j) + d_\rho^2(R_j)$. Indeed, as developed empirically and analytically in Appendix A.11, the shape of Pareto fronts for the two distances considered suggests the relevance of this aggregation. The real point closest to a simulated point is then the one in the sense of this distance.

The Fig. 58 summarizes calibration results. Morphological indicators are easier to approach than network indicators, for which a part of the simulated clouds does not superpose with observed points. We find again a certain complementarity between network heuristics. When considering the full set of indicators, few simulated points are situated far from the observed points, but a significant proportion of these is beyond the reach of simulation. Thus, the simultaneous capture of morphology and topology is obtained at the price of less precision.

We however obtain a good reproduction of correlation matrices as shown in Fig. 58 (histogram for d_ρ , bottom right). The worse heuristic for correlations is the biological one in terms of maximum, whereas the random produces rather good results: this could be due for example to the reproduction of very low correlations, which accompany a structure effect due to the initial addition of nodes which imposes already a certain correlation. On the contrary, the biological heuristic introduces supplementary processes which can possibly be beneficial to the network in terms of independence (or following the opposed viewpoint be detrimental in terms of correlations). In any case, this application shows that our model is able to resemble real configurations both for indicators and their correlations.

¹⁵ Each parameter being binned into $15/k$ equal segments, where k is the number of parameters: we empirically observed that this allowed to always have a minimal number of points in each area.

¹⁶ That are estimated in 4.1 as we recall, with a square window centered around the point, that we take here for $\delta = 4$.

Causality regimes

We furthermore study dynamical lagged correlations between the variations of the different explicative variables for cells (population, distance to the network, closeness centrality, betweenness centrality, accessibility). We apply the method of causality regimes introduced in 4.2. The Fig. 59 summarizes the results obtained with the application of this method on simulation results of the co-evolution model. The number of classes inducing a transition is smaller than for the RDB model, translating a smaller degree of freedom, and we fix in that case $k = 4$. Centroid profiles allow to understand the ability of the model to more or less capture a co-evolution.

The regimes obtained appear to be less diverse than the ones obtained in 4.2 or for the macroscopic co-evolution in 6.2. Some variables have naturally a strong simultaneous correlation, spurious from their definitions, such as closeness centrality and accessibility, or the distance to the road and the closeness centrality. For all regimes, population significantly determines the accessibility. The regime 1 corresponds to a full determination of the network by the population. The second is partly circular, through the effect of roads on populations. The regime 3 is more interesting, since closeness centrality negatively causes the accessibility: this means that in this configuration, the coupled evolution of the network and the population follow the direction of a diminution of congestion. Furthermore, as population causes the closeness centrality, there is also circularity and thus co-evolution in that case. When we locate it in the phase diagram, this regime is rather sparse and rare, contrary for example to the regime 1 which occupies a large portion of space for a low importance of the road ($w_{\text{road}} \leq 0.3$). This confirms that the co-evolution produced by the model is localized and not a characteristic always verified, but that it is however able to generate some in particular regimes.

7.2.3 Discussion

We have thus proposed a co-evolution model at the mesoscopic scale, based on a multi-modeling paradigm for the evolution of the network. The model is able to reproduce a certain number of observed situations at the first and second order, capturing thus a static representation of interactions between networks and territories. It also yields different dynamical causality regimes, being however less diverse than the simple model studied before: therefore, a more elaborated structure in terms of processes must be paid in flexibility of interaction between these. This suggests a tension between a “static performance” and a “dynamical performance” of models.

An open question is to what extent a pure network model with preferential attachment for nodes would reproduce results close to what we obtained. The complex coupling between aggregation and

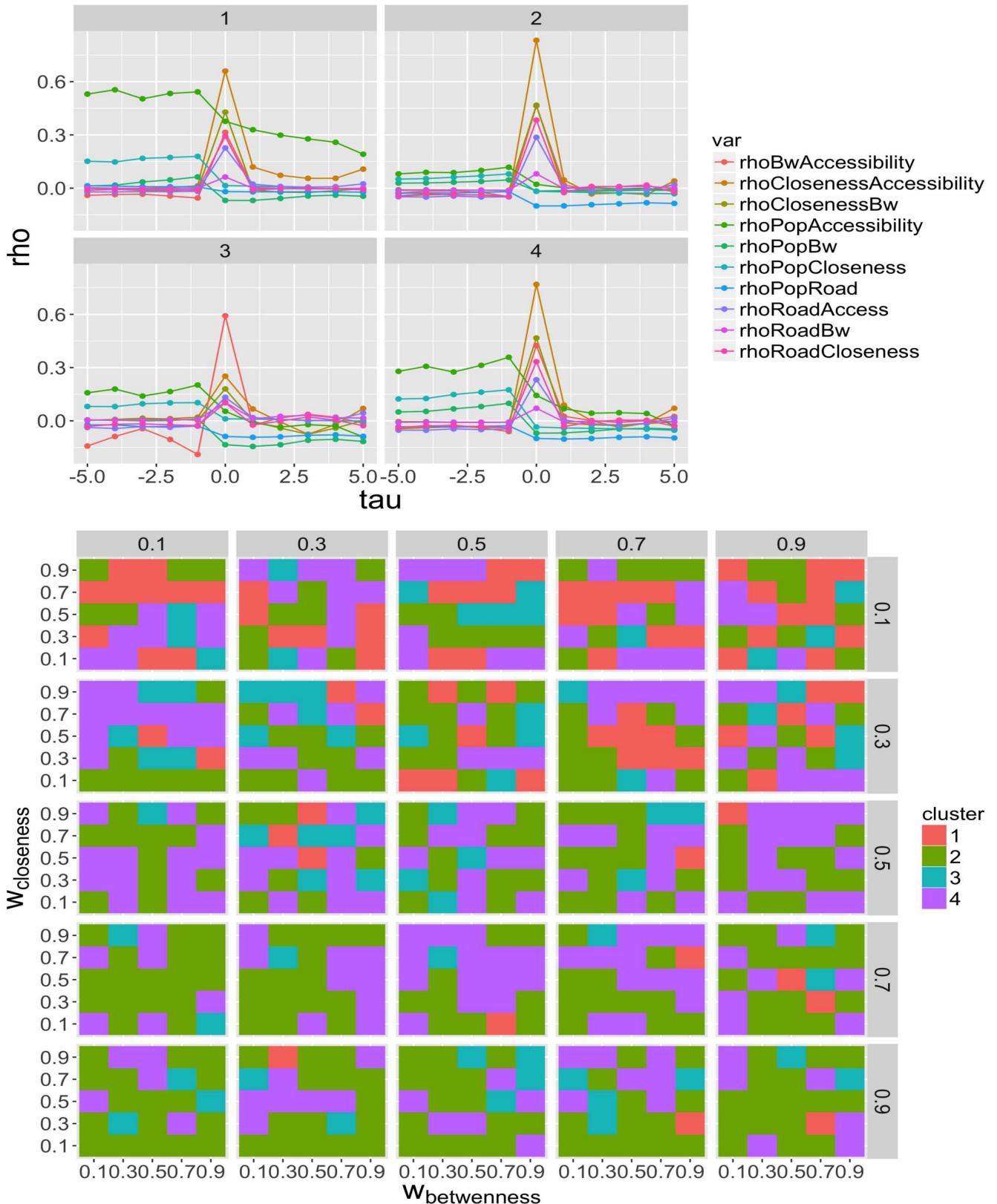


Figure 59: **Causality regimes for the co-evolution model.** (Top) Trajectories of classes centers in terms of $\rho[\tau]$ between the different explicative variables. (Bottom) Phase diagram of regimes in the parameter space for w_k , represented here as the variation of diagrams for (w_{bw}, w_{cl}) , along the variations of w_{road} (in rows) and of w_{pop} (in columns).

diffusion (shown in 5.2) could not be easily included, and the model could in any case not answer to questions on the coupling of the dynamics.

* * *

*

We have thus explored a co-evolution model based on morphogenesis that takes into account multiple processes for the evolution of the network. We studied its calibration on observed data at the first and the second order, and explored the causality regimes it produces.

We propose now a last entry into co-evolution at the mesoscopic scale, by developing a model that considerably complexifies the influence of the territory on the network, by taking into account governance processes.

* * *

*

7.3 CO-EVOLUTION AND GOVERNANCE

This section aims at giving directions for a more complex modeling of co-evolution, still at the mesoscopic scale. We have seen in 1.1 that governance processes correspond to a level that intrinsically couples networks and territories: collective decisions concern jointly transportation, territories, and their articulation. We have moreover studied the particular case of a Mega-city Region (MCR) in 1.2, and saw to what extent this context favoured a complexity of interactions. The emergence of MCR raises the question of the emergence of new governance modes, more or less easy to implement as show the examples of Stuttgart and the Rhin-Rhur metropolitan areas according to [Le Néchet, 2017].

We develop therefore here a co-evolution model at the scale of a MCR, which aims in particular at endogenizing some processes of governance of the transportation network. This model extends in particular the one introduced by [Le Néchet, 2010] which was then developed by [Le Néchet, 2011b].

7.3.1 Context

Mega-city regions and Gouvernance

We recall that a mega-city region is a network of highly connected cities in terms of economic and population flows, forming a polycentric region [Hall and Pain, 2006]. It is the last “urban regime” which emerged within systems of cities, and it could be a more plausible trajectory for large urban agglomerates than always larger monocentric cities. [Neuman and Hull, 2009] point out that the future sustainability of these MCR will be closely linked to their ability to *learn* new governance schemes, in the sense of an increased adaptability and flexibility of governance processes. [Innes, Booher, and Di Vittorio, 2010] suggest also that strategies implying self-organisation through the dialogue between stakeholders is a path to tackle the complexity of governing a MCR. We propose in the following to partly answer this question of the link between governance structure and evolution of the MCR, through the model we will develop.

Modeling co-evolution with governance processes

The role of governance processes in models coupling the evolution of transportation network with the evolution of land-use has already been investigated from different points of view in modeling approaches.

NETWORK GROWTH [Li et al., 2016] couples a network investment model with a traffic and localization model, and show that the ob-

tained steady state configurations outperform an operational research approach to network design in terms of overall accessibility.

Concerning network growth only, [Jacobs-Crisioni and Koopmans, 2016] proposes a simulation model in which alternatives between plausible investments (by different investors) are evaluated with a discrete choice model which utility function takes into account returns on investment but also variables to optimize such as accessibility. It is applied to the growth of the Dutch railway in the 19th century, and shown to reproduce quite accurately the historical network.

MODELING GOUVERNANCE [Xie and Levinson, 2011a] introduces a theoretical economic model of infrastructure investment. Two levels of governance, local and centralized are considered in the model. For the provision of new infrastructure that has to be split between two contiguous districts (space being one-dimensional), a game between governance agents determines both the level of decision and the attribution of the stock proportion to each district. Governments either want to maximize the aggregated utility (Pigovian government), or include explicit political strategies to satisfy a median voter. Numerical exploration of the model show that these processes are equivalent to compromises between cost and benefits, and that the level of governance depends on the state of the network.

[Xie and Levinson, 2011b] proposes a more simple version of this model on the governance side but coupled with a more realistic travel side : it couples on a synthetic growing network a traffic model with a pricing model and an investment model, and show that under the assumption of centralization, an equilibrium between demand and network performance can be reached, but that investments are not efficient on the long run, with a higher loss for decentralized investments.

We will be positioned in a logic close to the first model for the role of the governance structure, and close to the second for the precision of the inclusion of space.

GAME THEORY Some of these models, in particular [Xie and Levinson, 2011a], are based on game theory to model the behavior of stakeholders. It has already been widely applied for modeling in social and political sciences to questions dealing with cognitive interacting agents with individual interests [Ordeshook, 1986]. [Abler, Adams, and Gould, 1977] (p. 487) formulate a location decision problem for coffee farms on Kilimanjaro as a game combining a production strategy and a location strategy (fixing then the environmental conditions). This framework has furthermore already been used in transportation investment studies, such as e.g. in [Roumboutsos and Kapros, 2008] which use the notion of Nash equilibrium to understand choices of public or private operators concerning the integration of their system

in the broader mobility system. We will use game theory paradigms to integrate governance in a simple way in our model.

The aim of this section is thus to follow these different models, and to propose a co-evolution model in which network growth is integrated in an endogenous way, through the modeling of implied governance processes.

7.3.2 *The Lutecia Model*

We now describe the Lutecia model¹⁷, in its general structure, and then in the specification we will later develop.

Global model structure

The model couples in a complex way a module for land-use evolution with a module for transportation network growth. Submodels (or submodules), detailed in the following, include in particular a governance module that rules processes of network evolution. The most important feature of the Lutecia model is the inclusion of an endogenous infrastructure provision submodel, based on iterative increases in accessibility, within a Luti model.

The accessibility, that we will take here as a potential of access of actives to employments, is a cornerstone of the model. Indeed, micro-economic agents will relocate in order to maximize their accessibility, whereas new transportation infrastructure decisions will be taken by governance agents based on a criteria of maximization of accessibility increase in their area.

In its more general structure, the Lutecia model is composed by five sub-models, of which only three will be studied here for simplicity reasons. The sub-models are the following :

- LU stands for Land Use module : it proceeds to the re-localization of actives and employments given current conditions of accessibility.
- T stands for Transport module : it computes the transportation conditions such as flows and congestion in the urban region.
- EC stands for Evaluation of Cooperation module : it evaluates the agent or agents that will proceed to build a new infrastructure.

¹⁷ The name comes from an acronym linked to its structure which is detailed in the following. Naming models is a delicate operation since it induces a kind of reification or even personification, in any case can be seen as a kind of fetishism. It can potentially perturb the role of the model within the knowledge production process and make the model an end in itself. We are convinced that an endogenous naming through the uses of the model by the community is more appropriate. We make here an exception given the particular story of its genesis.

- I stands for Infrastructure provision module : it determines the localization of the new transportation infrastructure, based on a criteria of accessibility maximization.
- A stands for Agglomeration economies module : it evaluates the productivity of firms, depending on the accessibility to employments.

We will in the following study the coupling between the LU-EC-I sub-models: we assume at the first order no significant effect of congestion, and thus no role of transport modeling; and furthermore consider simple assumptions for economics and neglect agglomeration economies.

Different time scales are included in the model: a short scale, corresponding to daily mobility that yields flows in the transportation network and to firms productivity (modules T and A); an intermediate scale for residential and firms dynamics (module LU) ; and a long time scale for the evolution of the network (modules EC and I). Levels of stochasticity are considered accordingly: the smallest scales have deterministic dynamics whereas the longer exhibits randomness.

Detailed description of the model

DESCRIPTION OF THE ENVIRONMENT The mega-city region is modeled with a two level spatial zoning. The world is composed by a lattice of patches, that are the basic units to quantify land use. We assume that each patch k is characterized at time t by its resident actives $A_k(t)$ and number of employments $E_k(t)$. At a higher level, the MCR is decomposed into administrative areas that correspond to the city governance levels, to which we attribute M abstract agents called *mayors*: M_k gives thus the administrative area to which each patch belongs. We assume furthermore the existence of a global governance agent that correspond to a regional authority at the level of the MCR.

On top of this patch-level land-use and governance setup, we introduce a transportation network $G = (V, E)$ localized in space by its nodes coordinates (x_v, y_v) , and characterized by a speed v_G relative to movements in the euclidian space. Assuming that the network can be taken anywhere on each link, it unequivocally induces a geographical travel-time distance that we describe by the shortest path distance matrix between each patch $D = (d_{k,k'}(t))$. The accessibility of actives to employments is then defined for each patch as a Hansen accessibility with a decay of distance λ capturing typical commuting range, by

$$X_k^{(A)} = A_k \cdot \sum_{k'} E_{k'} \exp(-\lambda \cdot d_{k,k'}) \quad (18)$$

The accessibility of employments to actives is defined in a similar manner. Dynamics are taken in a discrete way: $t \in \{t_0 = 0, \dots, t_f\}$, with time ticks corresponding to a time scale at which land use typically evolves, i.e. 5 to 10 years. We take thus a slower speed for the evolution of the network which will be constructed by segments at each time step, whereas land-use will be considered as being in equilibrium at the scale of the decade, in consistence with the frame developed in chapter 1.

EVOLUTION OF LAND-USE For the land-use module, the model is based on the Lowry model [Lowry, 1964]. We assume that residential/employments relocations are at equilibrium at the time scale of a tick. In comparison, the evolution of transportation infrastructure is much slower (Wegener and Fürst, 2004)¹⁸. Actives and Employments relocate given some utilities that take into account both accessibility and the urban form. Indeed, one of the drivers of Urban Sprawl may be interpreted as a repulsion of residents by density. To aggregate both effects in a simple way, we take a Cobb-douglas function for utilities of actives and employments

$$U_k^{(A)} = X_k^{(A)^{\gamma_A}} \cdot F_k^{(A)^{1-\gamma_A}} \quad (19)$$

what is equivalent to have a linear aggregation of the logarithm of explicative variables. Employments follow an analog expression with a dedicated weight parameter γ_E . Here the utility is simply influenced only by accessibility and by an indicator of local urban form called *form factor*, given in the case of actives by $F_k^{(A)} = \frac{1}{A_k \cdot E_k}$, meaning that population is repulsed by density. The combination of the positive effect of accessibility to the negative effect of density produces a tension between contradictory objectives allowing a certain level of complexity already in the land-use sub-model alone. The form factor for jobs is taken as $F_k^{(E)} = 1$ for the sake of simplicity and following the fact that jobs can aggregate far more than dwellings.

Relocations are then done deterministically following a discrete choice model, which yields the value of actives at the next step as

$$A_i(t+1) = (1 - \alpha)A_i(t) + \alpha \cdot \left(\sum_j A_j(t) \right) \cdot \frac{\exp(\beta \tilde{U}_i(A))}{\sum_j \exp(\beta \tilde{U}_j(A))} \quad (20)$$

where β is the Discrete Choice parameter that can be interpreted as a “level of randomness”¹⁹ and \tilde{U}_i are the utilities normalized by

¹⁸ We do not consider land values, rents or transportation costs, that are the core of models in Urban Economics such as the Alonso and Fujita models for example (see [Lemoy, Raux, and Jensen, 2017] for a recent agent-based approach to these).

¹⁹ When $\beta \rightarrow 0$, all destination patches have an equal probability from any origin patch, whereas $\beta \rightarrow \infty$ gives fully deterministic behavior towards the patch with the best utility.

the maximal utility. α is the fixed fraction of actives relocating. Employments follow again a similar expression.

Network evolution : governance process

ASSUMPTIONS The governance part of the model has the following rationale :

- Three levels of governance are included, namely a central actor (the region, or regional government), local actors (municipalities) acting individually, and local actors cooperating what constitutes an intermediate level.
- Assuming a new infrastructure is to be built, the planning can be either from top-down decision (region) or from the bottom-up (local actors). We make the assumption that the processes behind the determination of the level of decision are far too complex (since they are generally political processes) to be taken into account in the model. This step is thus determined exogenously following an uniform law given a parameter.
- If the decision is taken at the local level, negotiations between actors occur. We assume that
 - the initiator of the new infrastructure can be any of the local actors, but richer cities will have more chance to built;
 - negotiations for possible collaboration are only done between neighbor cities, what is related to the medium range of infrastructure segments considered;
 - for this reason, and as n -players games have been shown to exhibit a chaotic behavior [Sanders, Doyne Farmer, and Galla, 2016] when n increases, we consider negotiations between two actors only. The probability of cooperation that are endogenously determined can be furthermore directly interpreted.
- For the sake of simplicity, the total stock of infrastructure built at one governance time step is constant, and decision times are also fixed²⁰.

NETWORK EVOLUTION The workflow for transportation network development is the following :

1. At each time step, 2 new road segments of length l_r are built. The choice between local and global is done by a uniform draw with probability ξ . In the case of local building, roads are attributed successively to mayors (one road maximum per mayor) with probabilities ξ_i which are proportional to the number of

²⁰ See the discussion for the implications of that hypothesis and possible relaxations.

employments of each, what means that richer areas will get more roads.

2. Areas building a road will enter negotiations. Possible strategies for players (negotiating areas, $i = 0, 1$, the strategies being written S_i) are to not collaborate (NC), i.e. develop his road segment alone, and to collaborate (C), i.e. wanting to develop jointly. Strategies are chosen simultaneously (non-cooperative game), in a random way according probabilities determined as detailed below. For (C, NC) and (NC, C) combinations, roads are built separately. For (NC, NC) both act as alone, and for (C, C) a common development is done.
3. Depending on the level of governance and the strategies chosen, the corresponding optimal infrastructures are build.

EVALUATION OF COOPERATION We detail now the way the co-operation probabilities are established. We denote $Z_i^*(S_0, S_1)$ the optimal infrastructure for area i with $(S_0, S_1) \in \{(NC, C), (C, NC), (NC, NC)\}$ which are determined by an heuristic in each zone separately (see implementation details), and Z_C^* the optimal common infrastructure computed with a 2 segments infrastructure on the union of both areas. It corresponds to the case where both strategies are C. Marginal accessibilities for area i and infrastructure Z is defined as $\Delta X_i(Z) = X_i^Z - X_i$. We introduce construction costs, noted I for a road segment, assumed spatially uniform. We furthermore introduce a cost of collaboration J that corresponds to a shared cost for building a larger infrastructure.

The determination of probabilities defining mixed strategies is based on the payoff matrix, which gives is the value of utility gains for each players and each possible decision configuration. The payoff matrix of the game is the following, with κ a normalization constant ("price of accessibility"), and the players being written $i \in \{0; 1\}$ (such that $1 - i$ denotes the player opposed to i)

$0 1$	C	NC
C	$U_i = \kappa \cdot \Delta X_i(Z_C^*) - I - \frac{J}{2}$	$\begin{cases} U_0 = \kappa \cdot \Delta X_0(Z_0^*) - I \\ U_1 = \kappa \cdot \Delta X_1(Z_1^*) - I - \frac{J}{2} \end{cases}$
NC	$\begin{cases} U_0 = \kappa \cdot \Delta X_0(Z_0^*) - I - \frac{J}{2} \\ U_1 = \kappa \cdot \Delta X_1(Z_1^*) - I \end{cases}$	$U_i = \kappa \cdot \Delta X_i(Z_i^*) - I$

To simplify, we assume the cost parameters dimensioned as an accessibility what is equivalent to have $\kappa = 1$. We will furthermore see that since only accessibility differentials are determining, the construction cost I does finally not play any role. This payoff matrix is used in two games corresponding to complementary processes:

- the coordination game in which players have a mixed strategy, and for which we consider the Nash equilibrium²¹ for corresponding probabilities, which implies a competition between players;
- an heuristic according to which players take their decision following a discrete choice model. It implies only a maximization of the utility gain and an indirect competition only.

We write $p_i = \mathbb{P}[S_i = C]$ the probability of each player to collaborate.

NASH EQUILIBRIUM We can solve the mixed strategy Nash Equilibrium for this coordination game in all generality. We detail the computation in Appendix A.12. By writing $U_i(S_i, S_{1-i})$ the full payoff matrix, we have the expression of probabilities

$$p_{1-i} = -\frac{U_i(C, NC) - U_i(NC, NC)}{(U_i(C, C) - U_i(NC, C)) - (U_i(C, NC) - U_i(NC, NC))}$$

What gives with the expression of utilities previously given,

$$p_i = \frac{J}{\Delta X_{1-i} Z_C^* - \Delta X_{1-i} Z_{1-i}^*} \quad (21)$$

This expression can be interpreted the following way: in this competitive game, the likelihood of a player to cooperate will decrease as the other player gain increases, and somehow counterintuitively, will increase as collaboration cost increases. The realism of this assumption must thus be moderated, and we can assume that in practice the equilibrium is never reached.

It also forces feasibility conditions on J and accessibility gains to keep a probability. These are

- $J \leq \Delta X_{1-i}(Z_C^*) - \Delta X_{1-i}(Z_{1-i}^*)$, what can be interpreted as a cost-benefits condition, i.e. that the gain induced by the common infrastructure must be larger than the collaboration cost;
- $\Delta X_{1-i}(Z_C^*) \leq \Delta X_{1-i}(Z_{1-i}^*)$, i.e. that the gain induced by the common infrastructure must be positive.

DISCRETE CHOICE DECISIONS Using the same utility functions, a random utility model for a discrete choice allows also to obtain expressions for probabilities. We have for player i the utility differential between the choice C and the choice NC given by

²¹ A Nash equilibrium is a strategy point in a discrete non-collaborative game for which no player can improve his gain by changing his strategy [Ordeshook, 1986].

$$U_i(C) - U_i(NC) = p_{i-1} (\Delta X_i Z_C^* - \Delta X_i Z_i^*) - J$$

Under the classical assumption of a model with a random utility distributed following a Gumbel law [Ben-Akiva and Lerman, 1985], we have $P[S_i = C] = \frac{1}{1 + \exp[-\beta_{DC}(U_i(C) - U_i(NC))]}$, where β_{DC} is the discrete choice parameter (that we will fix at a large value $\beta_{DC} = 400$, by supposing a certain determinism at this level, since there is then a second random level).

We substitute the expression of p_{i-1} in the expression of p_i , what leads p_i to verify the following equation

$$p_i = \frac{1}{1 + \exp \left(-\beta_{DC} \cdot \left(\frac{\Delta X_i Z_C^* - \Delta X_i Z_i^*}{1 + \exp(-\beta_{DC}(p_i \cdot (\Delta X_{i-1}(Z_C^*) - \Delta X_i(Z_{i-1}^*)) - J))} - J \right) \right)} \quad (22)$$

We demonstrate (see Appendix A.12) that there always exists a solution $p_i \in [0, 1]$, and we solve it numerically in the model to determine the probability to cooperate.

RANDOM DECISION We also consider a reference mechanism, which does not assume negotiations, but which in the case of a local decision draws randomly a mayor, following an uniform law with probabilities proportional to the number of employments of each.

Model implementation

All model parameters are recalled in Table 17. We give here only the parameters which have not been explicitly fixed previously, and these will be the privileged parameters on which the exploration and the application of the model will be done. The bound $\sqrt{2} \cdot K$ corresponds to the diagonal of the world, and the one for J has been empirically fixed according to the values of the bound given previously.

The model is implemented in NetLogo, for ergonomics reasons given its level of complexity, and also the possibilities of interactive exploration. A particular care has been given to the following points.

- Computation of distance matrices are necessary for each potential infrastructure segment, what makes the governance module very costly from the computational point of view. We use therefore a computation of shortest paths based on dynamic programming, inspired by [Tretyakov et al., 2011], updating directly the distance matrix instead of recomputing shortest paths each time.

Table 17: **Summary of Lutecia model parameters.** We also give the corresponding processes, typical bounds of the variation range and their default values.

Sub-model	Parameter	Name	Process	Domain	Default
Land-use	λ	Accessibility range	Accessibility	$]0; 1]$	0.001
	γ_A	Cobb-Douglas exponents actives		$[0; 1]$	0.85
	γ_E	Cobb-Douglas exponents employments	Utility	$[0; 1]$	0.85
	β	Discrete choices exponent		$[0; +\infty]$	1
	α	Relocation rate	Relocalization	$[0; 1]$	0.05
Transport	v_G	Network speed	Hierarchy	$[1; +\infty[$	5
Governance	J	Collaboration cost	Planning	$[0; 0.005]$	0.001
	l_r	Infrastructure length		$]0; \sqrt{2} \cdot K[$	2

- The network is for this reason represented in a dual way, under vector and raster forms. The correspondence between the two and their consistence is ensured.
- For the determination of the optimal infrastructure, the order of magnitude of the total number of infrastructures to explore is in $O(l_r \cdot N)$, if N is the number of patches and assuming that all potential infrastructures have their extremities in the center of a patch²². This considerably increases the operational computational cost, and we use an heuristic exploring a fixed number N_I of randomly chosen infrastructures.

More implementation details are given in Appendix A.12.

Model validation

Different experiments allow us to validate the model to a certain extent. We follow a modular strategy, i.e. by relatively independent tests of sub-models to begin with. The idea is to proceed to elementary experiments by making either land-use, or network, or both, evolve, and studying the consequences on the different aspects.

We work on synthetic systems. Population and employment configurations follow exponential mixtures. We give in Appendix A.12 details of initialization parameters.

LAND-USE Land-use dynamics always converge towards an asymptotic state when network does not evolve. We demonstrate the existence of the equilibrium in A.12. Furthermore, numerical experiments

²² For each patch, we will have an infrastructure for each other patch in a radius l_r , what asymptotically corresponds to the perimeter of the circle $2\pi l_r$. Furthermore, as detailed in A.12, we assume a snapping heuristic to existing infrastructures to keep a consistant network.

show that the model converge relatively quickly. Experiments targeting land-use only and which are detailed in A.12 give the following results.

- A large diversity of morphological trajectories in time, i.e. the evolution of morphological indicators for the distribution of population and employments, is obtained by playing on parameters $\gamma_A, \gamma_E, \lambda, \beta$, and also on the structure of a static network.
- Similarly, these trajectories do not converge towards the same forms and we have thus a diversity of final forms obtained.
- It is possible to minimize, at fixed $\alpha = 1$, the total quantity of relocation. We will however use this parameter to control the speed of urban sprawl, and will typically take values around 0.1, what corresponds to 10% of actives relocating at each time step, i.e. on a period of the order of the decade.

GOVERNANCE In order to understand the influence of governance parameters on forms produced by the model, we proceed to a simple experiment in the case of a bicentric system, without an initial network. Parameters for the land-use model are fixed at standard values $\gamma_A = \gamma_E = 0.8, \beta = 2; \lambda = 0.001, \alpha = 0.16$ and the length of infrastructure segments is fixed to $l_r = 2$. We consider uniquely the discrete choice game. The reference situation is given by a fully regional decision level, corresponding to $\xi = 1$. We compare it to two situations in which the level of decision is fully local ($\xi = 0$) but for which we force the possibility of collaboration to extreme values by the intermediate of the cooperation cost, taken respectively as $J = 0$ and $J = 0.005$.

The initial configuration together with three examples of network shapes obtained for each configuration are shown in Fig. 6o. Network shapes are visually²³ different and witness particular structural characteristics. In the case of the regional decision, a structuring arc links the two centers, from which extensions branch, first perpendicularly and then in parallel. The structure obtained in the case of a collaborative local is also tree-like but has less branches, the extensions being in majority following the existing branches. Finally, as we could have expected, the non-collaborative network seems to be less optimal in terms of covering than the first two, and shows redundancies. Concerning the urban structure, we obtain that the local levels better conserve the bicentric structure compared to the regional level (see the position of final centers compared to their initial position): through the network, the decision at a regional level has more potential to create new centralities.

²³ This preliminary experiment does not imply an intensive exploration, and it is thus impossible to translate these conclusions in a robust way in terms of indicators statistics.

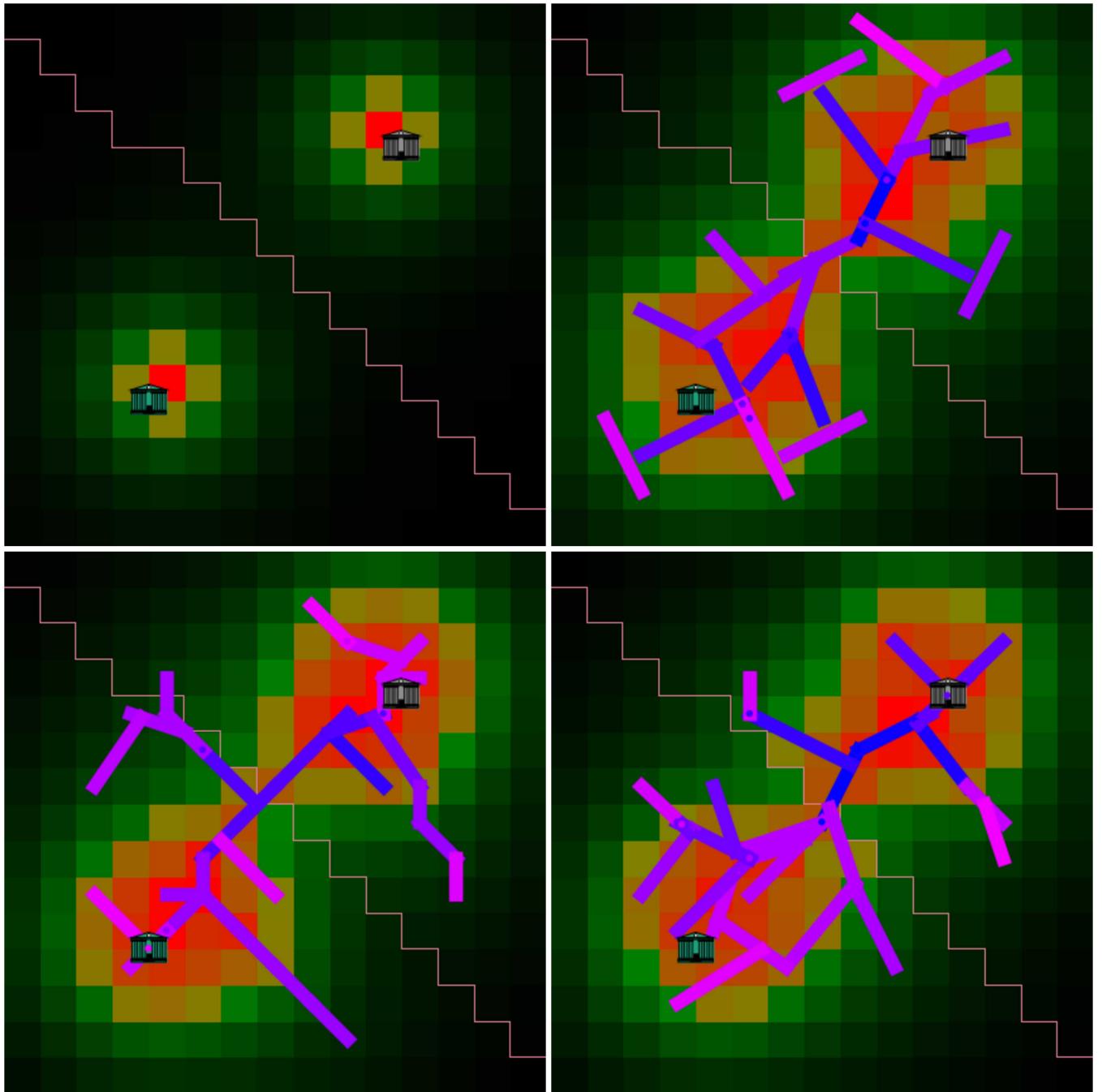


Figure 6o: Network topologies obtained for different levels of governance. The model is initialized on a symmetric synthetic configuration with two centers (*Top Left*). Parameters for the evolution of land-use are $\gamma_A = \gamma_E = 0.8; \beta = 2; \lambda = 0.001; \alpha = 0.16$, and for network evolution $l_r = 2$ and a discrete choices game. The evolution is stopped at fixed stock $S = 50$ and the heuristic exploration done for $N_I = 200$. (*Top Right*) Regional decision level ($\xi = 1$); (*Bottom Left*) Local decision level ($\xi = 0$) and low level of collaboration, obtained with a high cost of cooperation $J = 0.005$; (*Bottom Right*) Local level and high level of collaboration, with $J = 0$.

CO-EVOLUTION In a last stylized experiment, we propose to study more directly the effect of co-evolution, in particular on land-use variables. Therefore, we consider again the previous bi-centric configuration, with a disequilibrium of population and employments between the two centers (in practice with a rate of 2), and different proximities (close, at a distance of $0.4 \cdot K$, and far, at a distance of K). We fix a random local governance (choice of only one constructor with a probability proportional to employments) and the land-use and network parameters²⁴, and we study the influence of the decision level ξ on (i) the total accessibility gain between the initial and the final state, expressed as a rate $\frac{X(t_f)}{X(t_0)}$; and (ii) the evolution of relative accessibility between the two centers, given by $\frac{X_0(t_f)}{X_0(t_0)} / \frac{X_1(t_f)}{X_1(t_0)}$. The first indicators allows to understand the global benefit, whereas the second expresses the inequality between the centers (for example, is the weakest center drained by the more important, or does it benefit from it).

Results of the experiment are given in Fig. 61. The behavior of the accessibility gain unveil a direct effect of co-evolution processes: in the case of distant centers, the effect of ξ on it is inverted when we add the evolution of land-use. In the case of a network evolving alone, a local decision is optimal for total accessibility, whereas in the case of a co-evolution of processes, the optimal is at a fully regional decision. We interpret this stylized fact as the existence of a need for coordination for the success of a coupled evolution of the transportation network and land-use, what can be put in correspondence with the concept of TOD seen in chapter 1. In the case of close centers, the regional decision is always optimal, corresponding then to a more integrated metropolitan area. The variation of the relative accessibility are too low to conclude on the evolution of inequalities between the centers in the case of a coupled evolution.

Thus, this last experiment reveals indeed the existence of “co-evolution effects”, in the emergence of a need for regional coordination in the case of a coupled evolution.

7.3.3 Application to Pearl River Delta

It was suggested by [Liao and Gaudin, 2017] that a sort of multi-level governance recently emerged in China, in the context of economic activities. We try with our model to test the relevance of this paradigm regarding the urban structure of the MCR.

Model setup

We work on a simplified raster configuration (5km cells) for population in Pearl River Delta, and on the stylized freeway network. We choose to consider only the road network since, following [Hou and

²⁴ We take here $\gamma_A = 0.9, \gamma_E = 0.65, \lambda = 0.005, \beta = 1.8, \alpha = 0.1, l_r = 1, v_0 = 6$.

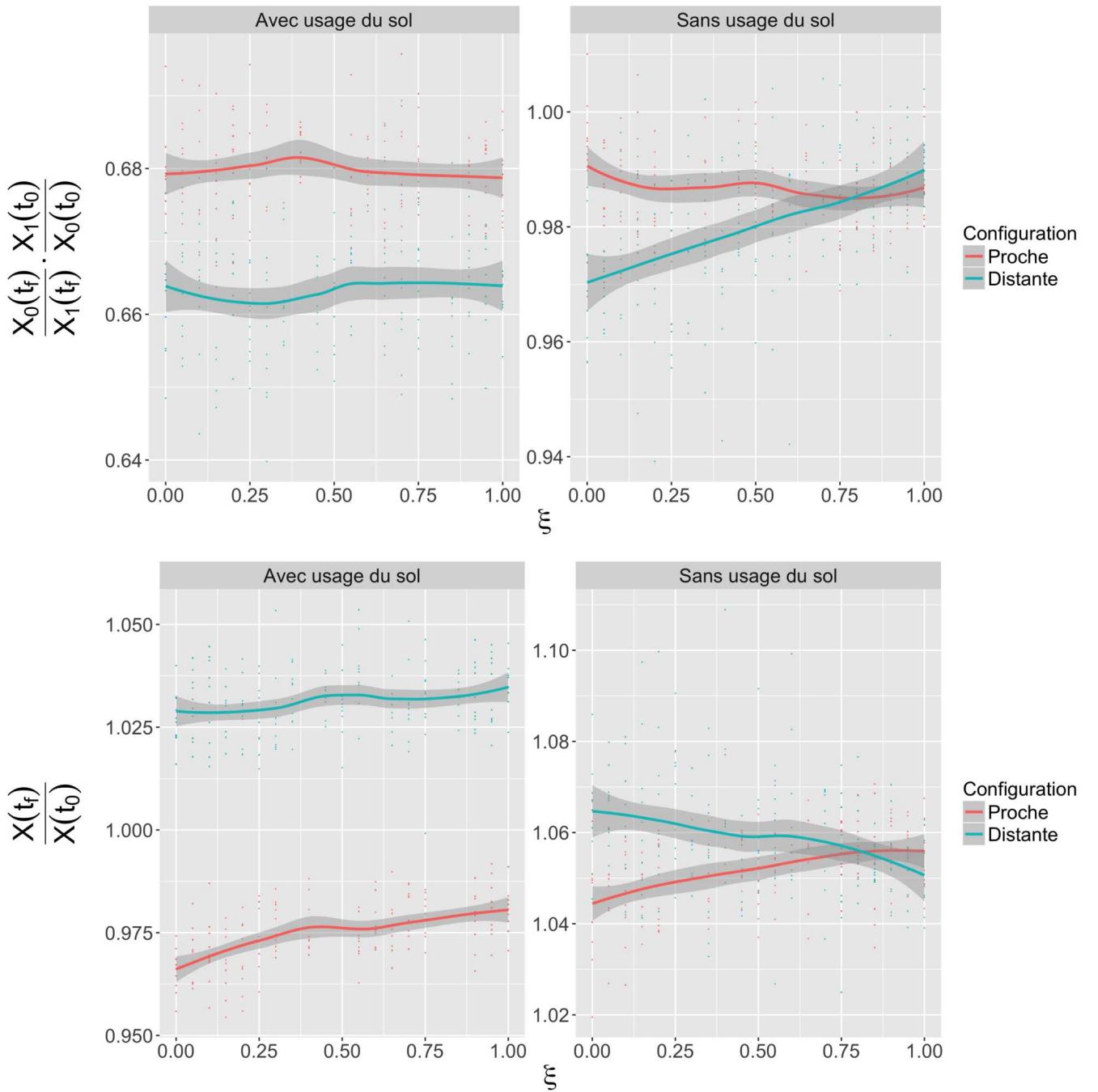


Figure 61: **Impact of co-evolution on accessibility in the Lutecia model.** We proceed to 10 repetitions with fixed parameters $\gamma_A = 0.9, \gamma_E = 0.65, \lambda = 0.005, \beta = 1.8, \alpha = 0.1, l_r = 1, v_0 = 6$, for a random local governance, and an evolution with constant stock $S = 20$. We compare the evolution with network only (without land-use) and with co-evolution, for the close and distant configurations. (Top) Evolution of the relative accessibility between centers, with and without land-use (columns) for the two configurations (colours); (Bottom) Total accessibility gain.

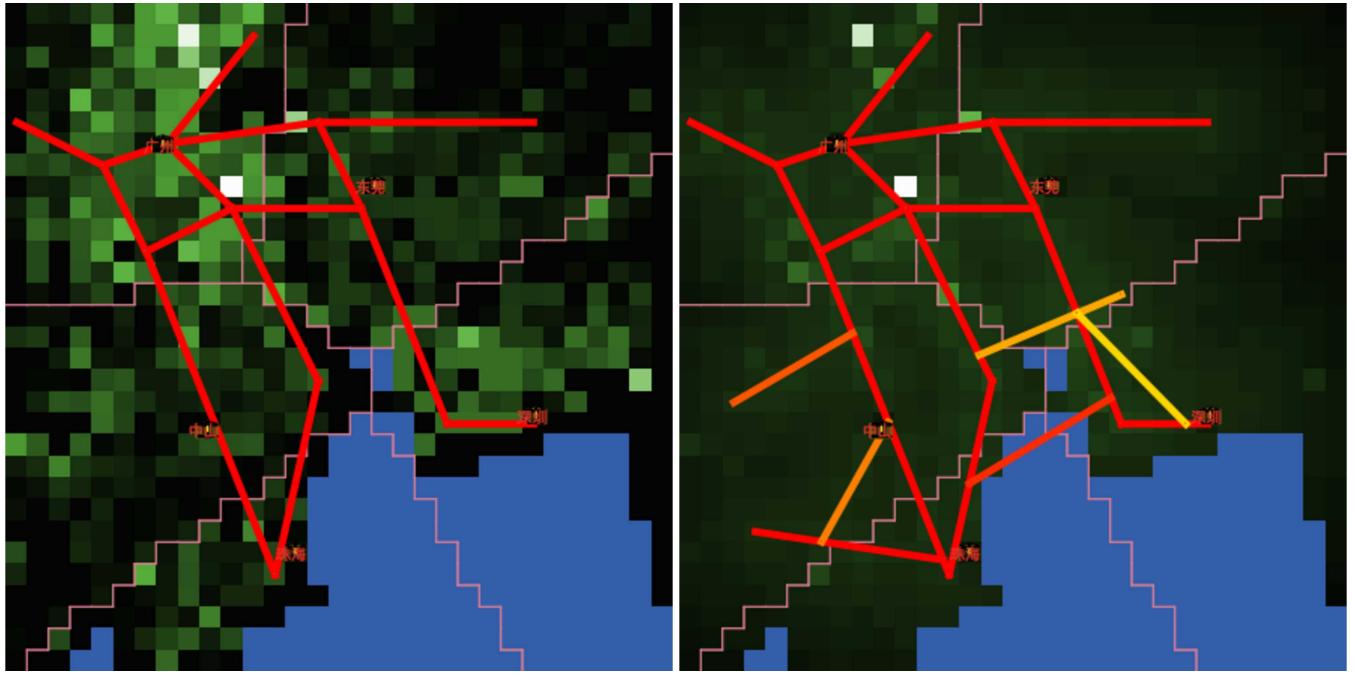


Figure 62: Example of application to Pearl River Delta. (*Left*) Initialization with the 2010 population raster, aggregated at the 5km resolution, and the simplified freeway network; (*Right*) State after 6 time steps ($\alpha = 1$).

Li, 2011], it has been the main driver of changes in accessibility patterns compared to railway network which accelerated development is recent. Networks are stylized from the plan given by [Hou and Li, 2011] which reproduces official documents of Guangdong province in 2010. We thus consider the freeway network in 2010 and the one planned at this time. Employment data are given for 2010 by [Swerts, 2017] at the level of cities. They are here uniformly distributed for each city in the simplified raster. The Fig. 62 illustrates the stylized configuration for Pearl River Delta.

Calibration procedure

To apply such a complex model to a semi-real situation, one must be particularly careful. It is important to choose the adequate processes and level of granularity to reproduce. In particular, our model is not aimed at producing particularly accurate land-use patterns, but uses their approximation as the basis of network growth, which qualitative evolution and the corresponding qualitative patterns in governance processes. We propose therefore to “calibrate” on the shape of a given infrastructure, in the sense of determining parameter configurations for which in probability the successive built pieces of infrastructure are the closest to pieces of the target infrastructure.

To calibrate on the network produced by the simulation, it must be compared to a reference network. This is however a difficult problem, as different proximity measures with different significations can be used. Geometrical measures focus on the spatial proximity of networks. For a network $(E, V) = ((e_j), (v_i))$, a node-based distance is given by $\sum_{i \neq i'} d^2(v_i, v_{i'})$. A more accurate measure which is not biased by intermediate nodes is given by the cumulated area between each pair of edges $\sum_{j \neq j'} A(e_j, e_{j'})$ (not a distance in the proper sense) where $A(e, e')$ is the area of the closed polygon formed by joining link extremities. We consider the latest for the calibration.

Calibration

The experiments we do are with a fixed land-use, since the required level of detail for more ancient or recent data, or even projections, for population and employments, is not allowed by the data we had access to.

We make governance parameters vary, including the type of game, with a fixed $l_r = 2$, and explore a Latin Hypercube Sampling of 4000 points in this parameter space, with 10 repetitions of the model for each point. The two experiments we performed correspond to different target configurations:

- no initial network and the 2010 network as a target, in the spirit of extrapolating the most probable governance configuration which led to the current configuration;
- initial network as the 2010 network, and planned network as target: extrapolation of the governance configuration for the planning.

We obtain qualitatively similar results for the two experiments, suggesting that there was no transition in the type of governance between the past network and the future network. Results are illustrated in Fig. 63. We obtain, by studying the graph of d_A as a function of ξ , that the regional level is the most realistic to reproduce network shape. However, discrete choices and competition games have a different behavior, and the competitive game is the closest to reality when ξ decreases: the relations between local actors would a priori be of a more competitive than an egoistic nature. When we study the variation of distance as a function of the observed collaboration level, we obtain an interesting inverted U-shape, i.e. that the most likely configurations are the ones where there is only collaboration, or the ones where there is no collaboration at all, but no intermediate situations. Finally, the comparison of statistical distributions of distances between target configurations and the types of games shows that the difference between the games is significant only for the real network but not for the planned network (what remains a conclusion difficult to interpret).

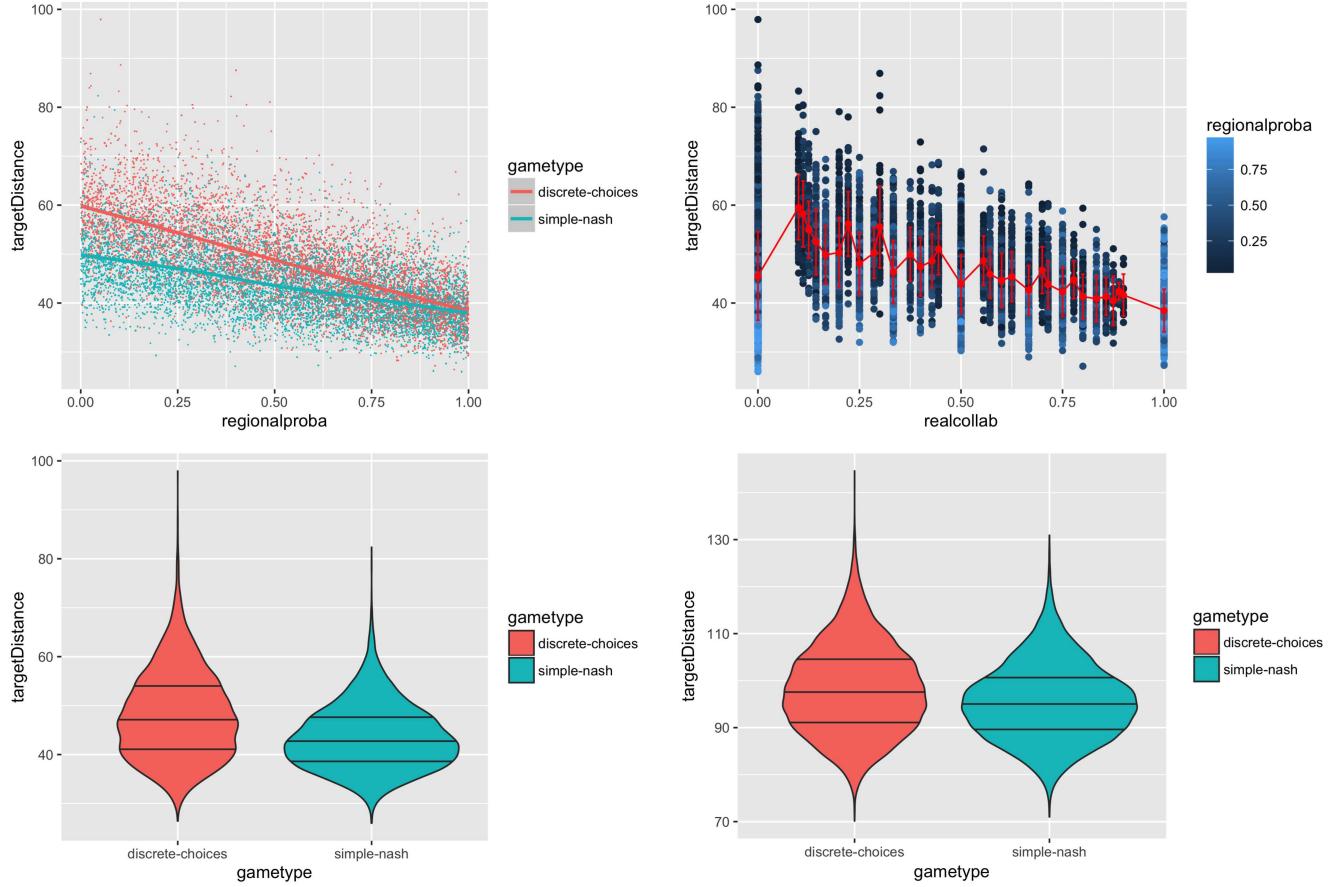


Figure 63: Model calibration with fixed land use. We take $\alpha = 0$ to make only network evolve, and sample the governance parameters space. (*Top Left*) Distance d_A to the target network (*targetDistance*), in the case of the real network, as a function of the regional decision probability ξ (*regionalproba*), for the two types of games (colour). (*Top Right*) Distance d_A as a function of the observed collaboration probability (*realcollab*); the red curve gives the averages with standard errors. (*Bottom Left*) Statistical distribution of distance as a function of the type of game, in the case of the real network; (*Top Right*) in the case of the planned network. The difference between the types of games is larger in the case of the real network in comparison to the planned network.

We thus draw from this experiment the following conclusions, to be naturally taken with caution.

- A competition between actors is less probables than an egoistic behavior in the case of local decisions, since the discrete choices game give better performances than the Nash for low values of ξ .
- Collaboration compromises correspond to less probable networks than situations with full collaboration or with no collaboration.

These conclusions can be put into perspective with the increased competition within the Delta revealed by [Xu and Yeh, 2005]. Thus, this application of the model allows to indirectly infer governance processes.

Discussion

Although the model must still be more deeply explored and for all its modules, some possible developments are worth of interest.

ENDOGENOUS LEVEL OF DECISION One relevant extension would be the study of the emergence of larger administrative zones by aggregation, i.e. the emergence of new levels of governance in polycentric metropolitan areas. The example of the *Métropole du Grand Paris* is a good illustration for it when considering it in a simplified way, since it is positioned between local collectivities and the Region but also the State [Gilli and Offner, 2009]. An extension of the model with rules to merge entities is a potential direction to study this question.

COMPETITION FOR AN EXTERNAL RESSOURCE The influence of external territories or of externalities on the evolution of a MCR is an open question. In the case of a common resource, localized within the spatial extent of the MCR, competition or collaboration dynamics can emerge between actors for its exploitation. This model is a solution to study this situation in a stylized way, and thus realize a controlled experiment on co-evolution dynamics, which would allow to answer more general questions concerning the role of territorial isolation in co-evolution processes.

* * *

*

We have thus build the first bricks of models aiming at a more complex integration of co-evolution processes, by developing the Lutecia

model which was then validated in a preliminary way and which potentialities have been demonstrated by the application to the case study of Pearl River Delta.

★ ★

★

CHAPTER CONCLUSION

This second entry on co-evolution models, at the mesoscopic scale, has been the occasion to explore the coupling between urban form and functions through the coupling between territory and network. In comparison with macroscopic models, processes that are taken here into account are much more varied and complementary.

A first morphogenesis model includes different heuristics for network growth, which are necessary and complementary to capture all the possible range of generated network configurations. We show that the model is able to resemble observed situations, for the territorial form, network topology, and also for static correlations between these indicators, while requiring a compromise between these different objectives. In terms of causality regimes, and thus of capturing co-evolutive dynamics, the model is able to capture some in some precise situations, but we learn from that experiment a fundamental lesson for co-evolutive models: a fidelity to processes or static configurations is obtained at the price of less flexibility in produced dynamical regimes. This could be a structural effect of models, or more interesting, a restriction of existing regimes in real situations.

We have then made the bet to introduce a more complex model, including an ontology for governance processes for the evolution of the transportation network. We carry out first experiments for model validation on synthetic data, and propose an application to the case of Pearl River Delta, renewing the view we gave in 1.2. We show for example that it is possible to extrapolate parameters linked to the level of collaboration between actors. This section allows thus to introduce a new approach to consider co-evolution, that takes into account the full conceptual frame developed in 1, and also opens numerous research directions.

* * *

*

CONCLUSION DE LA PARTIE III : UNE VUE COMPLÈTE DE LA CO-ÉVOLUTION

Cette partie a ainsi donné des premiers éléments d'exploration de différentes entrées sur la modélisation de la co-évolution. Nous avons exploré dans le chapitre 6 un modèle de co-évolution à l'échelle macroscopique, qui permet l'isolation de nombreux régimes de causalité, qu'on peut alors nommer régimes de co-évolution pour ceux présentant des causalités circulaires, et qui est calibré sur le système de villes français. Nous montrons ainsi que des mécanismes et une représentation simple permettent déjà de capturer synthétiquement et empiriquement la co-évolution à cette échelle.

Nous avons ensuite exploré des modèles à une échelle plus grande, impliquant une complexité croissante. Un modèle de co-évolution par morphogenèse permet de coupler la forme urbaine (distribution de la population et topologie du réseau) à une abstraction des fonctions urbaines (mesures de centralité et d'accessibilité dans le réseau). Les différentes heuristiques d'évolution du réseau qui ont été testées se révèlent complémentaires pour s'approcher de configurations réelles. Enfin, nous avons introduit des pistes pour la prise en compte des processus de gouvernance dans l'évolution des réseaux de transport.

Processes in models

Les modèles que nous avons développés l'ont été dans une logique de parcimonie, tout en cherchant à effectivement capturer des processus de co-évolution à différentes échelles et en s'encrant dans des disciplines variées : ces contraintes se paient par un prix en raffinement des mécanismes intégrés. Nous reviendrons sur ce compromis en 8.1.

A full view on co-evolution

Nous avons à ce stade apporté des éléments de réponse aux deux axes de notre problématique générale (comment définir et caractériser la co-évolution, et comment la modéliser). Il est remarquable de noter que ceux-ci s'articulent dans les trois domaines de connaissance du conceptuel (définition), de l'empirique (caractérisation) et de la modélisation (modèles). Ces trois aspects s'auto-génèrent l'un l'autre, et notre point de vue forme une véritable trinité, c'est-à-dire un concept à la fois unique et triple, dans lequel aucune des approches ne peut être ignorée (de la manière dont le fait [Morin, 2001] pour l'anthropologie complexe).

Ainsi, les modèles contiennent l'aspect individuel de la co-évolution (interactions réciproques entre entités), et dans certains cas l'aspect statistique au niveau d'une population. Cette conclusion est rendue possible par l'outil de caractérisation opérationnelle, celui-ci permettant par ailleurs de renforcer la pertinence de la définition.

Perspectives

Notre point de vue sur la co-évolution a bien entendu été réducteur et limité, puisque l'état actuel de nos modes de production de connaissance est encore loin d'une intégration paradigmatische de la complexité [Morin, 1991], et que toute tentative d'appréhension d'un système complexe combine habilement analyse et synthèse, réductionnisme et holisme, modularité et interdépendance. Afin d'enrichir notre point de vue, nous proposons finalement un chapitre d'ouverture.

CONCLUSION ET OUVERTURE

A building is never used the way it was designed, that is a reality which grasping makes the difference between good and excellent architects. The effective functional use give sense to any construction. So goes it for a knowledge edifice. We shall now take a look back on what we constructed and try to take a step back. This part develops first theoretical apparels emerging from the various aspects already tackled.

Une ouverture est principalement une mise en situation. Situation présente, situation future, situation passée. C'est en prenant ce recul qu'on s'imagine que cette trajectoire n'est pas fortuite, et qu'au fond, c'est peut être cet enfer qui nous délivrera, à l'image de l'ombre d'Euridice revisitée qui s'évade vers l'extase de la plume au dernier sous-sol. Il y a cette tradition incongrue de renseigner la profession souhaitée pour plus tard sur les fiches à chaque rentrée : finalement leur plus grand intérêt ne serait-il pas rétrospectivement, pour comprendre la dépendance au chemin de sa propre trajectoire. De conducteur de métro à cartographe, ce sera finalement un bon compromis. L'informatique qui passe par là est aussi crucial, les errances architecturales ont également joué leur rôle. Il sera sans doute impossible de dire si les systèmes urbains étaient là depuis le début, ou si l'histoire est réinterprétée à la lueur des idées triomphantes. Mais l'introspection illumine la position présente et la trajectoire future : finalement on est bien aux Enfers, mais on y est aussi pas si mal.

Une ouverture amène en effet des éléments de construction d'un méta point de vue et permet ainsi d'enrichir considérablement la connaissance produite. La nature des éléments suggérés conditionne la structure sous-jacente qu'on cherchera alors à faire émerger, qui permet en retour une réflexivité. Nous n'atteindrons pas des niveaux de réflexivité personnels à l'image de l'illustration ci-dessus, mais chercherons un certain niveau du point de vue disciplinaire et méthodologique.

Le dernier chapitre (8) apporte ainsi des éléments d'ouverture qui tiennent lieu de méta-synthèse lorsqu'on les articule dans le cadre global des recherches menées ici. Il propose ainsi à la fois une conclusion thématique et une ouverture théorique. Il met d'abord en perspective et synthétise nos contributions. Il élabore ensuite une articulation théorique des approches que nous avons prises, et enfin par la construction d'un cadre de connaissance permet une mise en perspective globale de l'ensemble du travail mené jusqu'à ce stade.

★ ★

★

8

CONCLUSION ET OUVERTURE THÉORIQUE

Theory is a key element of any scientific construction, especially in Human Sciences in which object definition and questioning are more open but also determining for research directions. We develop in this chapter a self-consistent theoretical background. It naturally emerges from thematic considerations of previous chapter, empirical explorations done in chapter ?? and modeling experiments conducted in chapter ??, as a linear structure of knowledge is not appropriate to translate the type of scientific enterprise we are conducting, typically in the spirit of SANDERS in [Livet et al., 2010] for which the simultaneous conjunction of empirical, conceptual and modeling domains is necessary for the emergence of knowledge. This theoretical construction is however presented to be understood independently, and is used as a structuring skeleton for the rest of the thesis.

We propose first to construct the *geographical theory* that will pose the studied objects and their meaning in the real world (their ontology), with their interrelations. This yields precise assumptions that will be sought to be confirmed or proven false in the following.

Staying at a thematic level appears however to be not enough to obtain general guidelines on the type of methodologies and the approaches to use. More precisely, even if some theories imply a more natural use of some tools¹, at the subtler level of contextualization in the sense of the approach taken to implement the theory (as models or empirical analysis), the freedom of choice may mislead into unappropriated techniques or questionings (see [Rimbault, 2016a] on the example of incautious use of big data and computation).

★ ★

★

The first section of this chapter is fully novel. The second uses elements from [Rimbault, 2018]. The third has been proposed by [Rimbault, 2017e] and then developed and applied in [Rimbault, 2017c], and its reflexive application has been presented by [Rimbault, 2017d].

¹ to give a rough example, a theory emphasizing the complexity of relations between agents in a system will conduct generally to use agent-based modeling and simulation tools, whereas a theory based on macroscopic equilibrium will favorise the use of exact mathematical derivations.

8.1 CONTRIBUTIONS AND PERSPECTIVES

Nous proposons à présent de passer en revue nos contributions au regard des différents cadres existants revus en première partie, et de suggérer des perspectives qu'elles ouvrent. Nous le faisons dans la logique de notre problématique générale, avec dans un premier temps nos apports sur la définition et la caractérisation de la co-évolution, et dans un second temps les différentes approches de modélisation de celle-ci.

8.1.1 *Definition and characterisation of co-evolution*

L'étape de définition et de caractérisation de la co-évolution se repose sur des résultats empiriques, théoriques et méthodologiques.

Conceptual definition

L'une de nos contributions principales est la construction d'une définition de la co-évolution au sein des systèmes territoriaux. Comme développé en 3.3, la géographie utilise ce concept de manière très floue, tandis que des disciplines où son usage pourrait sembler plus mature, comme dans le courant évolutionnaire de l'économie géographique (voir 3.3), ne s'accordent pas sur un usage précis [Schamp, 2010].

Nous précisons ainsi la définition qui en est prise dans la théorie évolutive des villes (voir par exemple [Paulus, 2004]), en gardant compatibilité. Notre définition repose en effet sur trois axes :

1. existence de processus de transformation des composantes du système territorial (*evolution*²) ;
2. modalités de co-évolution à différents niveaux : local, population, système³ ;
3. modularité en sous-systèmes territoriaux : les entités territoriales sont à la fois le support et l'objet de la co-évolution.

Notre apport par rapport à la littérature géographique mobilisant le concept est une clarification, qui permet par ailleurs la mise en place dans certains cas d'une caractérisation empirique. [Paulus, 2004] ou [Bretagnolle, Pumain, and Rozenblat, 1998] partent du postulat

² Sachant qu'on peut établir une correspondance faible avec reproduction et mutation, notamment dans le cas de composantes socio-économiques "simples" pour lesquelles les principes de l'évolution culturelle s'appliquent, mais que la correspondance devient conceptuelle quand les entités considérées sont plus complexes, comme justement dans notre cas des villes et des réseaux de transport.

³ Qui sont hiérarchiquement nécessaires : une relation au niveau de la population en implique une au niveau des individus, et la vue systémique implique une co-évolution au niveau des populations.

que la co-évolution existe nécessairement au sein des systèmes de villes, entre villes ou entre villes et réseaux de transport. Notre approche laisse une entrée à une vérification empirique et étend également l'application aux territoires de manière plus générale.

En positionnement par rapport à la littérature en économie géographique (voir [Schamp, 2010]), notre approche permet une vision fondamentalement multi-échelles, et donc plus facilement compatible avec les positionnements géographiques comme celui de la théorie évolutive des villes.

Enfin, nous avons étudié particulièrement le concept dans le cadre des interactions entre réseaux de transports et territoires : nous montrons qu'il s'agit d'un type de système territorial pour lequel il est particulièrement pertinent et opérationnel. Nous pouvons par là même revisiter le débat des effets structurants : la congruence de [Offner, 1993] peut être soit une corrélation fortuite, soit un vrai effet de co-évolution au niveau de la population. Une manifestation locale (lien local "attendu" entre deux entités) peut mais n'a pas de raison particulière de se manifester en tant que co-evolution au niveau de la population des entités (et donc il n'y a bien sûr aucun "effet systématique"). Mais qualifier les approches de cette question de "mystification scientifique" [Offner, 1993] relève du réductionnisme scientifique, que notre approche contribue à dépasser.

Spatial scales and non-stationarity

Une contribution empirique, permettant d'apporter des pistes pour la caractérisation de la co-évolution, est issue du travail mené en 4.1. L'existence de différentes échelles spatiales observables dans les corrélations statiques entre caractéristiques du territoire et celles du réseau, ainsi que la non-stationnarité spatiale de celles-ci, suggère la vérification du dernier point de notre définition, à savoir l'existence de sous-systèmes territoriaux au sein desquels la co-évolution pourrait se manifester.

Co-évolution régimes

Notre contribution fondamentale en termes de caractérisation de la co-évolution est la méthode des régimes de causalité développée en 4.2. Nous suggérons que selon les régimes observables, certains sont en effet des régimes de co-évolution, puisque présentant des relations causales circulaires observées statistiquement au niveau d'une population. Il s'agit ainsi d'une caractérisation empirique du niveau intermédiaire de co-évolution, qui est par ailleurs particulièrement intéressant puisque coïncidant avec les sous-systèmes territoriaux⁴.

⁴ Qui donne alors toute sa puissance à l'approche par la morphogenèse, en faisant le lien comme nous l'avons déjà suggéré et le développerons par la suite, avec la notion de niche écologique [Holland, 2012].

Nous pensons que notre mesure est un relativement bon proxy d'une co-évolution, puisque son application s'oriente vers l'étude des réseaux causaux [Seth, 2005], c'est-à-dire un ensemble de relations dirigées entre variables. [Castellacci and Natera, 2013] applique par exemple une méthode similaire à la notre, mais étendue à un réseau de variables, pour quantifier la co-évolution entre innovation et capacité d'absorption des territoires.

Notre approche est à remettre en perspective avec la vue de DIDEROT présentée en chapitre 1 : s'il existe une niche dans laquelle on isole des relations en effet circulaires, alors sur le temps long la dérive évolutionnaire (*drift*) par rapport à d'autres niches les entraînera sur des trajectoires bien différentes⁵. D'où l'importance de notre cadre général multi-scalaire, qui permet par ailleurs la considération du système plus globalement, et au sein duquel la mise en réseau des sous-systèmes complexifiera alors les relations de co-évolution⁶.

Empirical applicability

Nos différents cas d'étude empiriques témoignent toutefois de la difficulté de mettre en place les méthodes testées sur des données synthétiques ou uniquement théorique. L'application de la méthode des régimes de causalité donne des résultats très divers. Sur les données d'Île-de-France en 1.2, à une échelle temporelle courte et une portée spatiale restreinte, son application suggère l'existence de différents régimes. Sur les données sud-africaines en 4.2, on n'est pas capable de classifier les relations entre différentes variables, notamment à cause de l'autocorrélation de l'accessibilité, mais la méthode permet l'étude d'un sens de causalité entre croissance de population et croissance de temps moyen de trajet, ce qui donne toutefois des résultats concluants. Enfin, dans le cas de la France en 6.2, le signal obtenu est très faible, avec quasiment aucune corrélation significative pour la majorité des dates de 1836 à 1946. On dégage toutefois les résultats intéressant d'échelle intermédiaire de stationnarité spatiale, ainsi que d'une échelle de stationnarité temporelle pour les relations à longue distance. Ainsi en pratique, l'application de la méthode est à considérer au cas par cas, et les résultats peuvent provenir d'analyses annexes ou préliminaires.

Dans le cas des analyses des corrélations statiques, qui pourraient ouvrir une porte à une analyse fine et des corrélations significatives, on a déjà vu que l'absence de données temporelles empêche toute perspective d'analyse dans ce sens. Le développement de méthodes per-

⁵ Nous avons par ailleurs considéré ce cas de manière indirecte dans les modèles, lorsqu'ils sont calibrés sur le temps long sur des périodes successives : l'évolution des paramètres correspond à des dynamiques évolutives sur le temps long.

⁶ Il y aurait sur ce point une plus grande complexité des systèmes territoriaux par rapport aux systèmes biologiques "simples", c'est-à-dire ceux dans lesquels des niches écologiques sont clairement identifiables et isolables, dans le cas où la mise en réseau entre sous-systèmes est limitée.

mettant une caractérisation de la co-évolution (selon l'un des niveaux de notre définition ou selon une autre définition) à partir de données statiques reste une question ouverte.

En résumé, la co-évolution reste difficile à caractériser empiriquement, car (i) soit il n'y a effectivement aucune dynamique apparente, c'est-à-dire que les variables observables sont assimilables à du bruit (ce cas rejoint une grande partie de la littérature qui conclut à des dynamiques au cas par cas) ; (ii) les données sont très pauvres et malgré des indices suggérant l'existence de régimes de co-évolution, ceux-ci sont difficiles à caractériser.

Perspectives

The application of our approach must be lead carefully regarding the choice of scales, processes and objects of study. Typically, it will be not adapted to the quantification of spatio-temporal processes for which the temporal scale of diffusion if of the same order than the estimation window, as our stationarity assumption here stays basic. We could propose to proceed to estimations on moving windows but it would then require the elaboration of a spatial correspondence technique to follow the propagation of phenomena.

An example of concrete application that would have a strong thematic impact would be a characterization of a fundamental component of the Evolutive Urban Theory that is the hierarchical diffusion of innovation between cities [Pumain, 2010]. This would be done by analyzing potential spatio-temporal dynamics of patents classifications such as the one introduced by [Bergeaud, Potiron, and Rambault, 2017b]. We also underline that these are rather open methodological questions, for which a concretisation is the potential link between the non-ergodic properties of urban systems [Pumain, 2012b] and a wave-based characterization of these processes.

An other direction for developments and potential applications can be found when going to a more local scale, by exploring an hybridation with Geographically Weighted Regression techniques [Burdstone, Fotheringham, and Charlton, 1998]. The determination by cross-validation of Akaike criterion of an optimal spatial scale for the performance of these models, as done by [[2017arXiv170607467R](#)] in a multi-modeling fashion, could be adapted in our case to determine a local optimal scale on which lagged correlations would be the most significant, what would allow to tackle the question of non-stationarity by a mostly spatial approach.

8.1.2 *Modeling co-evolution*

Notre deuxième contribution fondamentale se situe dans la construction de modèles de co-évolution. Nous détaillons à présent nos con-

Table 18: Processus taken into account in our models.

Processus	Échelles	Concept	Modèles proposés
Attachement préférentiel/Gibrat	Meso/Macro	Croissance urbaine	Morphogenèse/Interactions
Diffusion/Etalement	Meso	Forme Urbaine	Morphogenèse
Centralité de proximité/Accessibilité	Meso/Macro	Accessibilité	Morphogenèse/Interactions
Flux direct	Macro	Interactions	Interactions
Flux indirect/Effet tunnel/Centralité de Chemin	Meso/Macro	Effet de réseau	Morphogenèse/Interactions
Proximité au réseau	Meso	Accessibilité	Morphogenèse
Relocalisations actifs/emplois	Meso	Mobilité résidentielle	Lutecia
Gouvernance des Transports	Meso	Gouvernance	Lutecia

tributions obtenues par l'intermédiaire de la modélisation, selon les deux axes complémentaires suivis.

Les processus pris en compte dans les modèles sont, comme nous l'avons déjà soulevé, voulus relativement simples pour permettre une certaine généralité et flexibilité, et n'incluent par exemple pas de processus économiques élaborés comme le modèle de [Levinson, Xie, and Zhu, 2007]. Ils remplissent toutefois leurs objectifs et couvrent un spectre assez large de processus. Ceux-ci sont synthétisés en Table 18.

Systems of Cities and the macroscopic scale

Considérons en particulier la co-évolution des territoires et des réseaux de transport au sein des systèmes de villes, à l'échelle macroscopique.

NETWORK EFFECTS Our results support the hypothesis that physical transportation networks are necessary to explain the morphogenesis of territorial systems, in the sense that some aspects are fully contained within networks and cannot be approximated by abstract proxies. We showed indeed on a relatively simple case that the integration of physical networks into some models effectively increase their explanatory power even when controlling for overfitting. This can be understood as a direction to expand Pumain's Evolutive Urban Theory (Pumain, 1997), that consider networks as carriers of interactions in systems of cities but do not put particular emphasis on their physical aspect and the possible spatial patterns resulting from

it such as bifurcations or network induced differentiations. The development of a sub-theory focusing on these aspect is an interesting direction suggested by our empirical and modeling results.

CO-EVOLUTION AT THE MACROSCOPIC SCALE Concernant la coévolution en elle-même, à l'échelle du système de villes, notre contribution principale est la compréhension globale des trajectoires et régimes possibles dans un modèle de co-évolution simple, c'est-à-dire se basant sur une ontologie abstraite pour le réseau et prenant en compte avec parcimonie des mécanismes d'évolution des villes et du réseau se basant sur les flux entre villes.

Nous retrouvons les faits stylisés typiques comme le renforcement de la hiérarchie pour certains paramètres d'auto-renforcement comme obtenu par [Baptiste, 2010]. Il s'agit à notre connaissance de la première fois qu'un modèle de co-évolution entre transport et villes dans un système de villes est exploré systématiquement, que ses régimes potentiels de co-évolution sont établis et interprétés. Notre modèle est mis en perspective avec celui de [Schmitt, 2014] : ce dernier est plus fidèle à la réalité en termes de processus microscopiques et de représentation du réseau, ce qui permet toutefois moins de flexibilité dans la production de régimes de co-évolution.

Pour l'application au cas réel du système de villes français, c'est également à notre connaissance la première fois qu'un tel modèle est calibré sur données observées. Il est difficile de dire si les processus de co-évolution sont effectivement observables, puisqu'au contraire de [Bretagnolle, 2003] nous ne trouvons pas de relation significative entre croissance des villes et accessibilité. La calibration permet toutefois d'extrapoler l'évolution de la valeur des paramètres de co-évolution dans le temps.

PERSPECTIVES The model has not yet been tested on other urban systems and other temporalities, and further work should investigate which conclusions we obtained here are specific to the French Urban System on this periods, and which are more general and could be more generic in system of cities. Applying the model to other system of cities also recalls the difficulty of defining Urban Systems. In our case, a strong bias should arise from considering France only, as Lille must be highly influenced by Brussels for example. The extent and scale of such models is always a delicate subject. We rely here on the administrative coherence and the consistence of the database, but sensitivity to system definition and extent should also be further tested.

Specifically-designed database of the highway networks containing its full genesis from 1950 to 2015).

Enfin, l'un de nos développements potentiels, la prise en compte plus fine du réseau physique, est l'objet de [Mimeur, 2016], qui pro-

duit des résultats intéressants quant à l'influence de la centralisation de la décision d'investissement dans le réseau sur les formes finales, mais garde des populations statiques et ne produit pas de modèle de co-évolution. De même, le choix des indicateurs pour quantifier la distance du réseau simulé à un réseau réel est un problème délicat dans ce contexte : des indicateurs comme le nombre d'intersections pris par [Mimeur, 2016] relève de la modélisation procédurale et non d'indicateurs de structure. C'est probablement pour la même raison que [Schmitt, 2014] ne s'intéresse qu'aux trajectoires de population et pas aux indicateurs de réseau : la conjonction et l'ajustage des dynamiques de population et de réseau à des échelles différentes semble être un problème difficile.

Territories and the macroscopic scale

Nous proposons à présent de développer nos contributions pour la modélisation de la co-évolution des territoires et des réseaux de transport à l'échelle mesoscopique.

URBAN MORPHOGENESIS Dans un premier temps, le cadre conceptuel de la morphogenèse développé en 5.1 est un apport thématique propre pour la modélisation urbaine : nous appuyons le rôle crucial de la forme urbaine, et de son lien fort avec la fonction urbaine. Ce cadre permet par ailleurs de mieux situer des modèles de morphogenèse comme celui de [Bonin and Hubert, 2014] (qui est à notre connaissance l'un des seuls modèles se présentant comme morphogénétiques ayant les fondements théoriques requis) dans un cadre interdisciplinaire.

Il permet également de considérer de façon cohérente des sous-systèmes territoriaux, puisque la recherche de règles morphogénétiques est conjointe à la définition de limites plus ou moins précises au sous-système considéré. Ce point rejoint remarquablement l'isolation géographique requise pour la co-évolution, et nous ferons la jonction théorique par la suite en 8.2.

MODELING CO-EVOLUTION WITH MORPHOGENESIS L'apport de notre modèle de co-évolution par morphogenèse est multiple et au moins les points suivants sont à noter :

- comparaison de multiples heuristiques de génération au sein d'un modèle de co-évolution ;
- calibration sur indicateurs morphologiques pour la distribution de la population et topologiques pour le réseau routier ;
- calibration au premier et au second ordre ;
- étude des régimes de co-évolution produit par un tel modèle.

L'ontologie couplée distribution de population et réseau permet le couplage fort entre forme et fonction, et justement de considérer des processus de co-évolution. En comparaison à [Barthelemy and Flammini, 2009] qui ne considèrent que le réseau, notre modèle permet plus de flexibilité dans les processus pris en compte, puisqu'il est alors possible par exemple d'ajouter des mécanismes propres à l'évolution de la population sans agir artificiellement sur la topologie du réseau, et réciproquement.

TOWARDS MODELING GOVERNANCE Enfin, le modèle Lutecia est également une contribution fondamentale vers la prise en compte de processus plus complexes impliqués dans la co-évolution, comme la gouvernance du système de transport. Comme nous l'avons déjà indiqué, [Xie and Levinson, 2011a] introduit un modèle économique théorique s'intéressant à une problématique similaire, et [Xie and Levinson, 2011b] développe une application simplifiée sur réseau synthétique. Nous allons plus loin en considérant une intégration à un modèle entièrement dynamique d'interaction entre transport et usage du sol, et implémentons une application stylisée au cas réel du Delta de la Rivière des Perles. Ce modèle ouvre la porte à une nouvelle génération de modèles, pouvant être potentiellement opérationnels dans le cas de systèmes régionaux à très grande vitesse d'évolution comme dans le cas Chinois.

PERSPECTIVES The question of the generic character of the model is also open: would it work as well when trying to reproduce Urban Forms on very different systems such as the United States or China. A first interesting development would be to test it on these systems and at slightly different scales (1km cell for example).

Finally, we believe that a significant insight into the non-stationarity of Urban Systems would be allowed by its integration into a multi-scale growth model. Urban growth patterns have been empirically shown to exhibit multi-scale behavior [Zhang et al., 2013]. Here at the meso-scale, total population and growth rates are fixed by exogenous conditions of processes occurring at the macro-scale. It is particularly the aim of spatial growth models such as the Favaro-pumain model [Favaro and Pumain, 2011] to determine such parameters through relations between cities as agents. One would condition the morphological development in each area to the values of the parameters determined at the level above. In that setting, one must be careful of the role of the bottom-up feedback: would the emerging urban form influence the macroscopic behavior in its turn ? Such multi-scale complex model are promising but must be considered carefully.

Table 19: Behavior of models regarding co-evolution.

Modèle	Effets structurants	Co-évolution individuelle	Co-évolution population	Co-évolution systémique
RBD 4.2	X	X	X	NA
Interactions 4.3	x	NA	NA	NA
Couplage faible 5.3	x	NA	NA	NA
SimpopNet 6.1	X	X	x	n.t.
Macro co-évolution 6.2	X	X	X	n.t.
Meso co-évolution 7.2	X	X	x	NA
Lutecia 7.3	n.t.	X	n.t.	NA
Empirique : Grand Paris 1.2	X	x	o	NA
Empirique : Afrique du Sud 4.2	X	x	o	n.t.
Empirique : France 6.2	o	x	o	n.t.

Position of models

Nous faisons une synthèse de la position des différents modèles au regard de la co-évolution en Table 19. Nous donnons les modèles qui ont été nouvellement introduits dans ce travail ainsi que les modèles extérieurs ayant été utilisés, et les études empiriques. Nous voyons ainsi qu'il est direct d'introduire une co-évolution au niveau individuel dans les modèles, mais que la co-évolution au niveau de la population, c'est-à-dire l'existence de causalités circulaires entre variables de réseau et variables de territoire, est plus difficile à obtenir de façon marquée. Les effets structurants (existence de relations causales dans un sens ou dans l'autre) sont quant à eux présents dans presque la totalité des modèles. Nous rappelons qu'il est difficile de mesurer une co-évolution sur données empiriques.

8.1.3 Approaches of coevolution

Finalement, nous proposons d'ouvrir des perspectives plus larges d'approches de la co-évolution différentes de la notre.

Alternative approaches

Nous avons fait le choix de caractéristiques élémentaires des territoires et des réseaux pour la modélisation de leur co-évolution : la plupart de nos modèles ne considère que des variables de population pour les territoires, et de nombreuses autres dimensions possibles (économique, politique, institutionnelle, sociale) sont occultées.

Des dimensions où il existe potentiellement des effets co-évolutifs et où une modélisation serait pertinente peuvent être regroupés de la manière suivante :

- problématiques liées au système de transport :
 - rôle de la tarification des transports et des investissements, déjà largement pris en compte dans les modèles économiques de LEVINSON comme [Levinson, Xie, and Zhu, 2007] ;
 - plus généralement rôle des acteurs de gouvernance dans l'évolution du système de transport, comme nous avons esquissé avec le modèle Lutecia en 7.3, et comme le fait de manière plus théorique [Xie and Levinson, 2011a] ;
 - rôle du changement technologique dans la relation entre forme urbaine et mobilité [Brotchie, 1984] ;
- problématiques liées aux acteurs faisant la ville :
 - rôle des différents acteurs de production de la ville (promoteurs immobiliers⁷ et collectivités locales par exemple [Le Goix, 2010]) et de leurs stratégies ;
 - en lien avec les approches de type Luti, approfondir le rôle des choix de localisation des acteurs (mobilités résidentielles ou acteurs économiques [Tannier, 2003]) dans la production territoriale, en relation aux réseaux ("échelle de l'accessibilité") identifiée au chapitre 1 ;
- enfin, à l'échelle des mobilités quotidiennes, les pratiques de mobilité selon les caractéristiques socio-économiques, est également une dimension territoriale pertinente à creuser pour l'étude de la co-évolution ([Cerqueira, 2017] montre par exemple les différenciations socio-économiques dans le lien entre accessibilité et mobilité), pour laquelle des pistes de modélisation ont par exemple été proposées par [Morency, 2005] qui construit par désagrégation une base de donnée intégrée couplant caractéristiques socio-économiques des ménages et données de mobilité.

Cette liste est bien évidemment loin d'être exhaustive, mais permet de se rendre compte des dimensions complémentaires qui permettraient également une entrée sur notre problématique générale.

Nous sommes donc loin d'avoir épuisé la problématique de la co-évolution, puisqu'il s'agirait alors de savoir : (i) dans quelle mesure notre définition est générale et s'applique à des dimensions qui n'ont pas été initialement conçues ; (ii) dans quelle mesure notre méthode de caractérisation s'applique aux différentes dimensions et quelles

⁷ Nous avons abordé en 1.2 brièvement des variables liées aux transactions immobilières, et montré les potentialités pour la mise en valeur de stratégies d'anticipation de desserte par le nouveau réseau, ce qui confirme ici la pertinence de ce point de vue.

méthodes alternatives sont envisageables ; (iii) si nos structures de modèles, relativement génériques, peuvent être étendues à ces problématiques connexes.

A Roadmap for an Operational Family of Models of Coevolution

Towards operational Models : what is possible ; what is desirable ; etc. As previously stated, one of our principal aims is the validation of the network necessity assumption, that is the differentiating point with a classic evolutive urban theory. To do so, toy-model exploration and empirical analysis will not be enough as hybrid models are generally necessary to draw effective and well validated conclusions. We briefly give an overview of planned work in the following, that will be the conclusion of this Memoire.

* * *

*

Nous avons ainsi pu dans cette section prendre du recul sur nos contributions et les mettre en perspective d'horizons plus vastes concernant la question de la co-évolution des réseaux de transport et des territoires.

La section suivante propose une articulation de nos différentes contributions d'un point de vue théorique, une synthèse permettant d'expliquer certaines connexions jusqu'alors relativement implicites.

* * *

*

8.2 A GEOGRAPHICAL THEORY

RAFFESTIN highlights in his preface of [Offner and Pumain, 1996] that a geographical theory that articulates spaces, networks and territories has never been formulated in a consistent way, since each approach has a vision reduced to some components only and does not aim at constructing an integrative theory. A research direction we propose to introduce here is the conjunction of approaches of the evolutive urban theory and of morphogenesis, to produce a theory that is both multi-scalar and fully integrates networks and territories.

8.2.1 Foundations

Our theoretical construction relies on four pillars that we will detail below⁸.

Networked human territories

Our first pillar corresponds to the theoretical construction elaborated in 1.1. We rely on the notion of *Human Territory* elaborated by RAFFESTIN as the basis for a definition of a territorial system. It allows to capture complex human geographical systems in all the extent of their concrete and abstract characteristics, and also their representations. For example, a metropolitan territory can be apprehended simply by the functional extent of daily commuting flows, or by the perceived or lived space for different populations, the choice depending on the precise question that is considered.

The territory of RAFFESTIN indeed corresponds to a consistent system of *synergetic inter-representation networks*, which are both a theory and a model for spatial cognition of individual and societies, constructed by PORTUGALI and HAKEN (see [Portugali, 2011] for a synthetic presentation). It postulates that representations are the product of a strong coupling between individuals of cognitions and their individual and collective behaviors. This approach to the territory is of course a particular choice and other entries, possibly compatible, can be taken [Murphy, 2012].

The concrete of this pillar is reinforced by the territorial theory of networks of DUPUY, yielding the notion of networked human territory, as a human territory in which a set of potential transactional networks have been realized, which is in accordance with visions of the territory as networked places [Champollion, 2006]. We will not use the implications of the development of the notion of *place*, these being too sparse (see the definition of [Hypergeo 2017]), and because of the redundancy with the territory in the vision of a complex link

⁸ Or more precisely a funding horizontal pillar which gives fundamental objects, i.e. foundations introduced in Chapter 1, two vertical pillars for the structure, and an horizontal synthesis pillar allowing to link these two.

between representations and the physical reality. We will assume for this first pillar the fundamental assumption, already introduced in Chapter 1, that real networks are necessary elements of territorial systems.

Evolutive urban theory

The second pillar of our theoretical construction is PUMAIN's evolutive urban theory, in close relation with the complex approach that we generally take. It has already been presented with details and its implications have been explored in Chapter 4. Here, this theory allows us to interpret territorial systems as complex adaptive systems and to introduce the co-evolution.

Urban morphogenesis

The notion of morphogenesis has been deeply explored and with an interdisciplinary point of view in 5.1. We recall here important axis and to what extent these contribute to the construction of our theory. Morphogenesis has been formalized especially by [Turing, 1952b] which proposes to isolate elementary chemical rules that could lead to the emergence of the embryo and its form.

The morphogenesis of a system consists in evolution rules that produce the emergence of its successives states, i.e. the precise definition of self-organization, with the additional property that an emergent architecture exists, in the sense of causal circular relations between the form and the function. Progresses towards the understanding of embryo morphogenesis (in particular the isolation of particular processes producing the differentiation of cells from an unique cell) have been made only recently with the use of complexity approaches in integrative biology [Delile, Doursat, and Peyriéras, 2016].

In the case of urban systems, the idea of urban morphogenesis, i.e. of self-consistent mechanisms that would produce the urban form, is more used in the field of architecture and urban design (as for example the generative grammar of "Pattern Language" of [Alexander, 1977]), in relation with theories of urban form [Moudon, 1997]. This idea can be pushed into very large scales such as the one of the building [Whitehand, Morton, and Carr, 1999] but we will use it more at a mesoscopic scale, in terms of land-use changes within an intermediate scale of territorial systems, with similar ontologies as the urban morphogenesis modeling literature (for example [Bonin and Hubert, 2012] describes a model of urban morphogenesis with qualitative differentiation, whereas [Makse et al., 1998] give a model of urban growth based on a mono-centric population distribution perturbed with correlated noises).

The concept of morphogenesis is important in our theory in link with modularity and scale. Modularity of a complex system consists

in its decomposition into relatively independent sub-modules, and the modular decomposition of a system can be seen as a way to disentangle non-intrinsic correlations [Kolchinsky, Gates, and Rocha, 2015] (to have an idea, think of a block diagonalisation of a first order dynamical system). In the context of large-scale cyber-physical systems design and control, similar issues naturally raise and specific techniques are needed to scale up simple control methods [Wang, Matni, and Doyle, 2017]. The isolation of a subsystem yields a corresponding characteristic scale. Isolating possible morphogenesis processes implies a controlled extraction (controlled boundary conditions e.g.) of the considered system, corresponding to a modularity level and thus a scale.

When local processes are not enough to explain the evolution of a system (with reasonable variations of initial conditions), a change of scale is necessary, caused by an underlying phase transition in modularity. The example of metropolitan growth is a good example: complexity of interactions within the metropolitan region will grow with size and the diversity of functions, leading to a change in the scale necessary to understand processes. The characteristic scales and the nature of processes for which these change occur can be precise questions investigated through modeling.

Finally, it is important to remark as we did in 5.1 that a territorial subsystem for which morphogenesis makes sense, which boundaries are well defined and which processes allow it to maintain itself as a network of processes, is close to an *auto-poietic system* in the extended sense of BOURGINE in [Bourgine and Stewart, 2004]⁹. These systems regulate their boundary conditions, what underlines the importance of boundaries that we will finally develop.

Co-evolution

Our last pillar consists in an approach to the concept of *co-evolution* complementary to the definition we already introduced. It is brought by HOLLAND which sheds a relevant light through an approach of complex adaptive systems (CAS) by a theory of CAS as agents which fundamental property is to process signals thanks to their boundaries [Holland, 2012].

In this theory, complex adaptive systems form aggregates at diverse hierarchical levels, which correspond to different level of self-organization, and boundaries are vertically and horizontally intricated in a complex way. That approach introduces the notion of *niche* as a relatively independent subsystem in which resources circulate

⁹ Which are however not cognitive, making these morphogenetic systems not alive in the sense of auto-poietic and cognitive. Given the difficulty to define the delineation of cities for example, we will leave open the issue of the existence of auto-poietic territorial systems, and will consider in the following a less restrictive point of view on boundaries.

(the same way as communities in a network as used in chapter 2): numerous illustrations such as economical niches or ecological niches can be given. Agents within a niche are then said to be *co-evolving*.

Empirically, results obtained witness a co-evolution at the mesoscopic scale such as in 4.2, confirming the existence of niches for some aspects of territorial systems. The co-evolution in that sense implies then strong interdependencies with circular causal processes (rejoining the definition we took) and a certain independence regarding the exterior of the niche.

The notion is naturally flexible as it will depend on ontologies, on the resolution, on thresholds, etc. that we consider to define the system. We postulate given the clues of existence obtained in empirical results, but also models reproducing processes in a credible manner under a reasonable independence assumption, that this concept can easily be transmitted to the evolutive urban theory and corresponds to the notion of co-evolution we took (and in particular at the level of a population of entities): co-evolving agents in a system of cities consist in a niche with their own flows, signals and boundaries and thus co-evolving entities in the sense of HOLLAND.

8.2.2 *A theory of co-evolutive networked territorial systems*

We synthesize the different pillars as a geographical theory of territorial systems in which networks play a central role in the co-evolution of system components.

Définition 1 - Territorial System. *A territorial system is a set of networked human territories, i.e. human territories in and between which real networks are materialized.*

The territory is indeed an element of the territorial system, which more generally connects different territories with networks. At this stage complexity and the evolutive and dynamical character of territorial systems are implied by the positions taken but not an explicit part of the theory. We will assume to simplify a discrete definition of temporal, spatial and ontological dimensions, under modularity and local stationarity assumptions. This aspect, both for the discrete and the stationarity, corresponds to an ontological simplification of the assumption of a “minimal scale” at which subsystems give a simple modular decomposition of the global system.

Hypothèse 1 - Discrete scales. *Assuming a discrete modular decomposition of a territorial system, the existence of a discrete set of temporal and functional scales for the territorial system is equivalent to the local temporal stationarity of a random dynamical system specification of the system.*

This proposition postulates a representation of system dynamics in time. Note that even in the absence of a modular representation, the system as a whole will verify the property. We will assume the case in which scales always exist, i.e. verifying one of the specifications of this assumption.

This definition of scales allows to explicitly introduce feedback loops, since we can for example condition the evolution of a scale to the evolution of another containing it, and thus emergence and complexity, making the theory compatible with the evolutive urban theory.

Hypothèse 2 - Intrication of scales and subsystems. *Complex networks of feedbacks exist both between and within scales [Bedau, 2002]. Furthermore, a horizontal and vertical imbrication of boundaries will not always be hierarchical.*

Within these complex subsystems intrications we can isolate co-evolving components using morphogenesis. The following proposition is a consequence of the equivalence between the independence of a niche and its morphogenesis. Morphogenesis provides the modular decomposition (under the assumption of local stationarity) necessary for the existence of scales, giving minimal vertically (scale) and horizontally (space) independent subsystems.

Hypothèse 3 - Co-evolution of components. *Morphogenesis processes of a territorial system are an equivalent formulation of the existence of co-evolutive subsystems.*

Finally we make a key assumption putting real networks at the center of co-evolutive dynamics, introducing their necessity to explain dynamical processes of territorial systems.

Hypothèse 4 - Necessity of networks. *The evolution of networks can not be explained only by the dynamics of other territorial components and reciprocally, i.e. co-evolving territorial subsystems include real networks. They can thus be at the origin of regime changes (transition between stationarity regimes) or more dramatic bifurcations in dynamics of the whole territorial system.*

8.2.3 Contextualization

Co-evolution is more or less easy to show empirically (see for example the debate on structuring effects) but we assume the existence of co-evolution processes at all scales of the system. Regional examples for the French system of cities may illustrate that aspect: Lyon has not the same interactions with Clermont-Ferrand than with Saint-Etienne, and network connectivity has probably a role in that (among

intrinsic interaction dynamics, and distance for example). At a even larger scale, we speculate that effects are even less observable, but precisely because of the fact that co-evolution is stronger and local bifurcations will occur with stronger amplitude and greater frequency than in macroscopic systems where attractors are more stable and stationarity scales smaller. It is for this reason that we tried to identify bifurcations and phase transitions in toy models, hybrid models, and empirical analyses, at different scales, on different case studies and with different ontologies.

One difficulty in our construction is the local stationarity assumption, which is essential to formulate models at the corresponding scale. Even if it seems a reasonable assumption on several scales and has already been observed in empirical data [Sanders, 1992], we were able to verify it more or less in our empirical studies.

Indeed, this question is at the center of current research efforts to apply deep learning techniques to geographical systems: PAUL BOURGINE¹⁰ has recently proposed a framework to extract patterns from complex adaptive systems. Using a representation theorem [Knight, 1975], any discrete stationary process is a *Hidden Markov Model*. Given the definition of a causal state as the set of states allowing an equivalent prediction of the future, the partition of system states induced by the corresponding equivalence relations allows to derive a *Recurrent Network* that is sufficient to determine the next state of the system, as it is a *deterministic* function of previous states and hidden states [Shalizi and Crutchfield, 2001]: $(x_{t+1}, s_{t+1}) = F[(x_t, s_t)]$ if x_t is the state of the system and s_t the hidden states. The estimation of hidden states and of the recurrent function thus captures entirely through deep learning dynamical patterns of the system, i.e. full information on its dynamics and internal processes.

The issues that raise then are if the stationarity assumptions can be tackled through augmentation of system states, and if heterogeneous and asynchronous data can be used to bootstrap long enough time-series necessary for a correct estimation of the neural network or any other estimator. These issue are related to the stationarity assumption for the first and to non-ergodicity for the second.

* * *

*

This section has thus given a theoretical opening, by proposing as hypothesis an articulation between the different complementary

¹⁰ Personal communication, January 2016.

approaches that we developed. This articulation allows a global perspective and reinforces our definition of co-evolution.

The next section concludes this opening from an epistemological point of view, by placing our work in the perspective of a knowledge framework, and opening thus reflexive approaches on it.

★ ★

★

8.3 AN APPLIED KNOWLEDGE FRAMEWORK

Nous proposons de monter encore en niveau de généralité et de nous placer à un niveau épistémologique, en introduisant un cadre théorique pour l'étude des processus de production de connaissance.

The complexity of knowledge production on complex systems is well-known, but there still lacks knowledge framework that would both account for a certain structure of knowledge production at an epistemological level and be directly applicable to the study and management of complex systems. We set a basis for such a framework, by first analyzing in detail a case study of the construction of a geographical theory of complex territorial systems, through mixed methods, namely qualitative interview analysis and quantitative citation network analysis. We can therethrough inductively build a framework that considers knowledge enterprises as perspectives, with co-evolving components within complementary knowledge domains. We finally discuss potential applications and developments.

The understanding of processes and conditions of scientific knowledge production are still mainly open questions, to which monuments of epistemology such as Kant's Critique of Pure Reason, and more recently Kuhn's study of "the structure of scientific revolutions" [Kuhn, 1970] or Feyerabend's advocacy for a diversity of viewpoints [Feyerabend, 1993], have brought elements of answer from a philosophical approach. A more empirical point of view was brought also recently with quantitative studies of science, in a way a *quantitative epistemology* that goes far beyond rough bibliometric indicators [Cronin and Sugimoto, 2014]. Contributions harnessing complexity, i.e. studying complex systems in a very broad sense, can be shown to have produced very diverse frameworks that can be counted as building bricks contributing to answers to the above high-level question. We will in the following use the term Knowledge Framework, for any such framework having an epistemological component tackling the question of nature of knowledge or knowledge production. To illustrate this, we can mention such frameworks in different domains, at different levels and with different purposes. For example, [Durantin et al., 2017] explores the potentialities of coupling engineering with design paradigms to enhance disruptive innovation. Also in Knowledge Management, using the constraint of innovation as an advantage to understand the complex nature of knowledge, [Carlile, 2004] introduces knowledge domains boundaries and production processes. Also introducing a meta-framework, but in the field of system engineering, [Gemino and Wand, 2004] recommends to use grammars to compare Conceptual Modeling Techniques. Meta-modeling frameworks can also be understood as Knowledge Frameworks. [Cottineau et al., 2015] describes a multi-modeling framework to test hypotheses in simulation of socio-technical complex systems. [Golden, Aiguier,

and Krob, 2012] postulates a unified formulation of systems, including necessarily different types of knowledge on a system on its different description components.

A possible explanation for this richness is the fundamental reflexive nature of the study of Complex Systems: because of the higher choice in methodology and what aspects of the system to put emphasis on, a significant part of a modeling or design entreprise is an investigation at a meta-level. Furthermore, studies of knowledge production are mainly rooted in complexity, implying a reflexive nature of theories accounting knowledge on complexity, as Hofstadter had well highlighted in [Hofstadter, 1980] by noticing the importance of “strange loops”, i.e. feedback loops allowing reflexivity such as a theory applying to itself, in what constitutes intelligence and the mind. Artificial intelligence is indeed a crucial field regarding our issues, as its progresses imply a finer understanding of the nature of knowledge. [Moulin-Frier et al., 2017] introduces a meta-framework for a general typology of approaches in Artificial Intelligence, what is a Knowledge Framework not in the proper sense but in a specific applied case.

The level of frameworks described above may be very general but is conditioned to a certain field or discipline, and to a certain approach or methodology. There exists to our knowledge no framework realizing a difficult exercise, that is to capture a certain structure of knowledge production at an epistemological knowledge, but conjointly is thought in a very applied perspective, with direct consequence in the design and management of complex systems. The contribution of this paper attempts to set a basis for a Knowledge Framework realizing this in the case of Complex Systems. To perform that, we postulate that the tension between these two contradictory objective is an asset to avoid on one side an impossible overarching generality and on the other side a too restraining domain-specific specificity. Based on the idea of complementary Knowledge Domains introduced by [Livet et al., 2010], its central aspect is a cognitive approach to science inducing co-evolutive processes of knowledge domains and their carriers. A first sketch of this framework was presented by [Raimbault, 2017e], in the specific case of complex territorial systems as studied by theoretical and quantitative geography. We choose to introduce it here with an inductive approach, i.e. starting from a concrete case study that has mainly inspired the construction of the framework to end with its generic description.

The rest of the section is organized as follows: the next section develops case studies, more precisely a detailed study of a geographical theory of complex urban systems, and a short example from engineering to illustrate the transferability of concepts. The third section specifies the definitions and formulates the epistemological framework.

We finally discuss issues on applicability, and potential developments such as a mathematical version of the framework.

8.3.1 Case Studies

Genesis of the Evolutive Urban Theory

The first case study relates the construction of the *Evolutive Urban Theory*, a geographical theory considering territorial systems from a complexity perspective, that have been developed for around 20 years. We analyse its genesis using mixed methods, namely semi-directed interviews with main contributors, and quantitative bibliometric analysis of main publications.

Interviews were done following methodological standards [Legavre, 1996] to ensure a limited interference of the interviewer's experiences but not make it fully disappear to ensure a precise context enhancing the fluency of the interviewed. We use here interviews¹¹ with Pr. D. Pumain who introduced and developed mainly the theory, and Dr. R. Reuillon, whose research on intensive and distributed computation and model exploration has been a cornerstone of latest developments.

Let first give an overview of its content. This theory was first introduced in [Pumain, 1997] which argues for a dynamical vision of city systems, in which self-organization is key. Cities are interdependent evolutive spatial entities whose interrelations produces the macroscopic behavior at the scale of city system. The city system is also described as a network of city what emphasizes its view as a complex system. Each city is itself a complex system in the spirit of [Berry, 1964], the multi-scale aspect being essential in this theory, since microscopic agents convey system evolution processus through complex feedbacks between scales. The positioning within Complex System Sciences was later confirmed [Pumain, 2003]. It was shown that this theory provide an interpretation for the origin of pervasive scaling laws, resulting from the diffusion of innovation cycles between cities [Pumain et al., 2006]. The aspect of resilience of system of cities, induced by the adaptive character of these complex systems, implies that cities are drivers and adapters of social change [Pumain, 2010]. Finally, path dependance yield non-ergodicity within these systems, making "universal" interpretations of scaling laws difficultly compatible with evolutive urban theory [Pumain, 2012b]. The construction of models of urban systems has been a key component for the theory, starting with the first Simpop model [Sanders et al., 1997]. Later example include for example the Simpop2 model, an agent-based model taking into account economic processes, that simulates growth pat-

¹¹ Both have a length of around 1h. Sound and transcript text are available under a CC Licence at <https://github.com/JusteRaimbault/Interviews> [*Entretiens vo.2 [Dataset]*]. Interviews are in French and translations here are done by the author.

terns on long time scales for Europe and the United States [Bretagnolle and Pumain, 2010b]. The latest accomplishment of the evolutive theory relies in the output of the ERC project GeoDiversity, presented in [Pumain and Reuillon, 2017d], that include both advanced technical (software OpenMole¹² [Reuillon, Leclaire, and Rey-Coyrehourcq, 2013]), thematic (knowledge from SimpopLocal [Schmitt, 2014] and Marius models [Cottineau, 2014]) and methodological (incremental modeling [Cottineau, Chapron, and Reuillon, 2015]) progresses.

The striking feature in the construction of all this is the balance between the different *types* of knowledge, of which a typology will be the starting point of our construction. The relation between theoretical considerations and empirical cases studies is fundamental. Indeed, the seminal article [Pumain, 1997] is already positioned as an “advocacy for a less ambitious theory, but that does not neglects the back-and-forth with observation”¹³.

We shall now turn to interviews to better understand the implications of the intrication of types of knowledge. D. Pumain traces back germinal ideas back to her graduate student work in 1968, when “everything started with a question of data”. The interest for cities, and *change in cities*, was driven by the availability of a refined migration flow dataset at different dates. Also rapidly, “[they] were frustrated that methods were missing”, but the access to the computation center (*technical tool*) allowed the test of newly introduced methods and models, linked to the Prigogine approach to complexity. Methods were however still limited to grasp the heterogeneity of spatial interactions. A progressively specified need and a chance encounter, with “a lady working on neural networks and agent-based modeling at the Sorbonne”, led to a bifurcation and a new level of interaction between modeling, theory and empirical knowledge: in 1997, two seminal articles, one stating the theoretical basis and the other introducing the first Simpop model, were published.

From this point, it was clear that all modeling enterprise was conditioned to empirical knowledge of geographical case studies and theoretical assumptions to test. Methods and technical tools took also a necessary role, when specific model exploration methods were developed together with the Software OpenMole. R. Reuillon relates that a qualitative shift of knowledge was rapidly made possible when systematic model exploration methods were introduced to understand the behavior of the SimpopLocal model. Initially, geographers were not sure if the model worked at all, in the sense that it produced expected stylized facts such as the emergence of hierarchy in a system of cities. Satisfying trajectories were found for some parameter values through genetic algorithm calibration, with distributed computation on grid [Schmitt et al., 2015]. The existence of multiple candidate so-

¹² <http://openmole.org/>

¹³ page 2, trad. author

lutions for parameter values is a barrier for concrete questions of necessity or sufficiency of a given mechanism of the agent-based model. This need, coming from the domain of empirical and theoretical geographical knowledge, led to the design of a specific algorithm the calibration profile, which is a methodological advance in model exploration [Reuillon et al., 2015].

This virtuous circle was continued with the Marius model family [Cottineau, 2014] and the Parameter Space Exploration algorithm [Chérel, Cottineau, and Reuillon, 2015]. R. Reuillon evaluates its impact from a Computer Scientist point of view: “I’m not sure if [geographers] were immediately conscious of the amplitude of the result, that was really heavy, people working with us directly saw it.” This positive vision is confirmed by D. Pumain, who highlights the benefits of these new methods for geographical knowledge, and that it was the first time that research led to publications at the edge of knowledge both in geography and computer science.

Taking a step back, emerges a typology of domains in which knowledge was created but also necessary for the other domains in the genesis of the Evolutive Urban Theory. The collection of data and construction of datasets is a first requirement for any further knowledge. From data are extracted empirical stylized facts, from which are induced theoretical hypotheses. Theory can then be tested for falsification, in the empirical domain but also through models, for example by doing targeted experiments in models of simulation. New methods are developed to better explore them. Tools are crucial at each step, to implement model, do data mining for example or collect and format data for example. The previous analysis reveals how these domains are interdependent, are in a sense *co-evolutive*.

We back up now this qualitative analysis with a modest quantitative bibliometric analysis. The idea is to investigate the structure of the core citation network of main publications constructing the Evolutive Urban Theory. We construct the citation network as described in Fig. ??, by using the data collection tool provided by [Raimbault, 2016c]¹⁴. Starting from the two seminal publications [Pumain, 1997] and [Sanders et al., 1997], the backward citation network is obtained at depth 2 (references citing these initial references, and the ones citing the citing), with filtering for the first step on authors to have at least one main contributor of the Theory (that we take as *Pumain, Sanders and Bretagnolle*, according to the full Pumain’s interview). We remove nodes of degree 1, to have the core structure only of the ego network. Note that we do not have missing links between nodes at the first level, because all citing links were retrieved.

Network has a density of 0.019, what is rather high for a citation network, and the signature of a high level of dependency between

¹⁴ all code and data are available at
<https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/QuantEpistemo>

publications. Starting from two separate nodes, we could have in theory distinct connected components, but as expected the network has only one because both aspects are strongly interconnected. To analyse the structure in a finer way, we detect communities using Louvain clustering algorithm, and evaluate the directed modularity of the partition as described by [Nicosia et al., 2009].

We show in Fig. ?? a visualization of the network. We obtain 7 communities with a modularity value of 0.39. To ensure the significance of modularity, we proceed to Monte Carlo simulations and randomize citation links 100 times, computing each time the modularity of communities within the randomized network. We obtain an average directed modularity of $\bar{m} = 0.002 \pm 0.015$, making the modularity of the real network highly significant (more than 200 standard deviations).

We analyse the content of communities by looking at publications of the first level. We find that communities are roughly consistent with the typology of domains: one on methods, three on spatio-temporal modeling of urban systems that mixes empirical and modeling, one conceptual, one on Simpop models, and a last on scaling laws that is fully empirical. Data papers are not yet current practice in geography and specific papers tackling the Data domain cant be found in the network. An increased citation rate between papers of the same domain could be expected because of the scientific standard to always situate a contribution regarding similar works. The significant value of modularity confirms that domains are consistent regarding an certain endogenous structure of knowledge production.

Engineering the Metropolitan

After the glance on domains of knowledge extracted in the previous case study, we propose to take the corresponding point of view on a rather different example more related to technology and engineering. We interpret thus issues of engineering related to Parisian metropolitan system through this prism of Knowledge Domains.

Taking the example of the progressive automatization of line 1, considered widely as a technical achievement, several integrated empirical and modeling studies were preliminary conducted [Belmonte et al., 2008]. The use and adaptation of particular methods such as agent-based modeling is crucial for the development of innovative autonomous transportation [Balbo, Adam, and Mandiau, 2016]. In this engineering problem, some technical solutions such as platform doors may be seen as tools that also evolve, and are necessary for a new conceptual approach (*automatic transportation*) to be implemented [Foot, 2005]. But they may also have interactions with other aspects of conceptual knowledge, such as management and organisation within the operator [Foot, 1994]. The complex multi-dimensional aspect of innovation for such systems was already highlighted for a while as [Hatchuel,

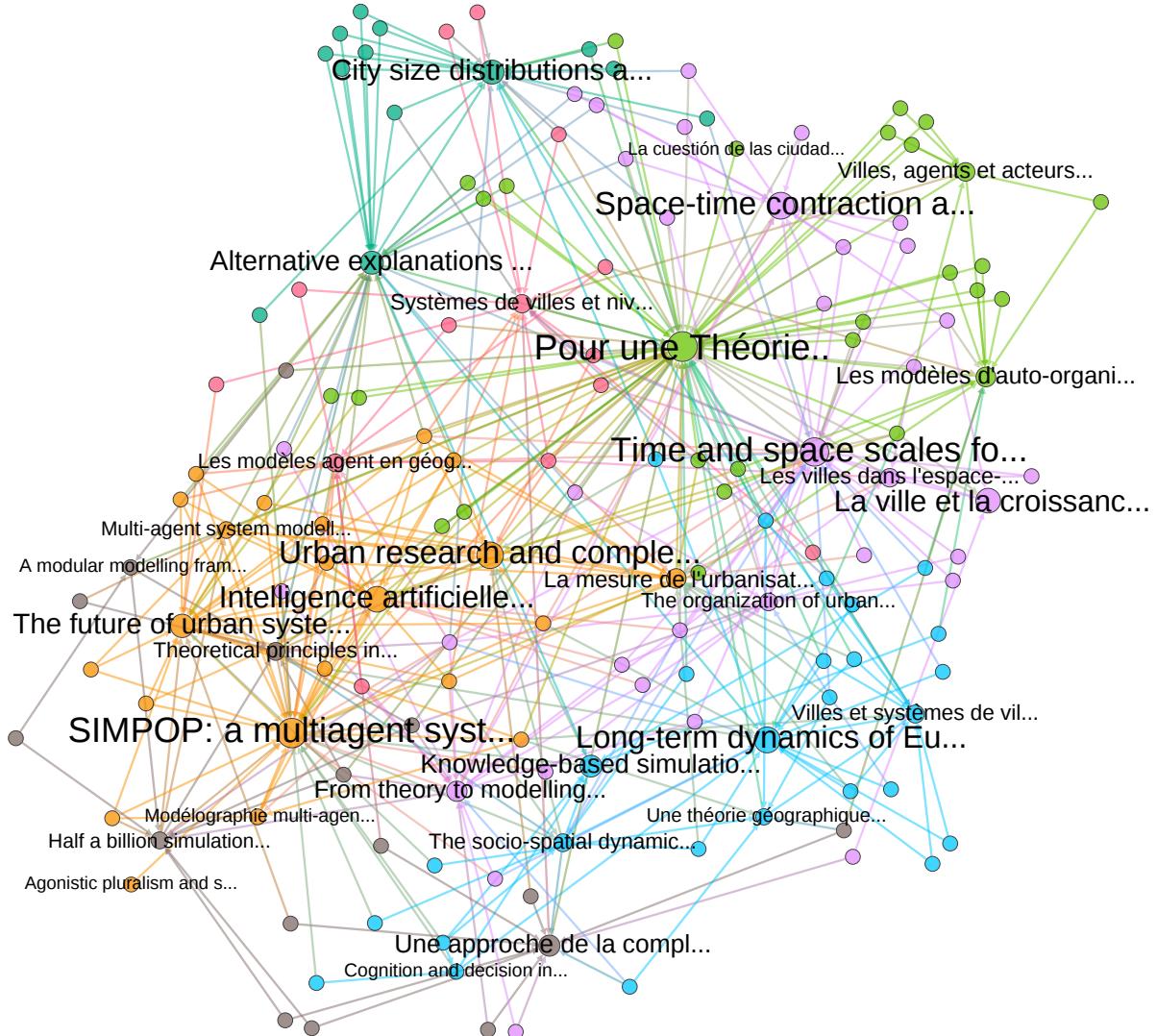


Figure 64: **Citation Network of main publications of Evolutive Urban Theory.** The network is constructed the following way: starting from the two seminal publications [Pumain, 1997] and [Sanders et al., 1997], we get citing publications, filter conditionally to one of the main contributors, get again citing publications and filter. Nodes are publications ($|V| = 155$), the size corresponding to eigenvector centrality, and edges are directed citation links ($|E| = 449$). Colors are communities obtained with Louvain clustering algorithm (7 communities, modularity 0.39).

Pallez, and Pény, 1988] shows. Other technical aspects, such as civil engineering issues [Moreno Regan, 2016], are also put in line when developing such a new approach, and they necessitate at least empirical and modeling, if not more, Knowledge Domains. This rather short example is an illustration of how the interpretation of knowledge domains can be applied to the engineering and management of a complex industrial systems.

Specific details would be needed for a more in-depth application, but we claim to have a proof-of-concept here. We summarize in Table ?? the engineering issues identified above, the corresponding knowledge domains, and the processes through which transferability may be achieved.

Table 20: Illustration of Knowledge Framework Application

Problème d'ingénierie	Domaines de connaissance	Transferabilité	Références
Transport autonome	Empirique, Modélisation	Modélisation intégrée	[Belmonte et al., 2008]
Modélisation innovante	Modélisation, Méthodes	Développement de méthodes	[Balbo, Adam, and Mandiau, 2016]
Spécifications fonctionnelles	Empiriques, Outils	Outils ergonomiques	[Foot, 2005]
Adaptation sociale	Théorique, Empirique	Implication des stakeholders	[Foot, 1994], [Hatchuel, Pallez, and Pény, 1988]
Contraintes techniques	Empirique, Modélisation	Modélisation intégrée	[Moreno Regan, 2016]

8.3.2 Knowledge Framework

We can formulate now inductively the knowledge framework. As mentioned, it takes the idea of interacting domains of knowledge from the framework introduced by [Livet et al., 2010], but extends these domains and takes a novel epistemological position, focusing on co-evolutive dynamics of agents and knowledge.

CONSTRAINTS To be particularly fitted for the study and management of complexity, we postulate that the framework must meet certain requirements, especially to take into account and even favor the *integrative nature of knowledge*, as illustrated by the importance of interdisciplinarity and diversity in the case studies. The framework must thus be favorable to the following:

- Integration of disciplines, as Complex Systems are by essence at the crossing of multiple fields
- Integration of knowledge domains, i.e. that no particular type of knowledge must be privileged in the production process¹⁵
- Integration of methodology types, in particular breaking the artificial boundaries between “quantitative” and “qualitative”

¹⁵ this is not incompatible with very strict system specifications, as multiple paths are possible to obtain the same fixed final state

methods, which are particularly strong in classical social sciences and humanities.

EPISTEMOLOGICAL FOUNDATIONS Le positionnement épistémologique du cadre est celui développé dans la première section de 3.3. Nous rappelons l'importance de la *perspective* [Giere, 2010c], composée des agents, des objets représentés, du but et du medium (le modèle). L'approche par agents est fondamentale pour la cohérence du cadre.

KNOWLEDGE DOMAINS We postulate the following knowledge domains, with their definitions:

- **Empirical.** Empirical knowledge of real world objects.
- **Theoretical.** More general conceptual knowledge, implying cognitive constructions.
- **Modeling.** The model is the formalized *medium* of the scientific perspective, as diverse as Varenne's classifications of models functions [Varenne, 2010b] (see below).
- **Data.** Raw information that has been collected.
- **Methods.** Generic structures of knowledge production.
- **Tools.** Proto-methods (implementation of methods) and supports of others domains.

We choose to keep separate Methods and Tools, to insist on the support role of tools, and because development of both are related but not identical. The same way, Data domain and Empirical Domain are distinct, as new datasets do not systematically imply new knowledge of empirical facts. The Modeling Domain has a central role as we postulate that *any knowledge on a complex system requires a model*.

CO-EVOLUTION OF KNOWLEDGE We can now formulate the central hypothesis of our framework, that is partially contained in the positioning within Perspectivism. We postulate that *any scientific knowledge construction on a complex system¹⁶* is a perspective in the sense of Giere. It is composed of knowledge contents within each domain, that

¹⁶ We believe that this intricate aspect of knowledge production is necessary present for Complex Systems, in echo of the remark on reflexivity in introduction. Even *simple models* of complex systems do imply a conceptual complexity that requires complexity of knowledge to be grasped. This last assumption may be related to the nature of complexity and to the relation between computational complexity and complexity in the sense of weak emergence, that is suggested for example by [Bolotin, 2014] that explains emergence and decoherence from the quantum level by the NP-completude of fundamental equations resolution. These considerations are far beyond the reach of this paper, and we take as an assumption that complex systems necessitate complex knowledge, whereas simple knowledge (in the sense of non co-evolving domains and agents) *can* exist for simple systems.

co-evolve between themselves and with the other elements of the perspective, in particular the cognitive agents. The notion of co-evolution is taken in the sense of [Holland, 2012], i.e. of co-evolving entities being within strongly interdependent niches with circular causal relations and that have a certain independence with the exterior within their boundaries. We note the importance of weak emergence in the sense of Bedau [Bedau, 2002] in the construction of the perspective from the co-evolution of its components, as it corresponds to an autonomous upper level that can be understood alone, as the scientific knowledge can be. Note that a perspective does not necessarily have components in all domains, but should generally have in most.

L'aspect social de la production de connaissance n'est pas inclus dans les domaines de connaissance, mais dans les agents et leur relation. [Roth and Cointet, 2010] montre une co-évolution des réseaux sociaux et des réseaux sémantiques avec l'exemple d'une communauté scientifique en biologie du développement et un environnements de blogs politiques, ce qui confirme dans notre cas la co-évolution entre les agents et les domaines.

APPLICATION The types of models to which our framework applies are supposed to be all possible models in a very loose sense, as Giere calls a model any medium of a perspective. A functional view of models as Varenne introduces [Varenne, 2010b] (introducing a typology of models through functions, e.g. explicative models, simulation models, predictive models, comprehensive models, interactive models, etc.) is a way to grasp the variety. We can also see it in terms of more classical classifications, and apply it to mathematical, statistical, simulation, data or conceptuel models for example. Concerning the constraints given before, as all knowledge are co-evolving no domain is particularly privileged. No discipline either as these will have their different aspects be contained within the domains, and finally qualitative and quantitative methods are present and necessary in most. We show in Fig. ?? a projection of knowledge domains as a complete network, to illustrate what relations between domains can be composed of.

8.3.3 Discussion

Application Range

We insist that our framework does not pretend to introduce a general epistemology of scientific knowledge, but far from that is rather targeted towards reflexivity in the understanding of complex systems. The level of generality is at a very different level, but the aim to practical implication in the handling of complexity contributes to a certain generic character in applications. It is furthermore particularly suited to study Complex Systems, since more reductionist approaches can

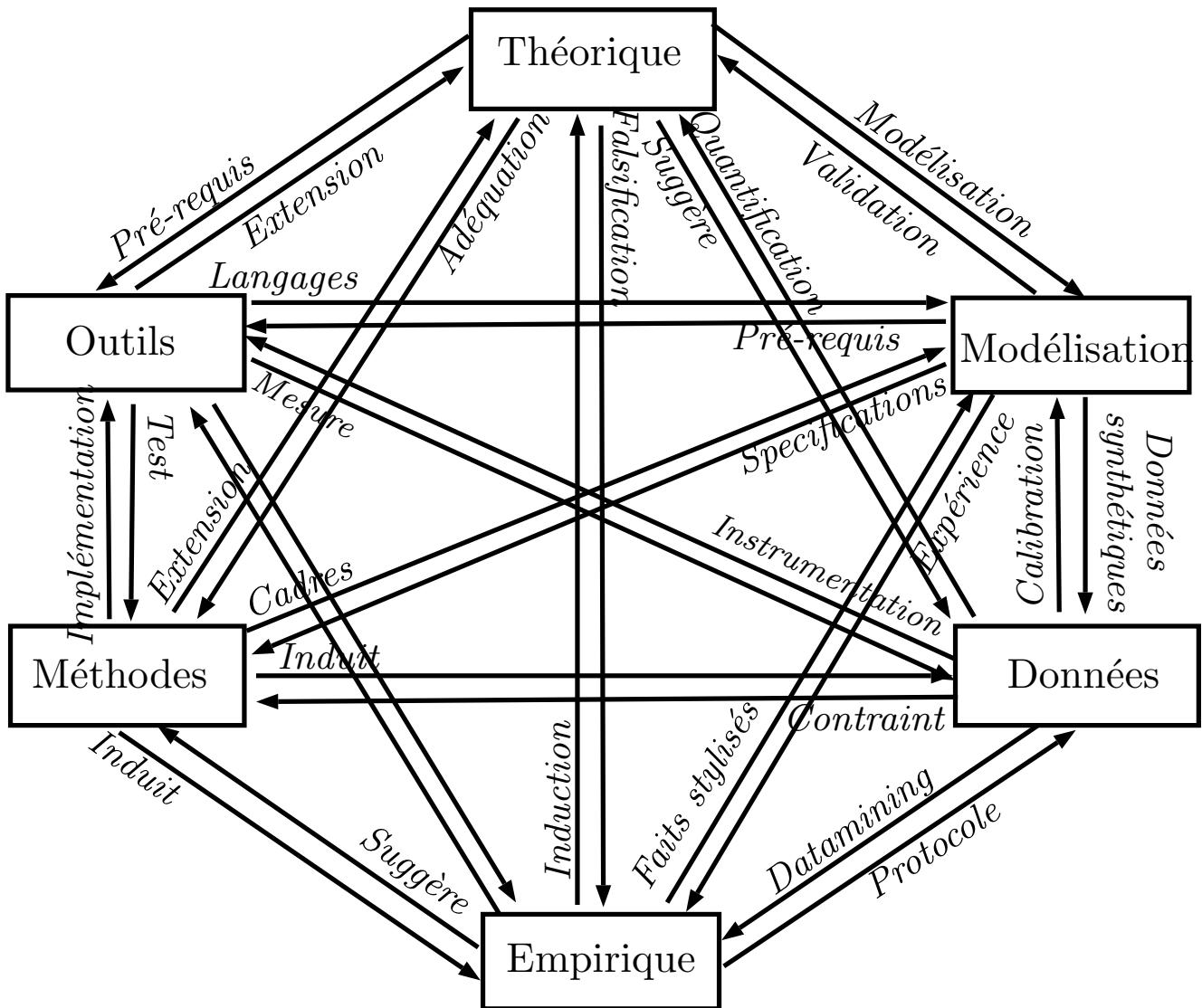


Figure 65: **Projection of a perspective into a full network of knowledge domains.** To illustrate the domains and the interaction processes between them, we do the exercise of trying to qualify all possible binary relations between two given domains. This does not reflect the real structure of the framework, but is an aid to consider what interactions can be. Note that the nature of relations is not always the same here, some being constraints, other knowledge transfer, other processes within other domains such as synthetic data which is a methodology. This shows that some domains act as catalysts for relations between others, in this network setting, what corresponds indeed to a situation of co-evolution.

handle more compartmented production of knowledge, whereas integration of disciplines and scales and therefore domains of knowledge has been emphasized as crucial to study complexity.

Towards a formalisation

The knowledge framework stays at an epistemological level, and its application could be formalized in a more systematic way. We give here a possible direction to achieve that, starting from the coupling of a formalization of the system model with one of the perspective. A perspective would be defined as a dataflow machine M in the sense of [Golden, Aiguier, and Krob, 2012] that gives a convenient way to represent it and to introduce timescales and data, to which is associated an ontology O in the sense of [Livet et al., 2010], i.e. a set of elements each corresponds to an entity (which can be an object, an agent, a process, etc.) of the real world. Purpose and carrier of the perspective are contained in the ontology if they make sense for studying the system. Decomposing the ontology into atomic elements $O = (O_j)_j$ and introducing an order relation between ontology elements based on weak emergence ($O_j \succ O_i$ if and only if O_j weakly emerges of O_i) should yield a canonical decomposition of the perspective containing the structure of the system. The challenge would be then to link this decomposition with the canonical decomposition of the dataflow machine postulated by [Golden, Aiguier, and Krob, 2012], and then define knowledge domains within this coupling: data is in flows of the machine, modeling in the machine, empirical and theoretical in ontologies, methods in the structure of the tree. Such an enterprise with consistent operations is however totally beyond the scope of this paper, but would be a powerful development.

We have studied with mixed method the construction of a scientific theory in theoretical and quantitative geography, and from that inductively introduced a knowledge framework aiming at understanding the production of knowledge on complex system as a complex system itself, namely a perspective with co-evolving components within interdependent knowledge domains. Note that the approach is fully reflexive as several components were necessary. We believe our framework is a useful tool to study complexity and manage complex systems, since it explicits some choices and directions of developments that may otherwise be unconscious.

Co-construction of theories and models: an synthesis of our contributions

Nous concluons ce chapitre d'ouverture par une mise en perspective cohérente des diverses contributions de la thèse, du point de vue de l'illustration de la co-évolution des connaissances dans différents domaines, et de boucler la boucle par un retour sur la construction de la théorie géographique. Comme précisé en préambule, un mode de lecture linéaire serait trop réducteur, puisque la plupart des travaux s'enrichissent mutuellement quel que soit leur domaine et leur portée, et un compte-rendu linéaire, au delà d'être intrinsèquement appauvrissant, est en quelque sorte un mensonge par omission

de l'ensemble des interactions complexes entre les pans de connaissance produite. Bien sûr l'exercice de synthèse et la capacité à faire rentrer dans un cadre formaté imposé, sont louables, voir souhaitables dans l'état actuel des conditions de production scientifiques.

Mais une posture fondamentale que nous prenons et défendons tout au long de ce travail est celle d'une science anarchiste proposée par FEYERABEND, qui sans être prise totalement littéralement et mise en contexte, est extrêmement fructifiante pour proposer des changements de paradigmes et s'émanciper de travaux *mainstream* dont les bases et la légitimité semblent s'enrichir malgré les critiques croissantes. L'écriture d'une monographie extrêmement formatée perd en intérêt de par le caractère contraint de l'exercice, et paraît relativement vaine vu la destinée de sous-utilisation pour une grande partie des travaux actuellement produits, sans être sauvée par la mise en ligne vu la langue imposée¹⁷.

Nous nous prenons à rêver de la possibilité d'une thèse entièrement digitale et dont le cheminement du lecteur tracé dans le support numérique serait à l'origine d'une multitude de visions possibles, traduisant effectivement la complexité du processus de construction, et des perspectives d'enrichissement innombrables par une rétroaction et une interaction avec les lecteurs, c'est-à-dire sortir du mode de présentation linéaire, comme déjà soutenu en introduction. L'invention de nouveaux modes de communication scientifiques¹⁸ est un défi urgent à part entière, et notre ébauche de réflexivité développée en Annexe F cherche à y contribuer.

La construction de théories géographiques, dans le cadre d'une géographie théorique et quantitative, s'effectue par itérations dans une dynamique de co-évolution avec les efforts empiriques et de modélisation [Livet et al., 2010]. Parmi les nombreux exemples, on peut citer la théorie évolutive des villes (co-construite par un spectre de travaux s'étendant par exemple des premières propositions de [Pumain, 1997] jusqu'aux résultats matures présentés dans [Pumain, 2012a]), l'étude du caractère fractal des structures urbaines (par exemple de [Frankhauser, 1998] à [Frankhauser, 2008]) ou plus récemment le projet Transmondyn [Sanders, 2017] visant à enrichir la notion de transition des systèmes de peuplement. Nous proposons ici une synthèse de différents travaux empiriques et de modélisation menés conjointement dans ce travail avec l'élaboration d'appareils théoriques

¹⁷ Ce qui relève bien sûr par ailleurs d'une problématique bien plus complexe que la simple audience [Tardy, 2004] et la richesse des pensées scientifiques permises par l'utilisation de différentes langues n'est pas discutable ainsi que la légitimité d'organisations comme l'ASRDLF. Mais c'est bien cette audience qui nous pose problème ici et dans ce cas il est quasiment aussi vieux jeu pour une école doctorale d'imposer le français comme langue d'écriture qu'un choix de consul à imposer un discours en français à une audience non-francophone.

¹⁸ La communication scientifique interne et externe est un défi à part entière, comme le rappelle [Martinez-Conde and Macknik, 2017] qui propose un véritable *storytelling* des résultats de la recherche scientifique.

visant à mieux comprendre les relations entre territoires et réseaux de transports.

FOR A THEORY AND MODELS OF COEVOLUTION Notre première entrée prend un point de vue d'épistémologie quantitative pour tenter d'expliquer le fait que, si la co-évolution entre territoires et réseaux a par exemple été prouvée par [Bretagnolle, 2009], la littérature est très pauvre en modèles de simulation endogenisant cette co-évolution. Une exploration algorithmique de la littérature a été menée dans [Raimbault, 2017d], suggérant un cloisonnement des domaines scientifiques s'intéressant à ce sujet. Des méthodes plus élaborées ainsi que les outils correspondants (collecte et analyse des données), couplant une analyse sémantique au réseau de citations, ont été développées pour renforcer ces conclusions préliminaires [Raimbault, 2016c], et les premiers résultats au second ordre semblent confirmer l'hypothèse d'un domaine peu défriché car à l'intersection de champs ne dialoguant pas nécessairement aisément. Ces premiers résultats d'épistémologie quantitative confirment l'intérêt d'une modélisation couplant des processus relevant de différentes échelles et domaines d'études, et surtout l'intérêt de l'élaboration d'une théorie propre.

EMPIRICAL STUDIES Le premier axe pour les développements en eux-mêmes consiste en des analyses empiriques. Une étude des corrélations spatiales statiques entre mesures de forme urbaine (indicateurs morphologiques calculés sur la grille de population eurostat) et mesures de forme de réseau (topologie du réseau routier issu d'OpenStreetMap), sur l'ensemble de l'Europe à différentes échelles, a pu révéler la non-stationnarité et la multi-scalarité spatiale de leurs interactions [Raimbault, 2016a]. Cet aspect a aussi été mis en évidence dans l'espace et le temps à une échelle microscopique lors de l'étude des dynamiques d'un système de transport [Raimbault, 2017b], conjointement avec l'hétérogénéité des processus pour un autre type de système [Raimbault, 2015]. Ces faits stylisés valident l'utilisation de modèles de simulation complexes, pour lesquels des premiers efforts de modélisation ont ouvert la voie vers des modèles plus élaborés.

MODELING A l'échelle mesoscopique, des processus d'agrégation-diffusion ont été montrés suffisant pour reproduire un grand nombre de formes urbaines avec un faible nombre de paramètres, calibrés sur l'ensemble du spectre des valeurs réelles des indicateurs de forme urbaine pour l'Europe. Ce modèle simple a pu, à l'occasion d'un exercice méthodologique explorant le possibilité de contrôle au second ordre de la structure de données synthétiques [Raimbault, 2016b], être couplé faiblement à un modèle de génération de réseau, démontrant une grande latitude de configurations potentiellement générées. L'exploration de différentes heuristiques autonomes de génération de

réseau a par ailleurs été menée, pour comparer par exemple des modèles de croissance de réseau routier basés sur l'optimisation locale à des modèles inspirés des réseaux biologiques : chacun présente une très grande variété de topologies générées. A l'échelle macroscopique, un modèle simple de croissance urbaine calibré dynamiquement sur les villes françaises de 1830 à 2000 (base Pumain-Ined) a permis de démontrer l'existence d'un effet réseau de par l'augmentation de pouvoir explicatif du modèle lors de l'ajout d'un effet des flux transitant par un réseau physique, tout en corrigeant le gain dû à l'ajout de paramètres par la construction d'un Critère d'Information d'Akaike empirique [Raimbault, 2016d]. Cet ensemble de modèles se positionne avec un objectif de parcimonie et dans une perspective d'application en multi-modélisation. Dans une démarche multi-agents plus descriptive et donc par un modèle plus complexe, [Le Néchet and Raimbault, 2015] décrivent un modèle de co-évolution à l'échelle métropolitaine (modèle Lutecia) qui inclut en particulier des processus de gouvernance pour le développement des infrastructures de transport. Pour ce dernier modèle, les premières études de la dynamique montrent l'importance du caractère multi-niveau du développement du réseau de transport pour obtenir des motifs complexes de réseaux et de collaboration entre agents. L'ensemble de ces efforts de modélisation supportent les fondements théoriques que nous avons proposé par la suite.

CONSTRUCTION OF A GEOGRAPHICAL THEORY Nous revoyons enfin sous l'oeil de la co-evolution des domaines la théories construite en 8.2. Nous insistons ici sur son caractère intégratif permettant de joindre théorie évolutive des villes et morphogenèse. En se basant sur les travaux précédents, nous proposons de joindre deux entrées pour la construction d'une théorie géographique ayant un focus privilégié sur les interactions entre territoires et réseaux. La première est par la notion de *morphogénèse*, qui a été explorée d'un point de vue interdisciplinaire dans [Antelope et al., 2016]. Pour notre part, la morphogénèse consiste en l'émergence de la forme et de la fonction, via des processus locaux autonomes dans un système qui exhibe alors une architecture auto-organisée. La présence d'une fonction et donc d'une architecture distingue les systèmes morphogénétiques de systèmes simplement auto-organisés (voir [Doursat, Sayama, and Michel, 2012]). De plus, les notions d'autonomie et de localité s'appliquent bien à des systèmes territoriaux, pour lesquels on essaye d'isoler les sous-systèmes et les échelles pertinentes. Les travaux sur la génération de forme urbaine calibrée par des processus autonomes, les premiers travaux sur la génération de réseaux par de multiples processus également autonomes, et des travaux plus anciens étudiant un modèle simple de morphogénèse urbaine qui suffisait à reproduire des motifs de forme stylisés [Raimbault, Banos, and Doursat, 2014], nous

suggèrent la possible existence de tels processus au sein des systèmes territoriaux.

D'autre part, le cadre de théorie évolutive des villes est plébiscité par nos résultats empiriques, qui montrent le caractère non-stationnaire, hétérogène, multi-scalaire des systèmes urbains. Pour rester le plus général possible, et comme nos résultats à la fois empiriques et de modélisation (génération de formes quelconques par le modèle d'agrégation-diffusion par exemple) s'appliquent aux systèmes territoriaux en général, nous nous plaçons dans le cadres de territoires humains de [Raffestin, 1987], c'est-à-dire "*la conjonction d'un processus territorial avec un processus informationnel*", qui peut être interprété dans notre cas comme le système complexe socio-techno-environnemental que constitue un territoire et les agents et artefacts qui y interagissent. L'importance des réseaux est soulignée par nos résultats sur la nécessité du réseau dans le modèle de croissance macroscopique : nous proposons alors de parler de *systèmes territoriaux complexes en réseau*, en ajoutant au plongement du territoire dans la théorie évolutive la particularité qu'il existe des composantes cruciales qui sont les réseaux (de transport en l'occurrence), dont l'origine peut être expliquée par la théorie territoriale des réseaux de [Dupuy, 1987].

Nous proposons alors l'hypothèse suivante afin de réconcilier nos deux approches : *l'existence de processus morphogénétiques dans lesquels les réseaux ont un rôle crucial est équivalente à l'existence de sous-systèmes dans les systèmes territoriaux complexes en réseaux, qu'on définit alors comme co-évolutifs*.

Cette proposition a de multiples implications, mais a typiquement guidé notamment les choix de modélisation vers une méthodologie modulaire et de multi-modélisation afin d'essayer d'exhiber des processus morphogénétiques, ainsi que les travaux empiriques vers une étude plus poussée des corrélations, causalités (dans le cas de séries temporelles) et recherche de décompositions modulaires des systèmes.

* * *

*

CHAPTER CONCLUSION

Ce chapitre nous a permis ainsi de prendre du recul sur nos contributions et de les mettre en perspective. Il ouvre en fait de nombreuses portes, et fait prendre conscience que la portée des connaissances reste embryonnaire.

Les questions soulevées par chacun des niveaux sont fondamentales pour l'étude des systèmes territoriaux complexes mais aussi des systèmes complexes en général. La théorie proposée en 8.2 pointe à nouveau la question de la non-stationnarité spatio-temporelle dans un contexte multi-échelle, que nous postulons cruciale mais peu explorée dans le cas des systèmes territoriaux. Nous distinguons également la difficulté d'intégration de théories existantes ce qui implique une compréhension des processus de couplage des modèles.

Ce problème est au cœur du cadre formel développé par la suite B.5, qui soulève aussi des questions d'imbrication d'échelles. Le problème d'obtenir une structure algébrique cohérente avec une action de monoïde sur les données implique une intégration de la théorie de KROB, ce qui questionne plus généralement l'intégration des approches d'ingénierie système (systèmes complexes "industriels") avec celles de systèmes complexes naturels.

La possibilité de théorie intégratives est soulevée par l'introduction du cadre de connaissance 8.3, qui pose également des problèmes plus généraux de production des connaissances et de nature de la complexité que nous avions brièvement abordé d'un point de vue épistémologique en 3.3.

Nous proposons de synthétiser une partie de ces diverses questions ouvertes dans un projet de recherche cohérent sur un long terme mais incluant des premières pistes concrètes immédiates, que nous présenterons en ouverture.

* * *

*

OUVERTURES GÉNÉRALES

Comme nous l'avons suggéré précédemment, l'ouverture permet en fait une prise de recul et dans notre cas une clarification du cadre global. Nous proposons donc ici l'exercice de recension des travaux d'ouverture déjà menés, celle des problèmes ouverts par notre travail, et leur synthèse dans un projet de recherche à long terme.

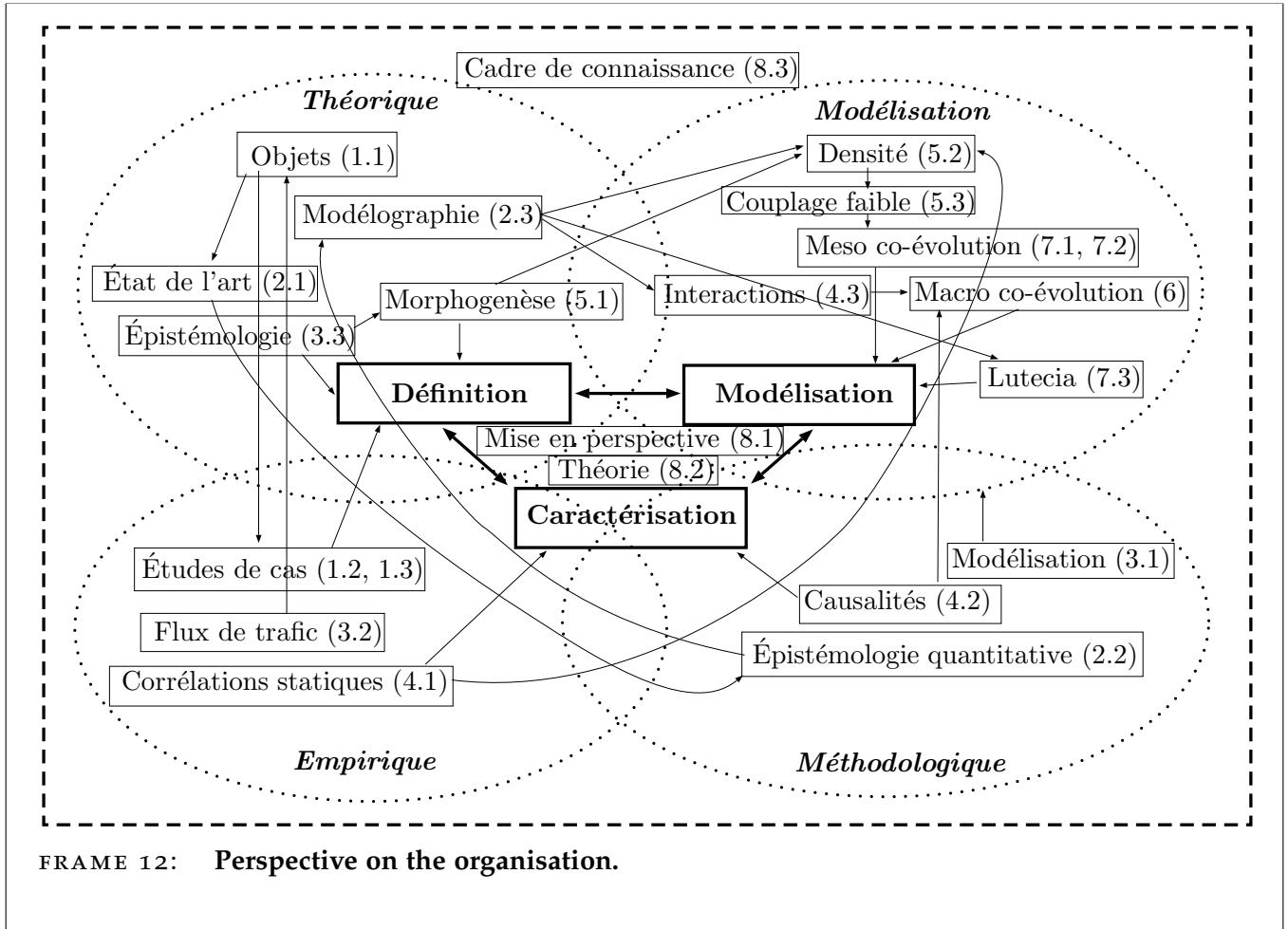
THEMATIC AND GENERAL PERSPECTIVES

Global perspective

Une relecture de la thèse à la lumière de l'articulation théorique proposée en 8.2 nous confirme que (i) l'approche morphogénétique était naturellement induite par la contrainte de niche écologique dans la définition de la co-évolution ; (ii) la théorie évolutive des villes est ainsi précisée pour le cas précis de la co-évolution ; (iii) les systèmes territoriaux doivent intrinsèquement induire de tels processus, puisqu'ils sont à la fois support et objets de ceux-ci. La question de la nécessité des réseaux pour représenter les systèmes territoriaux reste ouverte, et nous l'avons postulée dans notre construction théorique. Nos résultats suggèrent la pertinence de leur prise en compte, et ouvrent la question d'une démonstration de ce postulat.

Ensuite, une relecture par les domaines de connaissance permet de mieux comprendre l'articulation entre les différentes composantes : les constructions conceptuelles et empiriques de la première partie permettent une définition de la co-évolution, puis la mise en place de méthodes et de modèles en deuxième partie, qui en retour alimentent ces autres domaines en troisième partie. Nous proposons une analyse quantitative brève de ces dynamiques en F. Ainsi, l'interdépendance dans le cheminement, donnée par le diagramme en introduction (Encadré 1), est en fait bien plus complexe et pas forcément linéaire. Une deuxième lecture de notre monographie sera ainsi plus riche, par émergence des liens implicites. Nous proposons en Encadré ?? une relecture possible de l'organisation de notre travail, au regard de la problématique générale et des domaines de connaissance.

Notre travail peut également se placer dans une perspective plus large. Précisons la "méta-articulation" de notre travail, c'est-à-dire la structure implicite des divers développements et ouvertures et donc le cadre global dans lequel s'inscrit le cœur (trois premières parties). L'Encadré ?? schématise cette articulation. Le cœur, qui consiste en la réponse à la problématique, est constitué de trois axes en interaction forte : la définition, la caractérisation et la modélisation de la



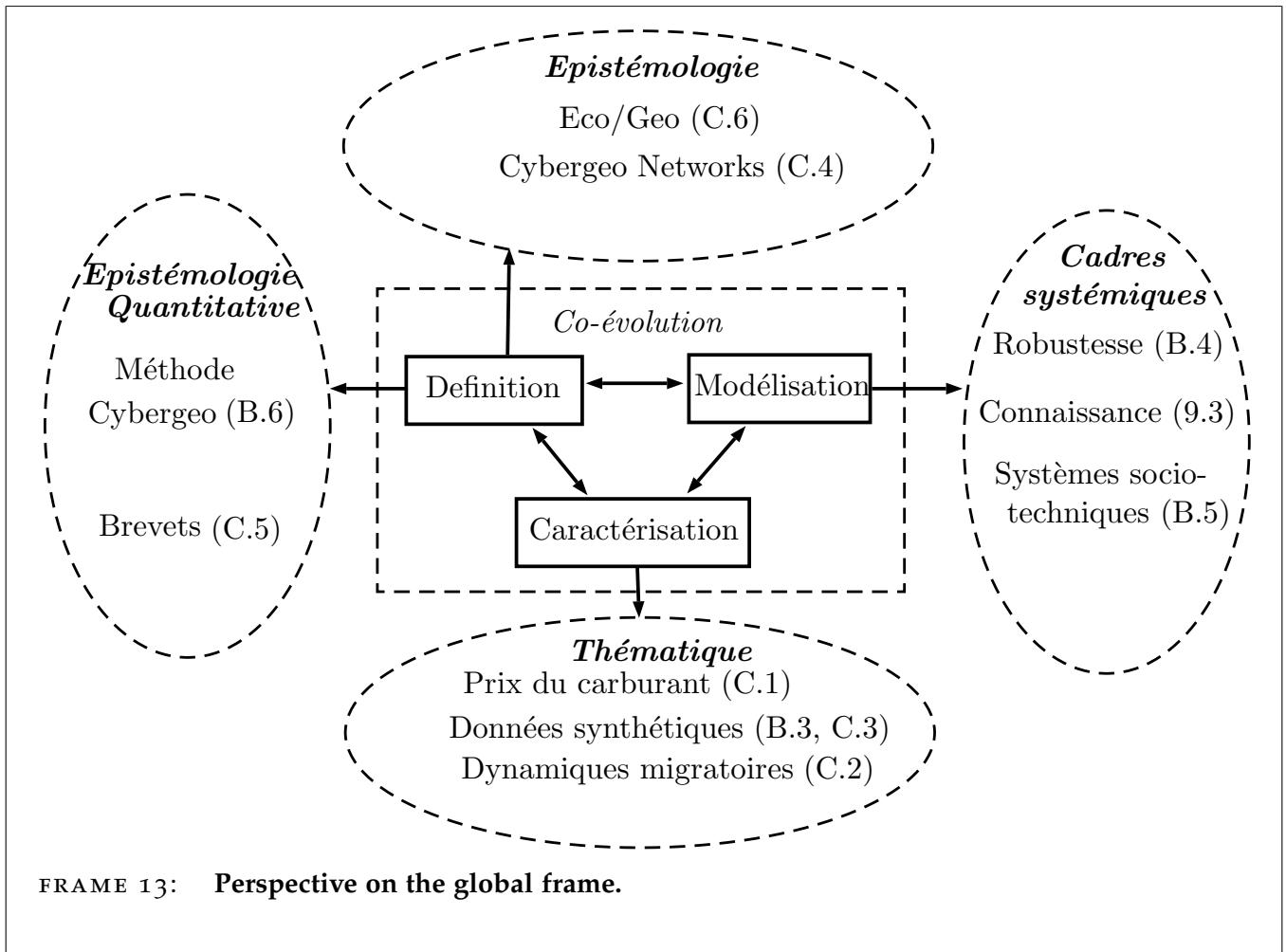
co-évolution des réseaux de transport et des territoires. Chacun appelle à sa manière des développements dans divers champs¹⁹ : des développements épistémologiques et en épistémologie quantitative, principalement liés à l'aspect de définition ; des développements de cadres systémiques, induits par les problématiques liées à la modélisation ; et des développements thématiques liés à la caractérisation.

Détaillons le contenu de chacun de ces développements, en les reliant au contenu correspondant principalement en Annexes :

1. Epistémologie quantitative : principalement en lien avec les méthodes et outils de revue systématique et d'exploration d'un paysage scientifique en 2, nous incluons le cas d'étude original qui a initié la méthode, le corpus du journal Cybergeo, en B.6, ainsi qu'une application à un corpus massif de brevets en C.5.
2. Epistémologie : la mise en contexte de l'étude de Cybergeo avec d'autres approches complémentaires permet une prise de recul épistémologique dans C.4 ; nous amorçons également une réflexion sur les liens entre économie et géographie en C.6.
3. Cadres systémiques : un cadre de connaissance, contribuant à organiser une connaissance complexe, a déjà été proposé en 8.3 ; un cadre formalisant le couplage des modèles des systèmes socio-techniques, suggérant des pistes de formalisation du cadre de connaissance, est développé dans B.5 ; un cadre pour l'étude de la robustesse des évaluations multi-attributs est développé dans B.4.
4. Thématique : les études de cas des systèmes de transport effectuées en 3.2 et en C.1 permettent en l'occurrence une confirmation des échelles pertinentes ; l'étude de la génération de données synthétiques, en lien avec la méthodologie développée en 3.1, est faite en B.3 pour la méthode et en C.3 pour un exemple d'application ; la modélisation des dynamiques migratoires au sein du Delta de la Rivière des Perles ébauchée en C.2, introduit une piste de modèles multi-échelle et raffine les interactions entre villes au niveau des flux individuels.

Ces différents champs sont bien sûr à intersections non vides (le cadre de connaissance de 8.3 relève par exemple à la fois du cadre systémique et de l'épistémologie, ou l'étude des brevets est un important aspect thématique en lien avec la théorie évolutive des villes) et en interactions : les études d'épistémologie quantitative informent l'épistémologie, qui guide les études thématiques, qui peuvent être mises en perspective dans les cadres systémiques, qui eux dépendent également du positionnement épistémologique.

¹⁹ Nous n'utilisons pas le terme domaine ici pour ne pas entraîner une confusion avec les domaines de connaissance, ceux-ci étant mobilisés différemment comme nous le verrons en Annexe F.



Ainsi, nous mettons en évidence une structure plus globale pour notre travail, qui dessine en partie la structure d'un projet de recherche que nous détaillerons par la suite.

Open questions

Nous développons à présent des questions fondamentales qui ont été abordées ou ouvertes tout au long de notre travail, que nous classons en trois axes : pratique scientifique (épistémologie appliquée), modélisation, et fondements des systèmes complexes spatiaux.

Applied Epistemology

A TRULY OPEN SCIENCE Un premier axe de développement crucial pour l'ensemble de l'écosystème de production de connaissance dans lequel nous nous inscrivons (voir chapitre 3) est la contribution à une ouverture maximale de la pratique scientifique, c'est-à-dire la combinaison de l'ensemble des approches résumées par [Fecher and Friesike, 2014], en particulier les aspects démocratique et public qui encouragent l'accès de tous à la production de connaissance et à ses résultats²⁰, et les aspects pragmatique et d'infrastructure qui appuient l'efficience augmentée dans un cadre ouvert.

La transparence et mise en disponibilité des données brutes ou au moins pré-traitées, et du code informatique produisant les sorties de simulation ou les figures, semble être plutôt l'exception que la règle en géographie. Comme le rappelle [Banos, 2013] qui y dédie l'un de ses principes, "*le modélisateur n'est pas le gardien de la vérité prouvée*", et comme rappelé en 3.2, une reproductibilité parfaite des résultats est nécessaire pour une reconnaissance d'une quelconque valeur par la communauté scientifique, comme une théorie qui ne fournit pas de possibilité de falsification ne peut être considérée comme scientifique au sens de POPPER. Des expériences de revue pour *Cybergeo* ont confirmé à l'unanimité ce problème fondamental. Rappelons que la revue *PNAS* exige les données brutes et tableau produisant toute figure, pour prévenir tout biais de visualisation qu'il soit volontaire (ce qui est rédhibitoire et conduit à un signalement) ou non.

Par ailleurs, la communication scientifique est un aspect important de la science ouverte. Le mode actuel de publication scientifique est loin d'être idéal. Un article n'est pas un format compréhensible ni vraiment reproductible, et pousse au biais. L'écriture d'un article en répondant aux normes de façon à être accepté peut être assimilé à "un jeu" dont les règles sont subtiles et qu'il faut maîtriser pour faire carrière. Selon notre positionnement, un tel mode de communication est contraire à l'honnêteté et l'intégrité intellectuelle néces-

²⁰ Sachant que l'ouverture des produits de la connaissance est évidente dans une perspective complexe, puisque comme le souligne [Morin, 1991], nos idées prennent une certaine indépendance dans la noosphère et ne nous appartiennent pas.

saires à une science éthique et ouverte. Les initiatives se multiplient pour proposer des modèles alternatifs : la revue post-publication en est une, l'utilisation de systèmes de contrôle de version et de dépôts publics une autre, ou la publication éclair de pistes de recherche²¹. Par exemple, [Michaël, Michael, and McDowell, 2017] décrit une expérience d'articles dynamiques évalués de manière ouverte par la communauté, avec des métriques associées permettant de faire émerger les travaux jugés intéressants.

De la même façon, nous soutenons qu'une présentation linéaire d'un projet de recherche est trop fortement réducteur, et que l'invention de modes de communication alternatifs est un enjeu futur pour la science ouverte. On peut par exemple imaginer des réseaux interactifs, traduisant la structure de la connaissance sous-jacente, et dans lesquels le lecteur peut naviguer entre les concepts et les analyses, être renvoyé directement vers les données, modèles et analyses. Les grilles de lecture principales en accord avec l'argument que prendrait une explication linéaire peuvent alors être superposées au réseau pour revenir à un mode plus classique de lecture. Une communication par le jeu est également une alternative crédible, notamment dans le cas d'une communication pour le public, et nous en donnons une illustration pour un problème d'écologie en Annexe C.7.

FOR AN EVIDENCE-BASED SCIENCE Nous postulons qu'une science entièrement *evidence-based*, quel que soit son objet, est possible et souhaitable en articulation avec la science ouverte. L'idée est de chercher à déconnecter la connaissance scientifique de tout dogmatisme, de tout *a priori* politique et de tout jugement de valeur²². Dans le cas de l'étude de sujets en lien avec des individus ou des sociétés (c'est-à-dire les sciences humaines), un tel positionnement n'est possible selon [Morin, 1991] que par le passage par l'établissement d'un "méta-point de vue", c'est-à-dire par une certaine réflexivité qui permet au connaisseur de comprendre sa position et sa propre démarche. Nous donnons des pistes pour la construction de tels points de vue,

²¹ Voir par exemple le *Journal of Brief Ideas* à <http://beta.briefideas.org/about>. Les descriptions courtes de pistes de recherche sont souvent reléguées à la discussion ou la conclusion des articles, qui s'écrivent de manière conventionnelle, souvent avec un biais pour justifier *a posteriori* l'intérêt de *sa nouvelle méthode* qu'il faut malheureusement vendre. On fait alors des plans sur la comète, propose des développements ayant peu de rapport, ou des domaines d'application *qui auront un impact* (lire qui sont à la mode ou qui reçoivent le plus de financements à la période de l'écriture). Ce manuscrit tombe bien évidemment partiellement sous ces critiques, comme les articles qui lui sont associés.

²² Sachant que par ailleurs ceux-ci doivent être plus que jamais développés et réfléchis pour articuler la science avec la société, mais doivent le moins possible interférer avec le processus de production de connaissance en lui-même. Suivant [Morin, 2004], une éthique de la connaissance et une pensée complexe induit naturellement une éthique plus large, permettant l'autonomie de la connaissance scientifique sans la rendre inhumaine.

sous la forme de ce que nous appelons *perspectivisme appliqué*, en Annexe C.4 ainsi qu'en Annexe B.5 pour une piste de formalisation.

Cette problématique est directement reliée à la question récurrente de la dichotomie “qualitatif-quantitatif”, que nous jugeons peu pertinente dans le cadre de sciences intégratives. En effet, si la dichotomie se base sur une différence entre objectif et subjectif, nous rappelons que toute connaissance est subjective, et que celles où le rôle du sujet est particulièrement déterminant peuvent “s’objectiver” par la prise du méta-point de vue, par exemple par le couplage avec d’autres approches, c’est-à-dire précisément par la prise d’une position intégrative. Si elle se base sur une question de nature des données, elle n’est que partiellement pertinente puisque la limite est floue : un texte d’interview peut très bien faire l’objet d’analyse textuelle alors qu’une régression doit être interprétée qualitativement. En fait, nous pensons qu’il existe différentes méthodes plus ou moins appropriées selon la connaissance à produire (voir par exemple [Gros, 2017] qui fustige l’utilisation de statistiques inférentielles pour un corpus ethnographique), mais qu’il n’y a pas “chasse gardée” de telle discipline sur telle méthode et que les couplages et transferts seront toujours plus nécessaires à l’avenir.

QUANTITATIVE EPISTEMOLOGY Les points précédents doivent être traités conjointement avec l’utilisation de méthodes d’épistémologie quantitative permettant une réflexivité accrue, comme par exemple la méthode par hyperréseau utilisée en 2.2, appliquée au corpus Cybergeo en B.6, à un corpus de brevets en C.5 et à notre propre travail en F. La plateforme CybergeoNetworks²³ est une collaboration dans cette direction, présentée en détails en C.4. Elle permet notamment la prise d’autonomie par les auteurs mais également par les journaux libres qui peuvent alors rivaliser avec les entreprises prédatrices d’édition qui valorisent à leur profit les analyses de corpus.

Modeling

Sur le plan de la méthodologie de la modélisation, nous donnons des axes précis complémentaires à ceux mis en place par [Pumain and Reuillon, 2017d] (multi-modélisation, exploration des modèles).

COUPLING MODELS La définition du couplage de modèles ou d’approches, et notamment du degré de couplage (couplage fort ou couplage faible) dépend des cadres utilisés et n’a pas forcément de fondement théorique. La construction de théories permettant une telle définition qui serait par ailleurs opérationnelle est une question ouverte. Une approche possible utilise par exemple les rapports entre complexités de Kolmogorov des différents modèles concernés. Une approche formelle

²³ Accessible à <http://shiny.parisgeo.cnrs.fr/cybergeonetworks/>.

est donnée en B.5 pour le couplage de perspectives. Cette approche est profondément liée aux questions épistémologiques, puisqu'il pourrait s'agir d'une manière de formaliser la logique du cadre de connaissance.

La question du couplage de modèles hétérogènes est bien sûr liée : dans quelle mesure est-il pertinent de choisir tel ou tel type de modèle et comment les coupler ? [Banos et al., 2015] l'illustrent pour un modèle épidémiologique, couplant un modèle classique par équations différentielles à un modèle de microsimulation. Le lien entre modèles agents et systèmes dynamiques peut être établi dans certaines configurations, comme nous l'avons fait pour le modèle de Simon et le modèle de Gibrat en B.1, mais la question de classes de problèmes pour lesquels des liens seraient systématiques ou non reste une question ouverte.

Enfin, la nécessité du benchmarking de modèles comparables a été soulevée depuis un certain temps [Axtell et al., 1996], mais reste très peu appliquée : le développement d'outils et de méthodes facilitant de telles comparaisons est également un point important.

EMPOWERING MODELS OF SIMULATION WITH VALIDATION AND ASSESSMENT TOOLS L'essentiel de l'entreprise d'OpenMole est orientée dans ce but de construction d'outils et de méthodes pour la validation des modèles. Nous contribuons à cet effort dans notre travail, par exemple en 4.3 par la construction d'un critère de sur-ajustement, ou en B.4 par l'élaboration d'une mesure de la robustesse d'un modèle aux données manquantes. L'étude du comportement des modèles par rapport au sur-ajustement, notamment dans le cadre de la multi-modélisation, est un enjeu fondamental pour le développement futur de ces approches.

Foundations of spatial complex systems

Certaines questions fondamentales ont été suggérées au sujet des systèmes complexes ayant une structure spatiale.

NON-STATIONARITY, NON-ERGODICITY AND PATH-DEPENDANCY

Le lien entre non-stationnarité spatiale et/ou temporelle et non-ergodicité, pouvant éclairer les propriétés de dépendance au chemin, n'a à notre connaissance pas été étudié systématiquement, au moins dans le cadre des systèmes territoriaux. Nous suggérons qu'un lien accru entre géosimulation, statistiques spatiales et économie géographique, contribuerait à la compréhension de ce type de question.

MULTI-SCALE MODELS Comme nous l'avons déjà amplement répété, il existe très peu de modèles des systèmes territoriaux effectivement multi-échelle, et leur développement à des échelles pertinentes et à

un degré de complexité raisonnable, est également un défi futur important.

METHODOLOGICAL STANDARDS Enfin, un effort considérable doit être fait, particulièrement en géographie, pour respecter des standards méthodologiques a minima : par exemple utilisation de classifications de séries temporelles appropriées [Liao, 2005], ajustement de loi puissances sur des données empiriques selon la méthode standard de [Clauset, Shalizi, and Newman, 2009] et non une simple régression des moindres carrés, utilisation de modèles non-linéaires si besoin.

★ ★

★

TOWARDS A RESEARCH PROGRAM

For an Integrated Geography

Research project

Nous détaillons finalement un projet de recherche à long terme qui (i) s'inscrit dans la continuité de cette monographie ; (ii) s'inscrit dans le cadre d'une géographie intégrée, et plus généralement d'une intégration verticale et horizontale, mais aussi des domaines et des types de connaissance ; (iii) s'attaque à un certain nombre de questions ouvertes mentionnées ci-dessus ; et (iv) est intrinsèquement réflexif et complexe.

The aim would be to solve a multi-scale geographical problem, that is to understand how and when interdependencies between cities have built regional systems of cities and to identify the most probable scenario of their potential coalescence as a consequence of globalisation processes. These high-level questions have direct practical implications for measuring global and local inequalities and managing urban growth.

The principal question we propose to investigate finds roots in the multi-scalar nature of territorial systems. Converging evidence suggest the relative independent historical development of regional urban systems across the world, and an increased interdependency between these in the processes of globalisation. Can we already quantify these at different scales ? How does the coupling and the opening of subsystems operate, and what are its most plausible conse-

quences, from convergence of dynamics to an increase of inter- and intra-subsystems inequalities ?

We postulate that a powerful entry to this research question is the construction of bridges between geographical theories of territorial systems in the spirit of the Evolutive Urban Theory and Scaling Theories of Cities. The first emphasize particularities of territorial entities whereas the second focuses on universal laws, and both provide credible explanations for scaling laws. A strategy to answer the question and combining both would consist in: (i) finding endogenous modular decompositions of territorial systems and corresponding scales, and quantifying their universality through inter and intra scaling; (ii) modeling this multi-scalar system by coupling models of urban growth, that would be validated through scaling properties. The models developed here are good candidates as sub-models, since co-evolution inside and between scales is a characteristic feature of complex urban systems.

An auxiliary research direction that I will conjointly tackle is the exploration of potential relations between territorial systems and artificial intelligence. It comes naturally as a corollary and is informative for the main question, for at least two very different reasons. The first is rather practical and linked to the emergence of ubiquitous information and computing in cities, that some observers design as "smart cities": the new large datasets available have been proven to be a powerful analysis tool as witness the numerous recent works by physicists on cities for example, and these new urban behavior may probably induce some regime changes partly because of of their self-fulfilling nature. The second is more difficult to grasp: the importance of morphogenesis in my current understanding of territorial systems and the possible application of this concept at different levels such as knowledge production. Morphogenesis can be used to conceptualize both the evolution of territories and of ideas: to what extent the emergence of territories contains an endogenous intelligence. The success of using slime mould network generation in my thesis, which have been shown otherwise to be powerful computation tools, is an other clue of a possible connexion.

A second auxiliary subject is the theoretical and applied study of knowledge production on Complex Systems. This axis will be necessary to the project, first to continue to enhance the reflexivity and interdisciplinarity through the further development of quantitative epistemology methods and tools such as more elaborated text-mining and meta-analysis tools, and secondly precisely because of reflexivity as concrete case studies such as the aforementioned language evolution precisely apply to territorial systems which main components are cognitive agents.

The strongly coupled elaboration of these different components, i.e. their co-evolution, in the exact spirit of what I achieved until now,

is necessary for the integrated nature of the project and achieve its objective of integrative theories.

★ ★

★

* * *

*

*Cet hiver a des airs de printemps
Des peuples ou de l'esprit, au diable l'âme.
Le vent se lève, ça faisait longtemps
Triste de s'enfermer pour quelques grammes.*

*Cet avenir des airs de passé
S'il fallait juste trouver le régime,
Assassinée la complexité
Maintes perspectives se cachent en les crimes.*

*Pour une morphogenèse politique
Adieu le coron, ses tristes briques
Murs qui s'érigent tuent votre espérance.*

*Perle de la mer, sirène hante la crique
Du haut des tours s'amuser du cirque
L'hiver d'idées qui peuple la France.*

* * *

*

CONCLUSION

*Explorer sans relâche les systèmes géographiques...
- ARNAUD BANOS*

Notre thèse est un système complexe qui exhibe une finalité auxiliaire déterministe : cette conclusion par un adage de BANOS. Les principes de son contexte, simples mais efficaces et profonds, traversent en effet ce travail : les "9 principes de Banos" sont implicitement présents dans la majorité des travaux menés et perspectives ouvertes. Même si une application idéale de ces principes relèverait d'un "Démon de Banos", à l'instar du Démon de Laplace ou de Maxwell, qui serait capable d'articuler interdisciplinaire et disciplinaire sans se perdre tout en respectant l'ensemble des principes, leur appréhension comme utopie scientifique, naturellement réflexive donc évolutive et adaptive, nous semble une entrée puissante pour de nouvelles approches intégratives des systèmes territoriaux.

Notre contribution épistémologique, méthodologique en lien avec ces points est essentielle, même si celle ci est difficile à expliciter et nécessitera un certain recul pour être effectivement cernée. D'une certaine manière, nous avons apporté une brique supplémentaire comme *proof-of-concept* du système de principes banosien, mais également comme implémentation et approfondissement de celui-ci sur certains points. Nous avons montré que leur application est loin d'être simple, et que toujours guette le risque de sombrer dans le réductionnisme malgré ces principes fondamentalement complexes. Le dixième commandement serait-il alors : *S'efforcer à appliquer ces principes en ne perdant jamais de vue la complexité et le rôle de la réflexivité ?*

Notre contribution thématique n'est pas forcément facile à situer et nécessitera un recul considérable pour appréhender ses implications. Avons-nous résolu le noeud gordien de la co-évolution ? L'avons-nous tranché ? La réponse la plus fidèle serait que nous en avons tranché une partie, celle naïve comprenant la définition dont nous sommes parti en introduction ou les positionnement de type "poule-et-oeuf" typique des débats des effets structurants, mais que nous avons noué un autre bien plus considérable, en révélant la complexité de ce concept et de ses manifestations.

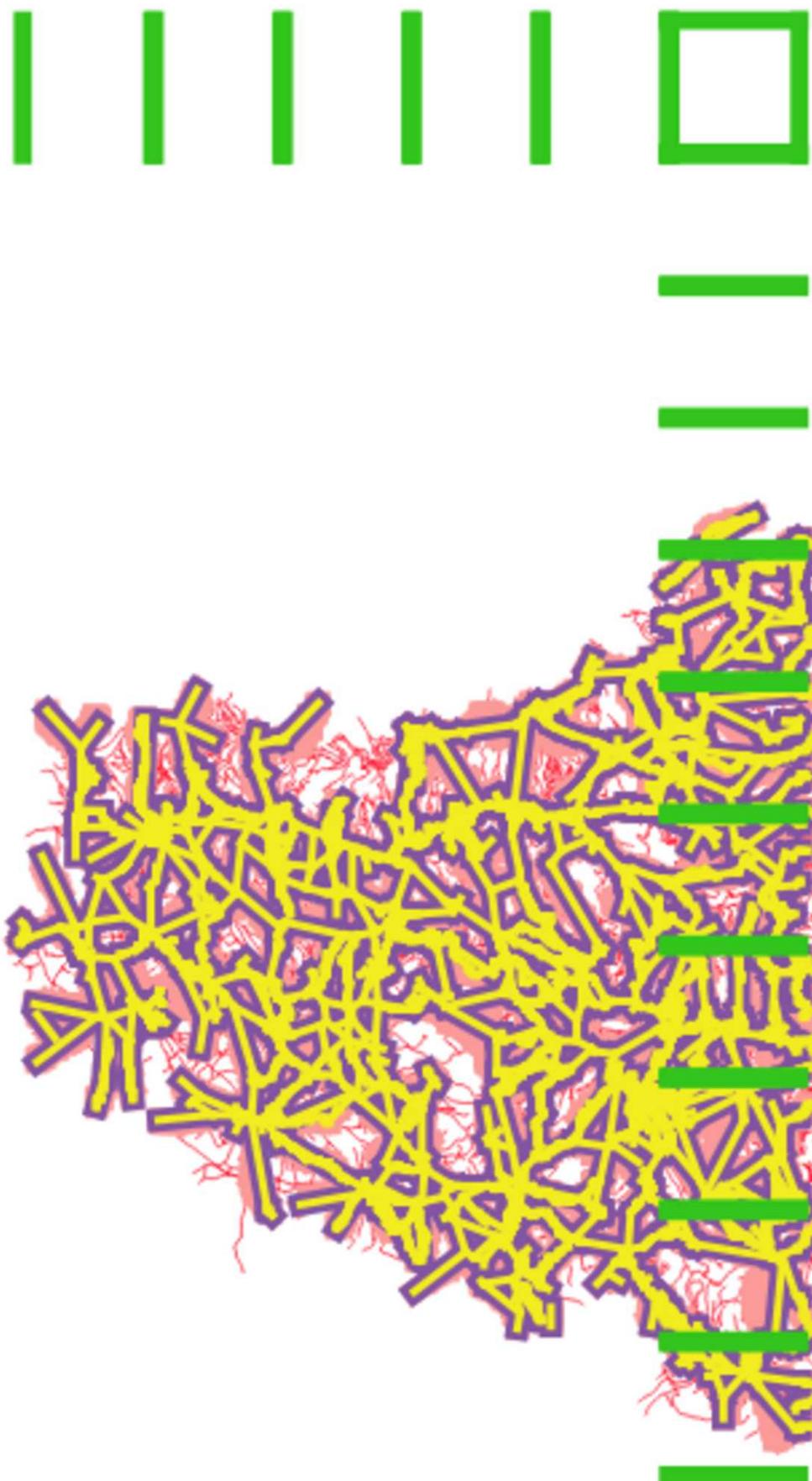
Revenant à notre problématique fondatrice, nous rappelons que (i) nous avons donné une définition de la co-évolution propre aux systèmes territoriaux ainsi qu'une méthode opérationnelle de caractérisation ; (ii) nous avons exploré des pistes de modélisation à différentes échelles, qui s'accordent avec un cadre théorique global. Répondre à

cette problématique nous a permis par ailleurs de progressivement dégager un cadre plus large et de vastes perspectives de recherche.

Notre modeste mission est accomplie, et un fantastique voyage commence tout juste. L'accomplissement passager devient les fondations de ceux à venir. La cumulativité des connaissances ne s'improvise pas, et nous espérons que le tissu complexe dont nous avons cousu les premières mailles sera assez robuste pour s'y insérer. *la route est longue mais la voie est libre.*

* * *

*



BIBLIOGRAPHY

- Aage, Niels et al. (2017). Giga-voxel computational morphogenesis for structural design. *Nature* 550.7674, pp. 84–86.
- Abadie, Alberto et al. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association* 105.490.
- Abbas, Assad et al. (2014). A literature review on the state-of-the-art in patent analysis. *World Patent Information* 37, pp. 3–13.
- Abercrombie, Michael (1977). Concepts in morphogenesis. *Proceedings of the Royal Society of London B: Biological Sciences* 199.1136, pp. 337–344.
- Abler, Ronald et al. (1977). *Spatial organization*. Prentice-Hall.
- Acemoglu, Daron et al. (2016). Innovation network. *Proceedings of the National Academy of Sciences* 113.41, pp. 11483–11488.
- Achibet, Merwan et al. (2014). A Model of Road Network and Buildings Extension Co-evolution. *Procedia Computer Science* 32, pp. 828–833.
- Ackermann, Gabriela et al. (2003). Analysis of built-up areas extension on the Petite Côte region (Senegal) by remote sensing. *Cybergeo: European Journal of Geography* 9.249.
- Adamatzky, Andrew and Jeff Jones (2010). Road planning with slime mould: if Physarum built motorways it would route M6/M74 through Newcastle. *International Journal of Bifurcation and Chaos* 20.10, pp. 3065–3084.
- Adams, Stephen (2010). The text, the full text and nothing but the text: Part 1 - Standards for creating textual information in patent documents and general search implications. *World Patent Information* 32.1, pp. 22–29.
- Aghion, Philippe and Peter Howitt (1992). A Model of Growth through Creative Destruction. *Econometrica* 60.2, pp. 323–51.
- Aghion, Philippe et al. (1998). *Endogenous growth theory*. MIT press.
- Aghion, Philippe et al. (2015). *Innovation and Top Income Inequality*. National Bureau of Economic Research.
- Akaike, Hirotugu (1998). Information theory and an extension of the maximum likelihood principle. *Selected Papers of Hirotugu Akaike*. Springer, pp. 199–213.
- Akcigit, Ufuk et al. (2013). *The Mechanics of Endogenous Innovation and Growth: Evidence from Historical US Patents*. Working Paper.
- Alexander, Christopher (1977). *A pattern language: towns, buildings, construction*. Oxford university press.
- Allen, Benjamin et al. (2017). Multiscale Information Theory and the Marginal Utility of Information. *Entropy* 19.6, p. 273.

- Allen, P. and M. Sanglier (1979). A dynamic model of growth in a central place system. *Geographical Analysis* 11, pp. 256–272.
- (1981). Urban evolution, self organisation and decision-making. *Environment and Planning* 13, pp. 168–183.
- Amar, Georges (1985). Essai de modélisation conceptuelle d'un réseau de circulation. *Cahier du Groupe Réseau* 3, pp. 61–72.
- Amsden, Alice H (1994). Why isn't the whole world experimenting with the East Asian model to develop?: Review of the East Asian miracle. *World Development* 22.4, pp. 627–633.
- Anas, Alex et al. (1998). Urban Spatial Structure. English. *Journal of Economic Literature* 36.3, pp. 1426–1464.
- Anderson, Philip W (1972). More is different. *Science* 177.4047, pp. 393–396.
- Andersson, Claes et al. (2006). A complex network approach to urban growth. *Environment and Planning A* 38.10, p. 1941.
- Andersson, Claes et al. (2002). Urban growth simulation from "first principles". *Physical Review E* 66.2, p. 026204.
- Angeletti, Thomas and Aurélien Berlan (2015). Les êtres collectifs en question. *Tracés. Revue de Sciences humaines* 29, pp. 7–22.
- Angrist, Joshua D et al. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association* 91.434, pp. 444–455.
- Antelope, Chenling et al. (2016). An Interdisciplinary Approach to Morphogenesis. *Working Paper, Santa Fe Institute CSSS* 2016.
- Antonioni, Alberto and Alessio Cardillo (2017). Coevolution of synchronization and cooperation in costly networked interactions. *Physical Review Letters* 118.23, p. 238301.
- Arcaute, Elsa et al. (2015). Constructing cities, deconstructing scaling laws. *Journal of The Royal Society Interface* 12.102, p. 20140745.
- Archer, Margaret S (2010). Morphogenesis versus structuration: on combining structure and action. *The British Journal of Sociology* 61.s1, pp. 225–252.
- Archibugi, Daniele and Mario Pianta (1992). Specialization and size of technological activities in industrial countries: The analysis of patent data. *Research Policy* 21.1, pp. 79–93.
- Arthur, W. Brian (2015). *Complexity and the Shift in Modern Science*. Conference on Complex Systems, Tempe, Arizona.
- Ashby, W Ross (1991). Requisite variety and its implications for the control of complex systems. *Facets of systems science*. Springer, pp. 405–417.
- Audretsch, David B and Maryann P Feldman (1996). R&D spillovers and the geography of innovation and production. *The American economic review* 86.3, pp. 630–640.
- Austin, Timothy R et al. (1996). Defining interdisciplinarity. *Publications of the Modern Language Association of America*, pp. 271–282.

- Axtell, Robert L (2016). 120 million agents self-organize into 6 million firms: a model of the US private sector. *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, pp. 806–816.
- Axtell, Robert et al. (1996). Aligning simulation models: A case study and results. *Computational & Mathematical Organization Theory* 1.2, pp. 123–141.
- Aziz-Alaoui, M. and C. Bertelle (2009). *From System Complexity to Emergent Properties*. Berlin: Springer.
- Badariotti, Dominique et al. (2007). Conception d'un automate cellulaire non stationnaire à base de graphe pour modéliser la structure spatiale urbaine: le modèle Remus. *Cybergeo: European Journal of Geography*.
- Baffi, Solène (2016). Railways and city in territorialization processes in South Africa : from separation to integration ? PhD thesis. Université Paris 1 - Panthéon Sorbonne.
- Bais, Sander et al. (2010). *In praise of science: curiosity, understanding, and progress*. MIT Press.
- Balbo, Flavien et al. (2016). Positionnement des systèmes multi-agents pour les systèmes de transport intelligents. *Revue des Sciences et Technologies de l'Information-Série RIA: Revue d'Intelligence Artificielle* 30.3, pp. 299–327.
- Baldwin, Timothy and Marco Lui (2010). Language identification: The long and the short of the matter. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 229–237.
- Ball, Stephen J (1990). Self-doubt and soft data: social and technical trajectories in ethnographic fieldwork. *International Journal of Qualitative Studies in Education* 3.2, pp. 157–171.
- Banos, Arnaud (2001). A propos de l'analyse spatiale exploratoire des données. *Cybergeo: European Journal of Geography*.
- (2013). Pour des pratiques de modélisation et de simulation libérées en Géographies et SHS. HDR. Université Paris 1.
- Banos, Arnaud and Cyrille Genre-Grandpierre (2012). Towards new metrics for urban road networks: Some preliminary evidence from agent-based simulations. *Agent-based models of geographical systems*. Springer, pp. 627–641.
- Banos, Arnaud et al. (2011). Christaller, toujours vivant! *Cybergeo: European Journal of Geography*.
- Banos, Arnaud et al. (2015). Coupling micro and macro dynamics models on networks: Application to disease spread. *International Workshop on Multi-Agent Systems and Agent-Based Simulation*. Springer, pp. 19–33.

- Baptiste, Hervé (1999). Interactions entre le système de transport et les systèmes de villes: perspective historique pour une modélisation dynamique spatialisée. PhD thesis. Centre d'études supérieures de l'aménagement (Tours).
- (2010). Modeling the Evolution of a Transport System and its Impacts on a French Urban System. *Graphs and Networks: Multilevel Modeling, Second Edition*, pp. 67–89.
- Barabasi, Albert-Laszlo (2002). Linked: How everything is connected to everything else and what it means. *Plume Editors*.
- Barndorff-Nielsen, Ole E et al. (2011). Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. *Journal of Econometrics* 162, pp. 149–169.
- Barrico, C. and C.H. Antunes (2006). Robustness Analysis in Multi-Objective Optimization Using a Degree of Robustness Concept. *Evolutionary Computation, 2006. CEC 2006. IEEE Congress on*, pp. 1887–1892.
- Barthelemy, Marc (2011). Spatial networks. *Physics Reports* 499.1, pp. 1–101.
- (2016). *The Structure and Dynamics of Cities*. Cambridge University Press.
- Barthelemy, Marc and Alessandro Flammini (2008). Modeling urban street patterns. *Physical review letters* 100.13, p. 138702.
- (2009). Co-evolution of density and topology in a simple model of city formation. *Networks and spatial economics* 9.3, pp. 401–425.
- Barthelemy, Marc et al. (2013). Self-organization versus top-down planning in the evolution of a city. *Scientific reports* 3.
- Bastani, O. et al. (2017). Interpretability via Model Extraction. *arXiv preprint arXiv:1706.09773*.
- Bastian, Hilda et al. (2010). Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS medicine* 7.9, e1000326.
- Batagelj, Vladimir (2003). Efficient algorithms for citation network analysis. *arXiv preprint cs/0309023*.
- Battiston, Federico et al. (2016). Emergence of multiplex communities in collaboration networks. *PloS one* 11.1, e0147451.
- Batty, Michael (1991). Generating urban forms from diffusive growth. *Environment and Planning A* 23.4, pp. 511–544.
- (2006). Hierarchy in cities and city systems. *Hierarchy in natural and social sciences*. Springer, pp. 143–168.
 - (2007). *Cities and complexity: understanding cities with cellular automata, agent-based models, and fractals*. MIT press.
 - (2013a). Big data, smart cities and city planning. *Dialogues in Human Geography* 3.3, pp. 274–279.
 - (2013b). *The new science of cities*. MIT Press.

- (2016). Theoretical filters: Reducing explanations in cities to their very essence. *Environment and Planning B: Planning and Design* 43.5, pp. 797–799.
- (2017). The Age of the Smart City. *CASA Working Paper*.
- Batty, Michael and Paul A Longley (1994). *Fractal cities: a geometry of form and function*. Academic Press.
- Batty, Michael and S Mackie (1972). The calibration of gravity, entropy, and related models of spatial interaction. *Environment and Planning A* 4.2, pp. 205–233.
- Batty, Michael and Yichun Xie (1994). From cells to cities. *Environment and planning B: Planning and design* 21.7, S31–S48.
- Bavoux, Jean-Jacques et al. (2005). *Géographie des transports*. Paris: Armand Colin.
- Bazin, Sylvie et al. (2007). L'évolution des marchés immobiliers résidentiels dans l'aire urbaine de Reims: un effet de la Ligne à Grande Vitesse Est-européenne? *Congress of the European Regional Science Association and ASRDLF, Paris*.
- Bazin, Sylvie et al. (2010). Lignes ferroviaires à grande vitesse et dynamiques locales : une analyse comparée de la littérature. *Transport et développement des territoires*. Le Havre, France, 21p.
- Bazin, Sylvie et al. (2011). Grande vitesse ferroviaire et développement économique local: une revue de la littérature. *Recherche Transports Sécurité* 27.3, pp. 215–238.
- Beaucire, Francis and Matthieu Drevelle (2013). «Grand Paris Express»: un projet au service de la réduction des inégalités d'accessibilité entre l'Ouest et l'Est de la région urbaine de Paris? *Revue d'Économie Régionale & Urbaine* 3, pp. 437–460.
- Bedau, Mark (2002). Downward causation and the autonomy of weak emergence. *Principia: an international journal of epistemology* 6.1, pp. 5–50.
- Beer, Randall D (2004). Autopoiesis and cognition in the game of life. *Artificial Life* 10.3, pp. 309–326.
- Bélizal, Édouard de et al. (2011). Quand l'aléa devient la ressource: l'activité d'extraction des matériaux volcaniques autour du volcan Merapi (Indonésie) dans la compréhension des risques locaux. *Cybergeo: European Journal of Geography*.
- Belmonte, Mylène et al. (2008). Automatisation intégrale de la ligne 1: étude et modélisation du trafic mixte. *Lambda-Mu*, Session-5B.
- Ben-Akiva, Moshe E and Steven R Lerman (1985). *Discrete choice analysis: theory and application to travel demand*. Vol. 9. MIT press.
- Benguigui, Lucien and Efrat Blumenfeld-Lieberthal (2007). A dynamic model for city size distribution beyond Zipf's law. *Physica A: Statistical Mechanics and its Applications* 384.2, pp. 613–627.
- Bennett, Jonathan (2010). *OpenStreetMap*. Packt Publishing Ltd.
- Bergeaud, Antonin et al. (2017a). Classifying patents based on their semantic content. *PloS one* 12.4, e0176310.

- Bergeaud, Antonin et al. (2017b). Classifying patents based on their semantic content. *PLOS ONE* 12.4, pp. 1–22.
- Berger, Thor and Kerstin Enflo (2017). Locomotives of local growth: The short-and long-term impact of railroads in Sweden. *Journal of Urban Economics* 98, pp. 124–138.
- Berne, Laurence (2008). Ouverture et fermeture de territoire par les réseaux de transports dans trois espaces montagnards (Bugey, Bauges et Maurienne). PhD thesis. Université de Savoie.
- Bernier, Xavier (2007). Les dynamiques réticulo-territoriales et la frontière en zone de montagne: approche typologique. *Flux* 4, pp. 8–19.
- Berroir, Sandrine et al. (2005). *La contribution des villes nouvelles au polycentrisme francilien*. UMR Géographie-cités.
- Berroir, Sandrine et al. (2017). Les systèmes urbains français: une approche relationnelle. *Cybergeo: European Journal of Geography*.
- Berry, Brian JL (1964). Cities as systems within systems of cities. *Papers in Regional Science* 13.1, pp. 147–163.
- Bettencourt, Luís MA and José Lobo (2016). Urban scaling in Europe. *Journal of The Royal Society Interface* 13.116, p. 20160005.
- Bettencourt, Luís MA et al. (2008). Why are large cities faster? Universal scaling and self-similarity in urban organization and dynamics. *The European Physical Journal B-Condensed Matter and Complex Systems* 63.3, pp. 285–293.
- Bettencourt, Luís MA et al. (2007). Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the national academy of sciences* 104.17, pp. 7301–7306.
- Biernacki, Christophe et al. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence* 22.7, pp. 719–725.
- Bigotte, João F et al. (2010). Integrated modeling of urban hierarchy and transportation network planning. *Transportation Research Part A: Policy and Practice* 44.7, pp. 506–522.
- Bird, Steven (2006). NLTK: the natural language toolkit. *Proceedings of the COLING/ACL on Interactive presentation sessions*. Association for Computational Linguistics, pp. 69–72.
- Bitbol, Michel and Pier Luigi Luisi (2004). Autopoiesis with or without cognition: defining life at its edge. *Journal of the Royal Society Interface* 1.1, pp. 99–107.
- Blanquart, Corinne and Martin Koning (2017). The local economic impacts of high-speed railways: theories and facts. *European Transport Research Review* 9.2, p. 12.
- Blei, David M et al. (2003). Latent dirichlet allocation. *Journal of machine Learning research* 3, pp. 993–1022.
- Block-Schachter, David (2012). Hysteresis and urban rail: The effects of past urban rail on current residential and travel choices. PhD thesis. Massachusetts Institute of Technology.

- Blondel, Vincent D. et al. (2008b). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 10, p. 10008.
- Blondel, Vincent D et al. (2008a). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008.10, P10008.
- Bloom, Nicholas et al. (2013). Identifying Technology Spillovers and Product Market Rivalry. *Econometrica* 81.4, pp. 1347–1393.
- Blumenfeld-Lieberthal, Efrat and Juval Portugali (2010). Network cities: A complexity-network approach to urban dynamics and development. *Geospatial Analysis and Modelling of Urban Structure and Dynamics*. Springer, pp. 77–90.
- Bohannon, John (2014). Scientific publishing. Google Scholar wins raves—but can it be trusted? *Science* 343.6166, p. 14.
- Bollen, Johan et al. (2014). From funding agencies to scientific agency. *EMBO reports* 15.2, pp. 131–133.
- Bolón-Canedo, Verónica et al. (2013). A review of feature selection methods on synthetic data. *Knowledge and information systems* 34.3, pp. 483–519.
- Bolotin, A. (2014). Computational solution to quantum foundational problems. *arXiv preprint arXiv:1403.7686*.
- Bonanno, G. et al. (2001). Levels of complexity in financial markets. *Physica A Statistical Mechanics and its Applications* 299, pp. 16–27.
- Bonin, Olivier and Jean-Paul Hubert (2014). Modélisation morphogénétique de moyen terme des villes: une schématisation du modèle théorique de RITCHOT et DESMARAIS dans le cadre du modèle standard de l'économie urbaine. *Revue d'Économie Régionale & Urbaine* 3, pp. 471–497.
- Bonin, Olivier, Jean-Paul Hubert, et al. (2012). Modèle de morphogénèse urbaine: simulation d'espaces qualitativement différenciés dans le cadre du modèle de l'économie urbaine. *49è colloque de l'ASRDLF*.
- Bonnafous, Alain (1987). The regional impact of the TGV. *Transportation* 14.2, pp. 127–137.
- (2014). Les observatoires permanents comme instruments d'évaluation ex post: Le cas français. International Transport Forum Discussion Paper.
- Bonnafous, Alain and François Plassard (1974). Les méthodologies usuelles de l'étude des effets structurants de l'offre de transport. *Revue économique*, pp. 208–232.
- Bonnafous, Alain et al. (1974). La detection des effets structurants d'autoroute: Application à la Vallée du Rhône. *Revue économique* 25.2, pp. 233–256.
- Bosch, F van den et al. (1990). The velocity of spatial population expansion. *Journal of Mathematical Biology* 28.5, pp. 529–565.
- Bouchaud, J. P. and M. Potters (2009). Financial Applications of Random Matrix Theory: a short review. *arXiv preprint arXiv:0910.1205*.

- Bouchaud, J-P et al. (2000). Apparent multifractality in financial time series. *The European Physical Journal B-Condensed Matter and Complex Systems* 13.3, pp. 595–599.
- Bourgine, P. et al. (2009). French Roadmap for complex Systems 2008–2009. *arXiv preprint arXiv:0907.2221*.
- Bourgine, Paul and Annick Lesne (2010). *Morphogenesis: origins of patterns and shapes*. Springer Science & Business Media.
- Bourgine, Paul and John Stewart (2004). Autopoiesis and cognition. *Artificial life* 10.3, pp. 327–345.
- Bouteiller, Catherine and Sybille Berjoan (2013). Open data en transport urbain: quelles sont les données mises à disposition? Quelles sont les stratégies des autorités organisatrices? *HALSHS preprint : halshs-00838632*.
- Bouveyron, Charles et al. (2016). The stochastic topic block model for the clustering of vertices in networks with textual edges. *Statistics and Computing*, pp. 1–21.
- Bracken, Louise J (2016). *Interdisciplinarity and Geography*. Wiley Online Library.
- Brand, Christian et al. (2013). Accelerating the transformation to a low carbon passenger transport system: The role of car purchase taxes, feebates, road taxes and scrappage incentives in the UK. *Transportation Research Part A: Policy and Practice* 49, pp. 132–148.
- Bretagnolle, Anne (2003). Vitesse et processus de sélection hiérarchique dans le système des villes françaises. *Données urbaines* 4.
- (2009). Villes et réseaux de transport : des interactions dans la longue durée, France, Europe, États-Unis. HDR. Université Panthéon-Sorbonne - Paris I.
- Bretagnolle, Anne et al. (2006). From theory to modelling: urban systems as complex systems. *CyberGeo: European Journal of Geography*.
- Bretagnolle, Anne et al. (2002). Time and space scales for measuring urban growth. *Cybergeo: European Journal of Geography*.
- Bretagnolle, Anne and Denise Pumain (2010a). Comparer deux types de systèmes de villes par la modélisation multi-agents. *Qu'appelle t-on aujourd'hui les sciences de la complexité? Langages, réseaux, marchés, territoires*, pp. 271–299.
- (2010b). Simulating Urban Networks through Multiscalar Space-Time Dynamics: Europe and the United States, 17th-20th Centuries. *Urban Studies* 47.13, pp. 2819–2839.
- Bretagnolle, Anne et al. (1998). Space-time contraction and the dynamics of urban systems. *Cybergeo: European Journal of Geography*.
- Bretagnolle, Anne et al. (2000). Long-term dynamics of European towns and cities: towards a spatial model of urban growth. *Cybergeo: European Journal of Geography*.
- Bretagnolle, Anne et al. (2016). La ville à l'échelle de l'Europe-Apports du couplage et de l'expertise de bases de données issues de l'imagerie satellitaire. *Revue Internationale de Géomatique* 26.1, pp. 55–78.

- Brotchie, John F (1984). Technological change and urban form. *Environment and Planning A* 16.5, pp. 583–596.
- Brown, Matthew J (2009). Models and perspectives on stage: remarks on Giere's scientific perspectivism. *Studies in History and Philosophy of Science Part A* 40.2, pp. 213–220.
- Bruck, Péter et al. (2016). Recognition of emerging technology trends: class-selective study of citations in the US Patent Citation Network. *Scientometrics* 107.3, pp. 1465–1475.
- Brunsdon, Chris et al. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis* 28.4, pp. 281–298.
- Brunsdon, Chris et al. (1998). Geographically weighted regression. *Journal of the Royal Statistical Society: Series D (The Statistician)* 47.3, pp. 431–443.
- Bulcke, Tim Van den et al. (2006). SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC bioinformatics* 7.1, p. 43.
- Bull, Larry et al. (2000). On meme–gene coevolution. *Artificial life* 6.3, pp. 227–235.
- Burgess, Ernest Watson et al. (1925). *The city*. University of Chicago Press.
- Burke, Nuala T (1972). Dublin 1600–1800: a study in urban morphogenesis. PhD thesis.
- Burnham, Kenneth P and David R Anderson (2003). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media.
- Cain, Jeff and Peggy Piascik (2015). Are Serious Games a Good Strategy for Pharmacy Education? *American journal of pharmaceutical education* 79.4.
- Callaway, Duncan S et al. (2000). Network robustness and fragility: Percolation on random graphs. *Physical review letters* 85.25, p. 5468.
- Camarero, Luis A and Jesús Oliva (2008). Exploring the social face of urban mobility: daily mobility as part of the social structure in Spain. *International Journal of Urban and Regional Research* 32.2, pp. 344–362.
- Camerer, Colin F et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, aaf0918.
- Campbell, John Y and Samuel B Thompson (2007). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies* 21.4, pp. 1509–1531.
- Carlile, Paul R (2004). Transferring, translating, and transforming: An integrative framework for managing knowledge across boundaries. *Organization science* 15.5, pp. 555–568.
- Carrignon, Simon et al. (2015). Modelling the co-evolution of trade and culture in past societies. *Winter Simulation Conference (WSC)*, 2015. IEEE, pp. 3949–3960.

- Carver, Stephen J (1991). Integrating multi-criteria evaluation with geographical information systems. *International Journal of Geographical Information System* 5.3, pp. 321–339.
- Castellacci, Fulvio and Jose Miguel Natera (2013). The dynamics of national innovation systems: A panel cointegration analysis of the coevolution between innovative capability and absorptive capacity. *Research Policy* 42.3, pp. 579–594.
- Ceccarini, Patrice (2001). Essai de formalisation dynamique de la cathédrale gothique: morphogénèse et modélisation de la Basilique Saint-Denis: les relations entre théologie, sciences et architecture au XIIIème siècle à Saint-Denis. PhD thesis. Paris, EHESS.
- Cerdeira, Eugênia Viana (2017). Les inégalités d'accès aux ressources urbaines dans les franges périphériques de Belo Horizonte (Brésil): quelles évolutions? *EchoGéo* 39.
- Chalidabongse, Junavit and CC Jay Kuo (1997). Fast motion vector estimation using multiresolution-spatio-temporal correlations. *Circuits and Systems for Video Technology, IEEE Transactions on* 7.3, pp. 477–488.
- Champollion, Pierre (2006). Territory and Territorialization: Present state of the Caenti thought. *International Conference of Territorial Intelligence*. INTI-International Network of Territorial Intelligence. Alba Iulia, Romania, p51–58.
- Chamussy, Henri et al. (1984). La dynamique de systèmes: une méthode de modélisation des unités spatiales. *Espace géographique* 13.2, pp. 81–93.
- Chang, Justin S (2006). Models of the Relationship between Transport and Land-use: A Review. *Transport Reviews* 26.3, pp. 325–350.
- Chapman, Michael J and Lynn Margulis (1998). Morphogenesis by symbiogenesis. *International Microbiology* 1.4.
- Chardonnell, Sonia (2007). Time-Geography: Individuals in Time and Space. *Models in Spatial Analysis*, pp. 97–126.
- Chasset, Pierre-Olivier et al. (2016). *cybergeo20 v1.0*. <http://dx.doi.org/10.5281/zenodo.53905>.
- Chaudhuri, G. and Clarke, Keith C. (2015). On the Spatiotemporal Dynamics of the Coupling between Land Use and Road Networks: Does Political History Matter? *Environment and Planning B: Planning and Design* 42.1, pp. 133–156.
- Chavalarias, David (2016). What's wrong with Science? *Scientometrics*, pp. 1–23.
- Chavalarias, David and Jean-Philippe Cointet (2013). Phylogenetic patterns in science evolution—the rise and fall of scientific fields. *Plos One* 8.2, e54847.
- Chavalarias, David et al. (2005). Nobel, Le Jeu De La Découverte Scientifique. *HALSHS preprint : halshs-00005009*.
- Chen, Duan-Rung and Khoa Truong (2012). Using multilevel modeling and geographically weighted regression to identify spatial

- variations in the relationship between place-level disadvantages and obesity in Taiwan. *Applied Geography* 32.2, pp. 737–745.
- Chen, Y. (2016). Normalizing and Classifying Shape Indexes of Cities by Ideas from Fractals. *arXiv preprint arXiv:1608.08839*.
- Chen, Yanguang (2009). Urban gravity model based on cross-correlation function and Fourier analyses of spatio-temporal process. *Chaos, Solitons & Fractals* 41.2, pp. 603–614.
- (2010). Characterizing growth and form of fractal cities with allometric scaling exponents. *Discrete Dynamics in Nature and Society* 2010.
- Chérel, Guillaume et al. (2015). Beyond Corroboration: Strengthening Model Validation by Looking for Unexpected Patterns. *PLoS ONE* 10.9, e0138212.
- Chicheportiche, Rémy and Jean-Philippe Bouchaud (2013). A nested factor model for non-linear dependences in stock returns. *arXiv preprint arXiv:1309.3102*.
- Chodrow, Philip S. (2017). Structure and information in spatial segregation. *Proceedings of the National Academy of Sciences* 114.44, pp. 11591–11596.
- Choi, Jinho and Yong-Sik Hwang (2014). Patent keyword network analysis for improving technology development efficiency. *Technological Forecasting and Social Change* 83, pp. 170–182.
- Chu, Dominique (2008). Criteria for conceptual and operational notions of complexity. *Artificial Life* 14.3, pp. 313–323.
- Clarke, Keith C and Leonard J Gaydos (1998). Loose-coupling a cellular automaton model and GIS: long-term urban growth prediction for San Francisco and Washington/Baltimore. *International journal of geographical information science* 12.7, pp. 699–714.
- Clarke, Keith C et al. (2007). A decade of SLEUTHing: Lessons learned from applications of a cellular automaton land use change model. *Classics in IJGIS: twenty years of the international journal of geographical information science and systems*, pp. 413–427.
- Clauset, Aaron et al. (2004). Finding community structure in very large networks. *Physical review E* 70.6, p. 066111.
- Clauset, Aaron et al. (2009). Power-law distributions in empirical data. *SIAM review* 51.4, pp. 661–703.
- Claval, Paul (1985). Causalité et géographie. *Espace géographique* 14.2, pp. 109–115.
- (1987). Réseaux territoriaux et enracinement. *Cahier/Groupe Réseaux* 3.7, pp. 44–60.
- Colander, David (2003). *The complexity revolution and the future of economics*. Working Paper. Middlebury College, Department of Economics.
- Colletis, Gabriel (2010). Co-évolution des territoires et de la technologie: une perspective institutionnaliste. *Revue d'Économie Régionale & Urbaine* 2, pp. 235–249.

- Combes, Pierre-Philippe and Miren Lafourcade (2005). Transport costs: measures, determinants, and regional policy implications for France. *Journal of Economic Geography* 5.3, pp. 319–349.
- Commenges, Hadrien (2013). The invention of daily mobility. Performative aspects of the instruments of economics of transportation. PhD thesis. Université Paris-Diderot - Paris VII.
- Cottet, Nathanaël et al. (2017). Observing a quantum Maxwell demon at work. *Proceedings of the National Academy of Sciences* 114.29, pp. 7561–7564.
- Cottineau, Clémentine (2014). L'évolution des villes dans l'espace post-soviétique. Observation et modélisations. PhD thesis. Université Paris 1 Panthéon-Sorbonne.
- (2015). *Urban scaling: What cities are we talking about?* Presentation of ongoing work at Quanturb seminar, April 1st 2015.
 - (2017). MetaZipf. A dynamic meta-analysis of city size distributions. *PLOS ONE* 12.8, pp. 1–22.
- Cottineau, Clémentine et al. (2015). An incremental method for building and evaluating agent-based models of systems of cities. *HAL-SHS preprint : halshs-01093426*.
- Cottineau, Clémentine et al. (2016). Back to the Future of Multimodeling. *Royal Geographical Society-Annual Conference 2016-Session: Geocomputation, the Next 20 Years* (1).
- Cottineau, Clémentine et al. (2015). A modular modelling framework for hypotheses testing in the simulation of urbanisation. *Systems* 3.4, pp. 348–377.
- Cottineau, Clémentine et al. (2015). Paradoxical Interpretations of Urban Scaling Laws. *arXiv preprint arXiv:1507.07878*.
- Cottineau, Clémentine et al. (2015). Revisiting some geography classics with spatial simulation. *Plurimondi. An International Forum for Research and Debate on Human Settlements*. Vol. 7. 15.
- Cottineau, Clementine et al. (2017). Initial spatial conditions in simulation models: the missing leg of sensitivity analyses? *Geocomputation Conference*.
- Couclelis, Helen (1985). Cellular worlds: a framework for modeling micro—macro dynamics. *Environment and planning A* 17.5, pp. 585–596.
- Courtat, Thomas et al. (2011). Mathematics and morphogenesis of cities: A geometrical approach. *Physical Review E* 83.3, p. 036106.
- Cronin, Blaise and Cassidy R Sugimoto (2014). *Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact*. MIT Press.
- Crosato, E. (2014). *Artificial Self-Assembly : Literature Review*. Working Paper. Vrije Universiteit Amsterdam.
- Crosato, Emanuele et al. (2017). Informative and misinformative interactions in a school of fish. *arXiv preprint arXiv:1705.01213*.
- Crucitti, Paolo et al. (2006). Centrality measures in spatial networks of urban streets. *Physical Review E* 73.3, p. 036125.

- Curran, Clive-Steven and Jens Leker (2011). Patent indicators for monitoring convergence—examples from NFF and ICT. *Technological Forecasting and Social Change* 78.2, pp. 256–273.
- Cussat-Blanc, Sylvain et al. (2012). A synthesis of the Cell2Organ developmental model. *Morphogenetic Engineering*. Springer, pp. 353–381.
- Cuthbert, Angela L et al. (2005). An empirical analysis of the relationship between road development and residential land development. *Canadian Journal of Regional Science* 28.1, pp. 49–76.
- Cuyala, Sylvain (2014). Analyse spatio-temporelle d'un mouvement scientifique. L'exemple de la géographie théorique et quantitative européenne francophone. PhD thesis. Université Paris 1 Panthéon-Sorbonne.
- D., Pumain et al. (1989). *Villes et auto-organisation*. Paris: Economica.
- Damm, David et al. (1980). Response of urban real estate values in anticipation of the Washington Metro. *Journal of Transport Economics and Policy*, pp. 315–336.
- De Domenico, Manlio et al. (2015). Ranking in interconnected multi-layer networks reveals versatile nodes. *Nature communications* 6.
- De Leon, FD et al. (2007). NetLogo Urban Suite-Tijuana Bordertowns model. *Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL*.
- De Nadai, Marco et al. (2016). The Death and Life of Great Italian Cities: A Mobile Phone Data Perspective. *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pp. 413–423.
- Deb, Kalyanmoy and Himanshu Gupta (2006). Introducing robustness in multi-objective optimization. *Evolutionary Computation* 14.4, pp. 463–494.
- Dechezleprêtre, Antoine et al. (2014). *Knowledge Spillovers from Clean and Dirty Technologies*. CEP Discussion Papers dp1300. Centre for Economic Performance, LSE.
- Deffuant, Guillaume et al. (2015). Visions de la complexité. Le démon de Laplace dans tous ses états. *Natures Sciences Sociétés* 23.1, pp. 42–53.
- Delile, Julien et al. (2016). Modélisation multi-agent de l'embryogenèse animale. *Modélisations, simulations, systèmes complexes*, pp. 581–624.
- Delons, Jean et al. (2008). PIRANDELLO an integrated transport and land-use model for the Paris area. *HALSHS preprint : hal-00319087*.
- Deng, Yi and Rongfang Liu (2007). Potential Impact of Housing Policy on Transportation Infrastructure in Chinese Cities. *Transportation Research Record: Journal of the Transportation Research Board* 2038, pp. 1–8.
- Depersin, J. and M. Barthelemy (2017). From global scaling to the dynamics of individual cities. *arXiv preprint arXiv:1710.09559*.

- Desjardins, Xavier (2010). la bataille du Grand Paris. *L'Information géographique* 74.4, pp. 29–46.
- (2016). Ce Grand Paris qui advient. Leçons pour la planification métropolitaine. *L'Information géographique* 80.4, pp. 96–114.
- Devaux, Nicolas et al. (2007). Extraction automatique d'habitations en milieu rural de PED à partir de données THRS. *Cybergeo: European Journal of Geography*.
- Di Meo, Guy (1998). De l'espace aux territoires: éléments pour une archéologie des concepts fondamentaux de la géographie. *L'information géographique* 62.3, pp. 99–110.
- Dick, Josef and Friedrich Pillichshammer (2010). *Digital nets and sequences: Discrepancy Theory and Quasi-Monte Carlo Integration*. Cambridge University Press.
- Diderot, Denis (1965). *Entretien entre d'Alembert et Diderot*. Garnier-Flammarion.
- Dietterich, Tom (1995). Overfitting and undercomputing in machine learning. *ACM computing surveys (CSUR)* 27.3, pp. 326–327.
- Ding, Rui et al. (2017). Heuristic urban transportation network design method, a multilayer coevolution approach. *Physica A: Statistical Mechanics and its Applications* 479, pp. 71–83.
- Dirk, Lynn (1999). A Measure of Originality The Elements of Science. *Social Studies of Science* 29.5, pp. 765–776.
- Dobbie, Melissa J and David Dail (2013). Robustness and sensitivity of weighting and aggregation in constructing composite indices. *Ecological Indicators* 29, pp. 270–277.
- Dodds, Peter Sheridan et al. (2017). Simon's fundamental rich-get-richer model entails a dominant first-mover advantage. *Physical Review E* 95.5, p. 052301.
- Dollens, Dennis (2014). Alan Turing's Drawings, Autopoiesis and Can Buildings Think? *Leonardo* 47.3, pp. 249–254.
- Dollfus, O and F Durand Dastès (1975). Some remarks on the notions of 'structure' and 'system' in geography. *Geoforum* 6.2, pp. 83–94.
- Dongguan Metro (2017). 东莞地铁规划线路 [Dongguan Metro Planning]. <http://jtapi.bendibao.com/ditie/inc/dg/guihuada.gif>.
- Doursat, René (2008). Programmable Architectures That Are Complex and Self-Organized-From Morphogenesis to Engineering. *ALIFE*, pp. 181–188.
- Doursat, René et al. (2012). *Morphogenetic engineering: toward programmable complex systems*. Springer.
- (2013). A review of morphogenetic engineering. *Natural Computing* 12.4, pp. 517–535.
- Dragomir, SS (1999). The Ostrowski's integral inequality for Lipschitzian mappings and applications. *Computers & Mathematics with Applications* 38.11, pp. 33–37.

- Drogoul, Alexis et al. (2013). Gama: multi-level and complex environment for agent-based models and simulations. *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, pp. 1361–1362.
- Drummond, Chris (2009). Replicability is not reproducibility: nor is it good science.
- Ducruet, César and Laurent Beauguitte (2014). Spatial science and network science: Review and outcomes of a complex relationship. *Networks and Spatial Economics* 14.3-4, pp. 297–316.
- Duda, John (2013). Cybernetics, anarchism and self-organisation. *Anarchist studies* 21.1, p. 52.
- Dupuy, Gabriel (1985). *Systèmes, réseaux et territoires: principes de réseautique territoriale*. Presses de l'école nationale des ponts et chaussées.
- (1987). Vers une théorie territoriale des réseaux: une application au transport urbain. *Annales de Géographie*. JSTOR, pp. 658–679.
- Dupuy, Gabriel and Lucien Gilles Benguigui (2015). Sciences urbaines: interdisciplinarités passive, naïve, transitive, offensive. *Métropoles* 16.
- Durand-Dastès, François (2003). Les géographes et la notion de causalité. *Enquête sur la notion de causalité*. PUF, pp. 145–160.
- Durantin, Arnaud et al. (2017). Disruptive Innovation in Complex Systems. *Complex Systems Design & Management*. Springer, pp. 41–56.
- Duranton, Gilles (1999). Distance, land, and proximity: economic analysis and the evolution of cities. *Environment and Planning a* 31.12, pp. 2169–2188.
- Duranton, Gilles and Matthew A Turner (2012). Urban growth and transportation. *The Review of Economic Studies* 79.4, pp. 1407–1440.
- Durham, William H (1991). *Coevolution: Genes, culture, and human diversity*. Stanford University Press.
- Dybdahl, Mark F and Curtis M Lively (1996). The geography of co-evolution: comparative population structures for a snail and its trematode parasite. *Evolution* 50.6, pp. 2264–2275.
- EUROSTAT (2014). *Eurostat Geographical Data*. <http://ec.europa.eu/eurostat/web/gisco>.
- Elman, Colin and Diana Kapiszewski (2018). The Qualitative Data Repository's Annotation for Transparent Inquiry (ATI) Initiative. *PS: Political Science & Politics* 51.1, pp. 3–6.
- Entretiens vo.2 [Data set]*. Zenodo. <http://doi.org/10.5281/zenodo.556331>.
- Epstein, Joshua M (2006). *Generative social science: Studies in agent-based computational modeling*. Princeton University Press.
- (2008). Why model? *Journal of Artificial Societies and Social Simulation* 11.4, p. 12.

- Epstein, Joshua M and Robert L Axtell (1996). *Growing artificial societies: Social science from the bottom up (complex adaptive systems)*. Brookings Institution Press MIT Press.
- Escobar, Francisco J et al. (2000). Distribution of Online Cartographic Products in Australia. *Cybergeo: European Journal of Geography*.
- Faivre, Emmanuel (2003). Infrastructures autoroutières, mobilité et dynamiques territoriales. PhD thesis. Université de Franche-Comté.
- Fan, C Cindy (2005). Modeling interprovincial migration in China, 1985–2000. *Eurasian Geography and Economics* 46.3, pp. 165–184.
- Farmer, J Doyne and Duncan Foley (2009). The economy needs agent-based modelling. *Nature* 460.7256, pp. 685–686.
- Fattori, Michele et al. (2003). Text mining applied to patent mapping: a practical business case. *World Patent Information* 25.4, pp. 335–342.
- Favarro, Jean-Marc and Denise Pumain (2011). Gibrat Revisited: An Urban Growth Model Incorporating Spatial Interaction and Innovation Cycles. *Geographical Analysis* 43.3, pp. 261–286.
- Febres, Gerardo et al. (2013). Complexity measurement of natural and artificial languages. *arXiv preprint arXiv:1311.5427*.
- Fecher, Benedikt and Sascha Friesike (2014). Open science: one term, five schools of thought. *Opening science*. Springer, pp. 17–47.
- Feyerabend, Paul (1993). *Against method*. Verso.
- Florida, Richard et al. (2008). The rise of the mega-region. *Cambridge Journal of Regions, Economy and Society* 1.3, pp. 459–476.
- Foot, Robin (1994). RATP, un corporatisme à l'épreuve des voyageurs. *Travail* 31, pp. 63–100.
- (2005). Faut-il protéger le métro des voyageurs? Ou l'appréhension du voyageur par les ingénieurs et les conducteurs. *Travailler* 2, pp. 169–206.
- Fotheringham, A. S. and D. W. S. Wong (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A* 23.7, pp. 1025–1044.
- Franco, Jessica et al. (2009). DiceDesign-package. *Designs of Computer Experiments*, p. 2.
- Frank, Morgan R et al. (2014). Constructing a taxonomy of fine-grained human movement and activity motifs through social media. *arXiv preprint arXiv:1410.1393*.
- Frankhauser, Pierre (1998). Fractal geometry of urban patterns and their morphogenesis. *Discrete Dynamics in Nature and Society* 2.2, pp. 127–145.
- (2008). Fractal geometry for measuring and modelling urban patterns. *The dynamics of complex urban systems*. Springer, pp. 213–243.
- Freud, Sigmund et al. (1989). *Totem and taboo: some points of agreement between the mental lives of savages and neurotics*. eng. The Standard

- edition of the complete psychological works of Sigmund Freud. New York: W.W. Norton.
- Frey, Rüdiger et al. (2001). Copulas and credit models. *Risk* 10.111114.10.
- Frigg, Roman and Ioannis Votsis (2011). Everything you always wanted to know about structural realism but were afraid to ask. *European journal for philosophy of science* 1.2, pp. 227–276.
- Fritsch, Bernard (2007). Infrastructures de transport, densification et étalement urbains: quelques enseignements de l'expérience nantaise. *Les Cahiers scientifiques du transport* 51, pp. 37–60.
- Fu, Jingying et al. (2014). 1 km grid population dataset of China (2005, 2010). *Global Change Research Data Publishing and Repository*. DOI: 10.3974/geodb.2014.01.06.v1.
- Fujita, Masahisa et al. (1999). On the evolution of hierarchical urban systems. *European Economic Review* 43.2, pp. 209–251.
- Fujita, Masahisa and Hideaki Ogawa (1982). Multiple equilibria and structural transition of non-monocentric urban configurations. *Regional science and urban economics* 12.2, pp. 161–196.
- Fujita, Masahisa and Jacques-François Thisse (1996). Economics of agglomeration. *Journal of the Japanese and international economies* 10.4, pp. 339–378.
- Fullerton, Don and Sarah E West (2002). Can taxes on cars and on gasoline mimic an unavailable tax on emissions? *Journal of Environmental Economics and Management* 43.1, pp. 135–157.
- Furlanello, C. et al. (2017). Towards a scientific blockchain framework for reproducible data analysis. *arXiv preprint arXiv:1707.06552*.
- Furman, Jeffrey L. and Scott Stern (2011). Climbing atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research. *American Economic Review* 101.5, pp. 1933–63.
- Fusco, Giovanni (2004). La mobilité quotidienne dans les grandes villes du monde: application de la théorie des réseaux bayésiens. *Cybergeo: European Journal of Geography*.
- Gabaix, Xavier (1999). Zipf's law for cities: an explanation. *Quarterly journal of Economics*, pp. 739–767.
- Gabaix, Xavier and Yannis M. Ioannides (2004). The evolution of city size distributions. *Cities and Geography*. Vol. 4. Handbook of Regional and Urban Economics. Elsevier, pp. 2341 –2378.
- Gabora, L. and M. Steel (2017). Autocatalytic networks in cognition and the origin of culture. *arXiv preprint arXiv:1703.05917*.
- Gallez, Caroline (2015). La mobilité quotidienne en politique. Des manières de voir et d'agir. HDR. Université Paris-Est.
- Gao, Zhong-Ke et al. (2017). Complex network analysis of time series. *EPL (Europhysics Letters)* 116.5, p. 50001.
- Gao, Zhong-Ke et al. (2015). Multiscale complex network for analyzing experimental multivariate time series. *EPL (Europhysics Letters)* 109.3, p. 30005.

- Gautier, Erwan and Ronan Le Saout (2015). The dynamics of gasoline prices: Evidence from daily French micro data. *Journal of Money, Credit and Banking* 47.6, pp. 1063–1089.
- Gell-Mann, Murray (1995). *The Quark and the Jaguar: Adventures in the Simple and the Complex*. Macmillan.
- Gell-Mann, Murray and James B Hartle (1996). Quantum mechanics in the light of quantum cosmology. *Foundations Of Quantum Mechanics In The Light Of New Technology: Selected Papers from the Proceedings of the First through Fourth International Symposia on Foundations of Quantum Mechanics*. World Scientific, pp. 347–369.
- Geman, Stuart and Donald Geman (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. Pattern Anal. Mach. Intell.* 6.6, pp. 721–741.
- Gemino, Andrew and Yair Wand (2004). A framework for empirical evaluation of conceptual modeling techniques. *Requirements Engineering* 9.4, pp. 248–260.
- Gerken, Jan M and Martin G Moehrle (2012). A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis. *Scientometrics* 91.3, pp. 645–670.
- Gershenson, Carlos (2015). Requisite variety, autopoiesis, and self-organization. *Kybernetes* 44.6/7, pp. 866–873.
- Giblin-Delvallet, Béatrice (2004). Lille métropole. Une eurométropole en devenir? *Vingtième siècle. Revue d'histoire*, pp. 69–80.
- Giere, Ronald N (2010a). An agent-based conception of models and scientific representation. *Synthese* 172.2, pp. 269–281.
- (2010b). *Explaining science: A cognitive approach*. University of Chicago Press.
- (2010c). *Scientific perspectivism*. University of Chicago Press.
- Gierer, Alfred and Hans Meinhardt (1972). A theory of biological pattern formation. *Kybernetik* 12.1, pp. 30–39.
- Gilbert, Scott F (2003). The morphogenesis of evolutionary developmental biology. *International Journal of Developmental Biology* 47.7–8, p. 467.
- Gilli, Frédéric (2005). Le Bassin parisien. Une région métropolitaine. *Cybergeo: European Journal of Geography*.
- (2014). *Grand Paris. L'émergence d'une métropole*. Presses de Sciences Po, coll. Nouveaux débats.
- Gilli, Frédéric and Jean-Marc Offner (2009). *Paris, métropole hors les murs: aménager et gouverner un Grand Paris*. Presses de Sciences Po.
- Giraut, Frédéric (2008). Conceptualiser le territoire. *Historiens et géographes* 403, pp. 57–68.
- Girres, Jean-François and Guillaume Touya (2010). Quality assessment of the French OpenStreetMap dataset. *Transactions in GIS* 14.4, pp. 435–459.

- Glaeser, Edward (2011). *Triumph of the city: How our greatest invention makes us richer, smarter, greener, healthier, and happier*. Penguin.
- Gleyze, Jean-François (2005). La vulnérabilité structurelle des réseaux de transport dans un contexte de risques. PhD thesis. Université Paris-Diderot-Paris VII.
- Goffman, Erving (1989). On fieldwork. *Journal of contemporary ethnography* 18.2, pp. 123–132.
- Golden, Boris et al. (2012). Modeling of complex systems ii: A minimalist and unified semantics for heterogeneous integrated systems. *Applied Mathematics and Computation* 218.16, pp. 8039–8055.
- Goldner, William (1971). The Lowry model heritage. *Journal of the American Institute of Planners* 37.2, pp. 100–110.
- Gollini, Isabella et al. (2013). GWmodel: an R package for exploring spatial heterogeneity using geographically weighted models. *arXiv preprint arXiv:1306.0413*.
- Goryachev, Andrew B and Alexandra V Pokhilko (2008). Dynamics of Cdc42 network embodies a Turing-type mechanism of yeast cell polarity. *FEBS letters* 582.10, pp. 1437–1443.
- Gottmann, Jean (1961). *Megalopolis: the urbanized northeastern seaboard of the United States*. Twentieth Century Fund.
- Gregg, Jay S et al. (2009). The temporal and spatial distribution of carbon dioxide emissions from fossil-fuel use in North America. *Journal of Applied Meteorology and Climatology* 48.12, pp. 2528–2542.
- Griliches, Zvi (1990). *Patent Statistics as Economic Indicators: A Survey*. NBER Working Papers 3301. National Bureau of Economic Research.
- Grimm, Volker et al. (2005). Pattern-oriented modeling of agent-based complex systems: lessons from ecology. *Science* 310.5750, pp. 987–991.
- Grimm, Volker et al. (2014). Towards better modelling and decision support: documenting model development, testing, and analysis using TRACE. *Ecological modelling* 280, pp. 129–139.
- Gros, Julien (2017). Quantifier en ethnographe. *Genèses* 3, pp. 129–147.
- Guangdong Province (2013). *Guangdong Statistical Yearbook, 2013*.
- Guangzhou Metro (2016). 广州地铁, 2016年年报 [Metro de Guangzhou, rapport annuel 2016]. <http://www.gzmtr.com/ygwm/gsgk/qynb/201705/P020170531671790326154.pdf>.
- Guérin-Pace, France and Denise Pumain (1990). 150 ans de croissance urbaine. *Economie et statistique* 230.1, pp. 5–16.
- Guérois, Marianne and Renaud Le Goix (2009). La dynamique spatio-temporelle des prix immobiliers à différentes échelles: le cas des appartements anciens à Paris (1990-2003). *Cybergeo: European Journal of Geography*.
- Guérois, Marianne and Fabien Paulus (2002). Commune centre, agglomération, aire urbaine: quelle pertinence pour l'étude des villes? *Cybergeo: European Journal of Geography*.

- Guérois, Marianne and Denise Pumain (2008). Built-up encroachment and the urban field: a comparison of forty European cities. *Environment and Planning A* 40.9, pp. 2186–2203.
- Guillot, C. and T. Lencuit (2013). Mechanics of Epithelial Tissue. *Science* 340.June, pp. 1185–1189.
- Guo, Xiaolei and Henry X Liu (2011). Bounded rationality and irreversible network change. *Transportation Research Part B: Methodological* 45.10, pp. 1606–1618.
- Gurciullo, S. et al. (2015). Complex Politics: A Quantitative Semantic and Topological Analysis of UK House of Commons Debates. *arXiv preprint arXiv:1510.03797*.
- Gutmann, Amy (2011). The ethics of synthetic biology: guiding principles for emerging technologies. *Hastings Center Report* 41.4, pp. 17–22.
- Hacking, Ian (1999). *The social construction of what?* Harvard university press.
- Haggett, Peter and Richard J Chorley (1970). *Network analysis in geography*. St. Martin's Press.
- Haken, Herman and Juval Portugali (2003). The face of the city is its information. *Journal of Environmental Psychology* 23.4, pp. 385–408.
- Haken, Hermann (1980). Synergetics. *Naturwissenschaften* 67.3, pp. 121–128.
- Haklay, Mordechai (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and planning B: Planning and design* 37.4, pp. 682–703.
- Hall, Bronwyn H et al. (2001). *The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools*. CEPR Discussion Papers 3094.
- Hall, C Michael (2005). Reconsidering the geography of tourism and contemporary mobility. *Geographical Research* 43.2, pp. 125–139.
- Hall, Peter Geoffrey and Kathy Pain (2006). *The polycentric metropolis: learning from mega-city regions in Europe*. Routledge.
- Hamerly, Greg, Charles Elkan, et al. (2003). Learning the k in k-means. *NIPS*. Vol. 3, pp. 281–288.
- Han, Dong (2010). Policing and racialization of rural migrant workers in Chinese cities. *Ethnic and Racial Studies* 33.1, pp. 593–610.
- Han, Sangjin (2003). Dynamic traffic modelling and dynamic stochastic user equilibrium assignment for general road networks. *Transportation Research Part B: Methodological* 37.3, pp. 225–249.
- Hansen, Walter G (1959). How accessibility shapes land use. *Journal of the American Institute of planners* 25.2, pp. 73–76.
- Haran, EGP and Daniel R Vining (1973). A modified Yule-Simon model allowing for intercity migration and accounting for the observed form of the size distribution of cities. *Journal of Regional Science* 13.3, pp. 421–437.

- Harris, Paul et al. (2011). Geographically weighted principal components analysis. *International Journal of Geographical Information Science* 25.10, pp. 1717–1736.
- Hart, Carolyn (2013). Held in mind, out of awareness. Perspectives on the continuum of dissociated experience, culminating in dissociative identity disorder in children. *Journal of Child Psychotherapy* 39.3, pp. 303–318.
- Harvey, David (1969). *Explanation in geography*. London: Edward Arnold.
- Hatchuel, Armand et al. (1988). Des stations de métro en mouvement: Station 2000, un scénario prospectif. *Les Annales de la recherche urbaine*. Vol. 39. 1. Persée-Portail des revues scientifiques en SHS, pp. 35–42.
- Heddebaut, Odile and Jean-Marie Ernecq (2016). Does the "tunnel effect" still remains in 2016? *3e Colloque du programme Vingt années sous la Manche, et au-delà: "Régions accessibles, régions en croissance ?"* Canterbury, United Kingdom.
- Heisenberg, Carl Philipp and Yohanns Bellaïche (2013). Forces in tissue morphogenesis and patterning. *Cell* 153.5.
- Hidalgo, C. A. (2015). Disconnected! The parallel streams of network literature in the natural and social sciences. *arXiv preprint arXiv:1511.03981*.
- Hijmans, Robert J (2015). raster : Geographic data analysis and modeling. *R Package*.
- Hillier, Bill (2016). The Fourth Sustainability, Creativity: Statistical Associations and Credible Mechanisms. *Complexity, Cognition, Urban Planning and Design*. Springer, pp. 75–92.
- Hillier, Bill and Julienne Hanson (1989). *The social logic of space*. Cambridge university press.
- Hofmann, Thomas (1999). Probabilistic latent semantic indexing. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 50–57.
- Hofstadter, Douglas H (1980). *Gödel, Escher, Bach: An Eternal Golden Braid;[a Metaphoric Fugue on Minds and Machines in the Spirit of Lewis Carroll]*. Penguin Books.
- Holland, John H (2012). *Signals and boundaries: Building blocks for complex adaptive systems*. MIT Press.
- Holmes, Caroline M et al. (2017). Luria-Delbrück, revisited: the classic experiment does not rule out Lamarckian evolution. *Physical biology* 14.5, p. 055004.
- Holstein, James A and Jaber F Gubrium (2004). The active interview. *Qualitative research: Theory, method and practice* 2, pp. 140–161.
- Holzinger, Andreas et al. (2014). Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions. *BMC bioinformatics* 15.6, p. I1.
- Homocianu, Marius (2009). Transport-land use interaction modeling - Residential choices of households in urban area of Lyon. PhD thesis. Université Lumière - Lyon II.

- Hopkins, Philip F et al. (2008). A cosmological framework for the co-evolution of quasars, supermassive black holes, and elliptical galaxies. I. Galaxy mergers and quasar activity. *The Astrophysical Journal Supplement Series* 175.2, p. 356.
- Hou, Quan and Si-Ming Li (2011). Transport infrastructure development and changing spatial accessibility in the Greater Pearl River Delta, China, 1990–2020. *Journal of Transport Geography* 19.6, pp. 1350–1360.
- Hussain, N et al. (2011). Hong Kong Zuhai Macao Link. *Procedia Engineering* 14, pp. 1485–1492.
- Huutoniemi, Katri et al. (2010). Analyzing interdisciplinarity: Typology and indicators. *Research Policy* 39.1, pp. 79–88.
- Hypergeo (2017). Hypergeo. <http://www.hypergeo.eu>.
- INRA (2013). *Issues in Neuroscience Research and Application: 2013 Edition*. ScholarlyEditions.
- Iacono, Michael et al. (2008). Models of transportation and land use change: a guide to the territory. *Journal of Planning Literature* 22.4, pp. 323–340.
- Iacovacci, Jacopo et al. (2015). Mesoscopic structures reveal the network between the layers of multiplex data sets. *Physical Review E* 92.4, p. 042806.
- Igel, Christian (2005). Multi-objective model selection for support vector machines. *International Conference on Evolutionary Multi-Criterion Optimization*. Springer, pp. 534–546.
- Innes, Judith E et al. (2010). Strategies for megaregion governance: Collaborative dialogue, networks, and self-organization. *Journal of the American Planning Association* 77.1, pp. 55–67.
- Jacobs-Crisioni, Chris and Carl C Koopmans (2016). Transport link scanner: simulating geographic transport network expansion through individual investments. *Journal of Geographical Systems* 18.3, pp. 265–301.
- Jacobs, Jane (2016). *The death and life of great American cities*. Vintage.
- Jacomy, Mathieu et al. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS one* 9.6, e98679.
- Janzen, Daniel H (1980). When is it coevolution. *Evolution* 34.3, pp. 611–612.
- Jarrow, Robert A (1999). In Honor of the Nobel Laureates Robert C. Merton and Myron S. Scholes: A Partial Differential Equation that Changed the World. *The Journal of Economic Perspectives*, pp. 229–248.
- Jégou, Anne et al. (2012). L'évaluation par indicateurs: un outil nécessaire d'aménagement urbain durable?. Réflexions à partir de la démarche parisienne pour le géographe et l'aménageur. *Cybergeo: European Journal of Geography*.

- Jelokhani-Niaraki, Mohammadreza and Jacek Malczewski (2012). A web 3.0-driven collaborative multicriteria spatial decision support system. *Cybergeo: European Journal of Geography*.
- Johansson, Börje (1993). Infrastructure, accessibility and economic growth. *International Journal of Transport Economics/Rivista internazionale di economia dei trasporti*, pp. 131–156.
- Josselin, Didier and Marc Ciligt-Travain (2013). Revisiting the optimal center location. A spatial thinking based on robustness, sensitivity, and influence analysis. *Environment and Planning B: Planning and Design* 40.5, pp. 923–941.
- Josselin, Didier et al. (2016). Straightness of rectilinear vs. radio-concentric networks: modeling simulation and comparison. *arXiv preprint arXiv:1609.05719*.
- Jun, Joseph K and Alfred H Hübner (2005). Formation and structure of ramified charge transportation networks in an electromechanical system. *Proceedings of the National Academy of Sciences of the United States of America* 102.3, pp. 536–540.
- Kallis, Giorgos (2007). When is it coevolution? *Ecological Economics* 62.1, pp. 1–6.
- Kaplan, Sarah and Keyvan Vakili (2015). The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal* 36.10, pp. 1435–1457.
- Kasraian, Dena et al. (2015). Development of rail infrastructure and its impact on urbanization in the Randstad, the Netherlands. *Journal of Transport and Land Use* 9.1.
- Kasraian, Dena et al. (2016). Long-term impacts of transport infrastructure networks on land-use change: an international review of empirical studies. *Transport Reviews* 36.6, pp. 772–792.
- Katz, Michael L (1996). Remarks on the economic implications of convergence. *Industrial and Corporate Change* 5.4, pp. 1079–1095.
- Kay, Luciano et al. (2014). Patent overlay mapping: Visualizing technological distance. *Journal of the Association for Information Science and Technology* 65.12, pp. 2432–2443.
- Ke, Yan et al. (2007). Spatio-temporal shape and flow correlation for action recognition. *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, pp. 1–8.
- Keersmaecker, Marie-Laurence et al. (2003). Using fractal dimensions for characterizing intra-urban diversity: The example of Brussels. *Geographical analysis* 35.4, pp. 310–328.
- Knight, Frank B (1975). A predictive view of continuous time processes. *The annals of Probability*, pp. 573–596.
- Koch, Christof and Gilles Laurent (1999). Complexity and the nervous system. *Science* 284.5411, pp. 96–98.
- Koch, Julian and Simon Stisen (2017). Citizen science: A new perspective to advance spatial pattern evaluation in hydrology. *PLOS ONE* 12.5, pp. 1–20.

- Kolchinsky, Artemy et al. (2015). Modularity and the spread of perturbations in complex dynamical systems. *Physical Review E* 92.6, p. 060801.
- Kondo, Shigeru and Takashi Miura (2010). Reaction-diffusion model as a framework for understanding biological pattern formation. *science* 329.5999, pp. 1616–1620.
- Koning, Martin et al. (2013). Dessertes ferroviaires à grande vitesse et dynamisme économique local : Une analyse économétrique exploratoire sur les unités urbaines françaises. *ASRDLF 2013, 50ème colloque de l'Association des sciences régionales de langue française*. Belgium, 32p.
- Kotelnikova-Weiler, N. and Florent Le Néchet (2017). Bricolage. *Dictionnaire passionnel de la modélisation urbaine*. L'oeil d'or, Paris.
- Krugman, Paul (1992). *A dynamic spatial model*. Working Paper. National Bureau of Economic Research.
- (1998). Space: the final frontier. *The Journal of Economic Perspectives* 12.2, pp. 161–174.
- Kryvobokov, Marko et al. (2013). Comparison of Static and Dynamic Land Use-Transport Interaction Models. *Transportation Research Record: Journal of the Transportation Research Board* 2344.1, pp. 49–58.
- Kuhn, Thomas S (1970). *The structure of scientific revolutions*. The University of Chicago Press.
- Kwan, M.P. (2012). The uncertain geographic context problem. *Annals of the Association of American Geographers* 102.5, pp. 958–968.
- Kwan, Mei-Po (1998). Space-time and integral measures of individual accessibility: a comparative analysis using a point-based framework. *Geographical analysis* 30.3, pp. 191–216.
- L'Hostis, Alain et al. (2014). Contribution de la future ligne ferroviaire à grande vitesse Tours-Bordeaux au développement des réseaux des villes, une évaluation par le potentiel de contact. *HALSHS preprint : hal-01163644*.
- L'Hostis, Alain et al. (2016). A Multicriteria approach for choosing a new public transport system linked to urban development : a method developed in the Bahn.Ville project for a tram-train scenario in the Saint-Étienne region. *Recherche Transports Sécurité* 2016.1-2, pp. 17–25.
- L'Hostis, Alain et al. (2012). La ville orientée vers le rail. *Ville et mobilité*.
- LU, Yi et al. (2012). The Chengdu-Guiyang High-Speed Rail Influence on the Location Advantage and Functional Positioning of Yibin City [J]. *Journal of Changsha University of Science & Technology (Social Science)* 5, p. 015.
- Lagesse, C. (2015). Read Cities through their Lines. Methodology to characterize spatial graphs. *arXiv preprint arXiv:1512.01268*.

- Larivière, Vincent and Yves Gingras (2010). On the relationship between interdisciplinarity and scientific impact. *Journal of the Association for Information Science and Technology* 61.1, pp. 126–131.
- (2014). Measuring Interdisciplinarity. *Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact*, p. 187.
- Larroque, Dominique et al. (2002). *Paris et ses transports: XIXe-XXe siècles, deux siècles de décisions pour la ville et sa région*. Recherches/Ipraus.
- Laughlin, Robert B (2006). *A different universe: Reinventing physics from the bottom down*. Basic Books.
- Launer, Robert L and Graham N Wilkinson (2014). *Robustness in statistics*. Academic Press.
- Le Goix, Renaud (2010). Acteurs, collectivités locales et contextes locaux dans la production des lotissements périurbains. *Les premières Journées du Pôle Ville-Ville, Transport et Territoire, Quoi de neuf?-20 au 22 janvier 2010*.
- Le Néchet, Florent (2009). Quantifier l'éloignement au modèle de Bussière: monocentrisme contre "Acentrisme". *Neuvièmes rencontres de Théo Quant*, 19–p.
- (2010). Approche multiscalaire des liens entre mobilité quotidienne, morphologie et soutenabilité des métropoles européennes: cas de Paris et de la région Rhin-Ruhr. PhD thesis. Université Paris-Est.
- (2011a). Consommation d'énergie et mobilité quotidienne selon la configuration des densités dans 34 villes européennes. *Cybergeo: European Journal of Geography*.
- (2011b). Urban dynamics modelling with endogeneous transport infrastructures, in a polycentric region. *17th European Colloquium on Quantitative and Theoretical Geography*. Athènes, Greece.
- (2015). De la forme urbaine à la structure métropolitaine: une typologie de la configuration interne des densités pour les principales métropoles européennes de l'Audit Urbain. *Cybergeo: European Journal of Geography*.
- (2017). De l'étalement urbain aux régions métropolitaines polycentriques : formes de fonctionnement et formes de gouvernance. *Peupler la terre - De la préhistoire à l'ère des métropoles*. Presses Universitaires François Rabelais.
- Le Néchet, Florent and Juste Raimbault (2015). Modeling the emergence of metropolitan transport authority in a polycentric urban region. *Plurimondi. An International Forum for Research and Debate on Human Settlements* 7.15.
- Le Texier, Marion and Geoffrey Caruso (2017). Assessing geographical effects in spatial diffusion processes: The case of euro coins. *Computer, Environment and Urban Systems* 61.A, pp. 81–93.
- Lechner, Thomas et al. (2004). Procedural modeling of land use in cities. *CiteSeer*.

- Lecuit, Thomas and Pierre-françois Lenne (2007). Cell surface mechanics and the control of cell shape, tissue patterns and morphogenesis. *Nat Rev Mol Cell Biol* 8.August, pp. 633–644.
- Lee, Minjin and Petter Holme (2015). Relating land use and human intra-city mobility. *PloS one* 10.10, e0140152.
- Lee, Minjin et al. (2017). Morphology of travel routes and the organization of cities. *Nature communications* 8.1, p. 2229.
- Lee, SeongWoo et al. (2009). Determinants of crime incidence in Korea: a mixed GWR approach. *World conference of the spatial econometrics association*, pp. 8–10.
- Leeuw, Sander van der et al. (2009). The Long-Term Evolution of Social Organization. *Complexity Perspectives in Innovation and Social Change*. Dordrecht: Springer Netherlands, pp. 85–116.
- Lefort, Isabelle (2012). Le terrain: l’Arlésienne des géographes? *Annales de géographie*. 5. Armand Colin, pp. 468–486.
- Legavre, Jean Baptiste (1996). La «neutralité» dans l’entretien de recherche. Retour personnel sur une évidence. *Politix* 9.35, pp. 207–225.
- Lemoy, Rémi and Geoffrey Caruso (2017). Scaling evidence of the homothetic nature of cities. *arXiv preprint arXiv:1704.06508*.
- Lemoy, Rémi et al. (2017). Exploring the polycentric city with multi-worker households: an agent-based microeconomic model. *Computers, Environment and Urban Systems* 62, pp. 64–73.
- Lerner, Josh and Amit Seru (2015). The use and misuse of patent data: Issues for corporate finance and beyond. *Booth/Harvard Business School Working Paper*.
- Leung, Yee et al. (2000). Statistical tests for spatial nonstationarity based on the geographically weighted regression model. *Environment and Planning A* 32.1, pp. 9–32.
- Leurent, Fabien and Houda Boujnah (2014). A user equilibrium, traffic assignment model of network route and parking lot choice, with search circuits and cruising flows. *Transportation Research Part C: Emerging Technologies* 47, pp. 28–46.
- Levinson, David M (2011). The coevolution of transport and land use: An introduction to the Special Issue and an outline of a research agenda. *Journal of Transport and Land Use* 4.2.
- Levinson, David Matthew et al. (2007). The co-evolution of land use and road networks. *Transportation and traffic theory*, pp. 839–859.
- Levinson, David (2008). Density and dispersion: the co-development of land use and rail in London. *Journal of Economic Geography* 8.1, pp. 55–77.
- Levinson, David and Wei Chen (2005). Paving new ground: a Markov chain model of the change in transportation networks and land use. *Access to destinations*. Emerald Group Publishing Limited, pp. 243–266.

- Levinson, David and Feng Xie (2011). Does first last? the existence and extent of first mover advantages on spatial networks. *Journal of Transport and Land Use*.
- Leydesdorff, Loet (2007). Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of the Association for Information Science and Technology* 58.9, pp. 1303–1319.
- L'horty, Yannick and Florent Sari (2013). Le Grand Paris de l'emploi: l'extension des infrastructures de transport peut-elle avoir des effets positifs sur le chômage local? *Revue d'Économie Régionale & Urbaine* 3, pp. 461–489.
- Li, Guan-Cheng et al. (2014). Disambiguation and co-authorship networks of the US patent inventor database (1975–2010). *Research Policy* 43.6, pp. 941–955.
- Li, J and U Wilensky (2009). *NetLogo Sugarscape 3 Wealth Distribution model*.
- Li, Tongfei et al. (2016). Integrated co-evolution model of land use and traffic network design. *Networks and Spatial Economics* 16.2, pp. 579–603.
- Li, Ye et al. (2014). Modeling Corridor and Growth Pole Coevolution in Regional Transportation Network. *Transportation Research Record: Journal of the Transportation Research Board* 2466, pp. 144–152.
- Liao, Liao and Jean Pierre Gaudin (2017). L'ouverture au marché en Chine (années 1980-2000) et le développement économique local: une forme de gouvernance multi-niveaux? *Cybergeo: European Journal of Geography*.
- Liao, T Warren (2005). Clustering of time series data—a survey. *Pattern recognition* 38.11, pp. 1857–1874.
- Liaw, Andy, Matthew Wiener, et al. (2002). Classification and regression by randomForest. *R news* 2.3, pp. 18–22.
- Lissack, Michael (2013). Subliminal influence or plagiarism by negligence ? The Slodderwetenschap of ignoring the internet. *Journal of Academic Ethics*.
- Liu, Liu and Alain L'Hostis (2014). Transport and Land Use Interaction: A French Case of Suburban Development in the Lille Metropolitan Area (LMA). *Transportation Research Procedia* 4. Sustainable Mobility in Metropolitan Regions. mobil.TUM 2014. International Scientific Conference on Mobility and Transport. Conference Proceedings., pp. 120 –139.
- Liu, Wei et al. (2011). Discovering spatio-temporal causal interactions in traffic data streams. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 1010–1018.
- Liu, Zhili et al. (2012). Solving the last mile problem: Ensure the success of public bicycle system in Beijing. *Procedia-Social and Behavioral Sciences* 43, pp. 73–78.

- Livet, Pierre et al. (2010). Ontology, a Mediator for Agent-Based Modeling in Social Science. *Journal of Artificial Societies and Social Simulation* 13.1, p. 3.
- Livingstone, David N (1995). The spaces of knowledge: contributions toward a historical geography of science. *Environment and planning D* 13 (1), pp. 13–42.
- Livingstone, David N. (2003). *Putting science in its place: geographies of scientific knowledge*. The University of Chicago Press.
- Livingstone, David N. and Charles W. J. Withers, eds. (2005). *Geography and revolution*. The University of Chicago Press.
- Loi, Daniel (1985). Une étude de la causalité dans la géographie classique française.[L'exemple des premières thèses régionales]. *Espace géographique* 14.2, pp. 121–125.
- Losavio, Cinzia and Juste Rimbault (2017). Modeling Residential Dynamics in a Mega-city Region: the Case of Pearl River Delta, China. *Urban China International Conference*.
- Louail, Thomas et al. (2017). Crowdsourcing the Robin Hood effect in cities. *Applied Network Science* 2.1, p. 11.
- Louf, Rémi and Marc Barthelemy (2014a). A typology of street patterns. *Journal of The Royal Society Interface* 11.101, p. 20140924.
- (2014b). How congestion shapes cities: from mobility patterns to scaling. *Scientific reports* 4, p. 5561.
- (2014c). Scaling: lost in the smog. *Environment and Planning B: Planning and Design* 41.5, pp. 767–769.
- (2016). Patterns of residential segregation. *PloS one* 11.6, e0157476.
- Louf, Rémi et al. (2013). Emergence of hierarchy in cost-driven growth of spatial networks. *Proceedings of the National Academy of Sciences* 110.22, pp. 8824–8829.
- Louf, Rémi et al. (2014). Scaling in Transportation Networks. *PLoS ONE* 9.7, e102007.
- Lowry, Ira S (1964). *A model of metropolis*. Rand Corporation Santa Monica, CA.
- Luo, Qiang et al. (2013). Spatio-temporal Granger causality: A new framework. *NeuroImage* 79, pp. 241–263.
- Lusso, Bruno (2009). Les musées, un outil efficace de régénération urbaine? Les exemples de Mons (Belgique), Essen (Allemagne) et Manchester (Royaume-Uni). *Cybergeo: European Journal of Geography*.
- Luzeaux, Dominique (2015). A formal foundation of systems engineering. *Complex Systems Design & Management*. Springer, pp. 133–148.
- Macharis, Cathy et al. (2010). A decision analysis framework for intermodal transport: Comparing fuel price increases and the internalisation of external costs. *Transportation Research Part A: Policy and Practice* 44.7, pp. 550–561.

- Mahmassani, Hani S and Gang-Len Chang (1987). On boundedly rational user equilibrium in transportation systems. *Transportation science* 21.2, pp. 89–99.
- Mainzer, Klaus and Leon O Chua (2013). *Local activity principle*. World Scientific.
- Maisonobe, Marion (2013). Diffusion et structuration spatiale d'une question de recherche en biologie moléculaire. *Mappe Monde* 110.2, p. 13202.
- Makse, Hernán A et al. (1998). Modeling urban growth patterns with correlated percolation. *Physical Review E* 58.6, p. 7054.
- Makse, Hernán A et al. (1995). Modelling urban growth. *Nature* 377.1912, pp. 779–782.
- Mangin, David (2013). *Paris/Babel. Une métropole européenne*. Editions de la Villette.
- Mangin, David and Philippe Panerai (1999). *Projet urbain*. Parenthèses.
- Manson, Steven M (2001). Simplifying complexity: a review of complexity theory. *Geoforum* 32.3, pp. 405–414.
- (2008). Does scale exist? An epistemological scale continuum for complex human–environment systems. *Geoforum* 39.2, pp. 776–788.
- Mantegna, Rosario N and H Eugene Stanley (1999). *Introduction to econophysics: correlations and complexity in finance*. Cambridge university press.
- Marchionni, Caterina (2004). Geographical economics versus economic geography: towards a clarification of the dispute. *Environment and Planning A* 36.10, pp. 1737–1753.
- Marler, R Timothy and Jasbir S Arora (2004). Survey of multi-objective optimization methods for engineering. *Structural and multidisciplinary optimization* 26.6, pp. 369–395.
- Martinez-Conde, Susana and Stephen L. Macknik (2017). Opinion: Finding the plot in science storytelling in hopes of enhancing science communication. *Proceedings of the National Academy of Sciences* 114.31, pp. 8127–8129.
- Masucci, A Paolo et al. (2013). Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows. *Physical Review E* 88.2, p. 022812.
- Mehaffy, Michael W (2007). Notes on the genesis of wholes: Christopher Alexander and his continuing influence. *Urban Design International* 12.1, pp. 41–49.
- Mendeley (2015). *Mendeley Reference Manager*. <http://www.mendeley.com/>.
- Mesoudi, Alex (2017). Pursuing Darwin's curious parallel: Prospects for a science of cultural evolution. *Proceedings of the National Academy of Sciences* 114.30, pp. 7853–7860.
- Michaël, Bon et al. (2017). Novel processes and metrics for a scientific evaluation rooted in the principles of science - Version 1. *Self Journal of Science*.

- Miller, Harvey J (1999). Measuring space-time accessibility benefits within transportation networks: basic theory and computational procedures. *Geographical analysis* 31.1, pp. 1–26.
- Mimeur, Christophe (2016). The traces of speed between space and network. PhD thesis. Université de Bourgogne Franche-Comté.
- Mimeur, Christophe et al. (2017). Revisiting the structuring effect of transportation infrastructure: an empirical approach with the French Railway Network from 1860 to 1910. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*.
- Min, Wanli and Laura Wynter (2011). Real-time road traffic prediction with spatio-temporal correlations. *Transportation Research Part C: Emerging Technologies* 19.4, pp. 606–616.
- Moeckel, Rolf et al. (2003). Creating a synthetic population. *Proceedings of the 8th International Conference on Computers in Urban Planning and Urban Management* (CUPUM).
- Monod, J. (1970). *Le hasard et la Nécessité*. Points.
- Moore, Christopher and Stephan Mertens (2011). *The nature of computation*. Oxford University Press.
- Moosavi, V. (2017). Urban morphology meets deep learning: Exploring urban forms in one million cities, town and villages across the planet. *arXiv preprint arXiv:1709.02939*.
- Morency, Catherine (2005). Contributions à la modélisation totalement désagrégée des interactions entre mobilité urbaine et dynamiques spatiales. PhD thesis. École Polytechnique de Montréal.
- Moreno Regan, Omar (2016). Etude du comportement des tunnels en maçonnerie du métro parisien. PhD thesis. Université Paris Est.
- Moreno, Diego et al. (2012). Un automate cellulaire pour expérimenter les effets de la proximité dans le processus d'étalement urbain: le modèle Raumulus. *Cybergeo: European Journal of Geography*.
- Moreno, María del Carmen Calatrava et al. (2016). On the uncertainty of interdisciplinarity measurements due to incomplete bibliographic data. *Scientometrics* 107.1, pp. 213–232.
- Morin, Edgar (1976). *La Méthode, tome 1. la nature de la nature*. Le Seuil.
- (1980). *La Méthode, tome 2. La Vie de la Vie*. Le Seuil.
- (1986). *La Méthode, tome 3. La connaissance de la connaissance*. Le Seuil.
- (1991). *La Méthode, tome 4. Les idées*.
- (2001). *La Méthode, tome 5. L'humanité de l'humanité*. Le Seuil.
- (2004). *La Méthode, tome 6. Éthique*. Le Seuil.
- Morris, Bradley J et al. (2013). Gaming science: the “Gamification” of scientific thinking. *Frontiers in psychology* 4.
- Moudon, Anne Vernez (1997). Urban morphology as an emerging interdisciplinary field. *Urban morphology* 1.1, pp. 3–10.
- Moulin-Frier, C. et al. (2017). Embodied Artificial Intelligence through Distributed Adaptive Control: An Integrated Framework. *arXiv preprint arXiv:1704.01407*.

- Murphy, Alexander B (2012). Entente territorial: Sack and Raffestin on territoriality. *Environment and Planning D: Society and Space* 30.1, pp. 159–172.
- NLTK (2015). *Natural Language Toolkit*, Stanford University.
- Nakamasu, Akiko et al. (2009). Interactions between zebrafish pigment cells responsible for the generation of Turing patterns. *Proceedings of the National Academy of Sciences* 106.21, pp. 8429–8434.
- Nature (2015). Interdisciplinarity, Nature Special Issue. *Nature* 525.7569, pp. 289–418.
- Nelson, Richard R and Sidney G Winter (2009). *An evolutionary theory of economic change*. Harvard University Press.
- Neuman, Michael and Angela Hull (2009). The futures of the city region. *Regional Studies* 43.6, pp. 777–787.
- Newman, MEJ (2011). Complex systems: A survey. *arXiv preprint arXiv:1112.1440*.
- (2014). Prediction of highly cited papers. *EPL (Europhysics Letters)* 105.2, p. 28002.
- Newman, Mark EJ (2003). The structure and function of complex networks. *SIAM review* 45.2, pp. 167–256.
- Newman, Peter WG and Jeffrey R Kenworthy (1996). The land use—transport connection: An overview. *Land use policy* 13.1, pp. 1–22.
- Nichols, Leah G (2014). A topic model approach to measuring interdisciplinarity at the National Science Foundation. *Scientometrics* 100.3, pp. 741–754.
- Nicosia, Vincenzo et al. (2009). Extending the definition of modularity to directed graphs with overlapping communities. *Journal of Statistical Mechanics: Theory and Experiment* 2009.03, P03024.
- Niederreiter, H (1972). Discrepancy and convex programming. *Annali di matematica pura ed applicata* 93.1, pp. 89–97.
- Niizato, Takayuki et al. (2010). A model of network formation by Physarum plasmodium: interplay between cell mobility and morphogenesis. *Biosystems* 100.2, pp. 108–112.
- Nilsson, Isabelle M and Oleg A Smirnov (2016). Measuring the effect of transportation infrastructure on retail firm co-location patterns. *Journal of Transport Geography* 51, pp. 110–118.
- Nitsch, Volker (2005). Zipf zipped. *Journal of Urban Economics* 57.1, pp. 86–100.
- Noruzi, Alireza (2005). Google Scholar: The new generation of citation indexes. *Libri* 55.4, pp. 170–180.
- OECD (2009). *OECD Patent Statistics Manual*.
- O’Sullivan, David and Steven M Manson (2015). Do Physicists Have ‘Geography Envy’? And What Can Geographers Learn From It? *Annals of the Association of American Geographers*.
- O’Brien, Oliver et al. (2014). Mining bicycle sharing data for generating insights into sustainable transport systems. *Journal of Transport Geography* 34, pp. 262–273.

- Offner, Jean-Marc (1993). Les "effets structurants" du transport: mythe politique, mystification scientifique. *Espace géographique* 22.3, pp. 233–242.
- (2000). 'Territorial deregulation': Local authorities at risk from technical networks. *International journal of urban and regional research* 24.1, pp. 165–182.
- Offner, Jean-Marc and Denise Pumain (1996). *Réseaux et territoires-significations croisées*. Editions de l'Aube.
- Offner, Jean-Marc et al. (2014). Les effets structurants des infrastructures de transport. *Espace Géographique* 42, p–51.
- Olsen, Sherry (1982). Urban metabolism and morphogenesis. *Urban Geography* 3.2, pp. 87–109.
- Omodei, Elisa et al. (2017). Evaluating the impact of interdisciplinary research: a multilayer network approach. *Network Science* 5.2, pp. 235–246.
- OpenStreetMap (2012). *OpenStreetMap*. <http://www.openstreetmap.org>.
- Openshaw, Stan (1983). *From data crunching to model crunching - the dawn of a new era*. Pion Ltd.
- (1984). *The Modifiable Areal Unit Problem*. Norwich, UK: Geo Books.
- Ordeshook, Peter C (1986). *Game theory and political theory: An introduction*. Cambridge University Press.
- Orfeuil, Jean-Pierre and Marc Wiel (2012). *Grand Paris: sortir des illusions, approfondir les ambitions*. Scrineo.
- Osmosis (2016). *OSMOSIS*. <http://wiki.openstreetmap.org/wiki/Osmosis>.
- Ostrowetsky, S. & al. (2004). Les Villes Nouvelles, 30 ans après. *Espaces et Sociétés* n°119, 4/2004.
- Otamendi, Javier et al. (2008). Selection of the simulation software for the management of the operations at an international airport. *Simulation Modelling Practice and Theory* 16.8, pp. 1103–1112.
- Ozer, Pierre and Florence De Longueville (2005). Tsunami en Asie du Sud-Est: retour sur la gestion d'un cataclysme naturel apocalyptique. *Cybergeo: European Journal of Geography*.
- Padeiro, Miguel (2009). The Underground Off the Walls : lines Extensions and Urban Evolution of Parisian Suburbs. PhD thesis. Université Paris-Est.
- (2013). Transport infrastructures and employment growth in the Paris metropolitan margins. *Journal of Transport Geography* 31, pp. 44–53.
- Páez, Antonio and Darren M Scott (2005). Spatial statistics for urban analysis: a review of techniques with examples. *GeoJournal* 61.1, pp. 53–67.
- Páez, Antonio et al. (2012). Measuring accessibility: positive and normative implementations of various accessibility indicators. *Journal of Transport Geography* 25, pp. 141–153.

- Palchykov, Vasyl et al. (2016). Ground truth? Concept-based communities versus the external classification of physics manuscripts. *EPJ Data Science* 5.1, p. 28.
- Paquot, Thierry (2010). *L'abc de l'urbanisme*. IAU - UPEC, pp. 91–94.
- Park, Inchae and Byungun Yoon (2014). A semantic analysis approach for identifying patent infringement based on a product–patent map. *Technology Analysis & Strategic Management* 26.8, pp. 855–874.
- Paulley, Neil J and F Vernon Webster (1991). Overview of an international study to compare models and evaluate land-use and transport policies. *Transport Reviews* 11.3, pp. 197–222.
- Paulus, Fabien (2004). Coévolution dans les systèmes de villes: croissance et spécialisation des aires urbaines françaises de 1950 à 2000. PhD thesis. Université Panthéon-Sorbonne-Paris I.
- Perez-Riverol, Yasset et al. (2016). Ten Simple Rules for Taking Advantage of Git and GitHub. *PLoS Comput Biol* 12.7, pp. 1–11.
- Perez, Pascal et al. (2016). Agent-Based Modelling for Urban Planning Current Limitations and Future Trends. *International Workshop on Agent Based Modelling of Urban Systems*. Springer, pp. 60–69.
- Pfaender, Fabien (2009). Spatialisation de l'information. PhD thesis. Université de Technologie de Compiègne.
- Pichon Rivière, Enrique (2004). *Le processus groupal*. Érès.
- Picon, Antoine (2013). *Smart cities: théorie et critique d'un idéal auto-réalisateur*. B2.
- Piers, Craig et al. (2007). *Self-Organizing Complexity in Psychological Systems*. Jason Aronson, Incorporated.
- Pigozzi, Bruce Wm (1980). Interurban linkages through polynomially constrained distributed lags. *Geographical Analysis* 12.4, pp. 340–352.
- Piketty, Thomas (2013). *Le capital au XXIe siècle*. Le Seuil.
- Pintea, Camelia-M. et al. (2017). The generalized traveling salesman problem solved with ant algorithms. *Complex Adaptive Systems Modeling* 5.1, p. 8.
- Plassard, François (1977). *Les autoroutes et le développement régional*. Presses Universitaires de Lyon.
- Pohle, J. et al. (2017). Selecting the Number of States in Hidden Markov Models - Pitfalls, Practical Challenges and Pragmatic Solutions. *arXiv preprint arXiv:1701.08673*.
- Porter, Alan and Ismael Rafols (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics* 81.3, pp. 719–745.
- Porter, Alan et al. (2007). Measuring researcher interdisciplinarity. *Scientometrics* 72.1, pp. 117–147.
- Portugali, J. (2000). *Self-Organization and the City*. Berlin: Springer-Verlag.

- Portugali, Juval (2011). SIRN–Synergetic Inter-Representation Networks. *Complexity, Cognition and the City*, pp. 139–165.
- Potiron, Yoann (2016). Estimating the integrated parameter of the locally parametric model in high-frequency data. PhD thesis. The University of Chicago.
- Potiron, Yoann and Per Mykland (2015). Estimation of integrated quadratic covariation between two assets with endogenous sampling times. *arXiv preprint arXiv:1507.01033*.
- Preschitschek, Nina et al. (2013). Anticipating industry convergence: Semantic analyses vs IPC co-classification analyses of patents. *Foresight* 15.6, pp. 446–464.
- Prigogine, Ilya and Isabelle Stengers (1997). *The end of certainty*. Simon and Schuster.
- Pritchard, David R and Eric J Miller (2009). Advances in agent population synthesis and application in an integrated land use and transportation model. *Transportation Research Board 88th Annual Meeting*. 09-1686.
- Pumain, Denise (1997). Pour une théorie évolutive des villes. *Espace géographique* 26.2, pp. 119–134.
- (2003). Une approche de la complexité en géographie. *Géocarrefour* 78.1, pp. 25–31.
 - (2005). Cumulativité des connaissances. *Revue européenne des sciences sociales*. *European Journal of Social Sciences* XLIII-131, pp. 5–12.
 - (2008). The socio-spatial dynamics of systems of cities and innovation processes: a multi-level model. *The Dynamics of Complex Urban Systems*, pp. 373–389.
 - (2010). Une théorie géographique des villes. *Bulletin de la Société géographie de Liège* 55, pp. 5–15.
 - (2012a). Multi-agent system modelling for urban systems: The series of SIMPOP models. *Agent-based models of geographical systems*. Springer, pp. 721–738.
 - (2012b). Urban systems dynamics, urban growth and scaling laws: The question of ergodicity. *Complexity Theories of Cities Have Come of Age*. Springer, pp. 91–103.
 - (2014). Les effets structurants ou les raccourcis de l'explication géographique. *Espace géographique* 43.1, pp. 65–67.
 - (2015). Adapting the model of scientific publishing. *Cybergeo: European Journal of Geography*.
- Pumain, Denise et al. (2009). Innovation cycles and urban dynamics. *Complexity perspectives in innovation and social change*, pp. 237–260.
- Pumain, Denise and Romain Reuillon (2017a). An Innovative and Open Toolbox. *Urban Dynamics and Simulation Models*. Springer, pp. 97–117.
- (2017b). Evaluation of the SimpopLocal Model. *Urban Dynamics and Simulation Models*. Springer, pp. 37–56.

- (2017c). The SimpopLocal Model. *Urban Dynamics and Simulation Models*. Springer, pp. 21–35.
- (2017d). *Urban Dynamics and Simulation Models*. Springer International.
- Pumain, Denise and Benoît Riandey (1986). Le Fichier de l’Ined. *Espace, populations, sociétés* 4.2, pp. 269–277.
- Pumain, Denise and Marie-Claire Robic (2002). Le rôle des mathématiques dans une «révolution» théorique et quantitative: la géographie française depuis les années 1970. *Revue d’histoire des Sciences Humaines* 6.1, pp. 123–144.
- Pumain, Denise and Lena Sanders (2013). Theoretical principles in interurban simulation models: a comparison. *Environment and Planning A* 45.9, pp. 2243–2260.
- Pumain, Denise et al. (2006). An evolutionary theory for interpreting urban scaling laws. *Cybergeo: European Journal of Geography*.
- Putman, Stephen H (1975). Urban land use and transportation models: A state-of-the-art summary. *Transportation Research* 9.2, pp. 187–202.
- Putra, Doni PE and Klaus Baier (2009). Der Einfluss ungesteuerter Urbanisierung auf die Grundwasserressourcen am Beispiel der indonesischen Millionenstadt Yogyakarta. *Cybergeo: European Journal of Geography*.
- Puzis, Rami et al. (2013). Augmented betweenness centrality for environmentally aware traffic monitoring in transportation networks. *Journal of Intelligent Transportation Systems* 17.1, pp. 91–105.
- QGis, DT (2011). Quantum GIS geographic information system. *Open Source Geospatial Foundation Project*.
- Querriaux, Xavier et al. (2004). Localisation optimale d’unités de soins dans un pays en voie de développement: analyse de sensibilité. *Cybergeo: European Journal of Geography*.
- R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.
- Raddick, M. J. et al. (2010). Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers. *Astronomy Education Review* 9.1, p. 010103.
- Raffestin, Claude (1987). Repères pour une théorie de la territorialité humaine. *Cahier/Groupe Réseaux* 3.7, pp. 2–22.
- Raimbault, J. and J. Gonzalez (May 2015). Application de la Morphogénèse de Réseaux Biologiques à la Conception Optimale d’Infrastructures de Transport. *Rencontres du Labex Dynamites*.
- Raimbault, Juste (2015). Hybrid Modeling of a Bike-Sharing Transportation System. *International Conference on Computational Social Science*.
- Raimbault, Juste (2016a). For a Cautious Use of Big Data and Computation. *Royal Geographical Society-Annual Conference 2016-Session: Geocomputation, the Next 20 Years* (1).

- Raimbault, Juste (2016b). Generation of Correlated Synthetic Data. *Actes des Journées de Rochebrune 2016*.
- Raimbault, Juste (2016c). Indirect Bibliometrics by Complex Network Analysis. *20e Anniversaire de Cybergeo*.
- Raimbault, Juste (2016d). Models of growth for system of cities: Back to the simple. *Conference on Complex Systems 2016*.
- (2016e). *TorPool v1.0, DOI : 10.5281/zenodo.53739*.
- (2017a). A Discrepancy-Based Framework to Compare Robustness Between Multi-attribute Evaluations. *Complex Systems Design & Management*. Springer, pp. 141–154.
- (2017b). A macro-scale model of co-evolution for cities and transportation networks. *Medium International Conference*. Sun-Yat Sen University.
- Raimbault, Juste (2017c). An Applied Knowledge Framework to Study Complex Systems. *Complex Systems Design & Management*, pp. 31–45.
- Raimbault, Juste (2017d). Co-construire Modèles, Etudes Empiriques et Théories en Géographie Théorique et Quantitative: le cas des Interactions entre Réseaux et Territoires. *Treizièmes Rencontres de ThéoQuant*.
- Raimbault, Juste (2017). Exploration of an Interdisciplinary Scientific Landscape. *arXiv preprint arXiv:1712.00805*.
- Raimbault, Juste (2017a). Identification de causalités dans des données spatio-temporelles. *Spatial Analysis and GEOmatics 2017*.
- Raimbault, Juste (2017b). Investigating the Empirical Existence of Static User Equilibrium. *Transportation Research Procedia* 22C, pp. 450–458.
- Raimbault, Juste (2017c). Modeling the Co-evolution of Urban Form and Transportation Networks. *Conference on Complex Systems 2017*.
- Raimbault, Juste (2017d). Models coupling urban growth and transportation network growth: An algorithmic systematic review approach. *Plurimondi 17*.
- Raimbault, Juste (2017e). Un Cadre de Connaissances pour une Géographie Intégrée. *Journée des jeunes chercheurs de l'Institut de Géographie de Paris*.
- Raimbault, Juste (2018a). An Urban Morphogenesis Model Capturing Interactions between Networks and Territories. *forthcoming in Mathematics of Urban Morphology*. D'Acci L., ed. Springer Nature - Birkhäuser Mathematics.
- Raimbault, Juste (2018b). Calibration of a Density-based Model of Urban Morphogenesis. *PLoS ONE, in revision*.
- Raimbault, Juste (2018). Co-evolution and morphogenetic systems. *arXiv preprint arXiv:1803.11457*.
- Raimbault, Juste (2018). Complexity, Complexities and Complex Knowledges. *forthcoming in Theories and models of urbanization*. Pumain D., ed. Springer Lecture Notes in Morphogenesis.

- Raimbault, Juste (2018). Indirect Evidence of Network Effects in a System of Cities. *Environment and Planning B, in revision*.
- Raimbault, Juste (2018a). Models for the Co-evolution of Cities and Networks. *forthcoming in Handbook on Cities and Networks, Rozenblat C., Neal Z., eds.*
- (2018b). Unveiling co-evolutionary patterns in systems of cities : systematic exploration of the SimpopNet model. *forthcoming in Theories and models of urbanization. Pumain D., ed. Springer Lecture Notes in Morphogenesis.*
- Raimbault, Juste and Solène Baffi (2017). Structural Segregation: Assessing the impact of South African Apartheid on Underlying Dynamics of Interactions between Networks and Territories. *European Colloquium in Theoretical and Quantitative Geography 2017.*
- Raimbault, Juste et al. (2014). A Hybrid Network/Grid Model of Urban Morphogenesis and Optimization. *4th International Conference on Complex Systems and Applications (ICCSA 2014)*, pp. 51–60.
- Raimbault, Juste and Antonin Bergeaud (2017). The Cost of Transportation: Spatial Analysis of Fuel Prices in the US. *European Working Group in Transportation 2017 Conference.*
- Ram, Karthik (2013). Git can facilitate greater reproducibility and increased transparency in science. *Source code for biology and medicine* 8.1, p. 7.
- Ramsey, James B (2002). Wavelets in economics and finance: Past and future. *Studies in Nonlinear Dynamics & Econometrics* 6.
- Rasmussen, Thomas Kjær et al. (2015). Stochastic user equilibrium with equilibrated choice sets: Part II-Solving the restricted SUE for the logit family. *Transportation Research Part B: Methodological* 77, pp. 146–165.
- Read, Dwight et al. (2009). The innovation innovation. *Complexity perspectives in innovation and social change*. Springer, pp. 43–84.
- Redner, Sidney (1998). How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems* 4.2, pp. 131–134.
- Reid, Chris R et al. (2016). Decision-making without a brain: how an amoeboid organism solves the two-armed bandit. *Journal of The Royal Society Interface* 13.119, p. 20160030.
- Rémy, Jean (2000). Métropolisation et diffusion de l'urbain: les ambiguïtés de la mobilité. *Les territoires de la mobilité*. Presses Universitaires de France, pp. 171–188.
- Renfrew, Colin (1978). Trajectory discontinuity and morphogenesis: the implications of catastrophe theory for archaeology. *American Antiquity*, pp. 203–222.
- Retaillé, Denis (2010). Au terrain, un apprentissage. *L'information géographique* 74.1, pp. 84–96.

- Reuillon, Romain et al. (2013). OpenMOLE, a workflow engine specifically tailored for the distributed exploration of simulation models. *Future Generation Computer Systems* 29.8, pp. 1981–1990.
- Reuillon, Romain et al. (2015). A New Method to Evaluate Simulation Models: The Calibration Profile (CP) Algorithm. *Journal of Artificial Societies and Social Simulation* 18.1, p. 12.
- Rey-Coyrehourcq, Sébastien (2015). Une plateforme intégrée pour la construction et l'évaluation de modèles de simulation en géographie. PhD thesis. Université Paris 1 Panthéon-Sorbonne.
- Reymond, Henri and Colette Cauvin (2013). La logique ternaire de Stéphane Lupasco et le raisonnement géocartographique bioculturel d'*Homo geographicus*. L'apport de la notion de couplage transdisciplinaire dans l'approche de l'agrégation morphologique des agglomérations urbaines. *Cybergeo: European Journal of Geography*.
- Reynard, Emmanuel et al. (2015). An Application for Geosciences Communication by Smartphones and Tablets. *Engineering Geology for Society and Territory-Volume 8*. Springer, pp. 265–268.
- Rietveld, Piet (1994). Spatial economic impacts of transport infrastructure supply. *Transportation Research Part A: Policy and Practice* 28.4. Special Issue Transport Externalities, pp. 329 –341.
- Rietveld, Piet et al. (2001). Spatial graduation of fuel taxes; consequences for cross-border and domestic fuelling. *Transportation Research Part A: Policy and Practice* 35.5, pp. 433–457.
- Rietveld, Piet and Stefan van Woudenberg (2005). Why fuel prices differ. *Energy Economics* 27.1, pp. 79–92.
- Rinia, Ed et al. (2002). Impact measures of interdisciplinary research in physics. *Scientometrics* 53.2, pp. 241–248.
- Ripoll, Fabrice (2017). Géographie de l'alternatif, Géographies alternatives ? Grand Témoin. *Journée des Jeunes Chercheurs de l'Institut de Géographie de Paris*.
- Robic, Marie-Claire (1982). Cent ans avant Christaller... une théorie des lieux centraux. *Espace géographique* 11.1, pp. 5–12.
- Rocca, Jean-Louis (2008). Power of knowledge: The imaginary formation of the Chinese middle stratum in an era of growth and stability. *Patterns of middle class consumption in India and China*, pp. 127–139.
- Rogers, Katherine W and Alexander F Schier (2011). Morphogen Gradients : From Generation to Interpretation. *Annu Rev Cell Dev Biol*.
- Romer, Paul M (1990). Endogenous Technological Change. *Journal of Political Economy* 98.5, S71–102.
- Ross-Hellauer, T (2017). What is open peer review? A systematic review [version 1; referees: 1 approved, 2 approved with reservations]. *F1000Research* 6.588.

- Roth, Camille (2009). Reconstruction Failures: Questioning Level Design. *Epistemological Aspects of Computer Simulation in the Social Sciences*. Springer, pp. 89–98.
- Roth, Camille and Jean-Philippe Cointet (2010). Social and semantic coevolution in knowledge networks. *Social Networks* 32.1, pp. 16–29.
- Rouleau, Bernard (1985). *Villages et faubourgs de l'ancien Paris: histoire d'un espace urbain*. Éditions du Seuil.
- Roumboutsos, Athena and Seraphim Kapros (2008). A game theory approach to urban public transport integration policy. *Transport Policy* 15.4, pp. 209–215.
- Rozenfeld, Hernán D et al. (2008). Laws of population growth. *Proceedings of the National Academy of Sciences* 105.48, pp. 18702–18707.
- Rubner, Yossi et al. (2000). The earth mover's distance as a metric for image retrieval. *International journal of computer vision* 40.2, pp. 99–121.
- Rucker, Gerta (2012). Network meta-analysis, electrical networks and graph theory. *Research Synthesis Methods* 3.4, pp. 312–324.
- Rui, Yikang and Yifang Ban (2011). Urban growth modeling with road network expansion and land use development. *Advances in Cartography and GIScience. Volume 2*. Springer, pp. 399–412.
- (2014). Exploring the relationship between street centrality and land use in Stockholm. *International Journal of Geographical Information Science* 28.7, pp. 1425–1438.
- Rui, Yikang et al. (2013). Exploring the patterns and evolution of self-organized urban street networks through modeling. *The European Physical Journal B* 86.3, pp. 1–8.
- Rushing Dewhurst, D. et al. (2017). Continuum rich-get-richer processes: Mean field analysis with an application to firm size. *arXiv preprint arXiv:1710.07580*.
- SDRIF (2013). *Île-de-France 2030. Orientations réglementaires et carte de destination générale des différentes parties du territoire*.
- STIF (2010). *ArcExpress, débat public sur le métro de rocade. Dossier du Maître d'Ouvrage*.
- Salton, Gerard and Michael J. McGill (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc.
- Samaniego, Horacio and Melanie E Moses (2008). Cities as organisms: Allometric scaling of urban road networks. *Journal of Transport and Land use* 1.1.
- Sanders, J. B. T. et al. (2016). The prevalence of chaotic dynamics in games with many players. *arXiv preprint arXiv:1612.08111*.
- Sanders, Lena (1992). *Système de villes et synergétique*. Economica.
- (2017). *Peupler la terre - De la préhistoire à l'ère des métropoles*. Presses Universitaires François Rabelais.
- Sanders, Lena et al. (1997). SIMPOP: a multiagent system for the study of urbanism. *Environment and Planning B* 24, pp. 287–306.

- Santamaria, Frédéric (2009). Le Schéma de développement de l'espace communautaire (SDEC): application défaillante ou élaboration problématique? *Cybergeo: European journal of geography*.
- Sarigöl, Emre et al. (2014). Predicting scientific success based on coauthorship networks. *EPJ Data Science* 3.1, p. 9.
- Sayama, Hiroki (2007). Decentralized control and interactive design methods for large-scale heterogeneous self-organizing swarms. *European Conference on Artificial Life*. Springer, pp. 675–684.
- Schamp, Eike W (2010). 20 On the notion of co-evolution in economic geography. *The handbook of evolutionary economic geography*, p. 432.
- Schlosshauer, Maximilian (2005). Decoherence, the measurement problem, and interpretations of quantum mechanics. *Reviews of Modern physics* 76.4, p. 1267.
- Schmid, Helmut (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of the international conference on new methods in language processing*. Vol. 12, pp. 44–49.
- Schmitt, Clara (2014). Modélisation de la dynamique des systèmes de peuplement: de SimpopLocal à SimpopNet. PhD thesis. Paris 1.
- Schmitt, Clara et al. (2015). Half a billion simulations: Evolutionary algorithms and distributed computing for calibrating the SimpopLocal geographical model. *Environment and Planning B: Planning and Design* 42.2, pp. 300–315.
- Schwarz, Nina (2010). Urban form revisited—Selecting indicators for characterising European cities. *Landscape and Urban Planning* 96.1, pp. 29–47.
- Seidl, David (2004). Luhmann's theory of autopoietic social systems. *Ludwig-Maximilians-Universität München-Munich School of Management*.
- Serra, Hélène and Juste Raimbault (2016). *Game-based tools to transmit freshwater ecology concepts*. SETAC 2016.
- Servais, Marc et al. (2004). Polycentrisme urbain: une réalité spatialement mesurable? *Cybergeo: European Journal of Geography*.
- Seth, Anil K (2005). Causal connectivity of evolved neural networks during behavior. *Network: Computation in Neural Systems* 16.1, pp. 35–54.
- Seto, Karen C et al. (2012). Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools. *Proceedings of the National Academy of Sciences* 109.40, pp. 16083–16088.
- Shalizi, Cosma Rohilla and James P Crutchfield (2001). Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of statistical physics* 104.3-4, pp. 817–879.
- Sheeren, David et al. (2015). Coévolution des paysages et des activités agricoles dans différents territoires d'élevage des montagnes françaises: entre intensification et déprise agricole. *Fourrages* 222, pp. 103–113.

- Shenzhen Planning Bureau (2016). 关于地铁2号线东延线、地铁3号线西延线线站位初步方案 [*À propos de la ligne est de la ligne 2 du métro, ligne 3 du métro de l'Ouest*]. Urban Planning Commission.
- Shibata, Naoki et al. (2008). Detecting emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation* 28.11, pp. 758–775.
- Silva, Filipe Batista e et al. (2013). A high-resolution population grid map for Europe. *Journal of Maps* 9.1, pp. 16–28.
- Simon, Herbert A. (1955). On a Class of Skew Distribution Functions. English. *Biometrika* 42.3/4, pp. 425–440.
- Skott, Peter and Paul Auerbach (1995). Cumulative causation and the “new” theories of economic growth. *Journal of Post Keynesian Economics* 17.3, pp. 381–402.
- Sorenson, Olav et al. (2006). Complexity, networks and knowledge flow. *Research policy* 35.7, pp. 994–1017.
- Souami, Taoufik (2012). *Ecoquartiers: secrets de fabrication*. Scrineo.
- Stanley, H Eugene et al. (1999). Econophysics: Can physicists contribute to the science of economics? *Physica A: Statistical Mechanics and its Applications* 269.1, pp. 156–169.
- Stevens, Forrest R. et al. (2015). Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. *PLoS ONE* 10.2, pp. 1–22.
- Stodden, Victoria (2010). The scientific method in practice: Reproducibility in the computational sciences. *MIT Sloan research paper*.
- Storper, Michael and Allen J Scott (2009). Rethinking human capital, creativity and urban growth. *Journal of economic geography* 9.2, pp. 147–167.
- Strauss, Sharon Y et al. (2005). Toward a more trait-centered approach to diffuse (co) evolution. *New Phytologist* 165.1, pp. 81–90.
- Sullivan, JL et al. (2010). Identifying critical road segments and measuring system-wide robustness in transportation networks with isolating links: a link-based capacity-reduction approach. *Transportation Research Part A: Policy and Practice* 44.5, pp. 323–336.
- Swerts, Elfie (2017). A data base on Chinese urbanization: ChinaCities. *Cybergeo: European Journal of Geography*.
- Swerts, Elfie and Eric Denis (2015). Megacities: The Asian Era. *Urban Development Challenges, Risks and Resilience in Asian Mega Cities*. Springer, pp. 1–28.
- Tadmor, Eitan (2012). A review of numerical methods for nonlinear partial differential equations. *Bulletin of the American Mathematical Society* 49.4, pp. 507–554.
- Tahamtan, Iman and Lutz Bornmann (2018). Core elements in the process of citing publications: Conceptual overview of the literature. *Journal of Informetrics* 12.1, pp. 203–216.
- Tan, Wei et al. (2013). Social-network-sourced big data analytics. *IEEE Internet Computing* 17.5, pp. 62–69.

- Tannier, Cécile (2003). Trois modèles pour mieux comprendre la localisation des commerces de détail en milieu urbain. *L'Espace géographique* 32.3, pp. 224–238.
- Tannier, Cécile et al. (2010). Simulation fractale d'urbanisation. MUP-city, un modèle multi-échelle pour localiser de nouvelles implantations résidentielles. *Revue Internationale de Géomatique* 20.3, pp. 303–329.
- Tardy, Christine (2004). The role of English in scientific communication: lingua franca or Tyrannosaurus rex? *Journal of English for academic purposes* 3.3, pp. 247–269.
- Taylor, Peter J (2016). A Polymath in City Studies. *Sir Peter Hall: Pioneer in Regional Planning, Transport and Urban Geography*. Springer, pp. 11–20.
- Ter Wal, Anne LJ and Ron Boschma (2011). Co-evolution of firms, industries and networks in space. *Regional studies* 45.7, pp. 919–933.
- Tero, Atsushi et al. (2006). Physarum solver: a biologically inspired method of road-network navigation. *Physica A: Statistical Mechanics and its Applications* 363.1, pp. 115–119.
- (2007). A mathematical model for adaptive transport network in path finding by true slime mold. *Journal of theoretical biology* 244.4, pp. 553–564.
- Tero, Atsushi et al. (2010). Rules for Biologically Inspired Adaptive Network Design. *Science* 327.5964, pp. 439–442.
- Thévenin, Thomas et al. (2013). Mapping the Distortions in Time and Space: The French Railway Network 1830–1930. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 46.3, pp. 134–143.
- Thom, René (1972). *Stabilité structurelle et morphogénèse*. InterÉditions.
- Thomas, Isabelle et al. (2018). City delineation in European applications of LUTI models: review and tests. *Transport Reviews* 38.1, pp. 6–32.
- Thompson, Darcy Wentworth (1942). *On growth and form*. Cambridge University Press.
- Tilman, David and Peter M Kareiva (1997). *Spatial ecology: the role of space in population dynamics and interspecific interactions*. Vol. 30. Princeton University Press.
- Tivadar, Mihai et al. (2014). OASIS—un Outil d'Analyse de la Ségrégation et des Inégalités Spatiales. *Cybergeo: European Journal of Geography*.
- Tolio, T et al. (2010). SPECIES—Co-evolution of products, processes and production systems. *CIRP Annals-Manufacturing Technology* 59.2, pp. 672–693.
- Torricelli, Gian Paolo (2002). Traversées alpines, ville et territoire: le paradoxe de la vitesse. *Revue de géographie alpine* 90.3, pp. 25–36.

- Tošić, Predrag T and Carlos Ordóñez (2017). Boolean Network Models of Collective Dynamics of Open and Closed Large-Scale Multi-agent Systems. *International Conference on Industrial Applications of Holonic and Multi-Agent Systems*. Springer, pp. 95–110.
- Trépanier, Martin et al. (2009). Calculation of transit performance measures using smartcard data. *Journal of Public Transportation* 12.1, p. 5.
- Tretyakov, Konstantin et al. (2011). Fast fully dynamic landmark-based estimation of shortest path distances in very large graphs. *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, pp. 1785–1794.
- Tsai, Yu-Hsin (2005). Quantifying urban form: compactness versus' sprawl'. *Urban studies* 42.1, pp. 141–161.
- Tsay, Ruey S. (2015). MTS: All-Purpose Toolkit for Analyzing Multivariate Time Series (MTS) and Estimating Multivariate Volatility Models. R package version 0.33.
- Tsekeris, Theodore and Nikolas Geroliminis (2013). City size, network structure and traffic congestion. *Journal of Urban Economics* 76.0, pp. 1 –14.
- Tseng, Yuen-Hsien et al. (2007). Text mining techniques for patent analysis. *Information Processing & Management* 43.5, pp. 1216–1247.
- Tumminello, Michele et al. (2005). A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences of the United States of America* 102, pp. 10421–10426.
- Turing, Alan M (1952a). The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London* 237, pp. 1–37.
- Turing, Alan Mathison (1952b). The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 237.641, pp. 37–72.
- Vallée, Julie (2009). Les disparités spatiales de santé en ville: l'exemple de Vientiane (Laos). *Cybergeo: European Journal of Geography*.
- Varenne, Franck (2010a). Framework for M&S with Agents in Regard to Agent Simulations in Social Sciences. *Activity-Based Modeling and Simulation*, pp. 53–84.
- (2010b). Les simulations computationnelles dans les sciences sociales. *Nouvelles Perspectives en Sciences Sociales* 5.2, pp. 17–49.
- (2017). *Théories et modèles en sciences humaines. Le cas de la géographie*. Editions Matériologiques.
- Varenne, Franck, Marc Silberstein, et al. (2013). *Modéliser & simuler. Epistémologies et pratiques de la modélisation et de la simulation, tome 1*.
- Varet, Suzanne (2010). Développement de méthodes statistiques pour la prédiction d'un gabarit de signature infrarouge. PhD thesis. Université Paul Sabatier-Toulouse III.

- Vattay, Gábor et al. (2015). Quantum criticality at the origin of life. *Journal of Physics: Conference Series*. Vol. 626. 1. IOP Publishing, p. 012023.
- Veiga, Allan Koch et al. (2017). A conceptual framework for quality assessment and management of biodiversity data. *PLOS ONE* 12.6, pp. 1–20.
- Verlinde, Erik (2017). Emergent gravity and the dark universe. *SciPost Physics* 2.3, p. 016.
- Viguié, Vincent et al. (2014). Downscaling long term socio-economic scenarios at city scale: A case study on Paris. *Technological forecasting and social change* 87, pp. 305–324.
- Visser, Hans and T De Nijs (2006). The map comparison kit. *Environmental Modelling & Software* 21.3, pp. 346–358.
- Volberda, Henk W and Arie Y Lewin (2003). Co-evolutionary dynamics within and between firms: From evolution to co-evolution. *Journal of management studies* 40.8, pp. 2111–2136.
- Von Bertalanffy, Ludwig (1972). The history and status of general systems theory. *Academy of Management Journal* 15.4, pp. 407–426.
- Wal, Anne L. J. Ter and Ron Boschma (2011). Co-evolution of Firms, Industries and Networks in Space. *Regional Studies* 45.7, pp. 919–933.
- Wang, Jiang-Jiang et al. (2009). Review on multi-criteria decision analysis aid in sustainable energy decision-making. *Renewable and Sustainable Energy Reviews* 13.9, pp. 2263–2278.
- Wang, Y.-S. et al. (2017). Separable and Localized System Level Synthesis for Large-Scale Systems. *arXiv preprint arXiv:1701.05880*.
- Ward, Douglas P et al. (2000). A stochastically constrained cellular model of urban growth. *Computers, Environment and Urban Systems* 24.6, pp. 539–558.
- Wardrop, John Glen (1952). Some theoretical aspects of road traffic research. *Proceedings of the institution of civil engineers* 1.3, pp. 325–362.
- Watson, Benjamin et al. (2008). Procedural urban modeling in practice. *IEEE Computer Graphics and Applications* 3, pp. 18–26.
- Watson, Mark W (1993). Measures of fit for calibrated models. *Journal of Political Economy* 101.6, pp. 1011–1041.
- Wegener, Michael and Franz Fürst (2004). Land-use transport interaction: state of the art. Available at SSRN 1434678.
- Wegener, Michael et al. (1991). One city, three models: comparison of land-use/transport policy simulation models for Dortmund. *Transport Reviews* 11.2, pp. 107–129.
- Weibull, Jörgen W (1976). An axiomatic approach to the measurement of accessibility. *Regional Science and Urban Economics* 6.4, pp. 357–379.
- Weidlich, W. and G. Haag (1988). *Interregional Migration*. Berlin: Springer-Verlag.

- West, Geoffrey (2017). *Scale: The Universal Laws of Growth, Innovation, Sustainability, and the Pace of Life in Organisms, Cities, Economies, and Companies*. Penguin.
- White, R. (1977). Dynamical central place theory. *Geographical Analysis* 9, pp. 226–243.
- (1978). The simulation of central place dynamics: Two sector systems and the rank size rule. *Geographical Analysis* 10, pp. 201–208.
- Whitehand, JWR et al. (1999). Urban morphogenesis at the microscale: how houses change. *Environment and Planning B: Planning and Design* 26.4, pp. 503–515.
- Whitney, D. E. (2012). Growth Patterns of Subway/Metro Systems Tracked by Degree Correlation. *arXiv preprint arXiv:1202.174*.
- Wicherts, Jelte M. (2016). Peer Review Quality and Transparency of the Peer-Review Process in Open Access and Subscription Journals. *PLoS ONE* 11.1, e0147913.
- Wiener, Norbert (1948). *Cybernetics*. Hermann Paris.
- Wilensky, Uri (1999). NetLogo.
- Wilson, A. (1981). *Catastrophe theory and bifurcation: Application to Urban and Regional System*. Croom Helm.
- Wilson, A.G. (2002). Complex spatial systems: Challenges for modellers. *Mathematical and computer modelling* 36.3, pp. 379–387.
- Wilson, G et al. (2017). Good enough practices in scientific computing. *PLoS Comput Biol* 13.6, e1005510.
- Withers, Charles W. J. (2009). Place and the spatial turn in geography and history. *Journal of the History of Ideas* 70 (4), pp. 637–658.
- Wolfram, Stephen (2002). *A new kind of science*. Wolfram media Campaign.
- Wolpert, L (2011). Positional information and patterning revisited. *J Theor Biol* 269.1, pp. 359–365.
- Wolpert, Lewis (1969). Positional Information and the Spatial Pattern of Cellular Differentiation. *J Theor Biol*, pp. 1–47.
- Wu, Fulong (2016). China's Emergent City-Region Governance: A New Form of State Spatial Selectivity through State-orchestrated Rescaling. *International Journal of Urban and Regional Research* 40.6, pp. 1134–1151.
- Wu, Jianjun et al. (2017). City expansion model based on population diffusion and road growth. *Applied Mathematical Modelling* 43, pp. 1–14.
- Wu, QT et al. (2012). The impact of Hong Kong-Zhuhai-Macao bridge on the traffic pattern of Pearl River Delta. *Acta Geographica Sinica* 67.6, pp. 723–732.
- Wu, Weiping (2006). Migrant intra-urban residential mobility in urban China. *Housing Studies* 21.5, pp. 745–765.
- Xie, Feng and David Levinson (2009a). How streetcars shaped suburbanization: a Granger causality analysis of land use and transit in the Twin Cities. *Journal of Economic Geography*, lbpo31.

- Xie, Feng and David Levinson (2009b). Modeling the growth of transportation networks: A comprehensive review. *Networks and Spatial Economics* 9.3, pp. 291–307.
- (2011a). Governance Choice - A Theoretical Analysis. *Evolving Transportation Networks*. Springer, pp. 179–198.
- (2011b). Governance Choice-A Simulation Model. *Evolving Transportation Networks*. Springer, pp. 199–221.
- Xie, Yichun (1996). A Generalized Model for Cellular Urban Dynamics. *Geographical Analysis* 28.4, pp. 350–373.
- Xie, Yichun et al. (2007). Simulating emergent urban form using agent-based modeling: Desakota in the Suzhou-Wuxian region in China. *Annals of the Association of American Geographers* 97.3, pp. 477–495.
- Xie, Yihui (2013). *knitr: A general-purpose package for dynamic report generation in R*. R package version 1.7.
- Xu, Jiang and Anthony GO Yeh (2005). City repositioning and competitiveness building in regional development: New development strategies in Guangzhou, China. *International Journal of Urban and Regional Research* 29.2, pp. 283–308.
- Xu, Xue-qiang and Si-ming Li (1990). China's open door policy and urbanization in the Pearl River Delta region. *International Journal of Urban and Regional Research* 14.1, pp. 49–69.
- Yamasaki, Kazuko et al. (2006). Preferential attachment and growth dynamics in complex systems. *Physical Review E* 74.3, p. 035103.
- Yamins, Daniel et al. (2003). Growing urban roads. *Networks and Spatial Economics* 3.1, pp. 69–85.
- Yang, Jiawen (2006). Transportation implications of land development in a transitional economy: Evidence from housing relocation in Beijing. *Transportation Research Record: Journal of the Transportation Research Board* 1954, pp. 7–14.
- Yang, Y. et al. (2017). Urban Dreams of Migrants: A Case Study of Migrant Integration in Shanghai. *arXiv preprint arXiv:1706.00682*.
- Yang, Yiming et al. (2000). Improving text categorization methods for event tracking. *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 65–72.
- Yasmin, Farhana et al. (2017). Macro-, meso-, and micro-level validation of an activity-based travel demand model. *Transportmetrica A: Transport Science* 13.3, pp. 222–249.
- Ye, Lin (2014). State-led metropolitan governance in China: Making integrated city regions. *Cities* 41, Part B.o. Chinese Cities in a Globalizing Context, pp. 200 –208.
- Ye, Xin (2011). Investigation of Underlying Distributional Assumption in Nested Logit Model Using Copula-Based Simulation and Numerical Approximation. *Transportation Research Record: Journal of the Transportation Research Board* 2254, pp. 36–43.

- Yerra, Bhanu M and David Levinson (2005). The emergence of hierarchy in transportation networks. *The Annals of Regional Science* 39.3, pp. 541–553.
- Yoon, Byungun and Yongtae Park (2004). A text-mining-based patent network: Analytical tool for high-technology trend. *The Journal of High Technology Management Research* 15.1, pp. 37–50.
- Yoon, Janghyeok and Kwangsoo Kim (2011). Detecting signals of new technological opportunities using semantic patent analysis and outlier detection. *Scientometrics* 90.2, pp. 445–461.
- Youn, Hyejin et al. (2015). Invention as a combinatorial process: evidence from US patents. *Journal of The Royal Society Interface* 12.106.
- Zahavi, Yacov and Antti Talvitie (1980). Regularities in travel time and money expenditures. *Transportation Research Board*.
- Zembri, Pierre (1997). Les fondements de la remise en cause du Schéma Directeur des liaisons ferroviaires à grande vitesse: des faiblesses avant tout structurelles. *Annales de géographie*. JSTOR, pp. 183–194.
- (2008). La contribution de la grande vitesse ferroviaire à l'interrégionalité en France.(High-speed rail and inter-regionality in France). *Bulletin de l'Association de géographes français* 85.4, pp. 443–460.
- (2010). The new purposes of the French high-speed rail system in the framework of a centralized network: a substitute to the domestic air transport market? *ERSA 2010 - 50th Congress of the European Regional Science Association*.
- Zhang, Junfu and Zhong Zhao (2013). Measuring the income-distance tradeoff for rural-urban migrants in China. *IZA Discussion Paper No. 7160*.
- Zhang, Kevin Honglin and Song Shunfeng (2003). Rural–urban migration and urbanization in China: Evidence from time-series and cross-section analyses. *China Economic Review* 14.4, pp. 386–400.
- Zhang, Kuilin et al. (2013). Dynamic pricing, heterogeneous users and perception error: Probit-based bi-criterion dynamic stochastic user equilibrium assignment. *Transportation Research Part C: Emerging Technologies* 27, pp. 189–204.
- Zhang, Lei and David Levinson (2007). The economics of transportation network growth. *Essays on transport economics*. Springer, pp. 317–339.
- Zhang, Tonglin and Bingrou Zhou (2014). Test for the first-order stationarity for spatial point processes in arbitrary regions. *Journal of agricultural, biological, and environmental statistics* 19.4, pp. 387–404.
- Zhang, Yingjia et al. (2015). Density and diversity of OpenStreetMap road networks in China. *Journal of Urban Management* 4.2, pp. 135–146.
- Zhang, Zhonghao et al. (2013). Identifying determinants of urban growth from a multi-scale perspective: A case study of the urban

- agglomeration around Hangzhou Bay, China. *Applied Geography* 45, pp. 193–202.
- Zheng, Shudan and Jianghua Zheng (2014). Assessing the completeness and positional accuracy of OpenStreetMap in China. *Thematic Cartography for the Society*. Springer, pp. 171–189.
- Zhou, Suhong (2016). The Development of the PRD and the New Pathways for Sustainable Urban Development of Zhuhai. *Medium Seminar - Urban Sustainable Development in Zhuhai*. Sun Yat-Sen University.
- Zhu, Liping et al. (2013a). Amoeba-based computing for traveling salesman problem: Long-term correlations between spatially separated individual cells of *Physarum polycephalum*. *Biosystems* 112.1, pp. 1–10.
- Zhu, Shanjiang and David Levinson (2015). Do people use the shortest path? An empirical test of Wardrop's first principle. *PloS one* 10.8, e0134322.
- Zhu, Yaojia et al. (2013b). Scalable text and link analysis with mixed-topic link models. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 473–481.
- Zhuhai Tramway (2016). 珠海现代有轨电车 [Tramway moderne de Zhuhai]. <http://www.zhmrt.com.cn/>.
- Zilsel (2015). *La position de la revue sociétés dans l'espace discursif de la sociologie française*. <http://zilsel.hypotheses.org/category/canular>.
- Ziman, John (2003). *Technological innovation as an evolutionary process*. Cambridge University Press.

Part IV

ANNEXES

Appendices are organized in the logic of Knowledge Domains, after a linear presentation of diverse supplementary materials for each section of main text.

A

INFORMATIONS SUPPLÉMENTAIRES

This appendix gathers various supplementary materials, necessary for the robustness but not necessary to the main argument. It includes for example more precise model explorations, generally needed to support conclusions in main text but too long or repetitive to be included.

Elle inclut notamment les points suivants :

- Relevés de terrain en Chine en A.1, pour les résultats qualitatifs présentés en Chapitre 1.
- Précisions pour l'épistémologie quantitative de 2.2 en A.2.
- Résultats complets pour la modélographie de 2.3 en A.3.
- Pour les corrélations statiques de 4.1 : résultats pour la Chine, analyses de sensibilité, algorithme de simplification de réseau, dérivation analytiques pour le caractère multi-échelle en A.4.
- Dérivations pour l'expression des corrélations retardées sur données synthétiques de 4.2 en A.5.
- Comportement du modèle et étude semi-analytique du modèle d'agrégation-diffusion de 5.2 en 2.3.
- Corrélation faisable pour le couplage faible de 5.3 en A.7.
- Figures étendues pour l'exploration du modèle SimpopNet de 6.1 en A.8.
- Figures étendues pour l'exploration du modèle macroscopique de co-évolution de 6.2 en A.8.
- Détails du modèle *slime mould* utilisé en 7.1, et figures étendues en A.10
- Processus de calibration au second ordre du modèle mesoscopique de co-évolution de 7.2 en A.11.
- Pour le modèle Lutecia de 7.3, étude du modèle d'usage du sol, dérivation de probabilités de coopération, détails d'implémentation et d'initialisation en A.12.

* * *

*

A.1 FIELDWORK ELEMENTS

A.1.1 Fieldwork in China

Nous précisons la localisation géographique des territoires et lieux évoqués en 1.2 et en 1.3 dans les cartes suivantes. Nous donnons :

- Une carte en Fig. 66 à l'échelle du sud de la Chine, qui permet de localiser le Delta de la Rivière des Perles (qui inclut Guangzhou et Zhuhai), Chengdu et Leshan, ainsi que Yangshuo.
- Une carte en Fig. 67 à l'échelle du Delta de la Rivière des Perles, qui permet de localiser les principales villes : Guangzhou/Foshan, Dongguan, Zhongshan, Zhuhai et Shenzhen (ZES), ainsi que Hong-Kong et Macao (ZAS).
- Une carte en Fig. 68 à l'échelle de Zhuhai, qui permet de localiser les différents quartiers de Zhuhai : Gongbei, Xiangzhou, Tangjia, ainsi que la gare de Zhuhai Bei, le pont HZMB et les *New Territories* à Hong-Kong (nous désignons par quartier ici non pas des districts administratifs, puisque par exemple Tangjia fait partie du district de Xinwan, mais des quartiers vécus).

A.1.2 Fieldwork Notebook

FIGURE 68: Nous rendons compte ici de manière synthétique les différentes sorties de terrain alimentant la section 1.3. S'il n'est a priori pas standard de fournir de manière brute et ouverte le contenu des carnets de terrain, [Goffman, 1989] souligne que celui-ci peut être un matériau de recherche en lui-même. Les compte-rendus bruts et les photos sont disponibles de manière ouverte à <https://github.com/JusteRaimbault/CityNetwork/tree/master/Data/Fieldwork>.

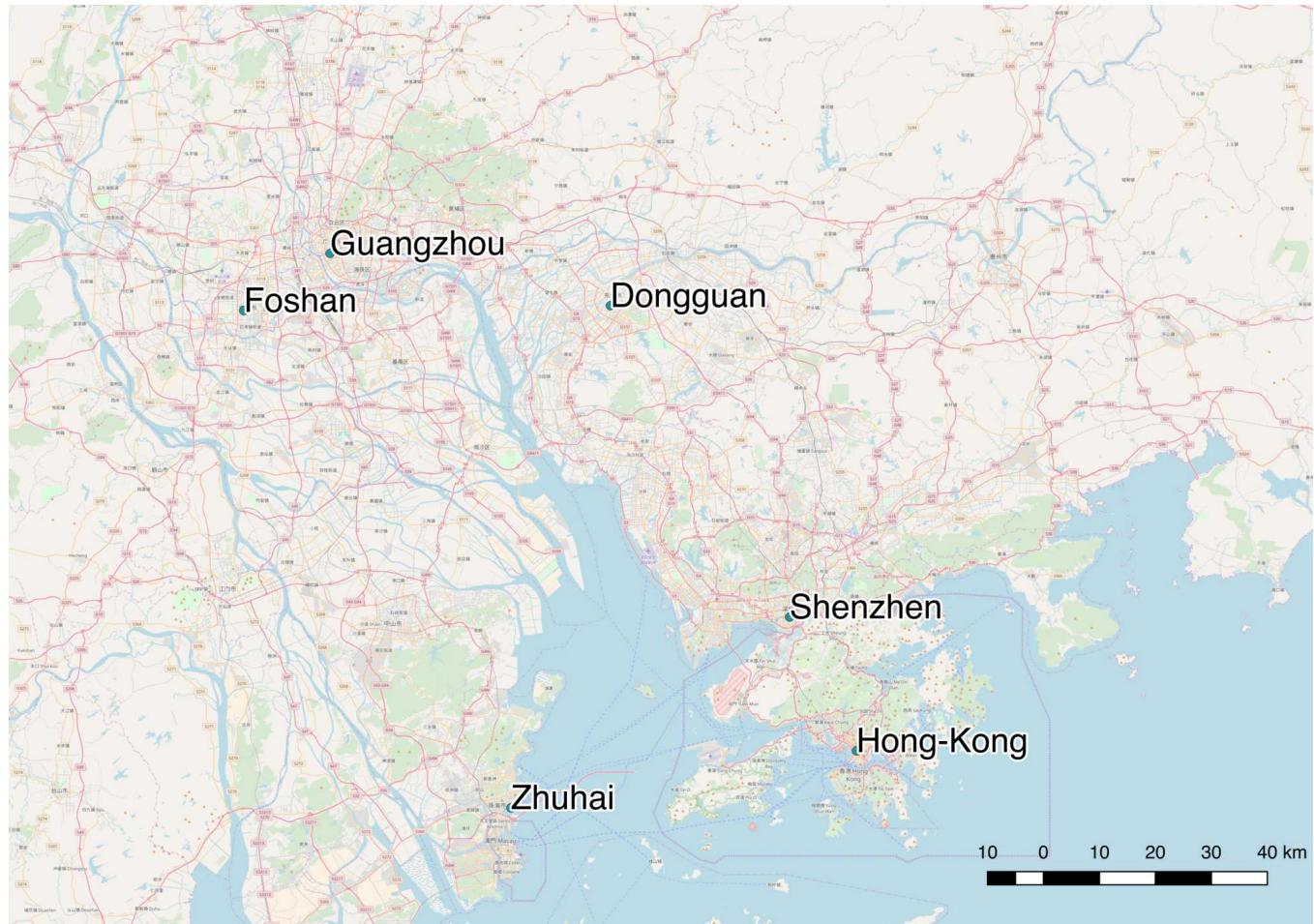
Ci-dessous sont résumés les contextes et observations principaux des sorties. Les lieux sont localisés dans les cartes de Fig. 66 à Fig. 68. Les sorties sont effectuées seul sauf si précisé pour certaines d'entre elles.

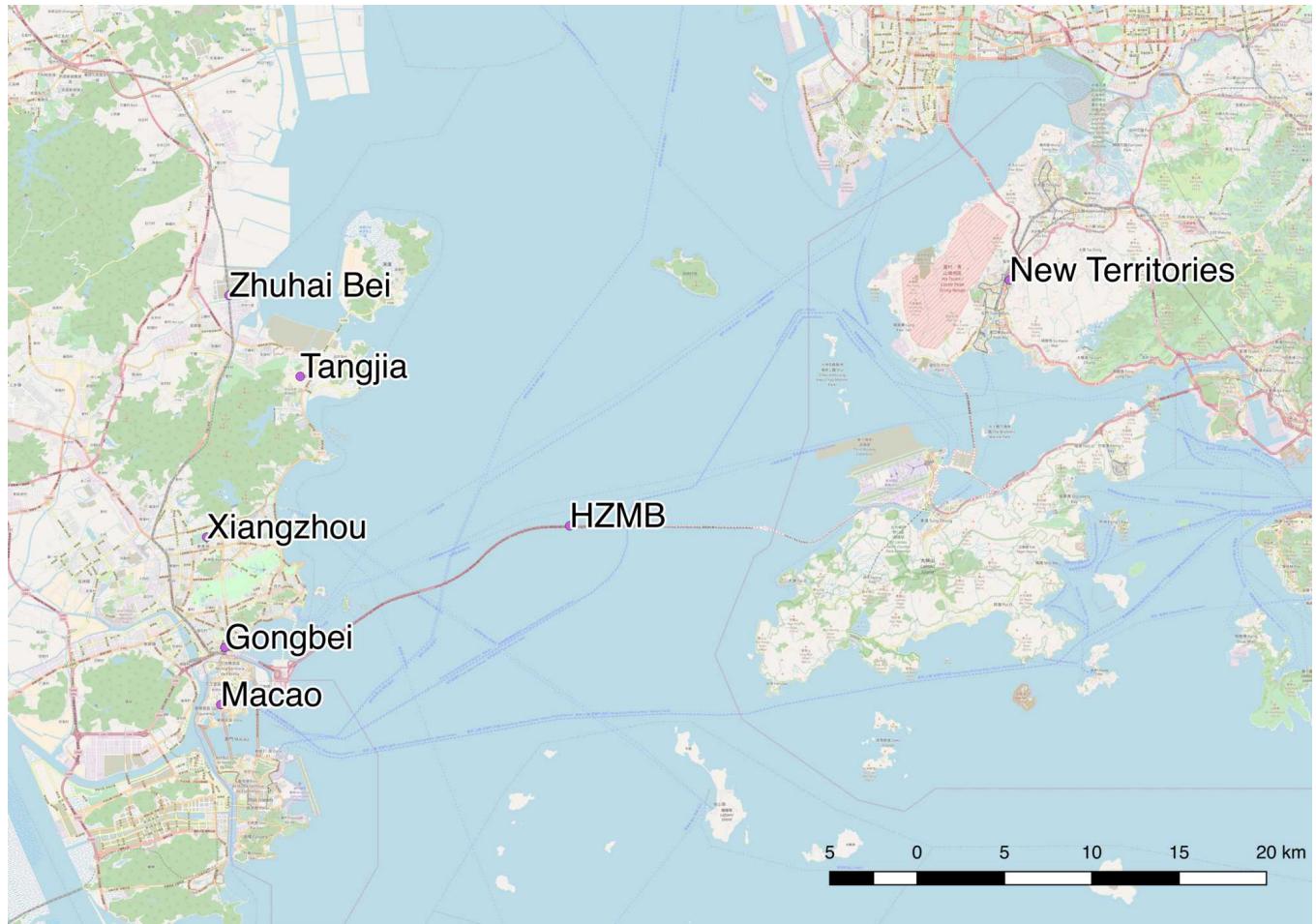
29 / 10 / 2016 Sortie à Zhuhai (Xiangzhou et Gongbei), avec C. Losavio pour guide et interprétation. Nature en ville et utilisation des parcs par les habitants.

06 / 11 / 2016 Sortie à Macao par Gongbei, avec C. Losavio. Flux journaliers par la frontière de la ZAS.

07 / 11 / 2016 Aller-retour Zhuhai-Hong-Kong. Relation apparente des habitants de Zhuhai à la ZAS.







16/01/2017 Tentative de relier Tangjia à Guangzhou par bus de ville, journée. Itinéraire final Tangjia-Zhongshan-Xiaolan-Zhuahaibei. Transports locaux et franges urbaines.

11/12/2016 De Pekin à Shenzhen par Guangzhou et Dongguan. Transports, difficultés d'accessibilité.

8/06/2017 De Hong-Kong à Tangjia par Zhuhai. Transports.

19/06/2017 Visite de terrain officielle dans le cadre de la Conférence Medium, Guangzhou, encadrée par guides et interprètes engagés par l'université SYSU. Rénovation Urbaine, projets urbains, patrimoine.

09/07/2017 Visite des New Territories à Hong-Kong : transport lourd depuis Kowloon puis différentes lignes de tramway sur place. Retour par le métro de Shenzhen puis par ferry jusqu'à Zhuhai.

11/07/2017 Aller-retour Tangjia-Guangzhou. Congestion des transports (routier et vélos libre-service).

24/07/2017 Sortie à Tangjia. Discontinuités socio-économiques locales.

31/07/2017 Sortie à Xiangzhou. Test du Tramway, Ligne 2.

09/08/2017 Sortie à Xiangzhou puis Tangjia. Opération de TOD : terminus ouest Tram ; bus pour la gare de Tangjia le long de la ligne à grande vitesse.

13/08/2017 De Yangshuo (Guanxi) à GuangzhouNan par le Train à Grande Vitesse.

17/08/2017 Bureau du Comité de Planification de la zone High Tech de Zhuhai. Administration et bureaucratie.

20/08/2017 Traversée de Leshan (Sichuan) en bus, aller-retour. Transports et Tourisme.

21/08/2017 De Guangzhou Baiyun à Zhongshan Daxue (campus sud de l'université SYSU) puis Tangjia. Transports, village urbain.

A.1.3 *Interviews*

Les "entretiens" menés relèvent de l'entretien actif non-structuré [Holstein and Gubrium, 2004] lors d'une mise en situation vécue conjoin-

tement. Les difficultés linguistiques de part et d'autre ont pu rendre compliqué les dialogues et nous donnons ici une synthèse des informations acquises. Les noms ont été modifiés lorsque l'accord explicite de l'interviewé n'a pas été obtenu. Dans cette synthèse narrative et subjective, la première personne désigne l'auteur.

12 / 08 / 2018 *Lin est une habitante de Guangzhou, originaire du Guanxi. Nous nous rencontrons au fond du dernier bus rentrant à Yangshuo après une visite à Pingxi. Un état d'ébriété facilite la prise de contact et la compréhension réciproque de mon très mauvais mandarin et de son mauvais anglais. Ils sont venus en week-end de team-building avec son équipe d'une start-up numérique. Une collègue aide à l'interprétation tandis que deux autres sont absolument absorbés dans une partie de Dota2 sur leur portable. Cette ville est la nouvelle destination tendance depuis qu'elle est à moins de deux heures de Guangzhou par la ligne à grande vitesse, elle est parait-il moins fréquentée que Guilin.*

Nous nous retrouvons plus tard dans le centre, après qu'elles se soient débarrassées de leur collègues qui cherchaient désespérément un poste internet fixe pour une nouvelle partie. Nous parlons de l'aspect touristique de ce centre-ville. Une foule de consommateurs se presse dans des ruelles pseudo-authentiques. Même les pics karstiques illuminés semblent faux à ce point. Des scouts communistes vendent des glaces aux lentilles, elles me disent qu'elles s'en méfient et que les glaces me donneront sûrement mal à l'estomac. Nous critiquons plus tard les bars à l'occidentale qui fleurissent dans ce genre de villes, elles me disent qu'ils sont fréquentés par "un certain type de personnes" (préjugé sociologique que je n'ai pas réussi à interpréter).

16 / 08 / 2016 *Je rencontre Zexian au restaurant en bas de la résidence Rencai Gongyu à Tangjia, où sont logés notamment les professeurs de l'université Zhongshan. Les commerces associés à ce complexe ne sont pas uniquement utilisés par les habitants locaux, et les gens (souvent des nouveaux riches vu le prix) viennent spécialement pour le tout nouveau KTV (karaoke). Elle me propose d'y aller à la suite. Elle me raconte qu'elle est étudiante dans un institut de langues à proximité de la port sud du campus. Elle étudie en particulier l'anglais, et voudrait qu'on reste en contact pour qu'elle puisse s'entraîner, nous échangeons alors les contact Weixin (Wechat).*

Elle m'explique que sa famille habite au sud de Zhongshan, à proximité donc, mais qu'il est très compliqué de rentrer. Le bus fait bien la connexion mais plusieurs changements sont nécessaires. Le train connecte Zhongshan à Zhuhai Bei ou Tangjia mais les gares sont peu accessibles, les horaires peu fréquentes à ces arrêts intermédiaires, et la réservation d'un billet compliquée. Elle prend le plus souvent un taxi sur demande via l'application Didi.

19-20 / 08 / 2018 *Xing est une jeune pékinoise d'une trentaine d'année rencontrée à l'entrée du Parc National d'Emeishan. Passé le délire de foule de la zone accessible aux voitures, peu de personnes souhaitent accomplir*

l'ascension initiatique intégralement, et nous nous parlons naturellement sur le chemin. Elle m'explique la signification de cette montagne et la portée symbolique de son ascension. Après la visite d'un ou deux temples, nous nous perdons.

Elle travaille à Pékin dans une entreprise de Design Industriel, c'est son premier emploi qu'elle a commencé il y a quelques mois. Son entreprise l'a envoyée passer un mois à Chengdu pour une formation. Elle a étudié à la Beijing Ligong Daxue (Université technologique de Beijing) et aurait souhaité partir étudier en Europe, mais les filières du domaine étaient trop sélectives. Elle parle allemand et y a fait une école d'été il y a quelques années. Elle est marathonienne mais confirme les difficultés à s'entraîner à Beijing, à cause de la pollution. Origininaire du Hebei, elle n'aime pas vivre à Beijing mais son travail l'y oblige. Le cadre de vie n'est pas particulièrement agréable et les problèmes de trafic sont pesants.

Elle me confirme l'aspect culturel du Jingye, l'une des Valeurs Centrales du Socialisme promues par la propagande du Parti qui se traduit par la dévouement au travail, mais se désole d'un manque d'ouverture d'esprit et d'inventivité.

★ ★

★

A.2 QUANTITATIVE EPISTEMOLOGY

A.2.1 Algorithmic systematic review

ALGORITHM DESCRIPTION Let A be an alphabet, A^* corresponding words and $T = \cup_{k \in \mathbb{N}} A^{*^k}$ texts of finite length on it. A reference is for the algorithm a record with text fields representing title, abstract and keywords. Set of references at iteration n will be denoted $\mathcal{C} \subset T^3$. We assume the existence of a set of keywords \mathcal{K}_n , initial keywords being \mathcal{K}_0 . An iteration goes as follows :

1. A raw intermediate corpus \mathcal{R}_n is obtained through a catalog request providing previous keywords \mathcal{K}_{n-1} .
2. Overall corpus is actualized by $\mathcal{C}_n = \mathcal{C}_{n-1} \cup \mathcal{R}_n$.
3. New keywords \mathcal{K}_n are extracted from corpus through Natural Language Processing treatment, given a parameter N_k fixing the number of keywords.

The algorithm stops when corpus size becomes stable or a user-defined maximal number of iterations has been reached. Fig. 7 shows the global workflow.

IMPLEMENTATION Because of the heterogeneity of operations required by the algorithm (references organisation, catalog requests, text processing), it was found a reasonable choice to implement it in Java. Source code is available on the Github repository of the project¹. Catalog request, consisting in retrieving a set of references from a set of keywords, is done using the Mendeley software API [Mendeley, 2015] as it allows an open access to a large database. Keyword extraction is done by Natural Language Processing (NLP) techniques, following the workflow given in [Chavalarias and Cointet, 2013], calling a Python script that uses [Bird, 2006].

CONVERGENCE AND SENSITIVITY ANALYSIS A formal proof of algorithm convergence is not possible as it will depend on the empirical unknown structure of request results and keywords extraction. We need thus to study empirically its behavior. Good convergence properties but various sensitivities to N_k were found as presented in Fig. 69. We also studied the internal lexical consistence of final corpuses as a function of keywords number. As expected, small number yields more consistent corpuses, but the variability when increasing stays reasonable.

Nous prenons l'hypothèse la plus faible pour le paramètre $N_k = 100$. En effet, plus N_k est grand, moins le domaine exploré sera restreint, ce qui augmente les chances de recouvrement de deux corpus

¹ at <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/QuantEpsitemo/AlgoSR>

TABLE 21: Composition of the initial corpus for the construction of the citation network.

Domaine	Titre	Référence
Sciences politiques	Les effets structurants du transport: mythe politique, mystification scientifique	[Offner, 1993]
Interdisciplinaire	Réseaux et territoires-significations croisées	[Offner and Pumain, 1996]
Géographie	Villes et réseaux de transport: des interactions dans la longue durée (France, Europe, Etats-Unis)	[Bretagnolle, 2009]
Transports	Land-use transport interaction: state of the art	[Wegener and Fürst, 2004]
Économie	The co-evolution of land use and road networks	[Levinson, Xie, and Zhu, 2007]
Économie	Modeling the growth of transportation networks: a comprehensive review	[Xie and Levinson, 2009b]
Physique	Co-evolution of density and topology in a simple model of city formation	[Barthelemy and Flammini, 2009]

provenant de requêtes initiales différentes. Dans ce cas, une faible distance finale entre corpus sera plus significative pour des valeurs de N_k grandes.

A.2.2 *Hypernetwork analysis*

INITIAL CORPUS Le tableau 21 donne la composition du corpus initial pour la construction du réseau de citation.

SENSITIVITY ANALYSIS L'analyse de sensibilité permettant de fixer les paramètres optimaux pour le réseau sémantique est montrée en Fig. 70.

SEMANTIC NETWORK Une visualisation du réseau sémantique est donnée en Fig. 71.



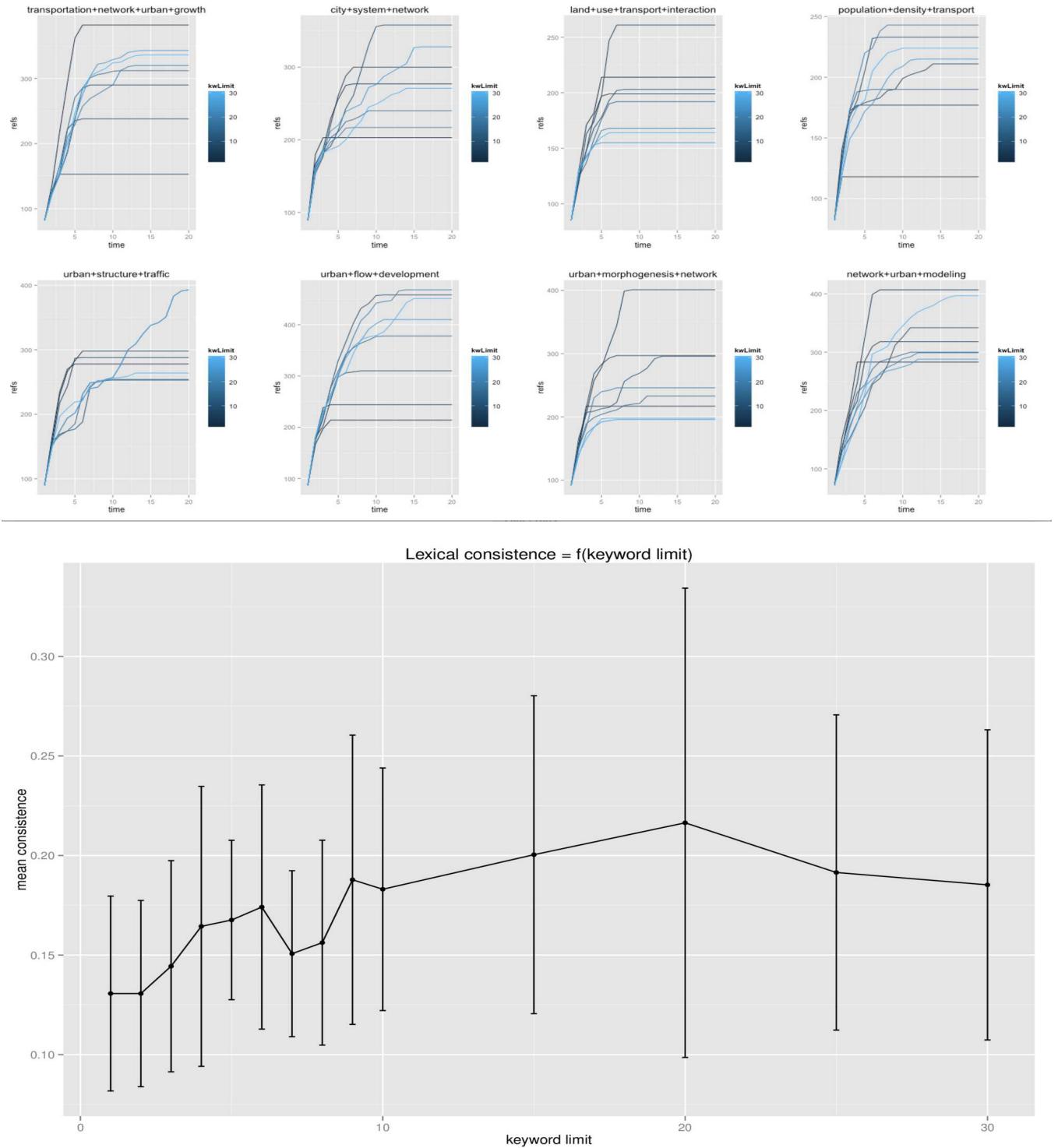


FIGURE 69: Convergence and sensitivity analysis. Left : Plots of number of references as a function of iteration, for various queries linked to our theme (see further), for various values of N_k (from 2 to 30). We obtain a rapid convergence for most cases, around 10 iterations needed. Final number of references appears to be very sensitive to keyword number depending on queries, what seems logical since encountered landscape should strongly vary depending on terms. Right : Mean lexical consistence and standard error bars for various queries, as a function of keyword number. Lexical consistence is defined though co-occurrences of keywords by, with N final number of keywords, f final step, and $c(i)$ co-occurrences in references, $k = \frac{2}{N(N-1)} \cdot \sum_{i,j \in \mathcal{K}_f} |c(i) - c(j)|$. The stability confirms the consistence of final corpuses.

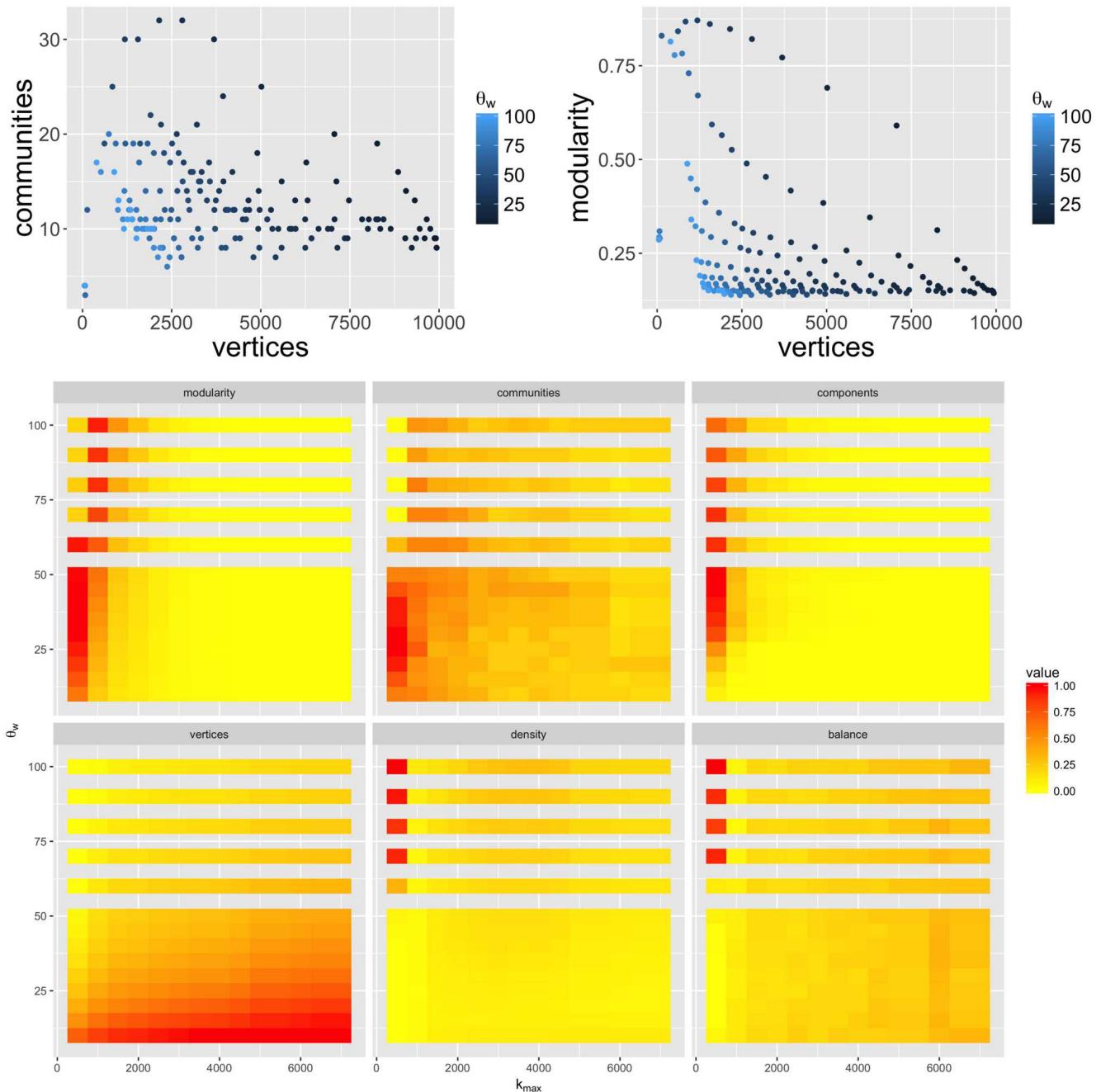


FIGURE 70: Sensitivity analysis of network indicators to filtering parameters.

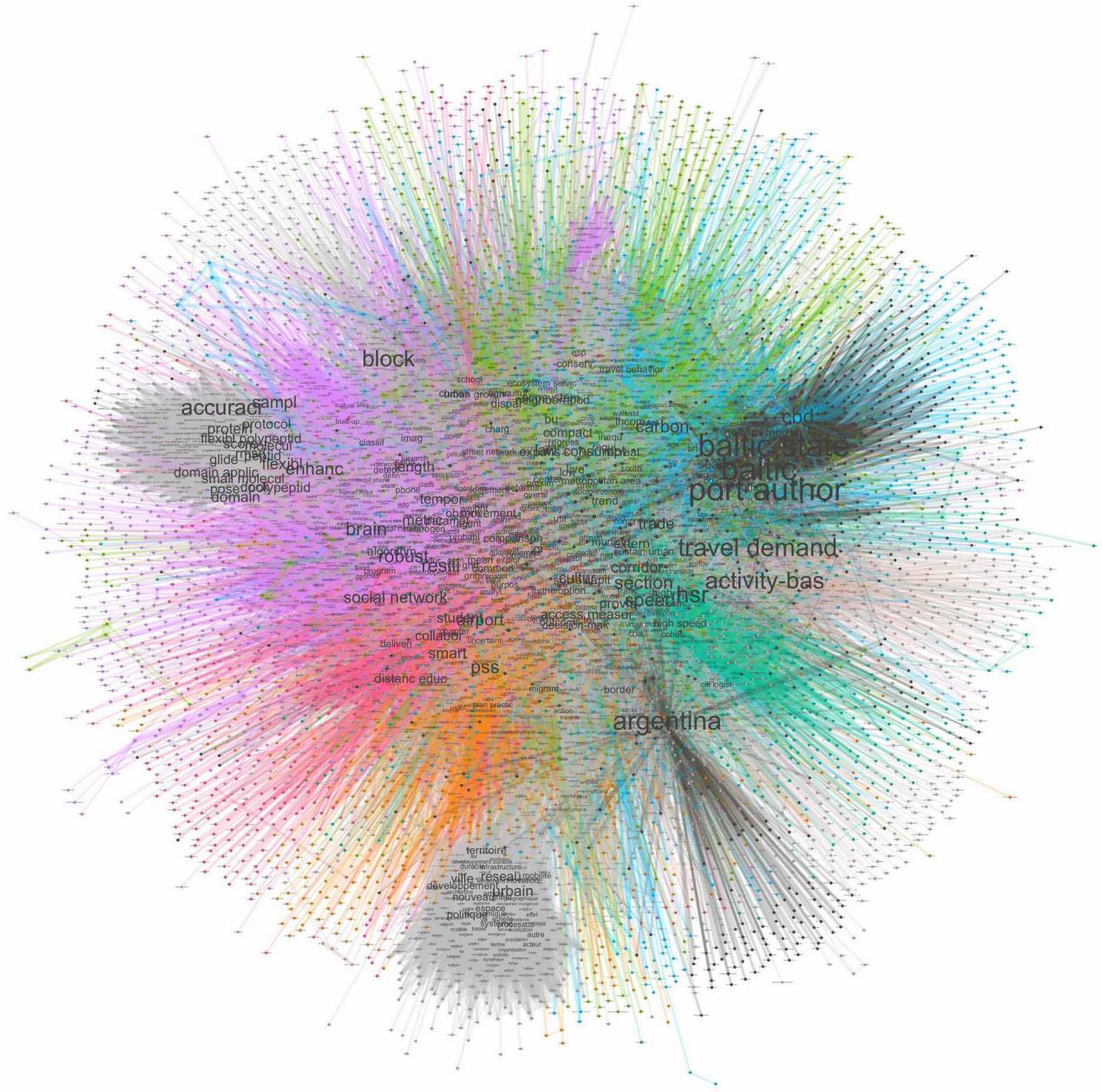


FIGURE 71: Semantic network of domains. Network is constructed by co-occurrences of most relevant keywords. Filtering parameters are here taken according to the multi-objective optimization done in Fig. ??, i.e. ($k_{\max} =$, $e_{\text{th}} =$, $f_{\min}, f_{\max} =$). The graph spatialization algorithm (Fruchterman-Reingold), despite its stochastic and path-dependent character, unveils information. A zoomable vectorial file (.svg) of the network is available as Supplementary Material.

A.3 MODELOGRAPHY

A.3.1 *Systematic Review Methodology*

Pour le choix des mots-clés initiaux pour la constitution indirecte (via requête sémantique), une alternative possible est d'extraire les mots-clés pertinents par sous-communautés du réseau de citations, puis sélectionner les plus pertinents ensuite pour chaque domaine. Nous faisons le choix de les extraire sur le corpus complet, puis de les récupérer par sous-communautés ensuite. Pour un petit corpus, la deuxième option est plus souhaitable, puisque la notion de pertinence moins importante que pour des très grands corpus, ou certains mots pertinents pourront être noyés et des moins pertinents ressortir de manière fortuite. En d'autre termes, la méthode de selection des mots-clés paraît plus robuste sur des petits corpus, comme le suggère la comparaison de cette application avec celle faite sur le journal Cybergeo et celle faite sur le corpus de brevets (voir C.5).

Article screening

Les méthodes utilisées ne permettent pas de s'affranchir d'un "bruit", c'est-à-dire d'article ne relevant a priori pas même de loin à la thématique. Nous avons obtenu par exemple des articles aussi divers qu'incongrus sur le genre et l'usage de la voiture, le cancer colorectal au Texas, la mécanique des vibrations au passage d'un train à grande vitesse, le transport des protéines dans la cellule, l'espace public à Beyrouth, les motifs spatiaux des *street gangs* à Los Angeles, la géologie urbaine à Bruxelles. Cela confirme que l'étape de filtrage manuel est essentielle.

Ce bruit peut être du par exemple à :

- Des citations effectives pour diverses raisons, mais n'ayant que peu de pertinence dans l'article citant.
- Du bruit intrinsèque à la recherche par mots-clés.
- Des erreurs de classification du catalogue.

Remarks on manual screening

Lors de la classification manuelle opérée lors de l'inspection des résumés, les points suivants ressortent :

- Les disciplines "a priori" sont jugées par le journal dans lequel l'article a été publié. En l'occurrence, nous opérons les choix particuliers suivants (pour d'autres journaux comme des journaux de physique il n'y a pas d'ambiguïté) : Journal of Transport Geography, Environment and Planning B : geography ; Journal of

Transport and Land-Use, Transportation Research : Transportation.

- La géographie en notre sens inclut l'urbanisme et les études urbaines si celles-ci ne sont pas trop proches de la planification (urbain durable par exemple).

A.3.2 *Meta-analysis*

Nous donnons ici les résultats numériques complets des analyses statistiques reliant caractéristiques de modèles et variables explicatives.

Variables values

Rappelons ici les variables utilisées dans la méta-analyse et leur modalités. Celles-ci sont :

- Type de modèle (TYPE) : strong, territory, network.
- Année de publication (YEAR), nombre entier.
- Communauté de citation (CITCOM), définies par le réseau de citations : Accessibility, Geography, Infra Planning, LUTI, Networks, TOD.
- Discipline a priori (DISCIPLINE) : biology, computer science, economics, engineering, environment, geography, physics, planning, transportation.
- Communauté sémantique (SEMCOM) : brt, complex networks, hedonic, hsr, infra planning, networks, tod.
- Méthodologie utilisée : ca (*Cellular Automaton*), eq (équations analytiques), map (cartographie), mas (*Multi-agent simulation*), ro (recherche opérationnelle), sem (*Structural Equation Modeling*), sim (simulation), stat (statistiques).
- Indice d'interdisciplinarité (INTERDISC) : réel dans [0, 1].
- Echelle temporelle (TEMPSCALE) : donnée en année, vaut 0 pour les analyses statiques.
- Echelle spatiale (SPATSCALE) : continent (10000), country (1000), region (100), metro (10). Ces modalités sont transformées numériquement en km par les valeurs données entre parenthèses (échelles stylisées).

Model selection

Concernant la sélection des modèles, celle-ci n'est pas opérée en critère unique, de par le faible nombre d'observations pour certains modèles, mais par l'optimisation au sens de Pareto des objectifs contradictoires de l'ajustement (R^2 ajusté, à maximiser) et du sur-ajustement (critère d'Akaike corrigé AICc, à minimiser), tout en contrôlant le nombre de points d'observation. La Fig. 72 donne pour chaque variable à expliquer la localisation de l'ensemble des modèles potentiels dans l'espace des objectifs, ainsi que le nombre d'observations correspondantes. Pour l'interdisciplinarité, deux nuages de points correspondent à des compromis différents, et nous sélectionnons les deux modèles optimaux (un pour chaque nuage). Pour l'échelle d'espace, nous postulons un R^2 positif, et un seul modèle optimal émerge alors. Pour l'échelle de temps, on a comme pour l'interdisciplinarité deux modèles compromis. Enfin, pour l'année, le gain en AICc entre les deux optimaux potentiels est négligeable en comparaison à la perte en R^2 , et nous sélectionnons donc le modèle optimal tel que $R^2 > 0.25$ et $AICc < 600$. Les résultats des modèles sont donnés par la suite.

Model fitting

INTERDISCIPLINARITY L'interdisciplinarité est ajustée selon les modèles linéaires présentés en Table 22.

TABLE 22:

Spatial scale L'échelle spatiale est ajustée selon le modèle linéaire dont l'ajustement est donné en Table 23.

TABLE 23:

Time scale L'échelle de temps est ajustée selon les modèles linéaires présentés en Table 24.

TABLE 24:

Year L'année de publication est ajustée selon le modèle linéaire dont l'ajustement est donné en Table 25.

TABLE 25:

★ ★

★

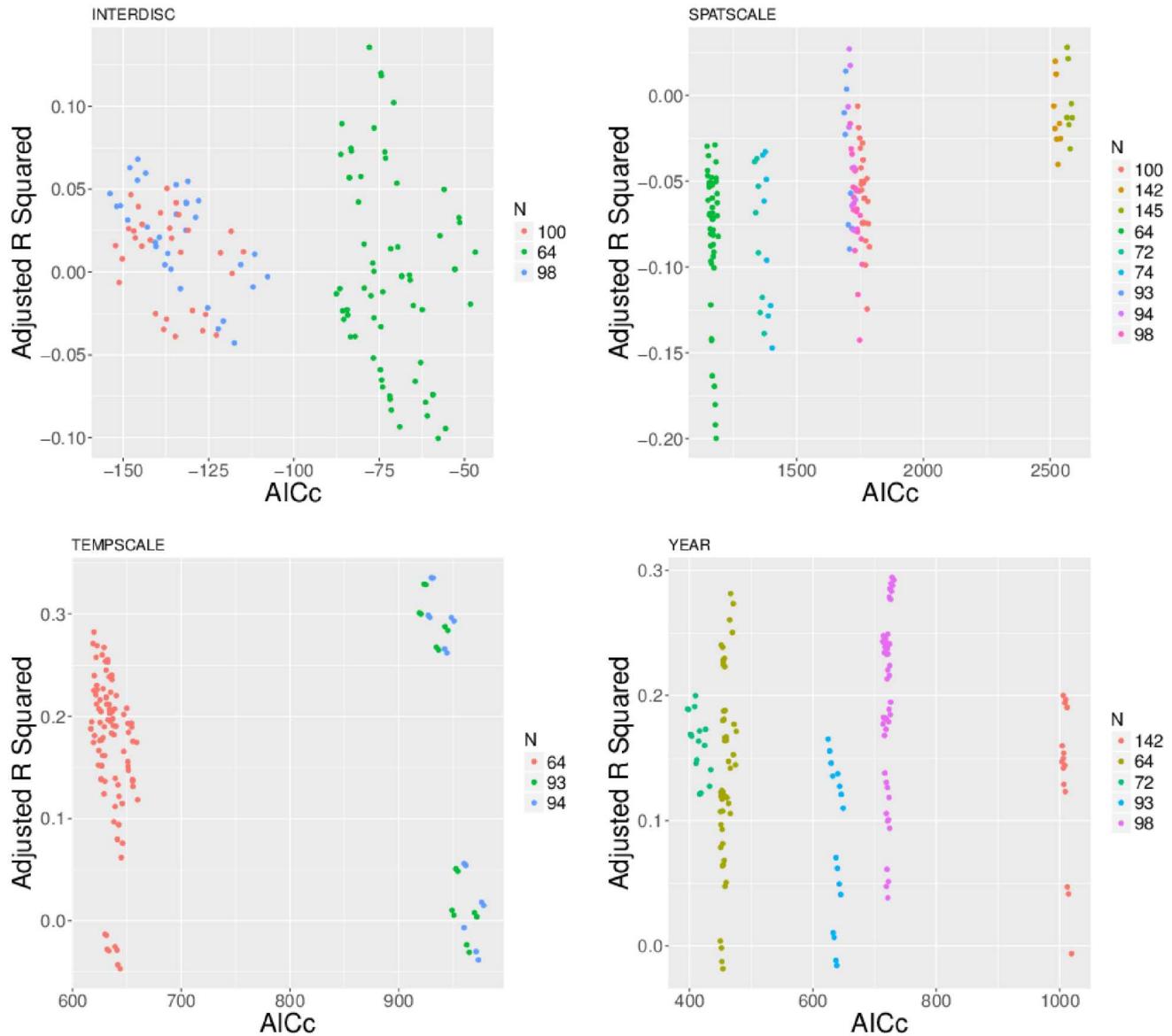


FIGURE 72: Multi-objective selection of linear models.

INTERDISC

	(1)	(2)
YEAR	-0.004 (-0.008, -0.00002), p = 0.055*	-0.002 (-0.005, 0.0001), p = 0.061*
TEMPSCALE	-0.0003 (-0.001, 0.001), p = 0.615	
DISCIPLINEengineering	0.144 (-0.082, 0.371), p = 0.218	
DISCIPLINEenvironment	0.092 (-0.132, 0.316), p = 0.425	
DISCIPLINEgeography	0.036 (-0.043, 0.114), p = 0.378	
DISCIPLINEphysics	-0.103 (-0.287, 0.080), p = 0.275	
DISCIPLINEplanning	-0.047 (-0.135, 0.041), p = 0.30	
DISCIPLINEtransportation	0.062 (-0.025, 0.149), p = 0.169	
TYPEstrong		-0.026 (-0.134, 0.081), p = 0.633
TYPERegion		0.044 (-0.026, 0.114), p = 0.222
SEMCOMcomplex networks		-0.217 (-0.522, 0.087), p = 0.166
SEMCOMhedonic	-0.179 (-0.407, 0.049), p = 0.130	-0.184 (-0.400, 0.032), p = 0.100*
SEMCOMhsr	-0.100 (-0.361, 0.162), p = 0.459	-0.122 (-0.357, 0.112), p = 0.309
SEMCOMinfra planning	-0.032 (-0.273, 0.209), p = 0.797	-0.096 (-0.321, 0.128), p = 0.404
SEMCOMnetworks	-0.038 (-0.272, 0.195), p = 0.750	-0.107 (-0.324, 0.109), p = 0.335
SEMCOMtod	-0.105 (-0.332, 0.121), p = 0.366	-0.152 (-0.364, 0.060), p = 0.165
Constant	8.962 (0.776, 17.147), p = 0.037**	5.531 (0.575, 10.487), p = 0.032**
Observations	64	98
R ²	0.314	0.155
Adjusted R ²	0.136	0.068
Residual Std. Error	0.109 (df = 50)	0.107 (df = 88)
F Statistic	1.761* (df = 13; 50)	1.789* (df = 9; 88)

Note:

*p<0.1; **p<0.05; ***p<0.01

SPATSCALE	
TEMPSCALE	-5.179 (-16.259, 5.901)
	p = 0.363
DISCIPLINEengineering	-154.461 (-3,003.326, 2,694.405)
	p = 0.916
DISCIPLINEenvironment	-5.878 (-3,977.974, 3,966.219)
	p = 0.998
DISCIPLINEgeography	1,445.457 (389.349, 2,501.565)
	p = 0.009***
DISCIPLINEphysics	292.559 (-2,717.659, 3,302.777)
	p = 0.850
DISCIPLINEplanning	-143.554 (-1,361.357, 1,074.249)
	p = 0.818
DISCIPLINEtransportation	568.329 (-606.167, 1,742.826)
	p = 0.346
Constant	235.357 (-458.201, 928.914)
	p = 0.508
<hr/>	
Observations	94
R ²	0.100
R ² ajusté	0.027
Erreur Std. Résiduelle	1,995.272 (df = 86)
Statistique F	1.369 (df = 7; 86)

Note:

*p<0.1; **p<0.05; ***p<0.01

TEMPSCALE

	(1)	(2)
YEAR	0.674 (-0.294, 1.643) p = 0.179	
TYPEstrong		100.271 (58.312, 142.230) p = 0.00002***
TYPEterritory	-38.933 (-64.249, -13.617) p = 0.004***	-14.988 (-37.411, 7.435) p = 0.194
DISCIPLINEengineering	-52.107 (-110.950, 6.735) p = 0.089*	-9.609 (-55.841, 36.624) p = 0.685
DISCIPLINEenvironment	17.110 (-37.350, 71.569) p = 0.541	17.886 (-45.319, 81.090) p = 0.581
DISCIPLINEgeography	3.640 (-15.364, 22.644) p = 0.709	9.126 (-7.590, 25.843) p = 0.288
DISCIPLINEphysics	46.879 (0.638, 93.120) p = 0.053*	77.897 (28.225, 127.570) p = 0.003***
DISCIPLINEplanning	1.304 (-19.336, 21.945) p = 0.902	4.553 (-14.865, 23.971) p = 0.648
DISCIPLINEtransportation	-14.718 (-34.978, 5.543) p = 0.161	8.753 (-9.864, 27.371) p = 0.360
INTERDISC	2.357 (-59.200, 63.915) p = 0.941	
Constant	-1,305.126 (-3,252.499, 642.247) p = 0.195	22.103 (-0.951, 45.156) p = 0.064*
Observations	64	94
R ²	0.385	0.393
Adjusted R ²	0.282	0.336
Residual Std. Error	26.984 (df = 54)	31.747 (df = 85)
F Statistic	3.755*** (df = 9; 54)	6.871*** (df = 8; 85)

Note:

*p<0.1; **p<0.05; ***p<0.01

	YEAR
TYPEterritory	10.898 (3.045, 18.750), p = 0.010***
TEMPSCALE	0.035 (-0.033, 0.103), p = 0.320
FMETHODeq	-6.224 (-20.162, 7.714), p = 0.387
FMETHODmap	4.747 (-7.595, 17.089), p = 0.456
FMETHODro	6.128 (-11.694, 23.950), p = 0.504
FMETHODsem	1.009 (-16.659, 18.676), p = 0.912
FMETHODsim	5.153 (-6.809, 17.114), p = 0.404
FMETHODstat	-0.357 (-10.925, 10.211), p = 0.948
DISCIPLINEengineering	13.486 (-7.238, 34.210), p = 0.210
DISCIPLINEenvironment	-3.668 (-21.605, 14.269), p = 0.691
DISCIPLINEgeography	1.121 (-4.528, 6.769), p = 0.700
DISCIPLINEphysics	3.392 (-8.461, 15.245), p = 0.578
DISCIPLINEplanning	-2.850 (-8.873, 3.173), p = 0.359
DISCIPLINEtransportation	5.503 (0.006, 11.000), p = 0.057*
INTERDISC	-12.876 (-29.567, 3.815), p = 0.138
SECOMhedonic	-5.769 (-19.931, 8.393), p = 0.430
SECOMhsr	6.135 (-9.889, 22.159), p = 0.458
SECOMinfra planning	-4.123 (-18.910, 10.663), p = 0.588
SECOMnetworks	4.711 (-9.736, 19.158), p = 0.527
SECOMtod	-1.653 (-15.837, 12.532), p = 0.821
Constant	2,004.945 (1,981.531, 2,028.359), p = 0.000***
Observations	64
R ²	0.510
Adjusted R ²	0.281
Residual Std. Error	6.617 (df = 43)
F Statistic	2.234** (df = 20; 43)

Note:

*p<0.1; **p<0.05; ***p<0.01

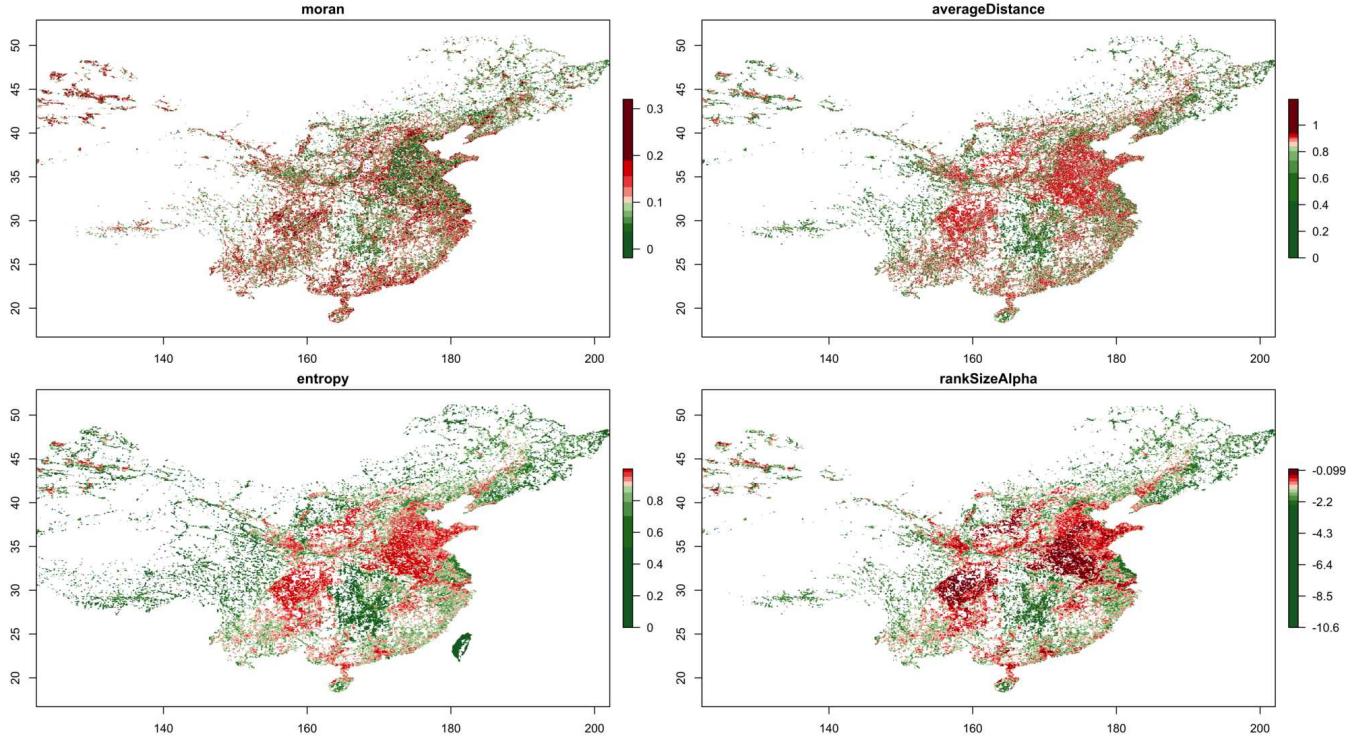


FIGURE 73: Morphological indicators for China. We give for areas where a population and the network are simultaneously defined, the Moran index I (`moran`), the average distance \bar{d} (`averageDistance`), the entropy \mathcal{E} (`entropy`) and the hierarchy γ (`rankSizeAlpha`).

A.4 STATIC CORRELATIONS

A.4.1 Morphological Measures

We compute for China, from the population grid with a 1km resolution [Fu, Jiang, and Huang, 2014], the morphological indicators. We consider areas of width 100km, in order to have a reasonable number of points for the estimation, with an offset of 50km. Corresponding maps are given in Fig. 73. The distribution of some indicators such as the entropy \mathcal{E} seems to be conditioned to province boundaries as in Sichuan, the uniformity of the dataset must possibly be questioned.

A.4.2 Road network

Network Simplification Algorithm

We detail here the road network simplification algorithm from OpenStreetMap data. The general workflow is the following: (i) data import by selection and spatial aggregation at the raster resolution; (ii) simplification to keep only the topological network, processed in parallel through *split/merge*.

OSM data are imported into a pgsql database (Postgis extension for the management of geometries and to have spatial indexes). The import is done using the software osmosis [Osmosis, 2016], from an image in compressed pbf format of the OpenStreetMap database². We filter at this stage the links (ways) which posses the tag highway, and keep the corresponding nodes.

The network is first aggregated at a 100m granularity in order to be consistently used with population grids. It furthermore allows to be robust to local coding imperfections or to very local missing data. For this step, roads are filtered on a relevant subset of tags³. For the set of segments of corresponding lines, a link is created between the origin and the destination cell, with a real length computed between the center of cells and a speed taken as the speed of the line if it is available.

The simplification is then operated the following way:

1. The whole geographical coverage is cut into areas on which computations will be partly done through parallel computation (*split* paradigm). Areas have a fixed size in number of cells of the base raster (200 cells).
2. On each sub-area, a simplification algorithm is applied the following way: as long as there still are vertices of degree 2, successive sequences of such vertices are determined, and corresponding links are replaced by a unique link with real length and speed computed by cumulation on the deleted links.
3. As the simplification algorithm keeps the links having an intersection with the border of areas, a fusion followed by a simplification of resulting graphs is necessary. To keep a reasonable computational cost, the size of merged areas has to stay low: we take merge areas composed by two contiguous areas. A paving by four sequences of independent merging allows then to cover the full set of joints between areas⁴, these sequences being executed sequentially. The Frame 14 shows the covering of joints by merging areas.

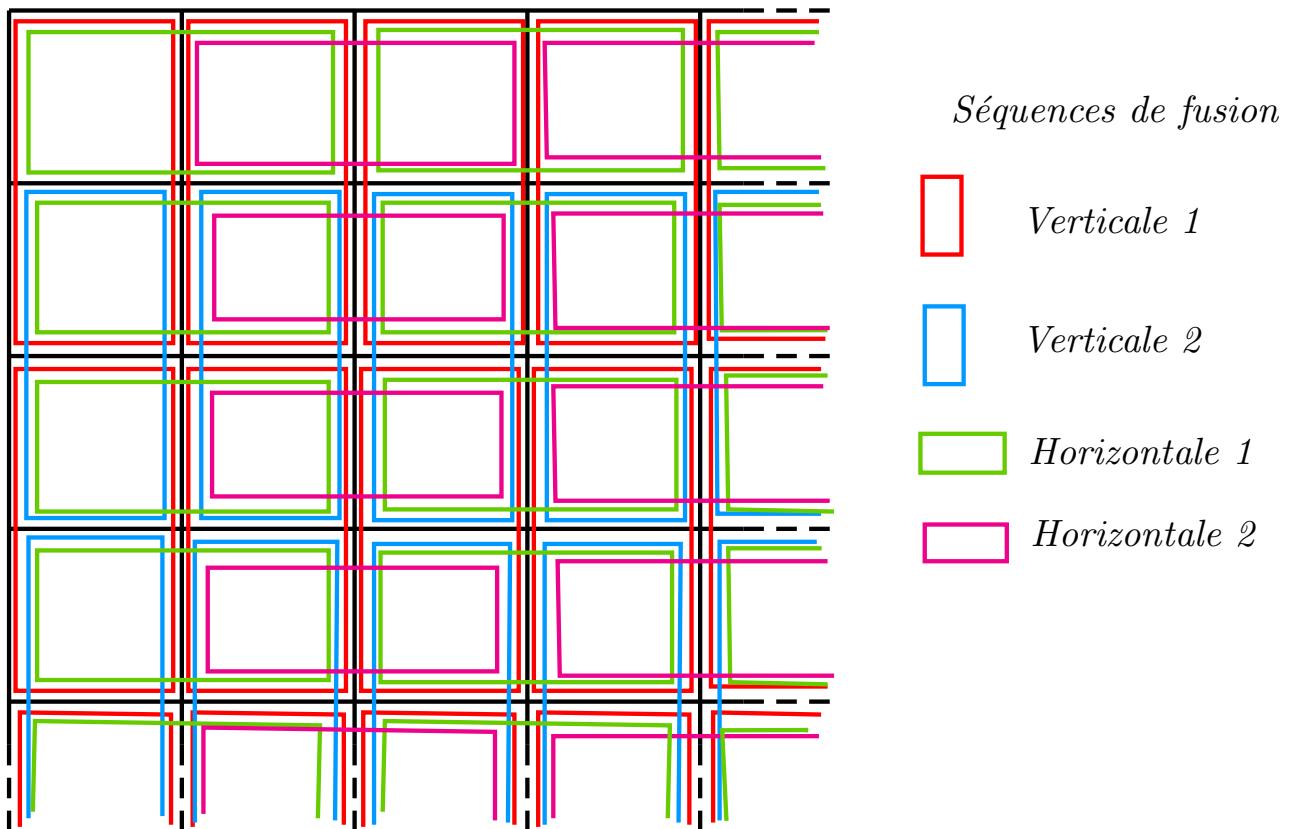
We have then at our disposition a topological graph given by the links between cells of the base raster, having distance and speed attributes corresponding to the underlying real links.

Graphs for Europe and China are available as open databases (see Appendix D).

² Dumps were retrieved from <http://download.geofabrik.de>, in July 2016 for Europe, and July 2017 for China.

³ That we take within `motorway`, `trunk`, `primary`, `secondary`, `tertiary`, `unclassified`, `residential`.

⁴ In the very rare cases of a link between two non-contiguous areas, the remaining link is not simplified. This case was not observed in practice in our data.



FRAME 14: **Illustration of merging sequences.** The four independent sequences (horizontally and vertically) allow the coverage of all joints between areas.

Network Indicators

We give in Fig. 74 a sample of network indicators for China.

A.4.3 *Sensitivity to resolution*

We evaluate here the sensitivity of indicators to grid size. We show in Fig. 75 morphological indicators and in Fig. 76 some network indicators, mapped for France, for different grid sizes. The sizes taken here, in correspondance to the 50km scale used in main results, are at similar magnitudes: we test windows of size 30km and 100km. The offsets are in each case half of the window (15km and 50km respectively). It is possible to see with eyeball validation that some indicators have a low sensitivity, the change in scale resembling a smoothing of the finer field: for example for morphology in the case of Moran, entropy and hierarchy. Average distance, which is indeed rather noisy at the smaller scale, is necessarily sensitive to aggregation, what is consistent with a sensitivity expected at smoothing. Network indicators are relatively robust to window size.

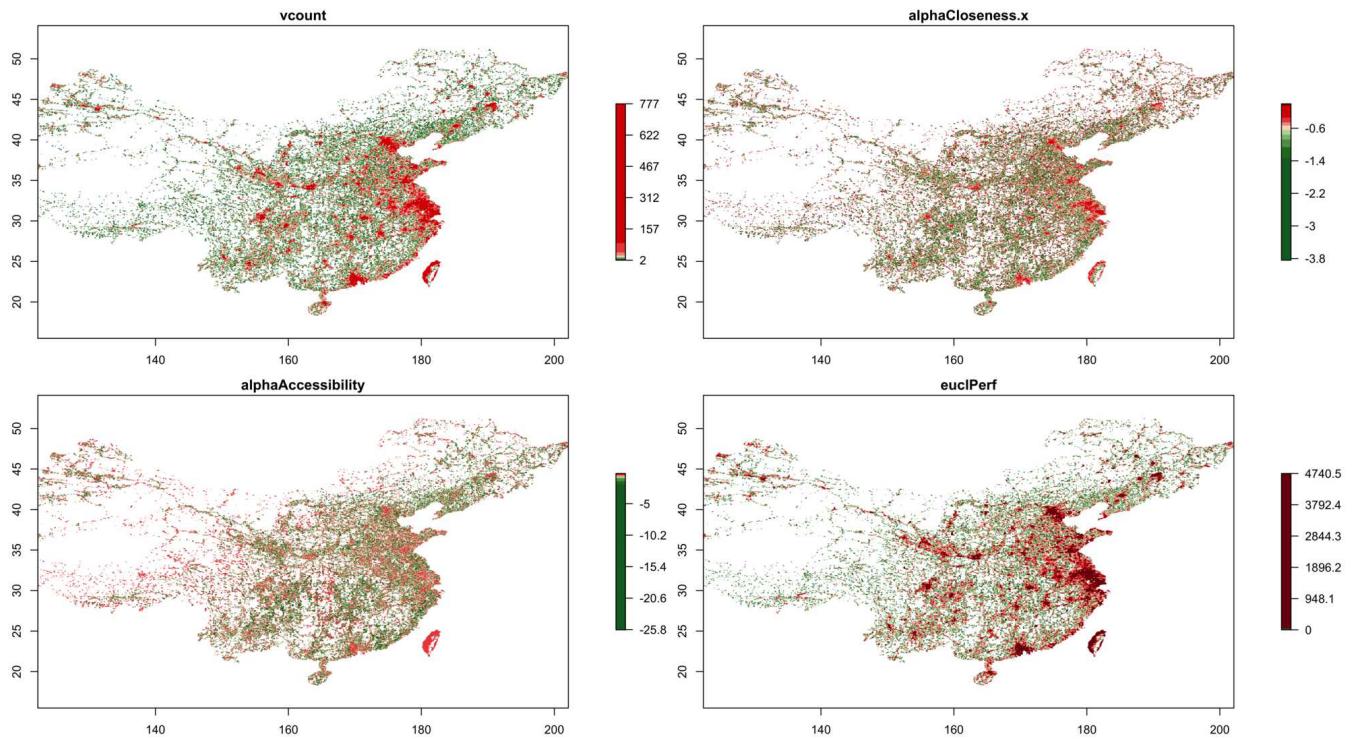


FIGURE 74: Network indicators for China. We show a selection of network indicators: number of nodes $|V|$ (vcount), closeness hierarchy α_{cl} (alphaCloseness.x), accessibility hierarchy α_Z (alphaAccessibility), euclidian performance v_0 (euclPerf).

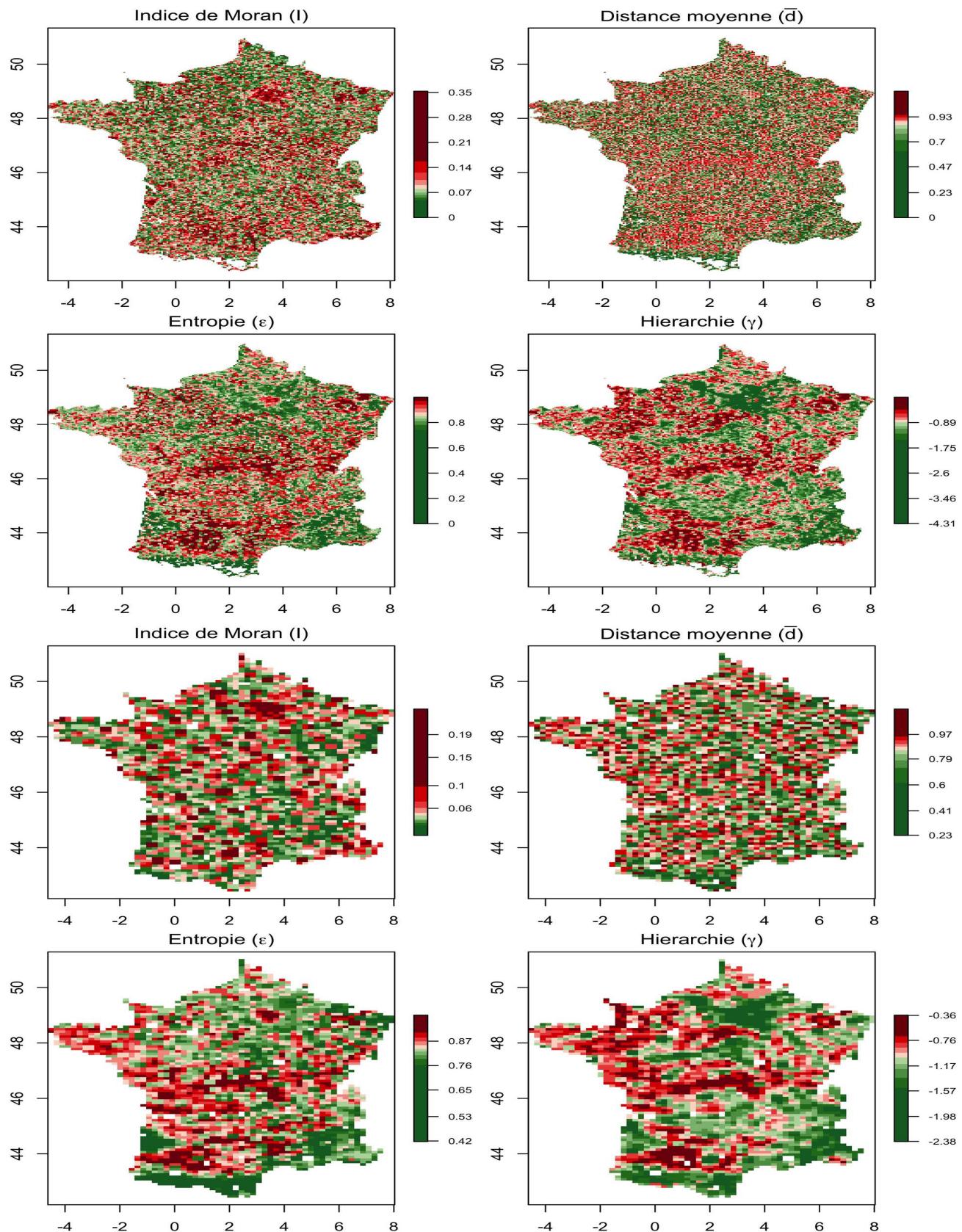


FIGURE 75: **Morphological indicators for different grid sizes.** The first four maps show the indicators computed on a window of size 30km, the last four maps with a window of size 100km.

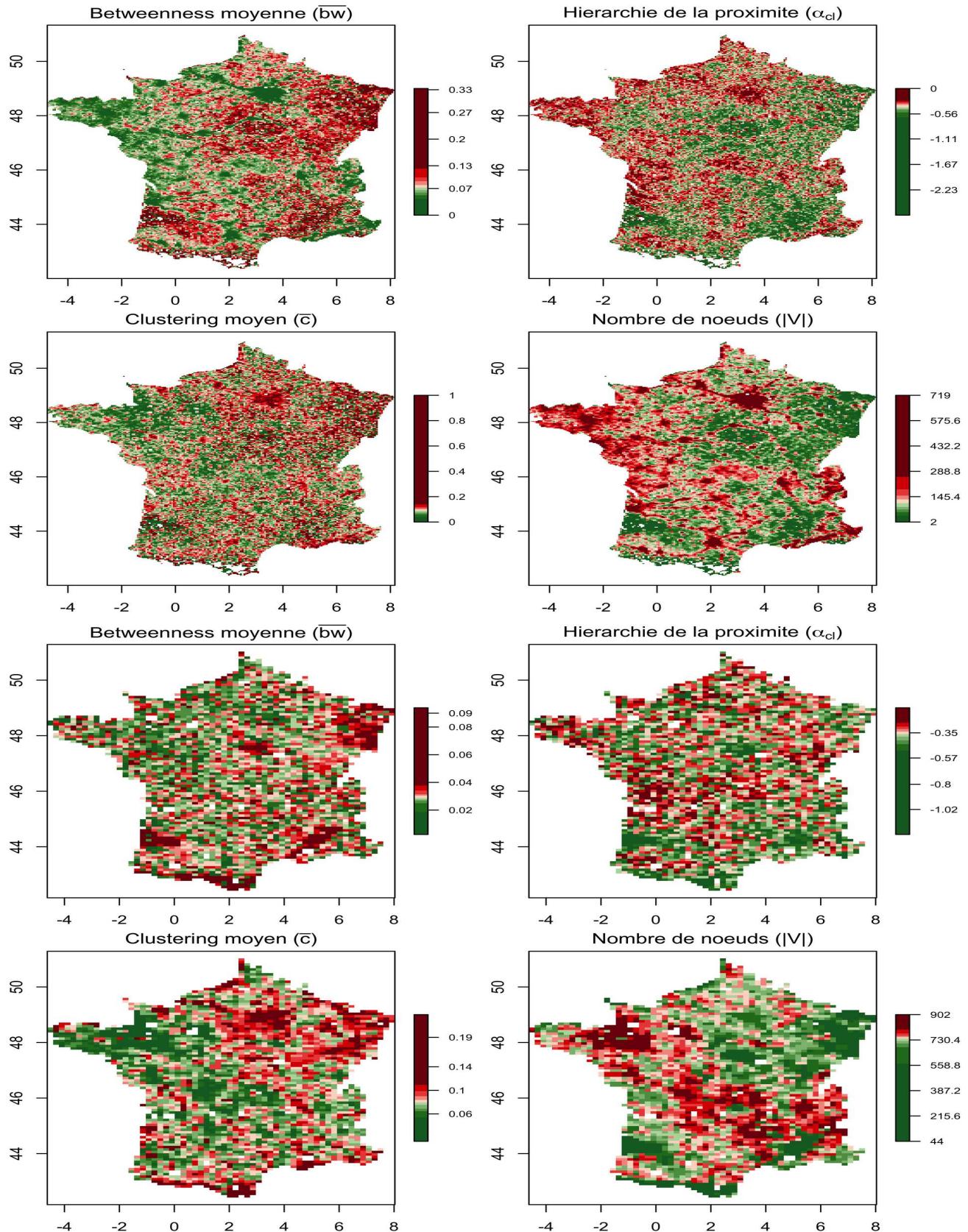


FIGURE 76: **Sample of network indicators for different grid sizes.** The first four maps give the indicators computed with a window of size 30km, the last four maps with a window of size 100km.

This comparison, on the one hand is to be taken cautiously because of the difficulty to directly compare scales for indicators, and on the other hand stays limited. We propose then a method to quantify the variability of indicators to window size. Let X_D and X_d two spatial fields corresponding to two spatial scales $D > d$ (that we take as characteristic distances). The fields are assumed discrete at points respectively denoted by $(\vec{x}_i^{(D)})_{1 \leq i \leq N_D}$ and $(\vec{x}_j^{(d)})_{1 \leq j \leq N_d}$. The idea is to compare a smoothing of the finer field to the field with the larger scale: if the correlation between these two values is high, it is possible to deduce one field from the other by aggregation and the scale of computation does not influence final results in an other way than the final resolution. Let $W_{ij} = (\exp - d_{ij}/d_0)_{ij}$ a matrix of spatial weights computed with euclidian distances d_{ij} between the points $\vec{x}_i^{(D)}$ and $\vec{x}_j^{(d)}$. Then with $W'_{ij} = W_{ij} / \sum_j W_{ij}$, we can compute the spatial smoothing of X_d at the points $\vec{x}_i^{(D)}$, with the matrix product

$$\tilde{X}_d(\vec{x}_i^{(D)}) = W' * \vec{x}_j^{(d)}$$

The correlation is then given by $\rho [\tilde{X}_d, X_D]$ estimated on all $\vec{x}_i^{(D)}$ points.

The Fig. 77 gives the variation of this correlation for all (D, d) couples, with a variable d_0 for smoothing. We generally observe the existence of a maximum, which corresponds to the optimal smoothing level to deduce the larger scale from the finer. The largest correlations on all indicators are obtained for $D = 50\text{km}$ and $d = 30\text{km}$, what means that indicators are not very sensitive to small variations in small window sizes. As expected, the lowest correlations are obtained for the largest scale difference ($100/30\text{km}$). Morphological indicators have the same qualitative behavior across combinations, and we find the behavior suggested by the previous maps (entropy and hierarchy being the less sensitive, Moran index and average distance a bit more sensitive). For the network, some indicators such as α_{bw} show a significant transition depending on $D - d$: there exists for this indicator a large sensitivity in small sizes. For all indicators, the sensitivity remains however reasonable. Finally, a smoothing of both fields yields asymptotic maximal correlations with very high values: the computation window size does not matter if we consider smoothed fields.

A.4.4 Spatial Correlations

The Fig. 78 gives the correlation matrix estimated for $\delta = \infty$. To have an idea of the robustness of the estimation, we investigate the relative size of confidence intervals at the 95% level (Fisher method) given by $\frac{|\rho_+ - \rho_-|}{|\rho|}$, for correlations such that $|\rho| > 0.05$. The median of this rate is at 0.04, the ninth decile at 0.12 and the maximum at 0.19, what

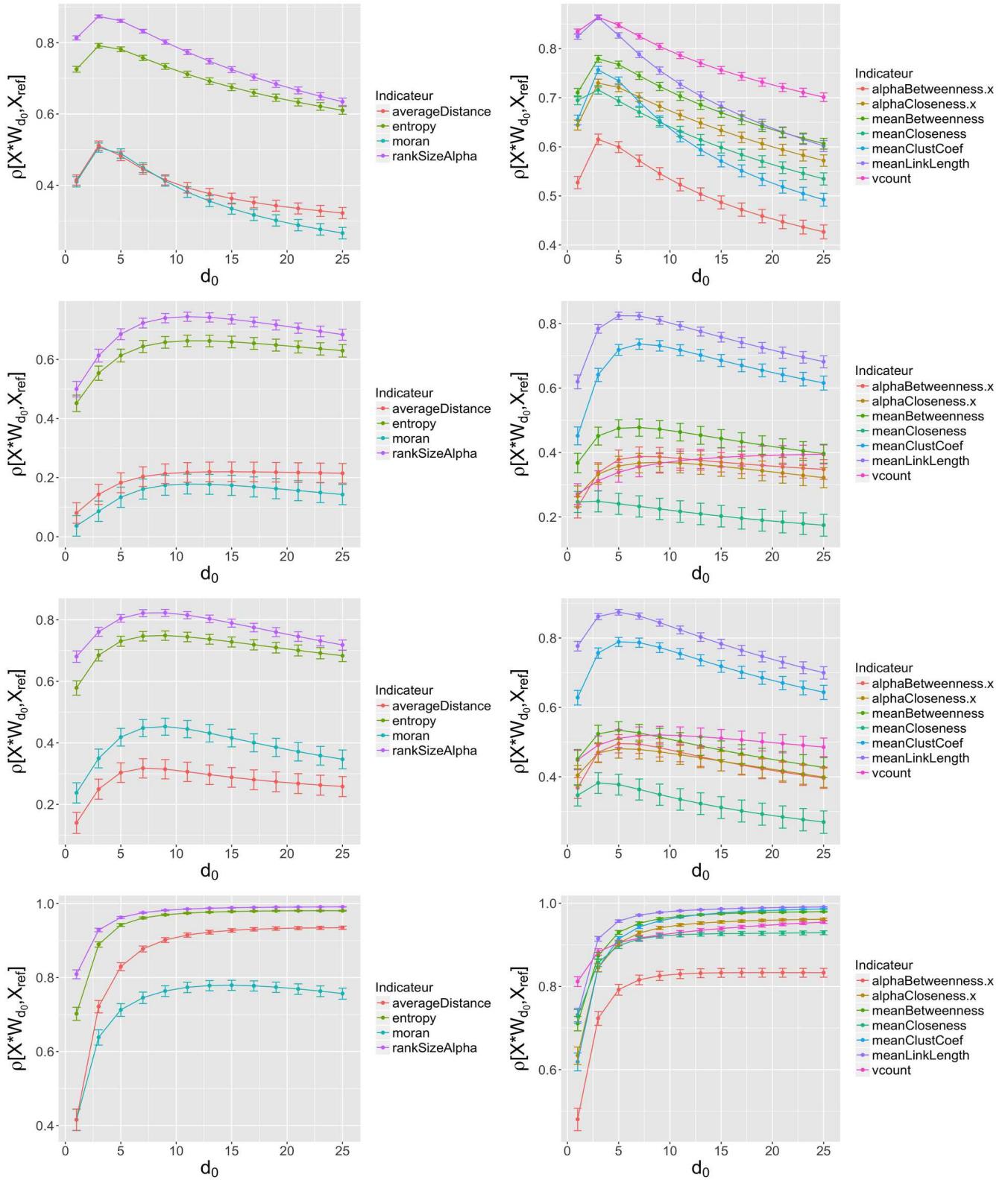


FIGURE 77: Correlations between indicators computed at different scales. From top to bottom (left column giving morphological indicators and right column network indicators), ($d = 30, D = 50$), ($d = 30, D = 100$), ($d = 50, D = 100$), and the last row gives the correlation between the two fields $d_1 = 30$ and $d_2 = 50$ both smoothed at the characteristic distance of d_0 .

means that the estimation is always relatively good compared to the value of correlations.

The Fig. 79 gives the spatial distribution for all Europe, of a sample of correlations between indicators: $\rho[\alpha_{cl}, l]$, $\rho[\gamma, \alpha]$, $\rho[bw, \gamma]$, $\rho[\alpha_{bw}, \alpha_{cl}]$, $\rho[|V|, l]$, $\rho[\gamma, r_\gamma]$ (with r_γ adjustment coefficient for γ). We see interesting structures emerging, such as the hierarchy and its adjustment which present an area of strong correlation in the center of Europe and negative correlation areas, or the number of nodes and the path length which correlate in mountains and along the coasts (what is expected since roads then do several detours) and have a negative correlation otherwise.

The Fig. 80 gives the statistical distributions of estimated correlations on all the areas, for different values of δ . We distinguish there the different blocks in the correlation matrix, i.e. correlations between morphological indicators, the ones between network indicators, and also cross-correlations. The latest have rather symmetrical distributions, whereas network and morphological correlations are dissymmetrical. We also give point clouds allowing to make a link between these different components.

A.4.5 Multiscalarity

Estimation of correlations for a multi-scalar process

We propose here to link the multi-scalar character of a spatio-temporal stochastic process with the estimation of its correlation matrix. To simplify and in the framework in which this result is used in main text, we consider static correlations estimated in space. To also simplify, we consider processes with two characteristic scales which linearly superpose, i.e. which can be written as

$$X_i = X_i^{(0)} + \tilde{X}_i$$

with $X_i^{(0)}$ trend at the small scales with a characteristic evolution distance d_0 , and \tilde{X}_i signal evolving at a characteristic distance $d \ll d_0$.

We can then compute the decomposition of the correlation between two processes, in a manner similar to what is done in C.3. Assuming

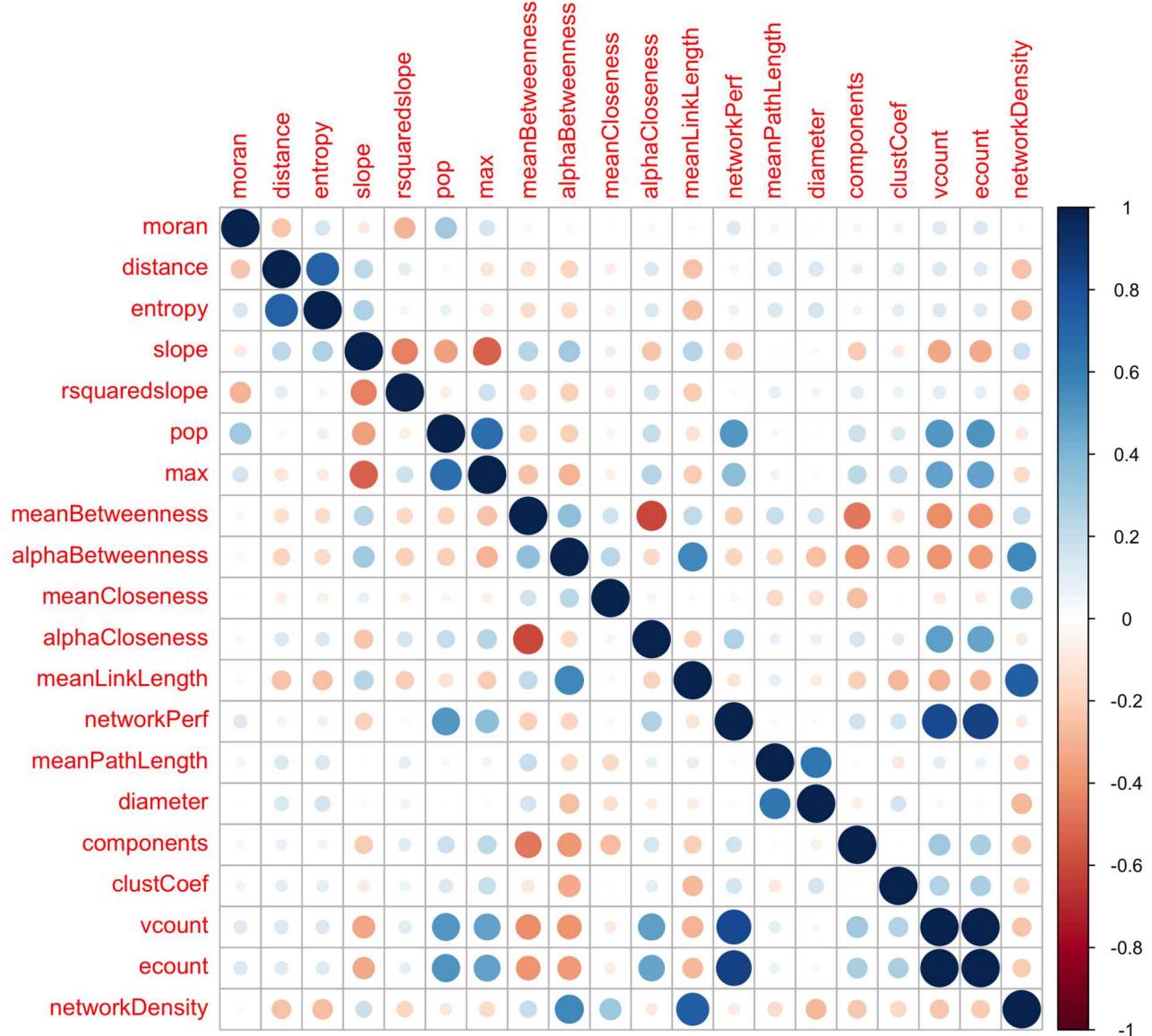


FIGURE 78: **Correlation matrix.** The matrix is estimated here on all indicator values for Europe, what is equivalent to take $\delta = \infty$.

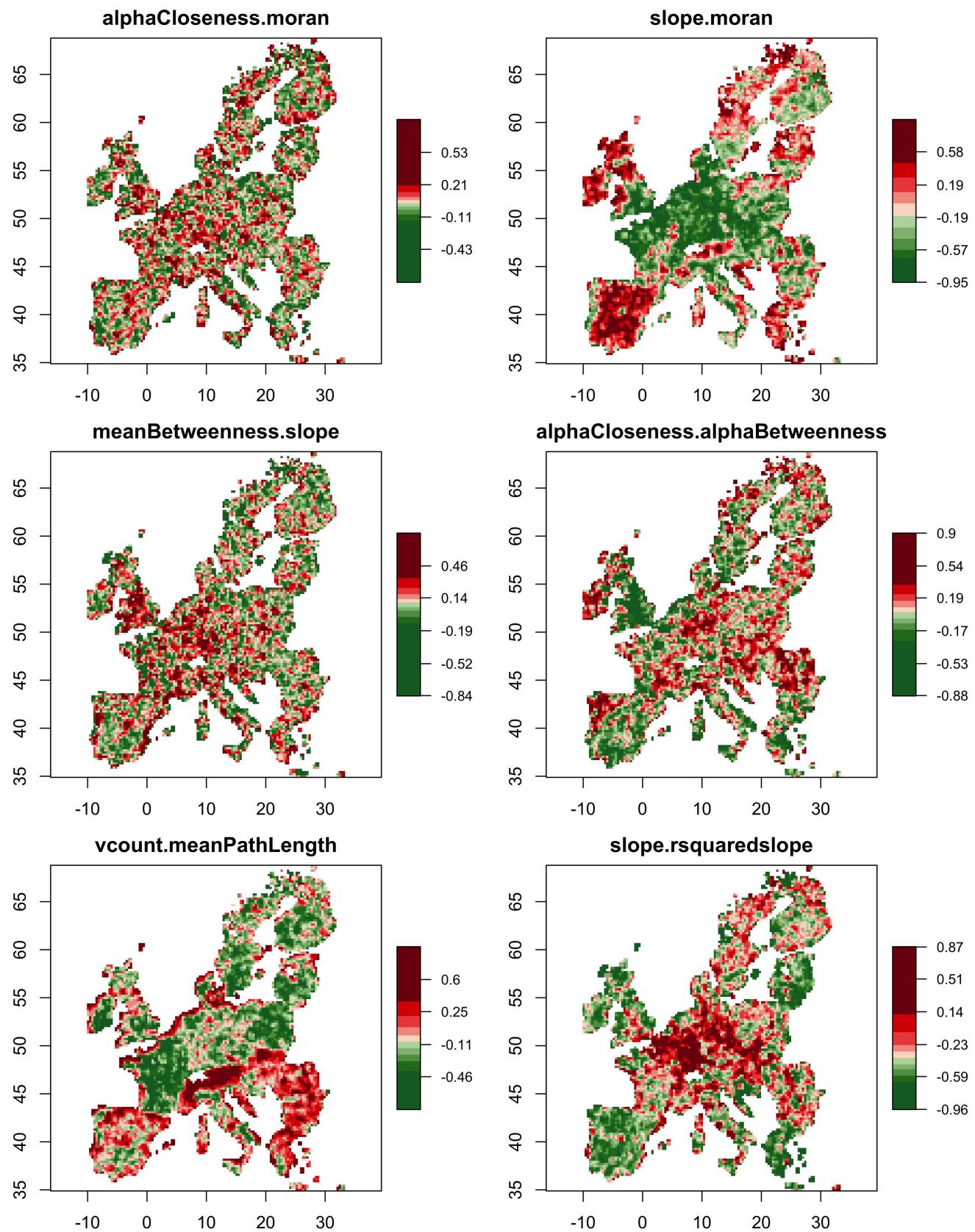


FIGURE 79: Spatial correlations for Europe. The estimation is done here with $\delta = 12$.

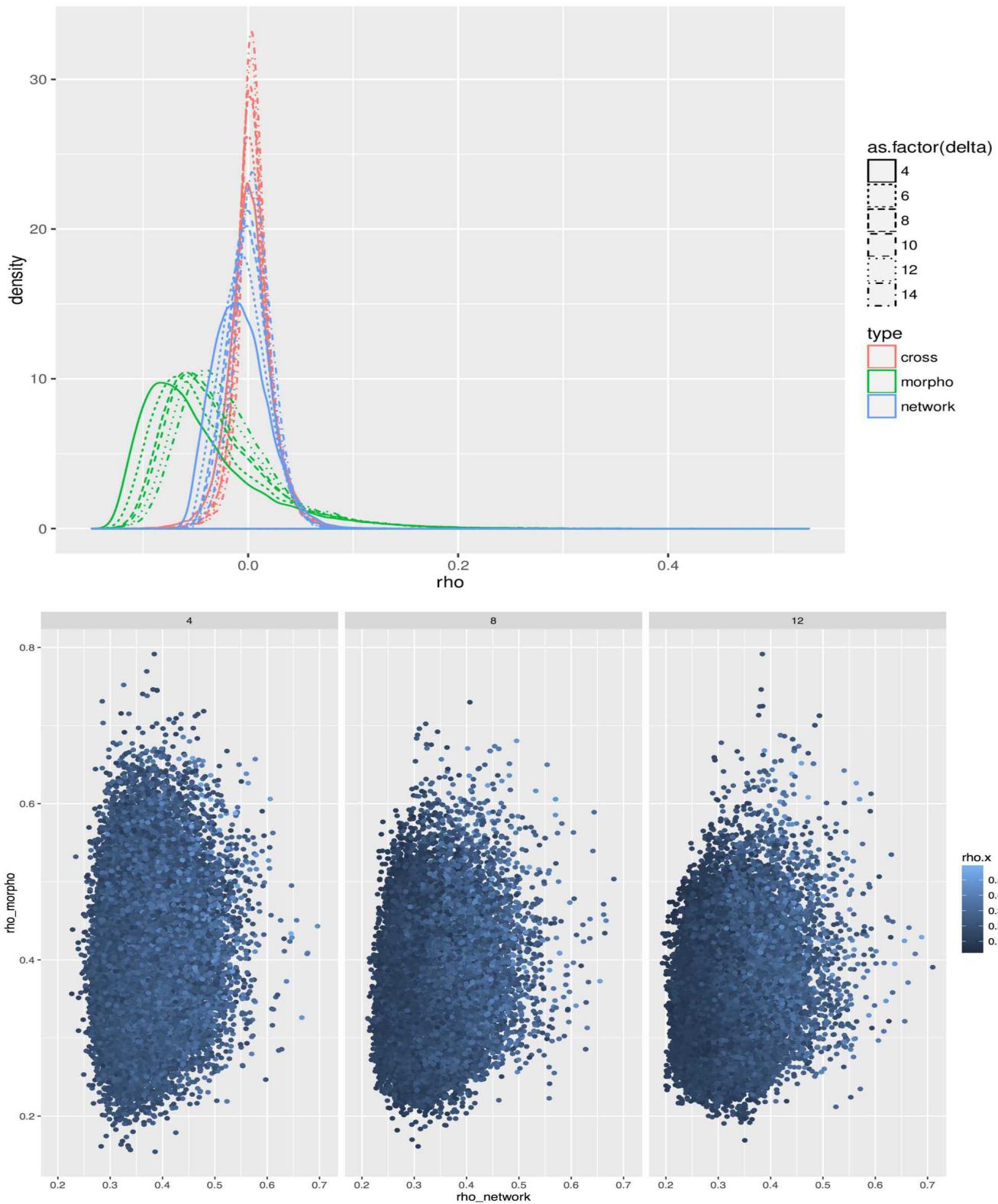


FIGURE 8o: Distribution of correlations. (Top) Statistical distribution of correlations, for the different morphological, network and cross-correlations blocks (color), for different values of δ (line type); (Bas) Average absolute correlations for the network as a function of correlations for morphology, the color level giving the cross-correlation, for different values of δ (columns).

that $\text{Cov}[X_i^{(0)}, \tilde{X}_j] = 0$ for all i, j , and denoting $\varepsilon_i = \frac{\sigma[X_i^{(0)}]}{\sigma[\tilde{X}_i]}$ the rate between standard deviations of trend and signal, we have

$$\begin{aligned}\rho[X_1, X_2] &= \rho[X_1^{(0)} + \tilde{X}_1, X_2^{(0)} + \tilde{X}_2] \\ &= \frac{\text{Cov}[\tilde{X}_1, \tilde{X}_2] + \text{Cov}[X_1^{(0)}, X_2^{(0)}]}{\sqrt{(\text{Var}[X_1^{(0)}] + \text{Var}[\tilde{X}_1]) (\text{Var}[X_2^{(0)}] + \text{Var}[\tilde{X}_2])}} \\ &= \frac{\varepsilon_1 \varepsilon_2 \rho[X_1^{(0)}, X_2^{(0)}] + \rho[\tilde{X}_1, \tilde{X}_2]}{\sqrt{(1 + \varepsilon_1^2)(1 + \varepsilon_2^2)}}\end{aligned}$$

By supposing $\varepsilon_i \ll 1$, we can develop this expression at the first order and obtain

$$\rho[X_1, X_2] = \left(\varepsilon_1 \varepsilon_2 \rho[X_1^{(0)}, X_2^{(0)}] + \rho[\tilde{X}_1, \tilde{X}_2] \right) \cdot \left(1 - \frac{1}{2}(\varepsilon_1^2 + \varepsilon_2^2) \right) \quad (23)$$

The addition of the trend to the signal thus introduces a correction on the correlation, on the one hand by the direct accounting of the attenuated correlation between trends, and on the other hand by the interference term as a multiplicative factor.

To apply this result to our problematic, we assume that $d \simeq l_0$, l_0 being the minimal distance to estimate correlations. We furthermore have the stationarity scale d_s which corresponds to the scale of variation of correlations, and according to the empirical results verifies $d_s > l_0$, significantly at least for some indicators (for example hierarchy and Moran, for which it is of the order of magnitude of the country). Finally, we denote by $\delta_0 = d_0/d$ the scale of the trend in terms of δ . We therefore assume

$$d < d_s < d_0$$

For δ values such that $\delta \cdot d < d_s$, we should have $\hat{\text{Cov}}_\delta[\tilde{X}_1, \tilde{X}_2] \simeq \hat{\text{Cov}}_{\delta=1}[\tilde{X}_1, \tilde{X}_2]$ if $\hat{\text{Cov}}_\delta$ is the estimator on the area of size δ .

Furthermore, we can reasonably assume that $\hat{\text{Var}}_{\delta=1}[X_i^{(0)}] \ll \hat{\text{Var}}_{\delta=d_s/d}[X_i^{(0)}]$, i.e. that the trend is constant at the largest scale in comparison to variation at the intermediate scales.

Under these assumptions, the estimator of ρ should vary as a function of δ according to the variations of ε_i as a function of δ . Finally, under the assumption that trends have a very low correlation (independent structural effects), we keep the correction by interferences in the expression of ρ , and thus that $\rho(\delta)$ decreases for low values of δ .

We have thus demonstrated that a simple multi-scalar structure of the process implies a variation of the estimated correlation as a function of δ , under a certain number of assumptions. The reciprocal has

no reason a priori to be true. The link we establish here is thus an illustration to reinforce an hypothesis, which is furthermore also sustained by results on the variation of the confidence interval described in the following.

Confidence interval for the correlation

We derive here the behavior of the correlation estimator as a function of the size of the sample. Under the assumption of a normal distribution of two random variables X, Y , then the Fisher transform of the Pearson estimator $\hat{\rho}$ such that $\hat{\rho} = \tanh(\hat{z})$ has a normal distribution. If z is the transform of the real correlation ρ , then a confidence interval for ρ is of size

$$\rho_+ - \rho_- = \tanh(z + k/\sqrt{N}) - \tanh(z - k/\sqrt{N})$$

where k is a constant. As $\tanh z = \frac{\exp(2z)-1}{\exp(2z)+1}$, we can develop this expression and reduce it, to obtain

$$\begin{aligned}\rho_+ - \rho_- &= 2 \cdot \frac{\exp(2k/\sqrt{N}) - \exp(-2k/\sqrt{N})}{\exp(2z) - \exp(-2z) + \exp(2k/\sqrt{N}) + \exp(-2k/\sqrt{N})} \\ &= 2 \cdot \frac{\sinh(2k/\sqrt{N})}{\cosh(2z) + \cosh(2k/\sqrt{N})}\end{aligned}$$

Using the fact that $\cosh u \sim_0 1 + u^2/2$ and that $\sinh u \sim_0 u$, we indeed obtain that $\rho_+ - \rho_- \sim_{N \gg 0} k'/\sqrt{N}$. ■

* * *

*

A.5 CAUSALITY REGIMES

A.5.1 Synthetic data

Time series

Calculons ici les valeurs théoriques des corrélations retardées pour un processus auto-régressif simple. Nous rappelons le cadre, à savoir $\vec{X}(t)$ qui est un processus stochastique suivant l'équation d'auto-régression

$$\vec{X}(t) = \sum_{\tau>0} \mathbf{A}(\tau) \cdot \vec{X}(t-\tau) + \vec{\varepsilon}(t)$$

et nous nous plaçons dans le cas où $\mathbf{A}(\tau) = 0$ pour $\tau \neq \tau_0$ et

$$\mathbf{A}(\tau_0) = \begin{pmatrix} 0 & \alpha \\ \alpha & 0 \end{pmatrix}$$

avec $-1 < \alpha < 1$. Nous supposons de plus $\vec{\varepsilon}$ bruit blanc et notons $\vec{\varepsilon} = (\varepsilon_X, \varepsilon_Y)$ et supposons $\text{Var}[\varepsilon_X] = \text{Var}[\varepsilon_Y] = \sigma^2$.

En notant $\vec{X} = (X, Y)$, le processus est spécifié par

$$\begin{cases} X(t) = \alpha \cdot Y(t - \tau_0) + \varepsilon_X \\ Y(t) = \alpha \cdot X(t - \tau_0) + \varepsilon_Y \end{cases}$$

En prenant la variance dans les deux équations et en faisant la différence, on obtient que nécessairement $\text{Var}[X] = \text{Var}[Y]$ car $\alpha^2 \neq 1$. La somme donne alors $\text{Var}[X] = \text{Var}[Y] = \frac{\sigma^2}{1-\alpha^2}$.

Nous calculons alors

$$\begin{aligned} \rho[X(t), Y(t - \tau_0)] &= \rho[\alpha Y(t - \tau_0) + \varepsilon_X, Y(t - \tau_0)] \\ &= \frac{\text{Cov}[\alpha Y(t - \tau_0) + \varepsilon_X, Y(t - \tau_0)]}{\sqrt{(\alpha^2 \text{Var}[Y] + \sigma^2) \text{Var}[Y]}} \\ &= \frac{\alpha \text{Var}[Y]}{|\alpha| \text{Var}[Y] \sqrt{1 + \frac{\sigma^2}{\alpha^2 \text{Var}[Y]}}} = \frac{\alpha}{|\alpha| \sqrt{1 + \frac{1-\alpha^2}{\alpha^2}}} \\ &= \alpha \end{aligned}$$

Il est en fait possible de calculer la corrélation retardée pour τ quelconque. Par stationnarité du processus, on a pour $\tau > 0$, $\rho[X(t), Y(t - \tau)] = \rho[X(\tau), Y(0)]$.

De la même manière que précédemment, nous développons pour $\tau > 0$

$$\begin{aligned}
\rho[X(\tau), Y(0)] &= \rho[aY(\tau - \tau_0) + \varepsilon_X, Y(0)] \\
&= \rho[a^2X(\tau - 2\tau_0) + a\varepsilon_Y + \varepsilon_X, Y(0)] \\
&= \frac{a^2 \text{Cov}[X(\tau - 2\tau_0), Y(0)]}{\sqrt{(a^4 \text{Var}[X] + (1 + a^2)\sigma^2) \text{Var}[Y]}} \\
&= \frac{\rho[X(\tau - 2\tau_0), Y(0)]}{\sqrt{1 + (1 + a^2)(1 - a^2)/a^4}} = a^2 \cdot \rho[X(\tau - 2\tau_0), Y(0)]
\end{aligned}$$

et donc par récurrence, pour $k \in \mathbb{N}$,

$$\rho[X(\tau), Y(0)] = a^{2k} \cdot \rho[X(\tau - 2k\tau_0), Y(0)]$$

Si $\tau \notin (2\mathbb{N} + 1)\tau_0$, on descend à $\rho[X(\tau'), Y(0)]$ tel que $\tau' < \tau_0$ et la corrélation est donc nulle.

Si $\tau \in (2\mathbb{N} + 1)\tau_0$, on a alors

$$\rho[X((2k + 1)\tau_0), Y(0)] = a^{2k+1}$$

Pour $\tau < 0$, le calcul est similaire en échangeant les variables.

Ce modèle simple auto-régressif permet ainsi de contrôler simplement les corrélations retardées à des ordres donnés.

Urban Morphogenesis

La Fig.81 donne, pour l'analyse non-supervisée menée sur les caractéristiques issues des corrélations retardées, le comportement des résultats du clustering en fonction du nombre de cluster k , qui permet de lire une transition en fonction de k . Nous donnons aussi que la répartition des clusters dans un plan principal pour $k = 6$.

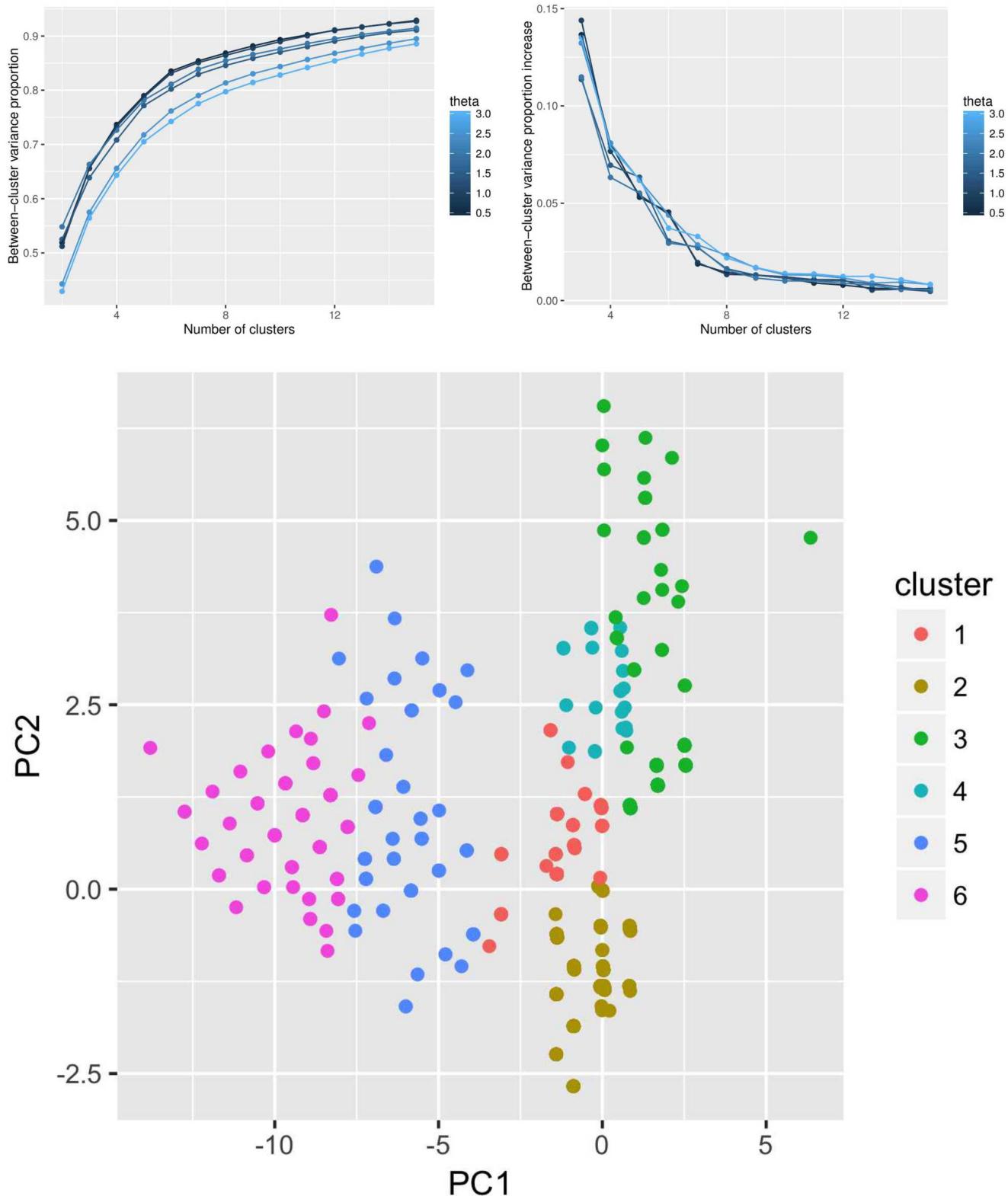
A.5.2 *South Africa*

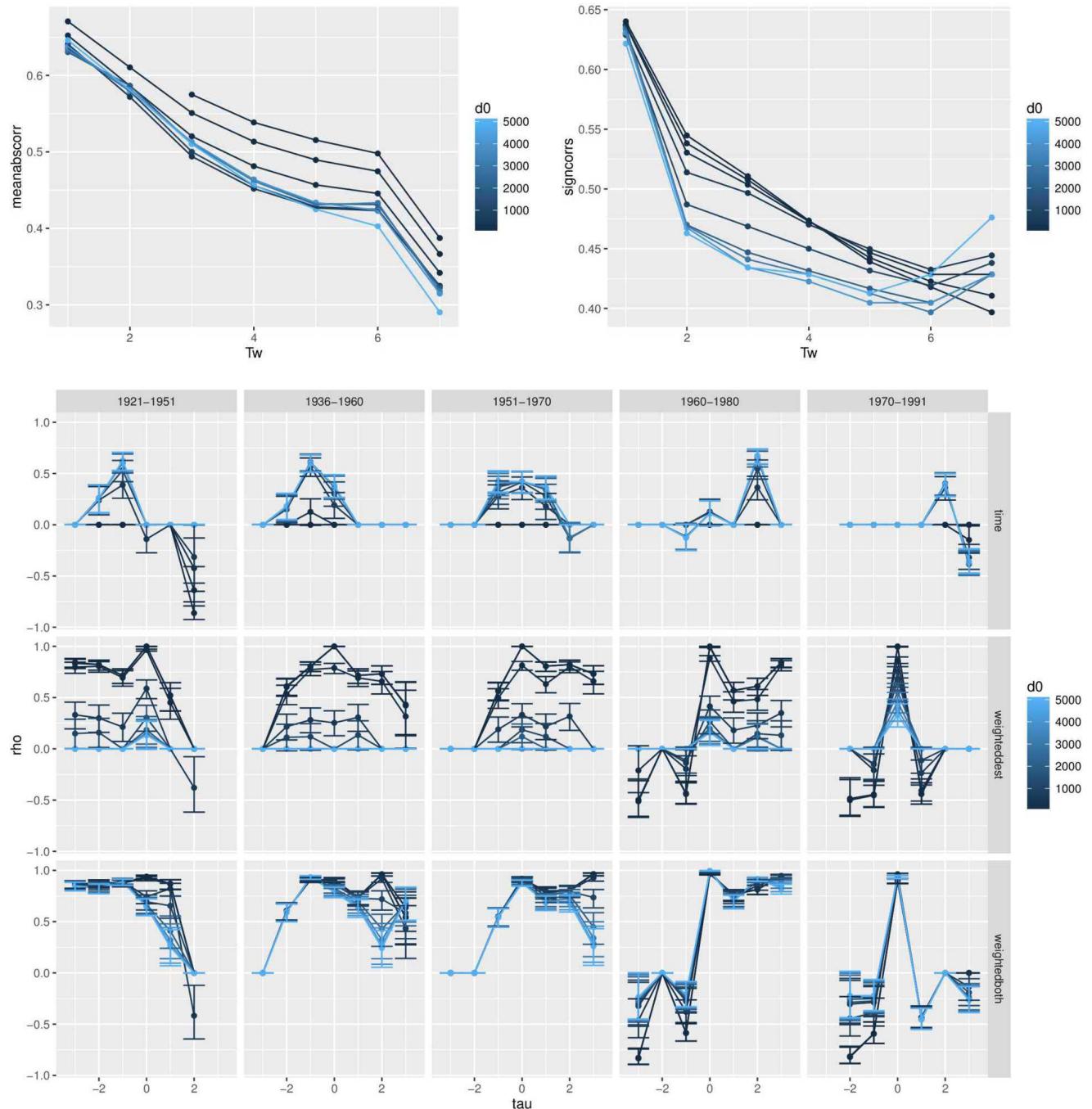
FIGURE 81: La Fig. 82 donne le comportement des corrélations estimées, en termes de corrélation absolue moyenne, et de proportion de corrélations significatives, en fonction de d_0 et de T_W . Elle donne également les profils de corrélations retardées pour les accessibilité pondérées, à l'origine et à la destination.

FIGURE 82:

★ ★

*





A.6 AGGREGATION-DIFFUSION MORPHOGENESIS

A.6.1 Extended Figures for Model Exploration

Convergence

Histograms for the 81 parameters points for which we did 100 repetitions are given in Fig. ??, for Moran index and slope indicators. Other indicators showed similar convergence patterns. The visual exploration of histograms confirms the numerical analysis done in main text for statistical convergence.

Indicators Behavior

We show in Fig. ?? to Fig. 87 the full behavior of all indicators, with all parameters varying, obtained through the extensive exploration, from which the plots in main text have been extracted. Because of the complex nature of emergent urban form, one can not predict output values without referring to this “exhaustive” parameter sweep.

Indicators Scatterplots

We show finally the full scatterplots of indicators, with real data points, in Fig. 88. These are preliminary step of the calibration on principal components, and we can see on these on which dimensions the model fails relatively to fit real data (in particular average distance).

A.6.2 Semi-analytical analysis of the simplified model

Partial Differential Equation

We propose to derive the PDE in a simplified setting. To recall the configuration given in main text, the system has one dimension, such that $x \in \mathbb{R}$ with $1/\delta x$ cells of size δx , and we use the expected values of cell population $p(x, t) = \mathbb{E}[P(x, t)]$. We furthermore take $n_d = 1$. Larger values would imply derivatives at an order higher than 2 but the following results on the existence of a stationary solution should still hold.

Denoting $\tilde{p}(x, t)$ the intermediate populations obtained after the aggregation stage, we have

$$\tilde{p}(x, t) = p(x, t) + N_g \cdot \frac{p(x, t)^\alpha}{\sum_x p(x, t)^\alpha}$$

since all populations units are added independently. If $\delta x \ll 1$ then $\sum_x p^\alpha \simeq \int_x p(x, t)^\alpha dx$ and we write this quantity $P_\alpha(t)$. We furthermore write $p = p(x, t)$ and $\tilde{p} = \tilde{p}(x, t)$ in the following for readability.

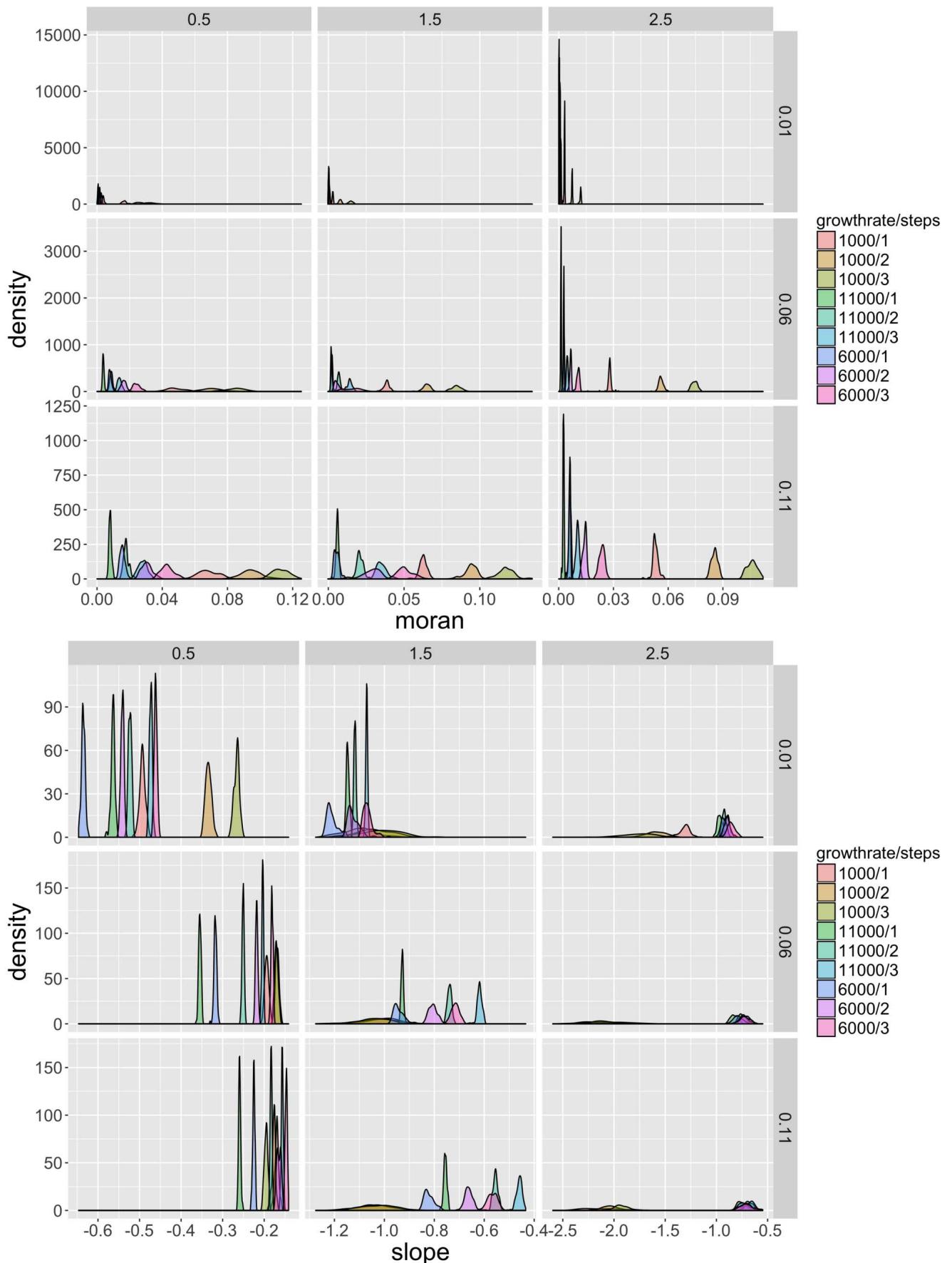


FIGURE 83: Histograms for Moran index (top) and slope (bottom), for varying α (columns), β (rows), N_G and n_d (colors).

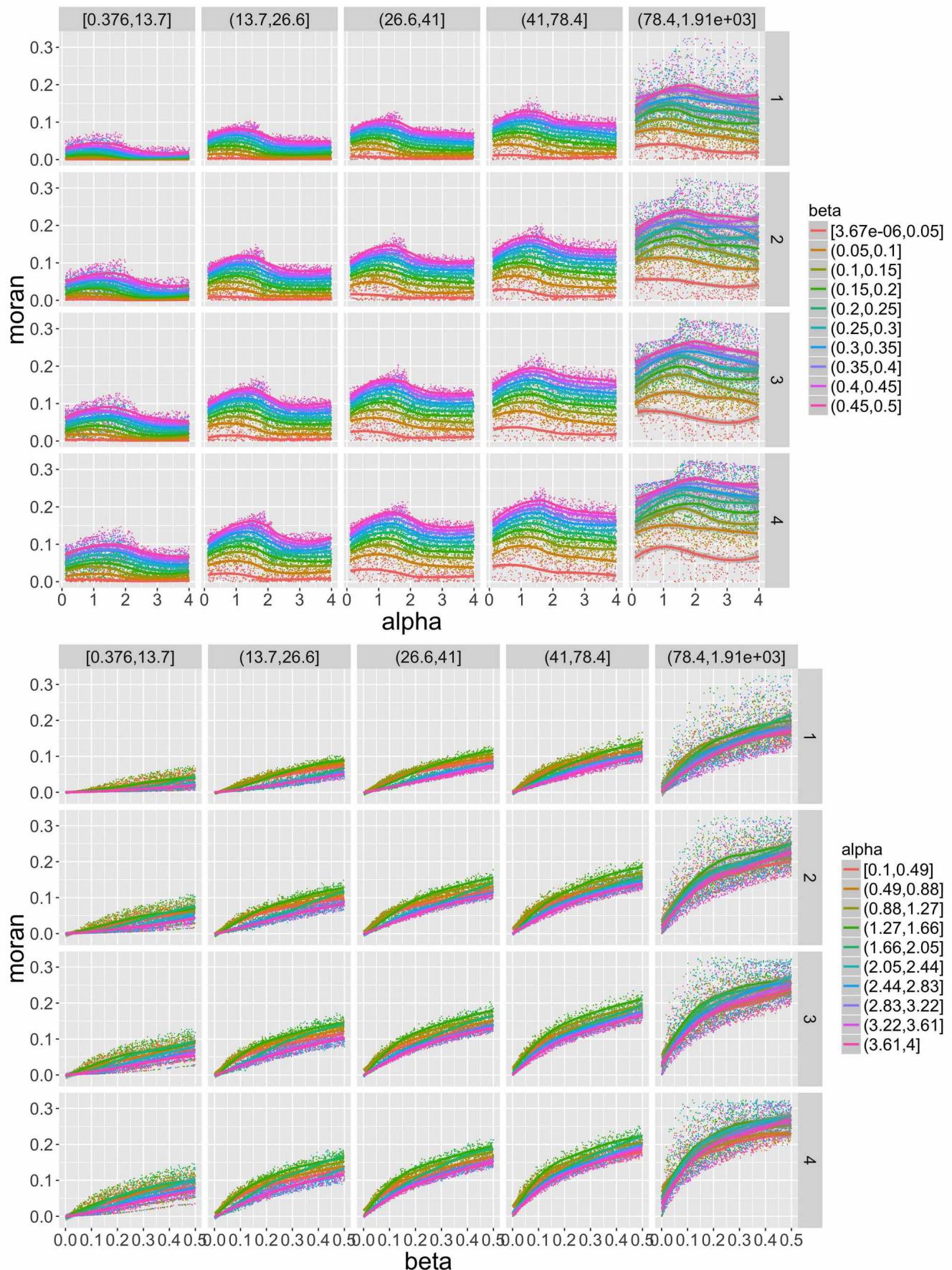


FIGURE 84: Moran index as a function of α (Top) and β (Bottom) for varying β (resp. α) given by color, and varying n_d (rows) and N_G (columns).

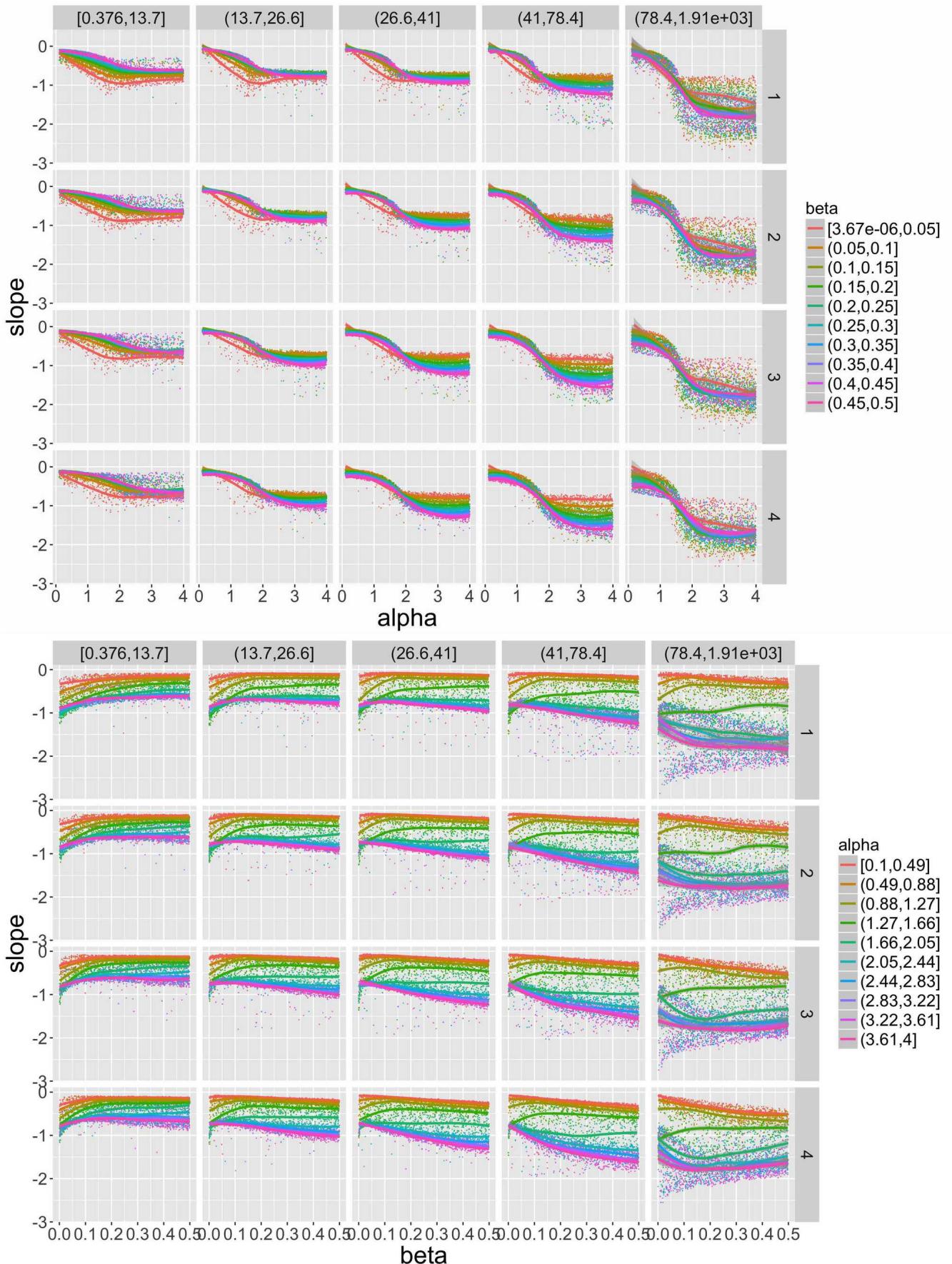


FIGURE 85: Slope as a function of α (Top) and β (Bottom) for varying β (resp. α) given by color, and varying n_d (rows) and N_G (columns).

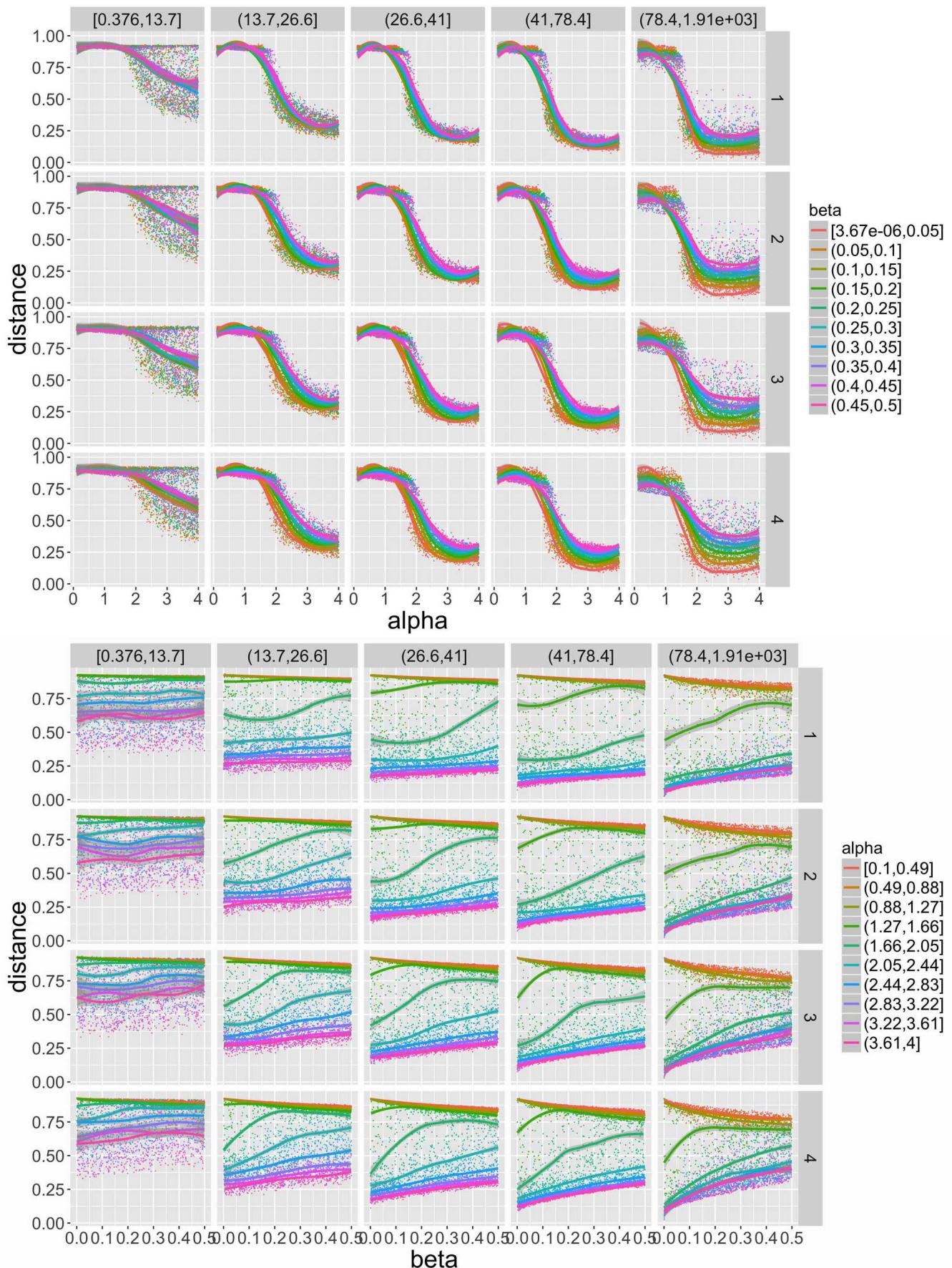


FIGURE 86: Average distance index as a function of α (Top) and β (Bottom) for varying β (resp. α) given by color, and varying n_d (rows) and N_G (columns).

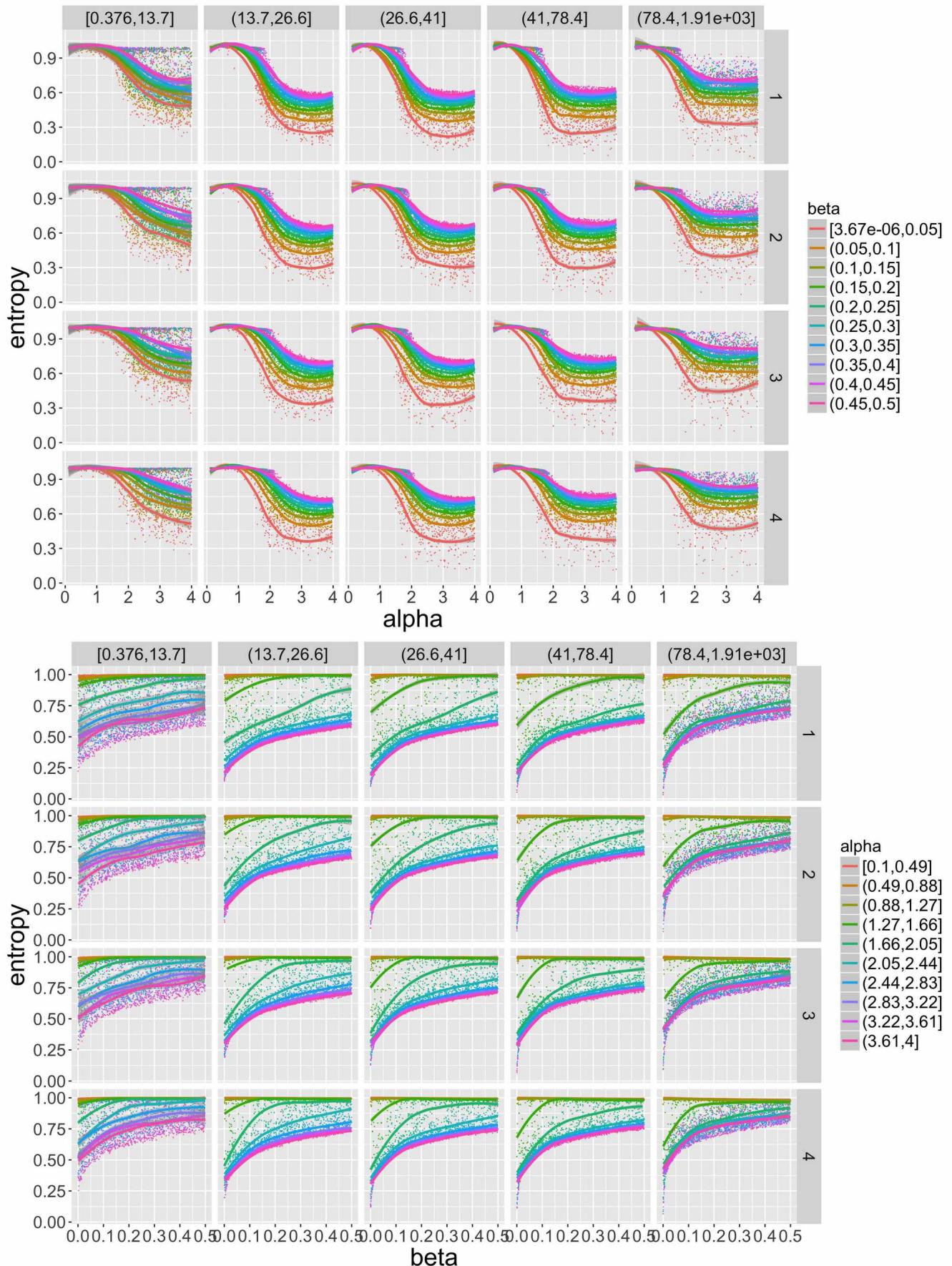


FIGURE 87: Entropy as a function of α (Top) and β (Bottom) for varying β (resp. α) given by color, and varying n_d (rows) and N_G (columns).

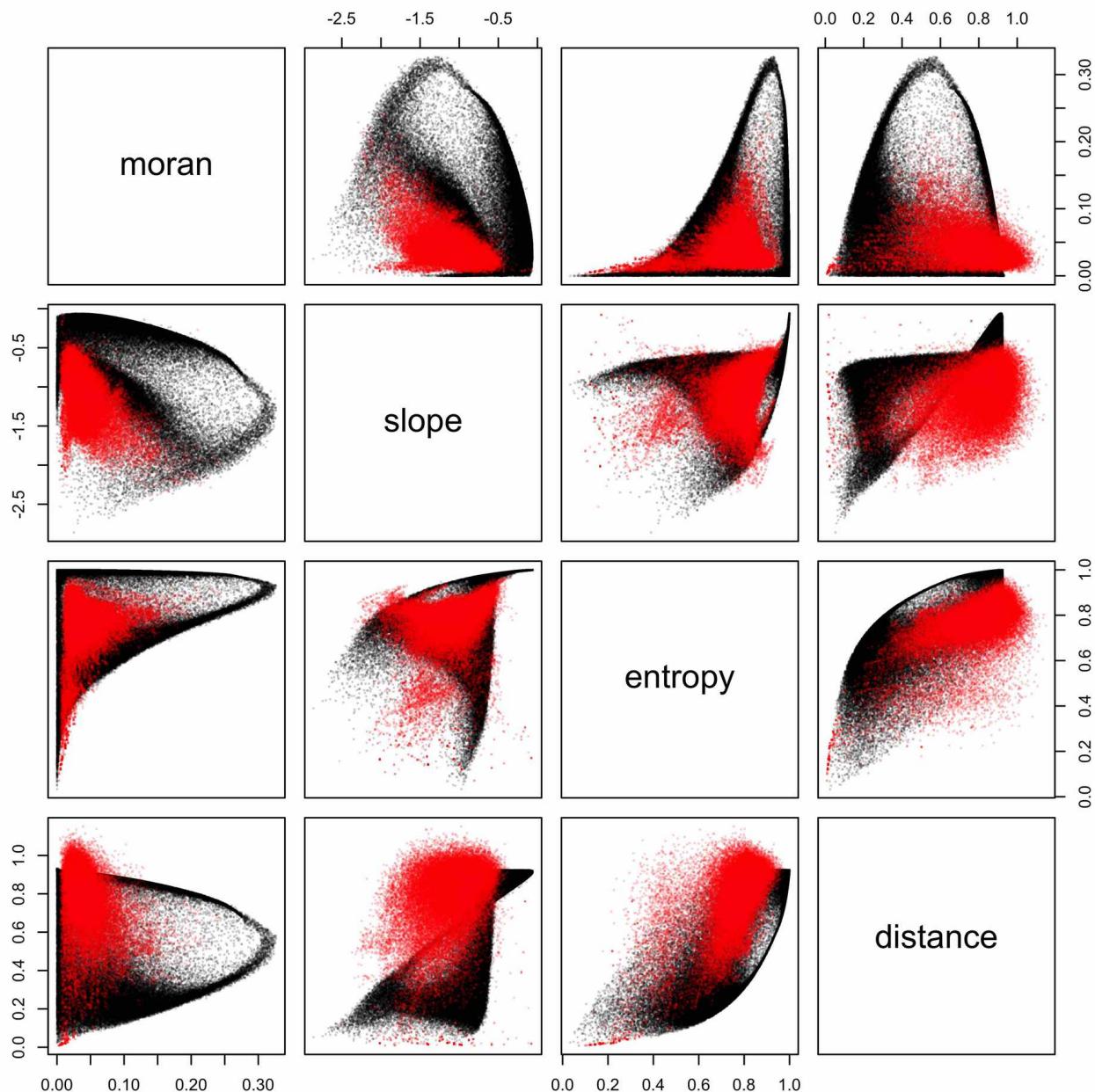


FIGURE 88: Scatterplots of indicators distribution in the sampled hypercube of the parameter space. Red points correspond to real data.

The diffusion step is then deterministic, and for any cell not on the border ($0 < x < 1$), if δt is the interval between two time steps, we have

$$\begin{aligned} p(x, t + \delta t) &= (1 - \beta) \cdot \tilde{p} + \frac{\beta}{2} [\tilde{p}(x - \delta x, t) + \tilde{p}(x + \delta x, t)] \\ &= \tilde{p} + \frac{\beta}{2} [(\tilde{p}(x + \delta x, t) - \tilde{p}) - (\tilde{p} - \tilde{p}(x - \delta x, t))] \end{aligned}$$

Assuming the partial derivatives exist, and as $\delta x \ll 1$, we make the approximation $\tilde{p}(x + \delta x, t) - \tilde{p} \simeq \delta x \cdot \frac{\partial \tilde{p}}{\partial x}(x, t)$, what gives

$$(\tilde{p}(x + \delta x, t) - \tilde{p}) - (\tilde{p} - \tilde{p}(x - \delta x, t)) = \delta x \cdot \left(\frac{\partial \tilde{p}}{\partial x}(x, t) - \frac{\partial \tilde{p}}{\partial x}(x - \delta x, t) \right)$$

and therefore at the second order

$$p(x, t + \delta t) = \tilde{p} + \frac{\beta \delta x^2}{2} \cdot \frac{\partial^2 \tilde{p}}{\partial x^2}$$

Substituting \tilde{p} gives

$$\begin{aligned} \frac{\partial^2 \tilde{p}}{\partial x^2} &= \frac{\partial^2 p}{\partial x^2} + \frac{N_G}{P_\alpha} \cdot \frac{\partial}{\partial x} \left[\alpha \frac{\partial p}{\partial x} p^{\alpha-1} \right] \\ &= \frac{\partial^2 p}{\partial x^2} + \alpha \frac{N_G}{P_\alpha} \left[\frac{\partial^2 p}{\partial x^2} p^{\alpha-1} + (\alpha-1) \left(\frac{\partial p}{\partial x} \right)^2 p^{\alpha-2} \right] \end{aligned}$$

By supposing that $\frac{\partial p}{\partial t}$ exists and that δt is small, we have $p(x, t + \delta t) - p(x, t) \simeq \delta t \frac{\partial p}{\partial t}$, what finally yields , by combining the results above, the partial differential equation

$$\delta t \cdot \frac{\partial p}{\partial t} = \frac{N_G \cdot p^\alpha}{P_\alpha(t)} + \frac{\alpha \beta (\alpha-1) \delta x^2}{2} \cdot \frac{N_G \cdot p^{\alpha-2}}{P_\alpha(t)} \cdot \left(\frac{\partial p}{\partial x} \right)^2 + \frac{\beta \delta x^2}{2} \cdot \frac{\partial^2 p}{\partial x^2} \cdot \left[1 + \alpha \frac{N_G p^{\alpha-1}}{P_\alpha(t)} \right] \quad (24)$$

Initial conditions should be specified as $p_0(x) = p(x, t_0)$. To have a well-posed problem similar to more classical PDE problems, we need to assume a domain and boundary conditions. A finite support is expressed by $p(x, t) = 0$ for all t and x such that $|x| > x_m$.

Stationary solution for density

The non-linearity and the integral terms making the equation above out of the scope for analytical resolution, we study its behavior numerically in some cases. Taking a simple initial condition $p_0(0) = 1$ and $p_0(x) = 0$ for $x \neq 0$, we show that on a finite domain, density

$d(x, t)$ always converge to a stationary solution for large t , for a large set of values of (α, β) with fixed $N_G = 10$ ($\alpha \in [0.4, 1.5]$ varying with step 0.025 and $\log \beta \in [-1, -0.5]$ with step 0.1). We show in Fig. 89 the corresponding trajectories on a typical subset. The variation of the asymptotic distribution as a function of α and β are not directly visible, as they depend on very low values of the outward flows at boundaries. We give in Fig. 90 their behavior, by showing the value of the maximum of the distribution. Low values of β give an inversion in the effect of α , whereas high values of β give comparable values for all α .

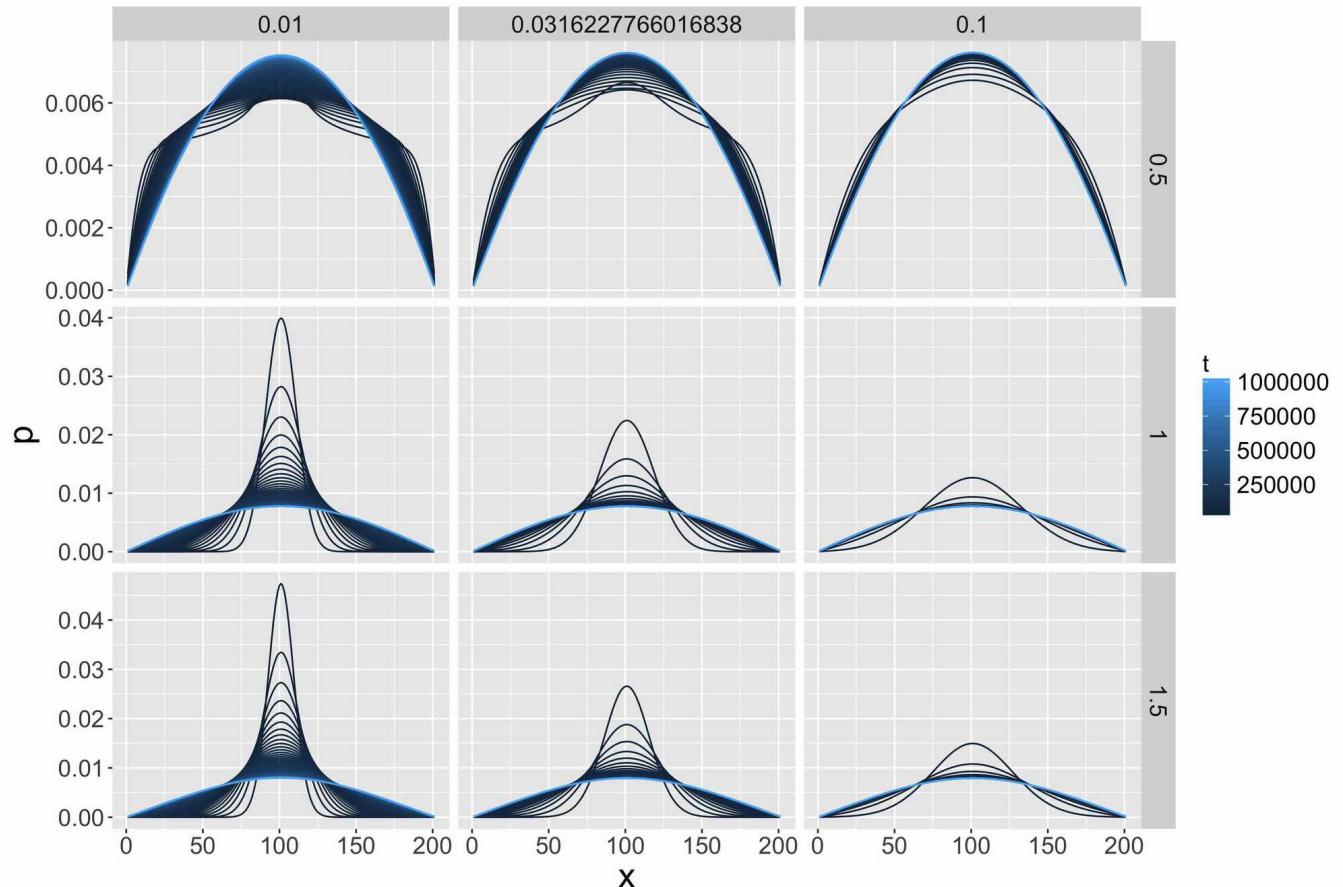


FIGURE 89: Trajectories of densities as a function of the spatial dimension, for varying β (columns) and α (rows). Color gives time.

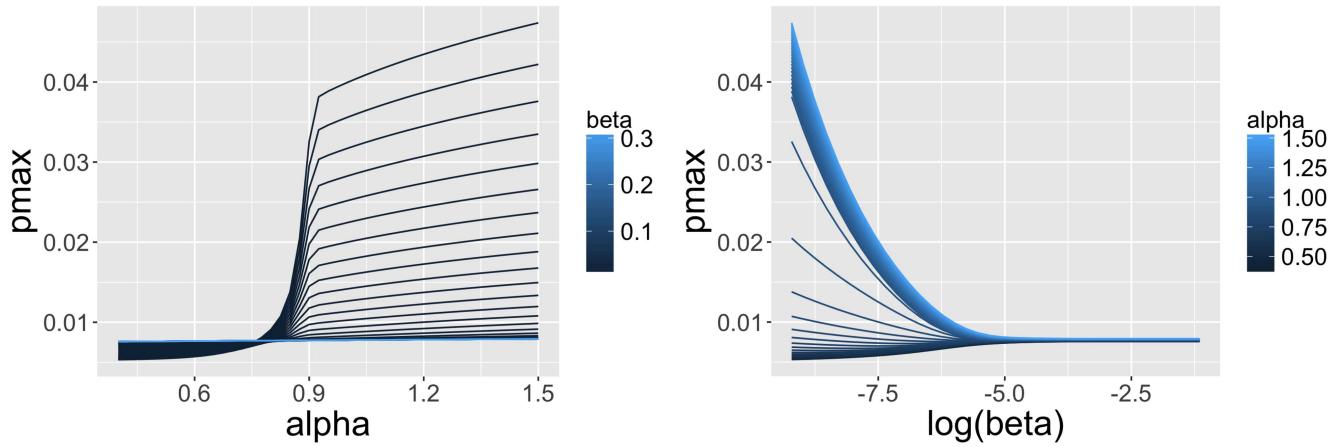


FIGURE 90: Dependency of $\max d(t \rightarrow \infty)$ to α and β .

A.7 CORRELATED SYNTHETIC DATA

Pour la simulation du couplage faible entre génération d'une configuration de densité et modèle de génération de réseau, la Fig. ?? donne les erreurs sur les corrélations faisables montrées en Fig. 41, ainsi que l'amplitude des corrélations pour l'ensemble de la matrice, c'est-à-dire à la fois la corrélation absolue maximale $c_{ij} = \max_k |\rho_{ij}^k|$ et l'amplitude totale $a_{ij} = \max_k \rho_{ij}^{(k)} - \min_k \rho_{ij}^{(k)}$.

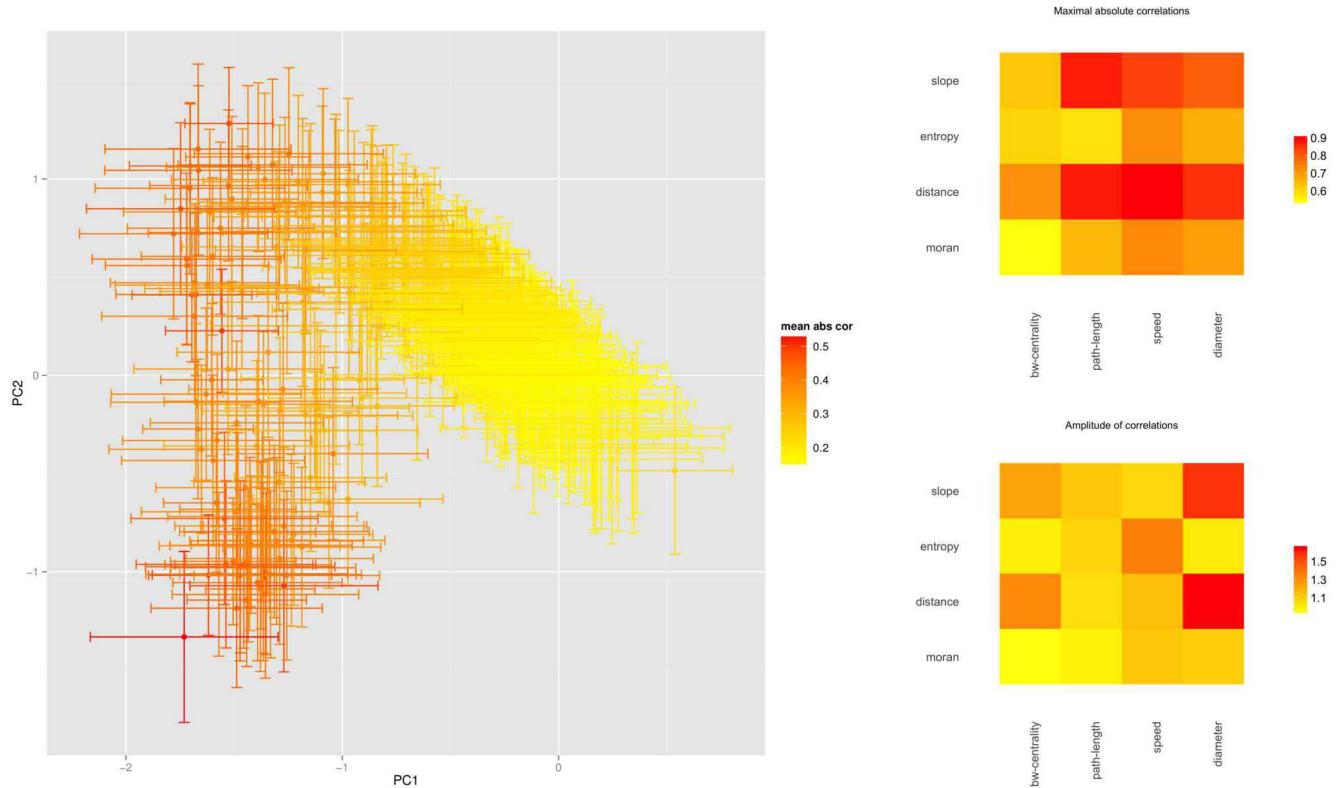


FIGURE 91: (b) Heatmaps for amplitude of correlations, defined as $a_{ij} = \max_k \rho_{ij}^{(k)} - \min_k \rho_{ij}^{(k)}$ and maximal absolute correlation, defined as $c_{ij} = \max_k |\rho_{ij}^{(k)}|$. (c) Projection of correlation matrices in a principal plan obtained by Principal Component Analysis on matrix population (cumulated variances: PC1=38%, PC2=68%). Error bars are initially computed as 95% confidence intervals on each matrix element (by standard Fisher asymptotic method), and upper bounds after transformation are taken in principal plan. Scale color gives mean absolute correlation on full matrices.

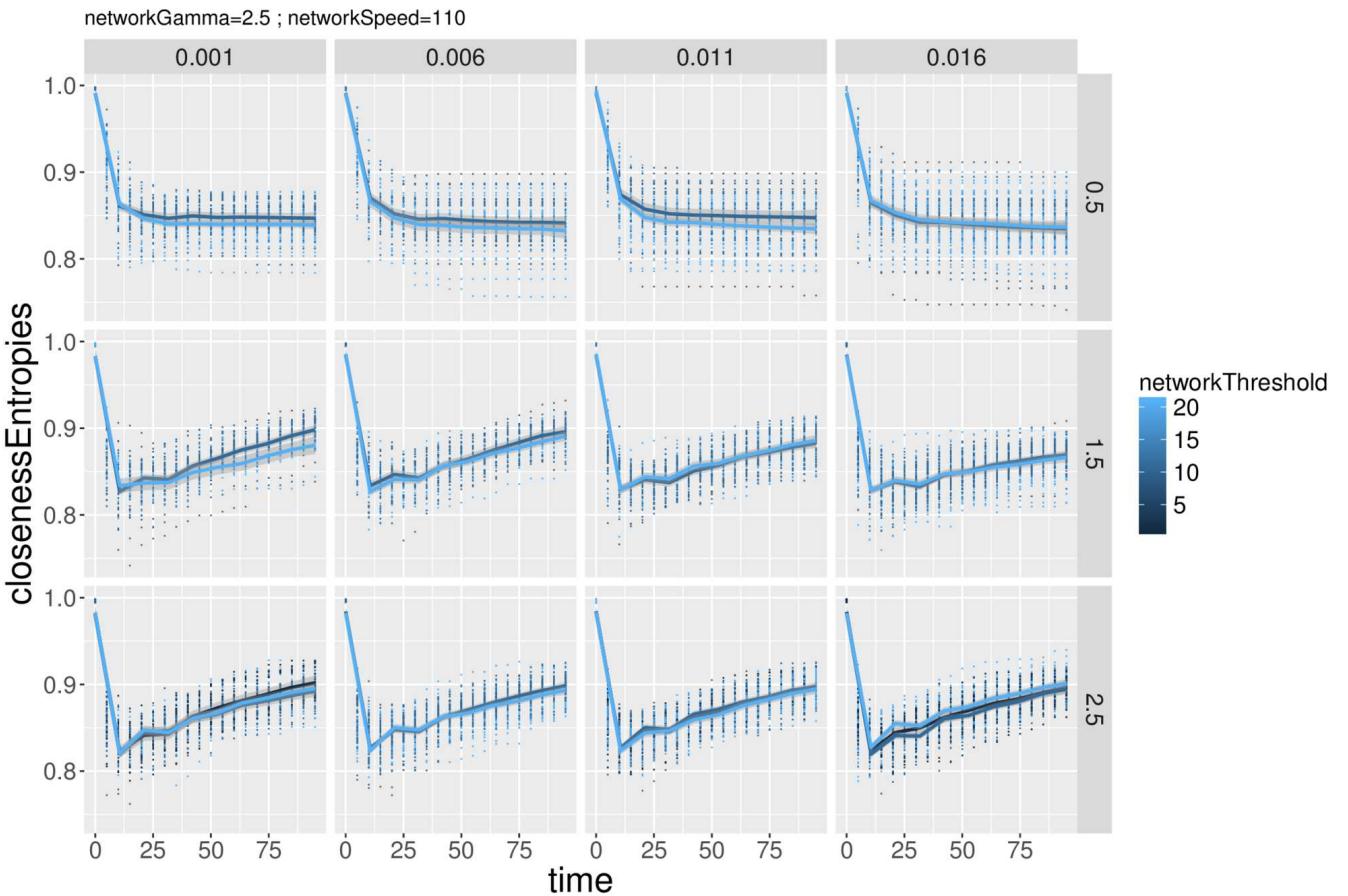


FIGURE 92: Entropy of closeness centralities.

A.8 EXPLORATION OF THE SIMPOPNET MODEL

Nous donnons ici des figures supplémentaires permettant de se rendre compte de la sensibilité des résultats aux paramètres non présentés en texte principal.

La Fig. 92 permet de visualiser la sensibilité de l'entropie des centralités $\varepsilon[\mu_i]$ en fonction de d_G , θ_N et γ_G . La forme des courbes temporelles est principalement sensible à γ_G .

La Fig. 93 donne les variations de ρ_r en fonction de d_G et γ_G , pour des valeurs variables de θ_N et de γ_N . Nous constatons que la régularité observée en fonction de d_G et de γ_G n'est pas visiblement sensible aux variations de θ_N et de γ_N .

La Fig. 94 donne les corrélations ρ_d en fonction de la distance pour l'ensemble des couples de variables, pour d_G et γ_G variables. Nous retrouvons qualitativement les mêmes comportements que avec $d_G = 0.016$, à l'exception d'une très légère croissance pour les plus grande distances, pour la corrélation entre la population et l'accessibilité, à $d_G = 0.001$ et $\gamma_G = 0.5$, qui reste difficile à interpréter.

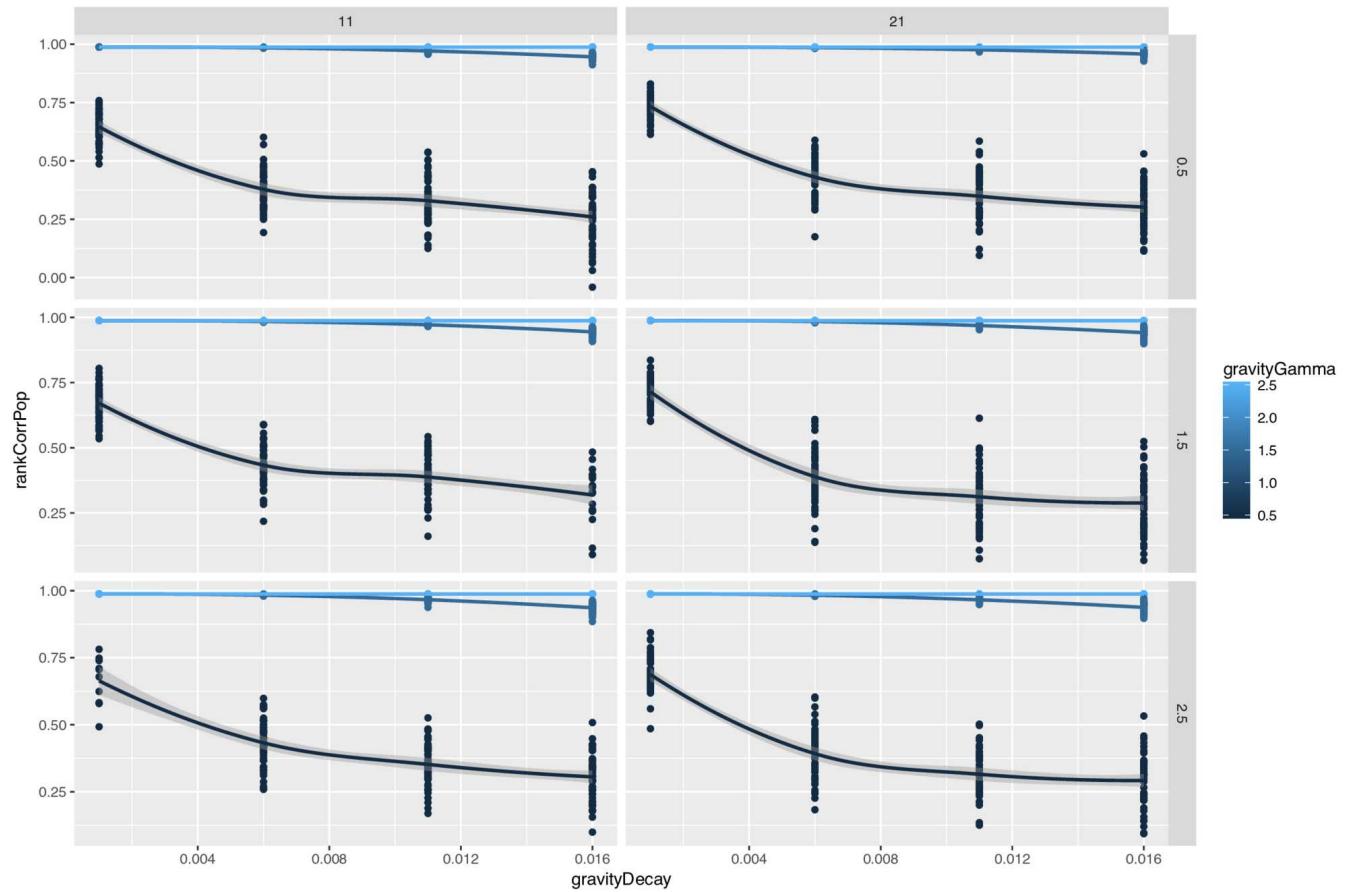


FIGURE 93: Population rank correlations.

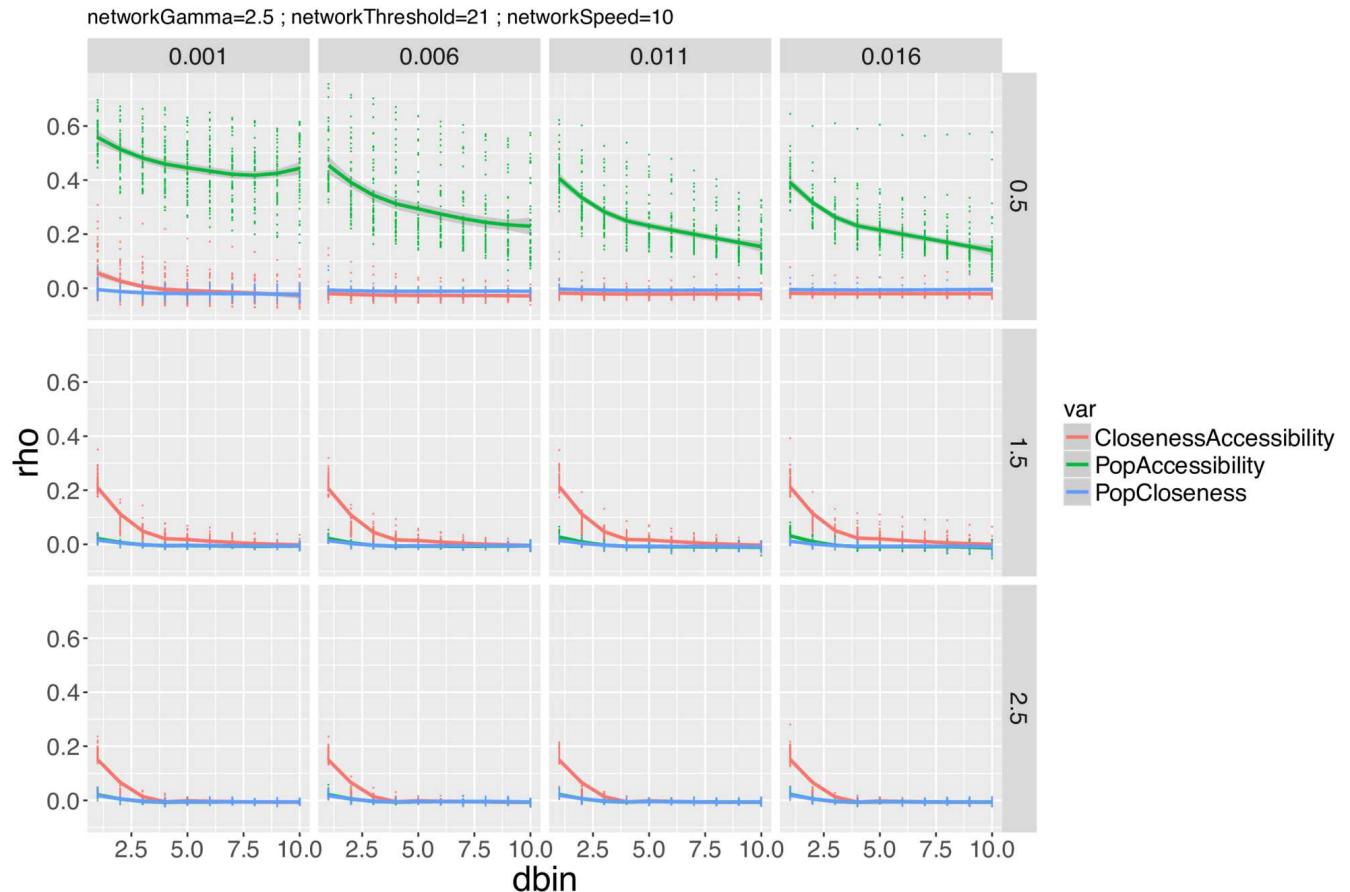


FIGURE 94: Distance correlation.

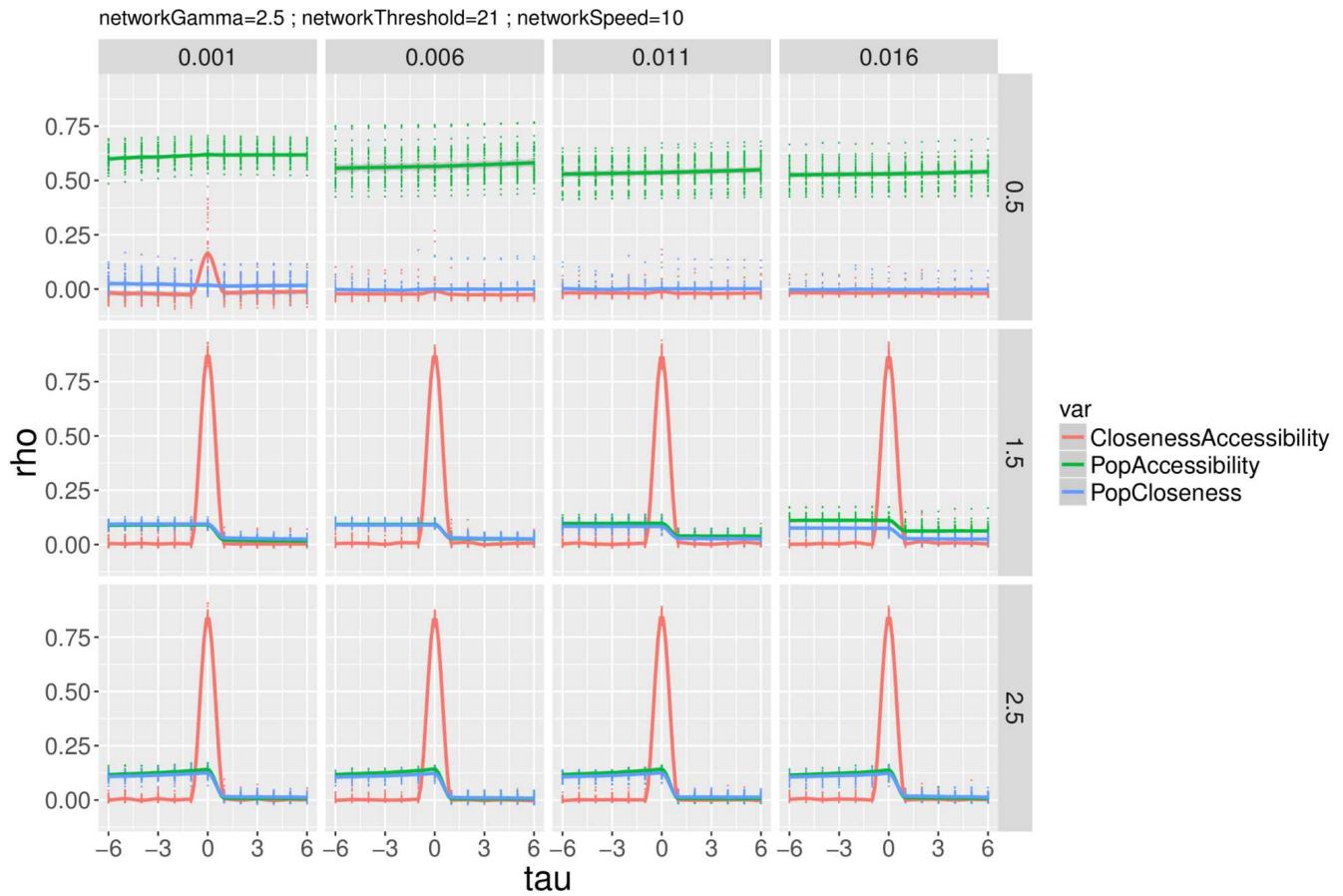


FIGURE 95: Lagged correlations.

Enfin, nous donnons en Fig. 95 les corrélations retardées ρ_τ entre l'ensemble des couples de variables, pour d_G et γ_G variables. De même, les comportements qualitatifs sont globalement stables pour les paramètres autres que γ_G .

A.9 MACROSCOPIC CO-EVOLUTION MODEL

A.9.1 Synthetic data

Exploration

Nous donnons en Fig. ?? la sensibilité des indicateurs temporels pour le modèle de co-évolution sur données synthétiques, en particulier $\bar{c}_i(t)$ et $\varepsilon[\mu_i](t)$, pour des variations de d_G , γ_G et ϕ_0 . Le comportement de \bar{c}_i est sensible à γ_G et ϕ_0 mais très peu à d_G . Celui de $\varepsilon[\mu_i]$ ne dépend que de γ_G pour son comportement moyen, et de d_G pour sa dispersion dans les faibles valeurs de d_G .

Nous donnons en Fig. 97 le comportement d'indicateurs agrégés, à savoir $C[Z_i]$ et $\rho_r[Z_i]$. La complexité des trajectoires d'accessibilité varie principalement selon d_G , γ_G et ϕ_0 pour les faibles valeurs. La corrélation de rang des accessibilités est quant à elle uniquement sensible à d_G et γ_G , ce qui veut dire que des différences d'évolution du réseau ne perturbent pas la dynamique de la hiérarchie des accessibilités.

La Fig. 99 donne les corrélations ρ_d en fonction des déciles de distance pour l'ensemble des couples de variables. Les fortes valeurs de d_G donnent des corrélations nulles pour l'ensemble des valeurs de la distance, tandis que $d_G = 10$ témoigne de régimes locaux. Une corrélation constante entre centralité et accessibilité émerge pour une valeur intermédiaire $d_G = 60$, qui est éventuellement à mettre en correspondance avec le maximum de complexité pour les accessibilités obtenu précédemment.

Enfin, la Fig. ?? donne les corrélations retardées ρ_τ pour l'ensemble des couples de variables. Les variations de γ_G influencent peu les régimes obtenus, contrairement à d_G , pour lequel on observe une variation continue de la forme qualitative des profils.

Plus précisément, nous observons que la corrélation entre population et accessibilité est globalement constante, probablement du fait de l'auto-corrélation, et n'entre pas en jeu dans la définition des régimes. Pour des grandes valeurs de d_G , on observe une déviation positive des corrélations pour les délais positifs et négatifs pour accessibilité et centralité. Il y a dans ce cas causalité circulaire et le modèle capture une co-évolution dans ce sens. L'accessibilité cause fortement la centralité pour $d_G = 10$, puis la tendance s'inverse pour les grands d_G . Pour $d_G = 10$, nous observons une relation à sens unique de la population vers le réseau. Pour les régimes intermédiaires, il y a circularité directement entre population et centralité. Enfin, pour $d_G > 110$ il y a "circularité indirecte" entre population et accessibilité, puisque accessibilité cause centralité qui cause population.

Cette exploration visuelle est préliminaire et est continuée par la validation statistique des différents régimes en texte principal.

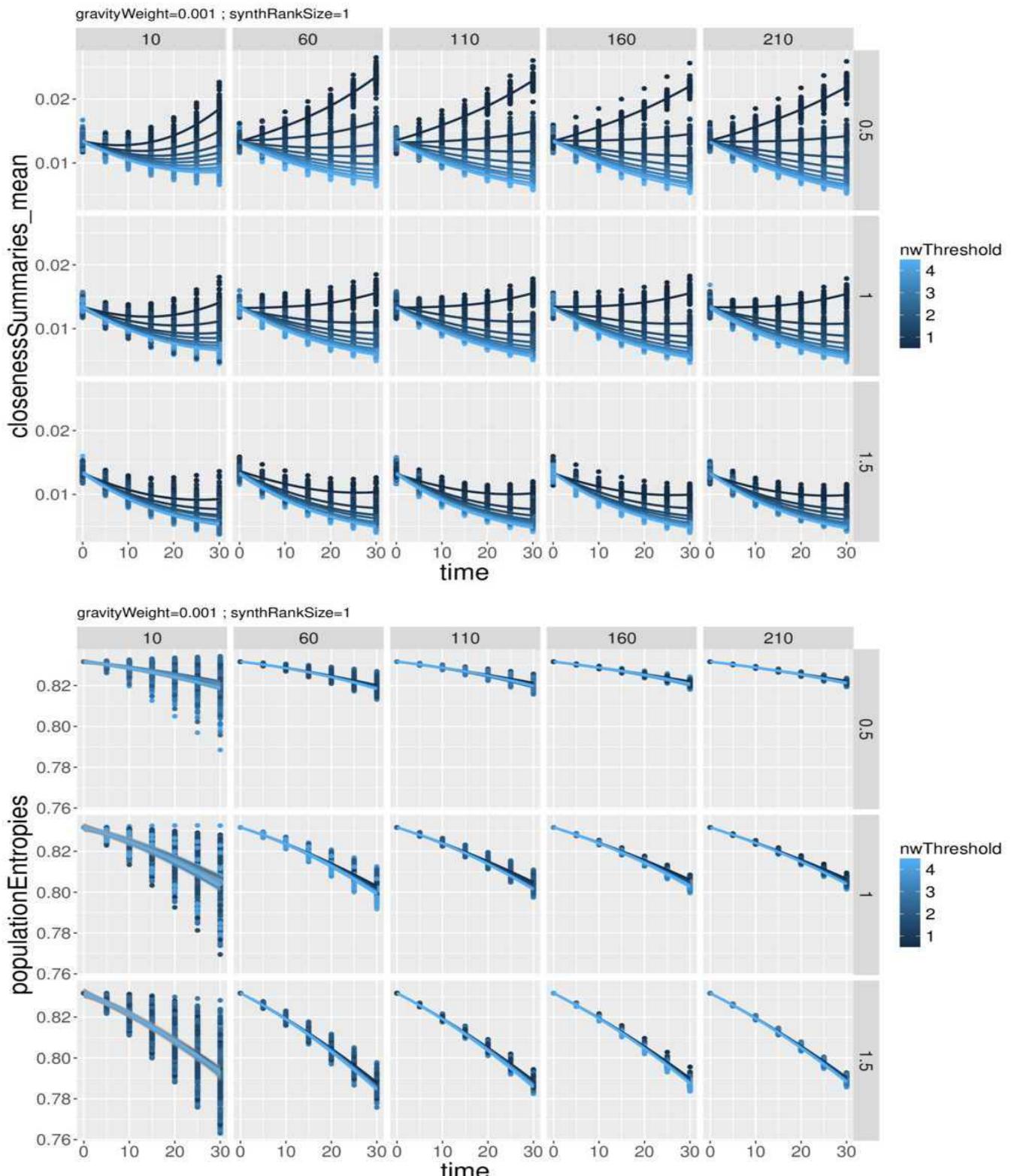


FIGURE 96: Behavior of the co-evolution model.

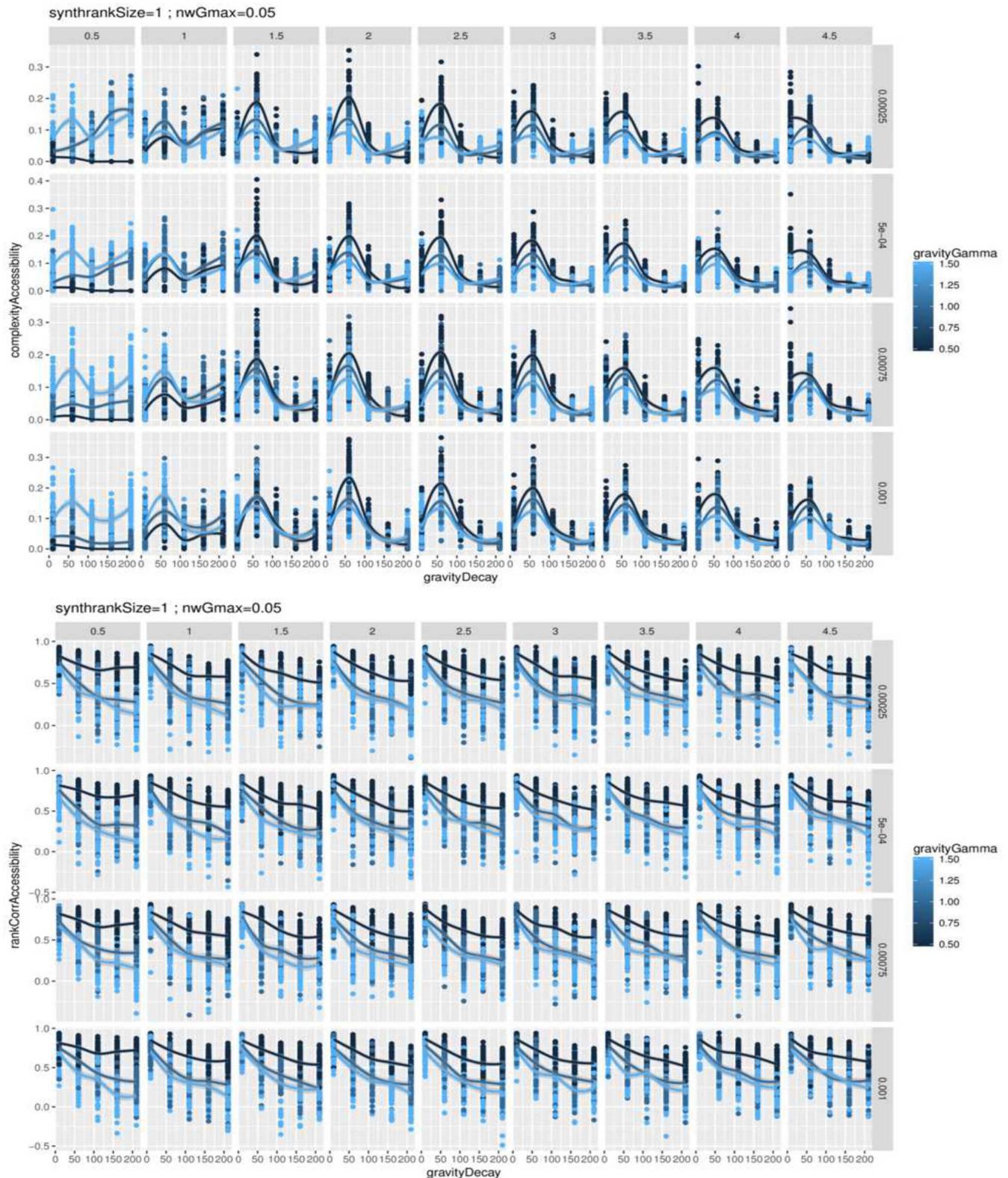
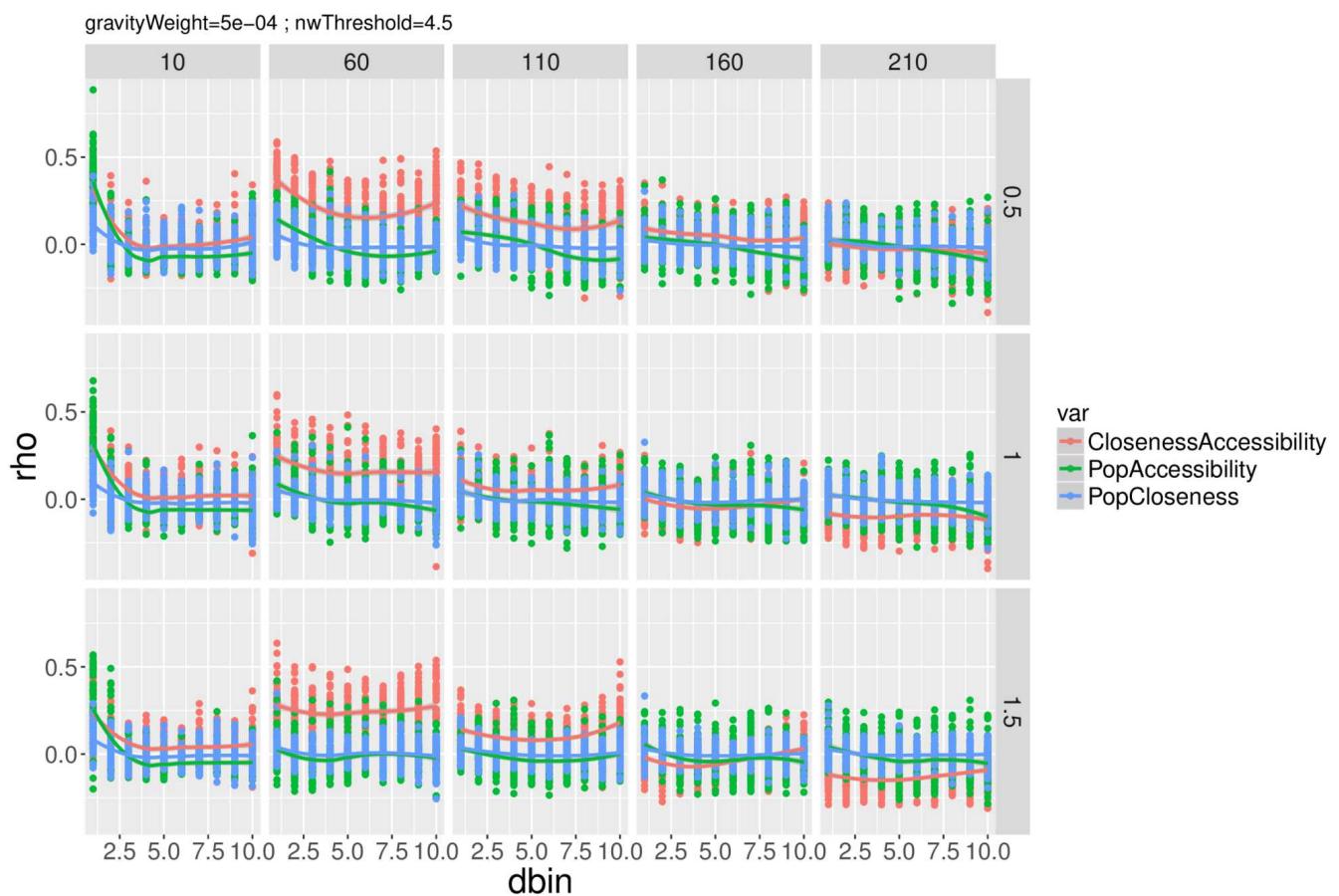


FIGURE 97: Aggregated indicators behavior for the model of coevolution at the macroscopic scale.



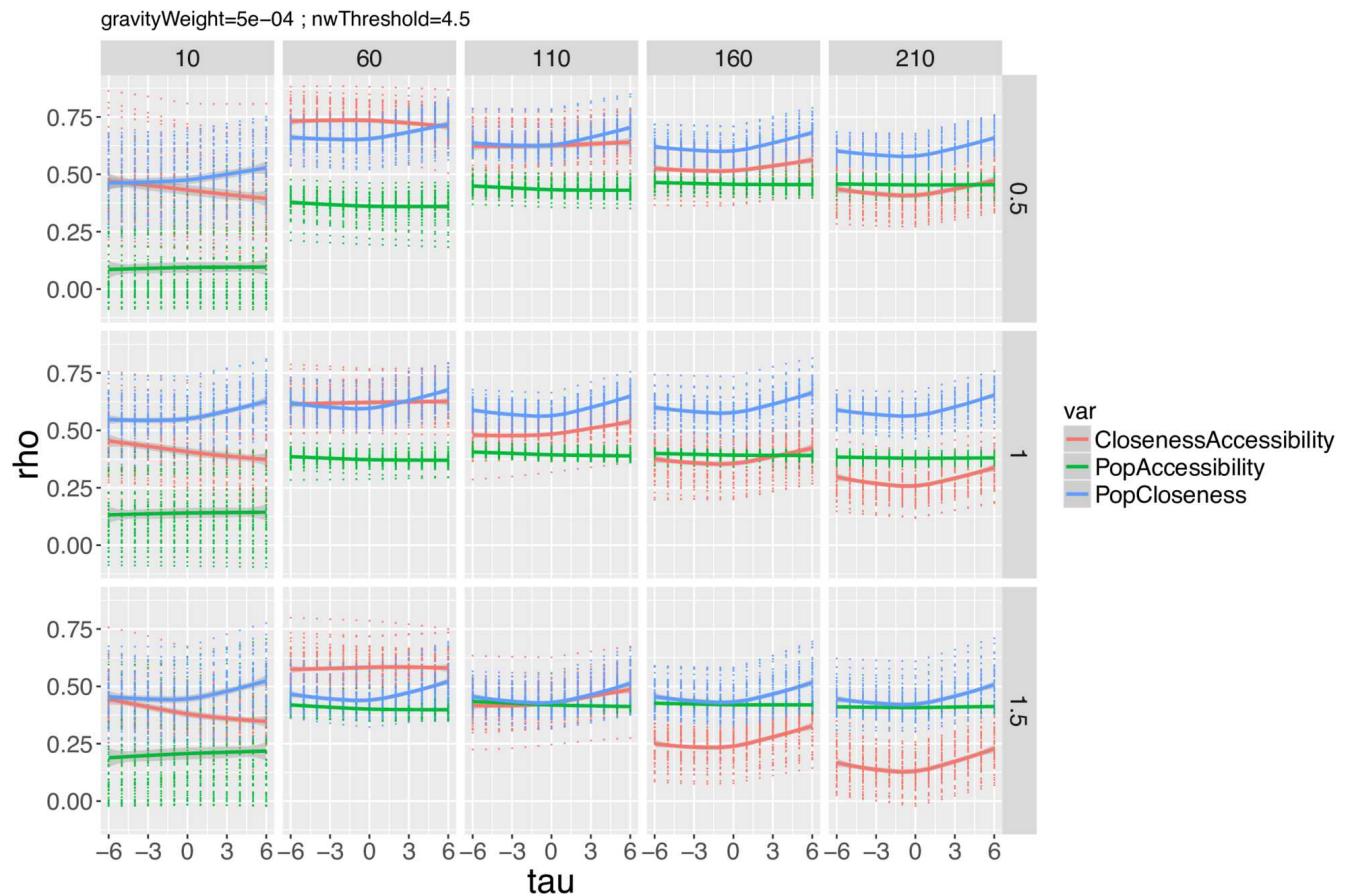


FIGURE 99: Lagged correlations.

Application of the PSE algorithm

L'algorithme a été précisément appliqué avec les objectifs $\rho_{\tau_{\pm}}[x_i, x_j] - \rho_0$ avec x_i les 3 variables considérées et $i < j$, la corrélation estimée étant nulle si non significative ou moins forte que ρ_0 . Les objectifs varient dans $[-0.2, 0.2]$ avec un pas de 0.01 (en pratique, la quasi-totalité des valeurs obtenues sont inférieures en valeur absolue à 0.1, puisque les premiers et derniers centiles y sont inférieurs, sauf deux exceptions à 0.12 et 0.16).

L'algorithme est lancé sur grille avec 300 îles en parallèle, chaque île ayant une durée de vie de 2 heures, pour un total de 616 générations.

Les résultats de la population obtenue sont montrés sous forme de nuage de points en Fig. 100. Nous constatons que la corrélation dont la distribution est la plus dispersée est $\rho_{\tau_+}[\mu_i, c_i]$. Par ailleurs, chaque couple de corrélations possède des quadrants impossibles à atteindre, suggérant des comportements impossibles du modèle : par exemple, il n'y a quasiment aucun point avec une causalité négative entre population et centralité et une causalité négative entre centralité et accessibilité, ces deux liens étant alors incompatibles. Le couple avec lequel il semble le plus dur d'étendre les circularités directes est accessibilité et centralité, ce qui suggère une domination de la centralité par rapport à la population dans l'expression de l'accessibilité puisque le lien avec population possède une plus grande étendue de liberté.

Principalement, l'algorithme révèle une richesse de comportements étendant encore celle obtenue par l'exploration simple.

A.9.2 *Real data*

Nous donnons en Fig. ?? les fronts de Pareto pour la calibration du modèle sur données réelles selon (ϵ_G, ϵ_L) , similaires à ceux donnés en 101, mais ici avec la couleur donnant la valeur du paramètre d_G . Nous constatons une dichotomie entre des grandes valeurs de d_G et des faibles, par exemple au sein de la période 1946, la diminution correspondant à un gain considérable pour la population. Dans ce cas, les interactions lointaines correspondent mieux à un ajustement de la distance, tandis que la population suit plutôt une logique locale.

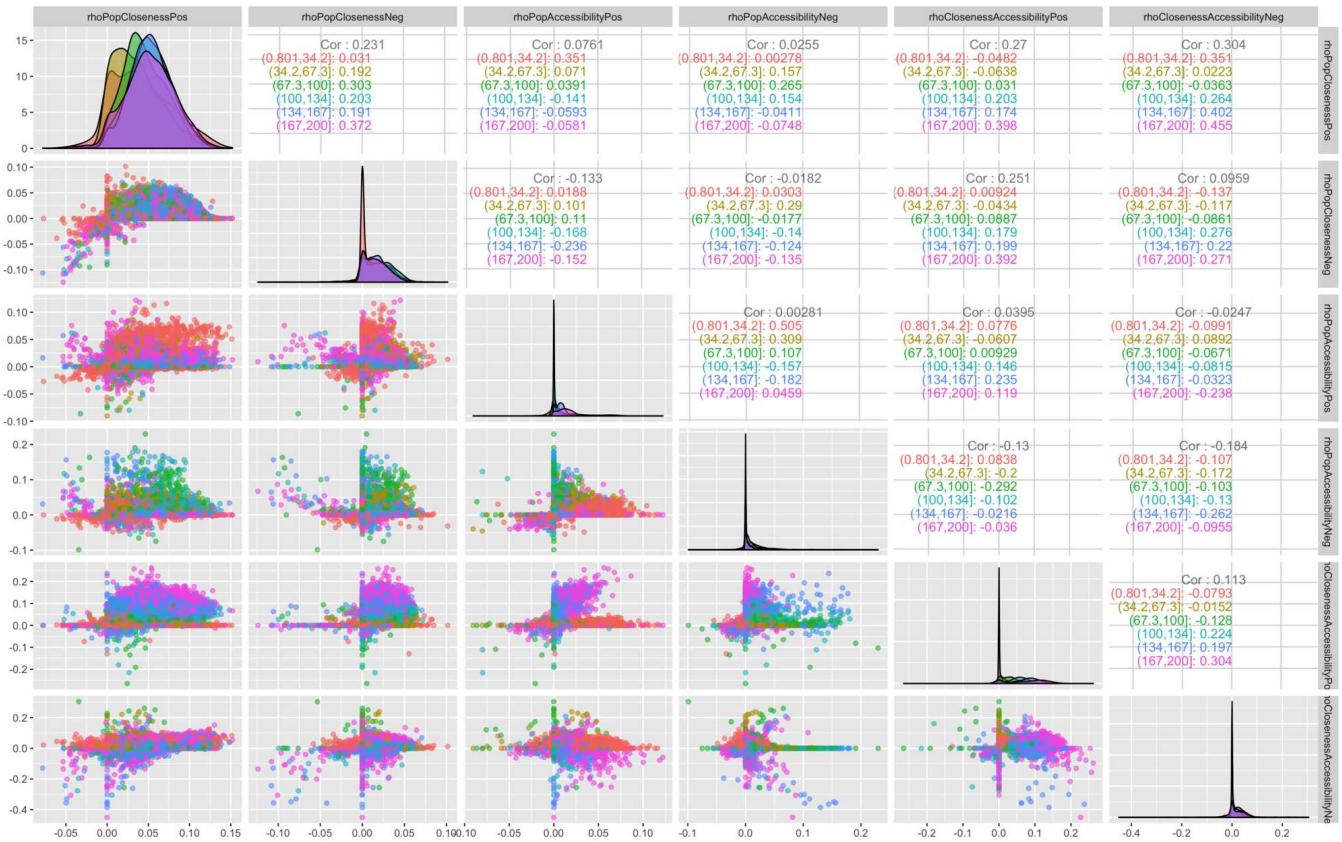


FIGURE 100: Application of the PSE algorithm

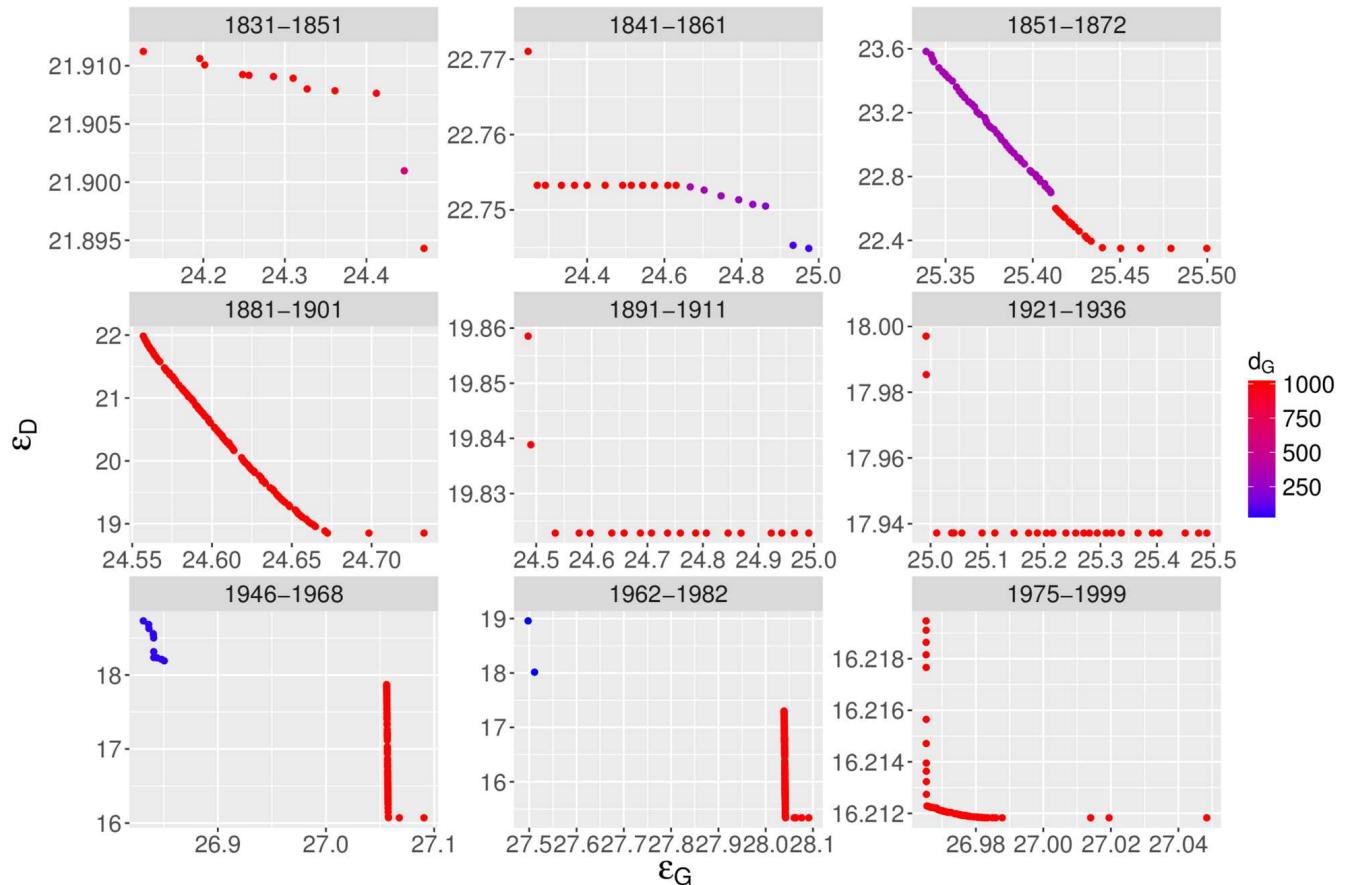


FIGURE 101: Pareto fronts for the bi-objective calibration with population and distance.

A.10 NETWORK GENERATION HEURISTICS

A.10.1 Slime mould model

We recall here the procedure of type *slime mould* to evolve the biological network, based on [Tero, Kobayashi, and Nakagaki, 2007]. The network is composed by nodes characterized by their pressure p_i and by links characterized by their length L_{ij} , their diameter D_{ij} , an impedance Z_{ij} and the flow traversing them ϕ_{ij} . The relation analogous to Ohm's law for links writes

$$\phi_{ij} = \frac{D_{ij}}{Z_{ij} \cdot L_{ij}} (p_i - p_j)$$

Furthermore, the conservation of flows at each node (Kirchoff's law) imposes

$$\sum_i \phi_{ij} = 0$$

for all j except the source and the sink, that we assume at indices j_+ and j_- , such that $\sum_i \phi_{ij_+} = I_0$ and $\sum_i \phi_{ij_-} = -I_0$ with I_0 initial flow parameter.

The combination of above constraints gives for all j

$$\sum_i \frac{D_{ij}}{Z_{ij} \cdot L_{ij}} (p_i - p_j) = \mathbb{1}_{j=j_+} I_0 - \mathbb{1}_{j=j_-} I_0$$

what simplifies into a matrix equation, by denoting $\mathbf{Z} = \left(\frac{\frac{D_{ij}}{Z_{ij} \cdot L_{ij}}}{\sum_i \frac{D_{ij}}{Z_{ij} \cdot L_{ij}}} \right)_{ij}$,

and also $\vec{k} = \frac{\mathbb{1}_{j=j_+} I_0 - \mathbb{1}_{j=j_-} I_0}{\sum_i \frac{D_{ij}}{Z_{ij} \cdot L_{ij}}}$ and $\vec{p} = p_i$, what simplifies into

$$(Id - \mathbf{Z}) \vec{p} = \vec{k}$$

The system admits a solution when $(Id - \mathbf{Z})$ is invertible. The space of invertible matrices being dense in $\mathcal{M}_n(\mathbb{R})$, by multilinearity of the determinant, an infinitesimal perturbation of the position of nodes allows to invert the matrix if it is indeed singular. We obtain thus the pressures p_i and as a consequence the flows ϕ_{ij} .

The evolution of the diameter D_{ij} between two equilibrium stages is a function of the flow at equilibrium, through the equation

$$D_{ij}(t+1) - D_{ij} = \delta t \left[\frac{\phi_{ij}(t)^\gamma}{1 + \phi_{ij}(t)^\gamma} - D_{ij}(t) \right]$$

TABLE 26: Morphological indicators for centers of classes for initial density grids.

Class	Moran I	Distance \bar{d}	Entropy \mathcal{E}	Hierarchy γ
1	0.23	0.66	0.76	0.62
2	0.47	0.50	0.75	0.53
3	0.21	0.42	0.57	0.65
4	0.24	0.75	0.90	0.87
5	0.15	0.76	0.84	0.72

We take to simplify $\gamma = 1.8$, following the configuration used by [Tero et al., 2010] for the generation of a network in a real configuration. We furthermore take $\delta t = 0.05$ and $I_0 = 10$.

The generation of a network can be achieved from an initial network, until reaching a convergence criteria, for example $\sum_{ij} \Delta D_{ij}(t) < \varepsilon$ with ε fixed threshold parameter. We will use this model with a criteria of a number of iterations, and proceed to an iteration to obtain final networks with a reasonable number of links.

A.10.2 Results

In the experiment exploring the distance to real networks, the initialization of the density is done according to 50 density grids classified into 5 morphological classes (10 grids per class). The Table 26 gives the composition of centers of classes in terms of morphological indicators. Classes can be interpreted the following way:

- Class 5: lowest Moran, high distance, hierarchy and entropy; numerous population centers that are localized and dispersed.
- Class 4: highest entropy and hierarchy; a small number of localized centers.
- Class 3: lowest distance and entropy; diffuse population.
- Class 2: highest Moran; one or a few centers with consequent size.
- Class 1: intermediate values for all indicators; a certain number of centers of intermediate size.

Topological spaces of networks generated in 7.1 can be conditioned to morphological classes for initial density distribution. This conditioning is shown in Fig. 102. We also give feasible spaces with real points. Classes 1 and 5 seem to be the ones for which being close to real points is the easiest, in terms of extreme points.

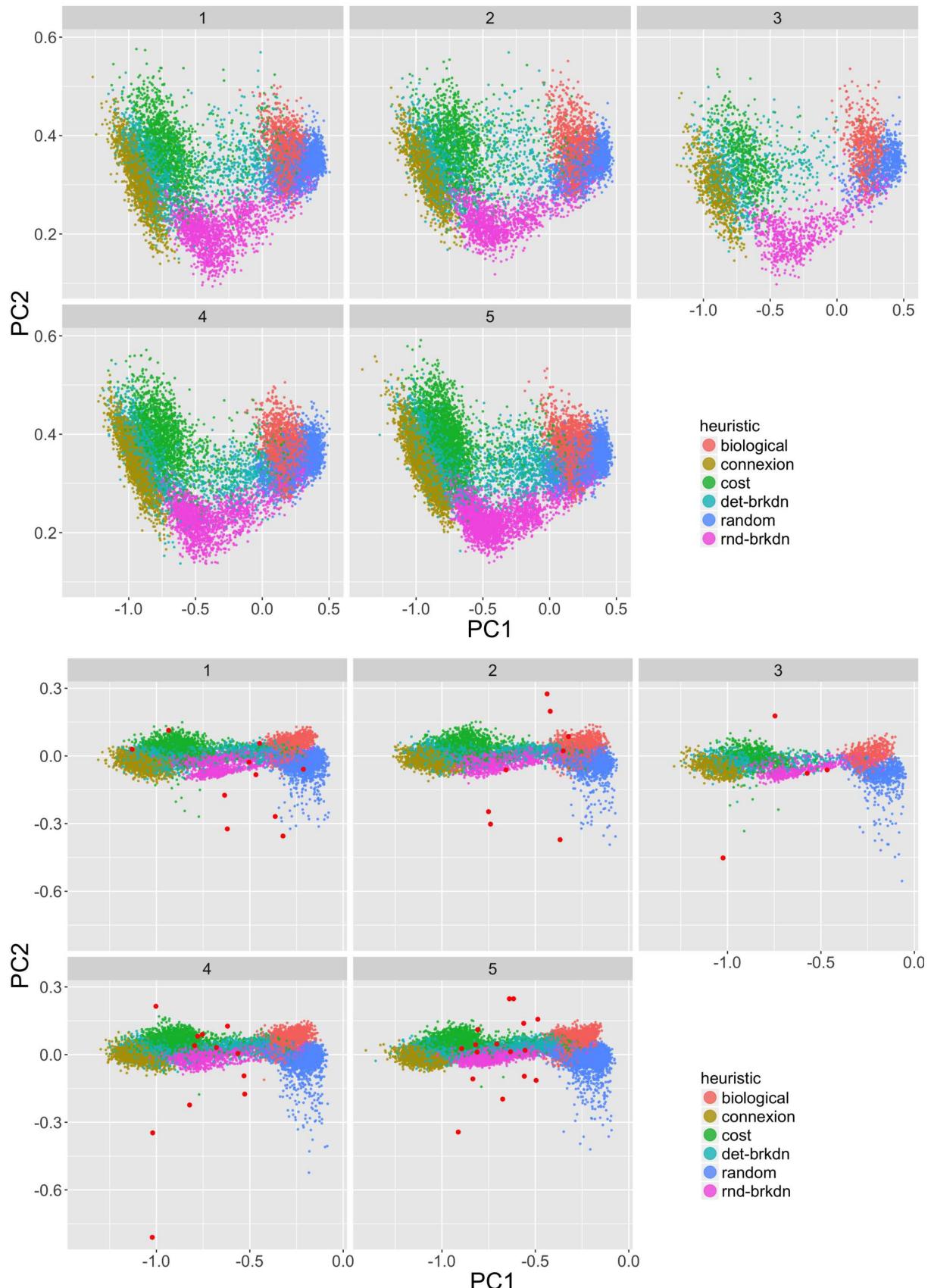


FIGURE 102: Conditioning of results to morphological classes for density. (Top) Topological feasible space for the different generation heuristics, conditioned to the morphological density class. (Bottom) Same plots with real points in red.

A.11 CO-EVOLUTION AT THE MESOSCOPIC SCALE

A.11.1 Calibration

In order to justify the aggregation of distances for indicators and for correlations, we have visually controlled the shape of Pareto fronts for these two objectives for around twenty simulated points. An example for two points is given in Fig. 103. It appears that these fronts are close to be not existing, i.e. that there almost exist a global optimum.

Let illustrate to what extent a linear aggregation with equal coefficients can be relevant in the case of a Pareto front which is close to being vertical/horizontal. The function

$$f_\alpha : x \mapsto \frac{1}{(x+1)^\alpha}$$

takes this form in a neighborhood of 0 when α becomes large. We then consider the two objectives $o_1(x) = x$ and $o_2(x) = f_\alpha(x)$, which can either be considered for a bi-objective minimization, or in the frame of a linear aggregation through the minimization of $o(x) = \beta x + (1-\beta)\frac{1}{(x+1)^\alpha}$. That latest is minimal in $x = \left(\frac{\beta}{\alpha(1-\beta)}\right)^{\frac{1}{\alpha+1}} - 1$, term which can be developed into

$$x = \frac{\ln(\beta(1-\beta))}{\alpha+1} + \frac{\ln \alpha}{\alpha+1} + o\left(\frac{1}{\alpha}\right)$$

Furthermore, let consider that in the frame of a bi-objective optimization, we take the compromise at which the variations of o_1 equalize the ones of o_2 , what is equivalent to take x such that $\frac{\partial f}{\partial x} = \frac{\partial f^{-1}}{\partial x}$. This equation leads to $\frac{x^{\frac{1}{\alpha}}}{x+1} = \frac{1}{\alpha^{1/\alpha+1}}$. We can then develop at the second order on each side to obtain

$$\frac{\ln x}{\alpha} = x \left[1 - 2 \frac{\ln \alpha}{\alpha+1} + o\left(\frac{1}{\alpha}\right) \right] - 2 \frac{\ln \alpha}{\alpha+1} + o\left(\frac{1}{\alpha}\right)$$

We indeed necessarily have $x \rightarrow_{\alpha \rightarrow \infty} 0$, since if $x \rightarrow K \neq 0$, we have a contradiction in the previous equation since $1/(1+K) \neq 0$. It implies that $\frac{\ln x}{\alpha} = o\left(\frac{1}{\alpha}\right)$, and thus that

$$x = 2 \frac{\ln \alpha}{\alpha+1} + o\left(\frac{1}{\alpha}\right)$$

In order thus to have the same order of magnitude for the solutions to the two approaches, we need to eliminate the term in $1/(\alpha+1)$ in the first, what is equivalent to take $\ln(\beta(1-\beta)) = 0$ and therefore $\beta = 1/2$.

Thus, there is equivalence of orders of magnitude in α for the two approaches if and only if $\beta = 1/2$. Given the shape of our Pareto fronts, we consider that the solution is analogous and consider thus the sum of the two distances.

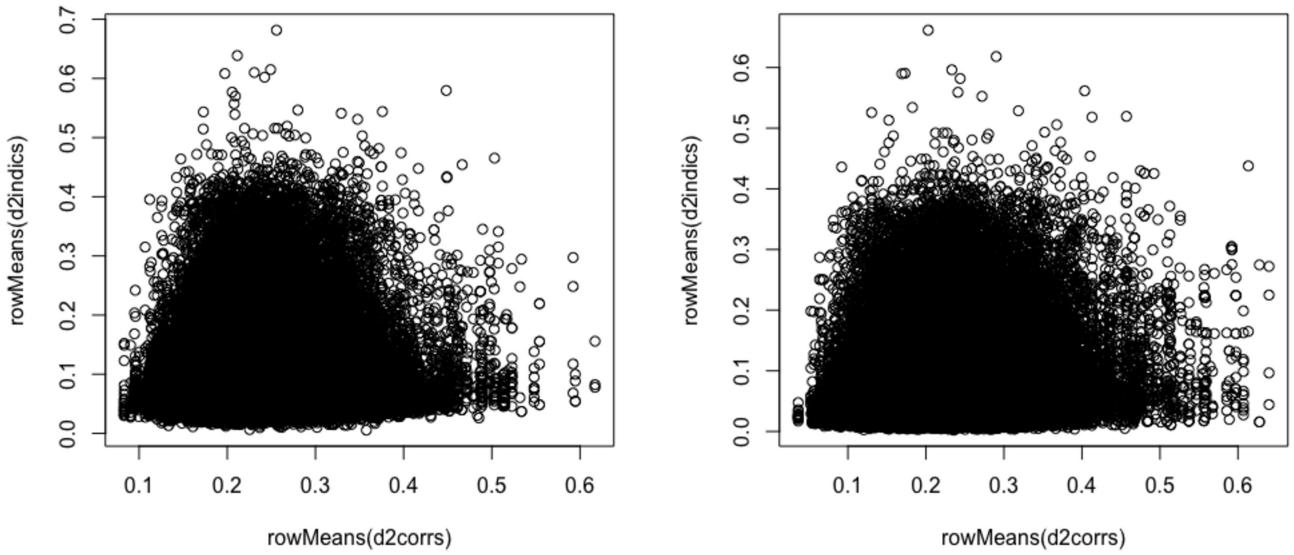


FIGURE 103: Example of Pareto fronts for the calibration at the first and second order. We give for two particular simulation points, the distances to indicators d_I^2 and the distances to correlations d_C^2 for all the real points.

A.12 TRANSPORTATION SYSTEM GOVERNANCE MODELING

A.12.1 Land-use model

Convergence

We study here the issue of the convergence in time of the distribution of activities, with a fixed infrastructure.

Let consider a very simple case: by taking $\lambda = 0$ the problem is made not spatial and by taking $\gamma_A = 1$ we achieve the decoupling between population and employments. By denoting $\beta' = \sum_j E_j \cdot \beta$ and $P_0 = \alpha \cdot \sum_i P_i$, the existence of a fixed point for populations is equivalent to the resolution of

$$P_i = P_0 \cdot \frac{\exp(\beta' \cdot P_i)}{\sum \exp(\beta' \cdot P_i)}$$

The function is indeed continuous in P_i and variation ranges for population are $[0, \sum_i P_i]$, it therefore admits a fixed point through the Brouwer fixed point theorem.

Indeed, in all generality, if we write

$$(\vec{P}(t+1), \vec{E}(t+1)) = f(\vec{P}(t), \vec{E}(t))$$

for arbitrary parameter values, the function f is also continuous in each component, and takes its values with a bounded closed interval

(employments being also limited) therefore a compact. The same way that [Leurent and Boujnah, 2014] establishes it for a model of traffic flows, we also have a fixed point in our case, what corresponds to an equilibrium point. The unicity is however not trivial and there is no reason for it to be a priori verified. We empirically verify the systematic convergence at fixed infrastructure (see below the exploration of the parameter space).

Exploration

We proceed to an exploration of the behavior of the land-use model alone, i.e. at fixed infrastructure, in order to understand the influence of parameters on the urban form. We fix $\alpha = 1$ here to study the model in an extreme case.

We follow the urban form indicators defined in 4.1, for the distribution of population and employments, in time and until the model has converged. We reduce the morphological space of the spatial distribution of actives in a principal plan, such that $PC_1 = -0.98 \cdot I - 0.13 \cdot E + 0.05\bar{d} - 0.13 \cdot \gamma$ and $PC_2 = -0.19 \cdot I + 0.57 \cdot E - 0.16\bar{d} + 0.77 \cdot \gamma$. The first component expresses a level of dispersion and the second a hierarchical aggregation.

The Fig. 104 gives temporal trajectories in the plan (PC_1, PC_2) for $\gamma_A = 0.9, \gamma_E = 0.6, v_0 = 6$, for different values of λ and β and also for different initial networks. We observe that increasing β has the tendency to make trajectories uniform. For $\beta = 1$, the shape of the network strongly conditions trajectories conjointly to λ : we switch for example from a decreasing dispersion and a u-shaped hierarchy to a stable dispersion and an increasing hierarchy for low values of λ , between no network and a spider network.

The Fig. 105 gives the value of PC_1 for the final configuration on all the space of explored parameters. We thus observe the variability of forms (here in terms of dispersion) as a function of all parameters: for example, for large β values, complex diagrams emerge. For low β values, we have a diagonal privileged for dispersion within concentrated configurations.

Finally, in order to understand the influence of parameters on total mobility within a complete trajectory, we study in Fig. 106 the cumulated variation of actives given by $\tilde{\Delta} = \sum_t \sum_k |\Delta A_k(t)|$. We see that high values of γ_A , for a high β , allow to minimize the total quantity of relocalization, which have a very low dependence in γ_E . It is therefore possible to optimize, even at fixed α , the total quantity of urban sprawl.

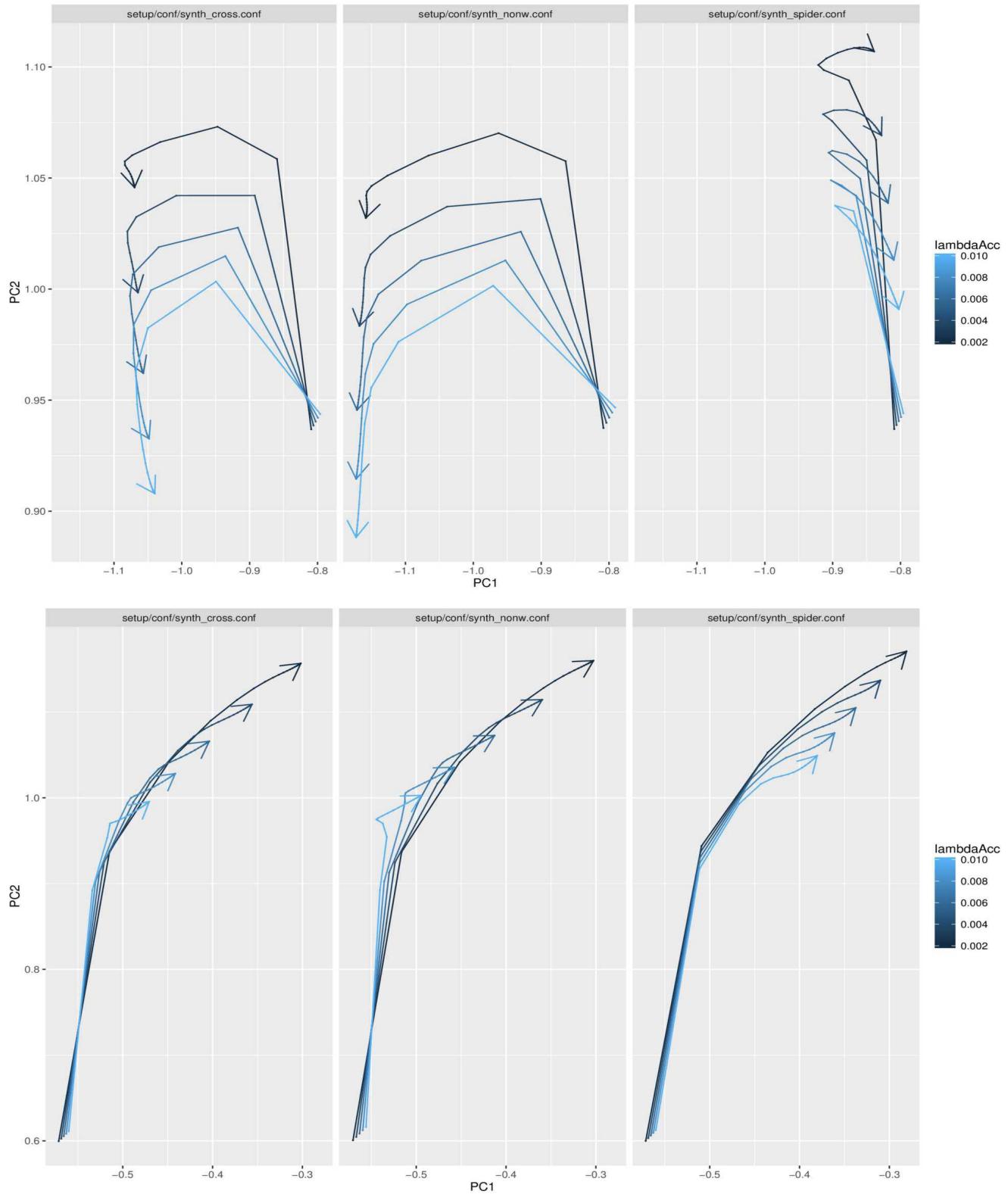


FIGURE 104: Morphological trajectories for the distribution of population. We fix here $\gamma_A = 0.9$ and $\gamma_E = 0.6$. (Top) Trajectories in the space (PC_1, PC_2) for $\beta = 1$, with variable λ (color), and for three different network configurations (columns): cross network, no network, cross network with branches (spider). (Bottom) Same plots, for $\beta = 2$.

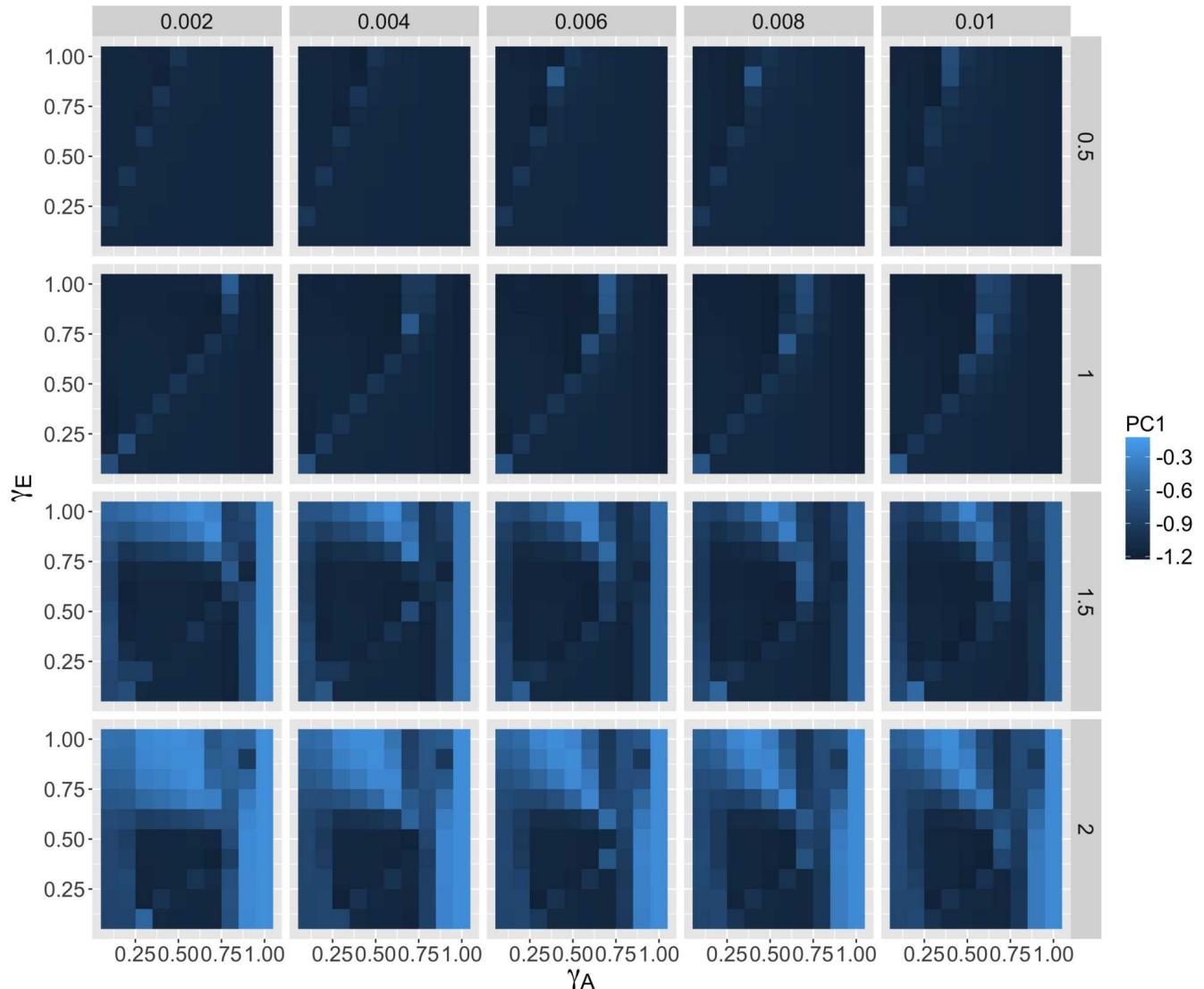


FIGURE 105: Sensitivity of the urban form. For the distribution of populations, without initial network, value of PC_1 as a function of (γ_A, γ_E) , with variable λ (columns) and variable β (rows).

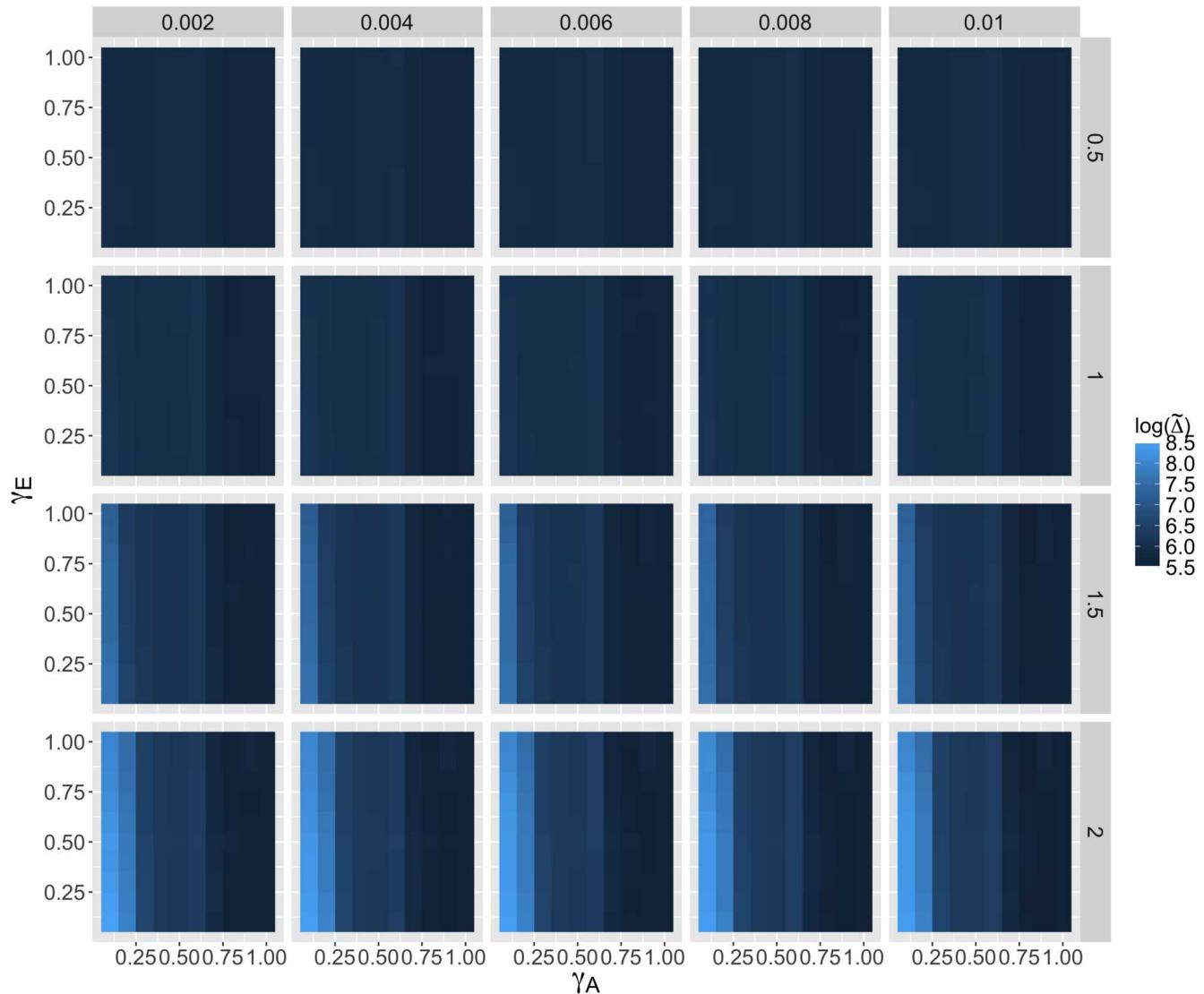


FIGURE 106: Cumulated variability of urban configurations. Value of $\ln \tilde{\Delta}$, without initial network, as a function of (γ_A, γ_E) , with variable λ (columns) and variable β (rows).

A.12.2 Transportation model

We did not take into account transportation flows in our implementation of the model, assuming that the constructed infrastructures have a sufficient capacity to be significantly not sensitive to congestion.

For the computation of flows between cells, the operation is the following: flows ϕ_{ij} are computed by solving on p_i, q_j through a fixed point method (Furness algorithm), of the system of gravity flows:

$$\begin{cases} \phi_{ij} = p_i q_j A_i E_j \exp(-\lambda_{tr} d_{ij}) \\ \sum_k \phi_{kj} = E_j \\ \sum_k \phi_{ik} = A_i \\ p_i = \frac{1}{\sum_k q_k E_k \exp(-\lambda_{tr} d_{ik})} \\ q_j = \frac{1}{\sum_k p_k A_k \exp(-\lambda_{tr} d_{kj})} \end{cases}$$

where λ_{tr} is a parameter giving the spatial reach of daily flows. The iteration of the last two equations rapidly converges starting from equal weights, by maintaining at each stage normalized weights.

In order to implement the stage of flows distribution within the network, when flows between cells are known, we should for example determine flows of the Static User Equilibrium with an appropriated algorithm. An assignment by shortest paths is implemented with the computation of flows in the model, but we deactivate this process in order to simplify the study of the model.

Congestion can be computed as a ratio to capacity, as c/c_{max} if c is the flow and c_{max} the capacity. The speed is obtained with a BPR function of the form $v(c) = v_0 \left(1 - \frac{c}{c_{max}}\right)^{\gamma_c}$. Our configuration os equivalent to assuming an infinite capacity $c_{max} = \infty$.

A.12.3 Probabilities to cooperate

The equilibrium assumption implies that conditional expectancies of each player are equal given their two choices, i.e. that

$$\mathbb{E}[U_i | S_i = C] = \mathbb{E}[U_i | S_i = NC]$$

It is indeed equivalent in that case to maximize $\mathbb{E}[U_i]$ as a function of p_i , since by conditioning we have $\mathbb{E}[U_i] = p_i \mathbb{E}[U_i | S_i = C] + (1 - p_i) \mathbb{E}[U_i | S_i = NC]$,

On a alors

$$\mathbb{E}[U_i | S_i = C] = p_{1-i} U_i(S_i = C, S_{1-i} = C) + (1 - p_{1-i}) U_i(S_i = C, S_{1-i} = NC)$$

et donc

$$\begin{aligned} p_{1-i} U_i(S_i = C, S_{1-i} = C) + (1 - p_{1-i}) U_i(S_i = C, S_{1-i} = NC) \\ = p_{1-i} U_i(S_i = NC, S_{1-i} = C) + (1 - p_{1-i}) U_i(S_i = NC, S_{1-i} = NC) \end{aligned}$$

ce qui donne

$$p_{1-i} = -\frac{U_i(C, NC) - U_i(NC, NC)}{(U_i(C, C) - U_i(NC, C)) - (U_i(C, NC) - U_i(NC, NC))}$$

En substituant les expressions des utilités à partir de la matrice de gain, on obtient l'expression de p_i en fonction du coût de collaboration J et de la différence des différentiels d'accessibilité.

Discrete choice coordination

Pour déterminer la probabilité de coopération dans le cas des choix discrets, il s'agit de résoudre $f(p_i) = 0$ avec

$$f(x) = \frac{1}{1 + \exp\left[-\beta_{DC} \frac{\Delta_i}{1 + \exp(-\beta_{DC}(x\Delta_{1-i} - J))} - J\right]} - x$$

où nous avons noté $\Delta_i = \Delta X_i(Z_C^*) - \Delta X_{\bar{i}}(Z_{\bar{i}}^*)$.

On a immédiatement $f(0) > 0$ et $f(1) < 0$ et f est continue, il existe donc toujours une solution $x \in [0, 1]$ par le théorème des valeurs intermédiaires.

Concernant l'unicité, il est possible de la montrer sous certaines conditions. Un calcul de $\frac{\partial f}{\partial x}$ donne

$$\frac{\partial f}{\partial x} = 2(\cosh u(x) - 1) + \beta^2 \Delta_i \Delta_{1-i} \frac{\exp(-\beta_{DC}(x\Delta_{1-i} - J))}{(1 + \exp(-\beta_{DC}(x\Delta_{1-i} - J)))^2}$$

où $u(x) = -\beta_{DC} \left(\frac{\Delta_i}{1 + \exp(-\beta_{DC}(x\Delta_{1-i} - J))} - J \right)$.

Comme $\cosh u \geq 1$, on a $\frac{\partial f}{\partial x} > 0$ si $\Delta_i \Delta_{1-i} > 0$. La fonction est dans ce cas strictement croissante et on a une unique solution.

En pratique, la solution est déterminée par algorithme de Brent, avec les bornes $[0, 1]$ et une tolérance de 0.01.

A.12.4 *Implementation details*

DISTANCE MATRIX Distance via network are updated in a dynamical programming fashion for efficiency purposes (because of the numerous network updates), the following way :

1. Euclidian distance matrix $d(i, j)$ is computed analytically

NETWORK GROWTH Les infrastructures potentielles, au nombre de N_I lors de la recherche heuristique d'une infrastructure optimale, sont tirées aléatoirement parmi l'ensemble des infrastructures possibles ayant une extrémité au centre d'une cellule. Si l'extrémité est à une distance inférieure à un seuil θ_I d'un lien déjà existant du réseau, celle-ci est remplacée par sa projection sur le lien correspondant. Il s'agit de l'étape d'accrochage permettant d'obtenir un réseau de forme raisonnable localement. En cohérence avec la représentation raster du réseau, nous prenons $\theta_I = 1$, ce qui correspond à la taille d'une cellule.

A.12.5 *Setup*

Synthetic setup

This section describes the setup in the case of synthetic configurations of MCR.

Initial distribution of Actives and Employments is done around governance centers at positions \vec{x}_i using exponential kernels by

$$A(\vec{x}) = A_{\max} \cdot \exp\left(\frac{\|\vec{x} - \vec{x}_i\|}{r_A}\right); E(\vec{x}) = E_{\max} \cdot \exp\left(\frac{\|\vec{x} - \vec{x}_i\|}{r_E}\right)$$

Setup on a real configuration

Nous montrons en Fig. 107 la population et les réseaux sur lesquels les expériences sur données réelles sont menées : à usage du sol fixe,

- une expérience sans réseau initial, et avec pour réseau cible de calibration le réseau de 2010 ;
- une expérience avec le réseau initial de 2010, et pour réseau cible le réseau planifié.

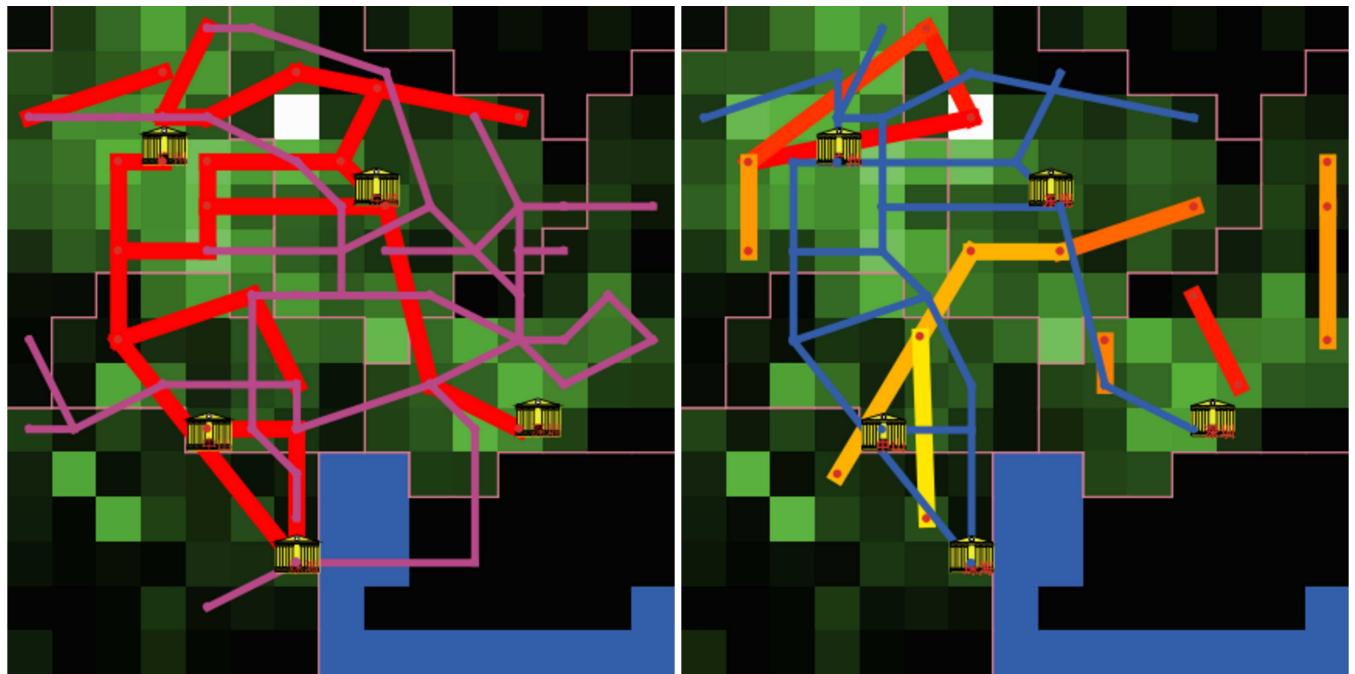


FIGURE 107: Setup on real data used for model application.

B

DÉVELOPPEMENTS MÉTHODOLOGIQUES

Cette annexe rassemble différents développements méthodologiques qui sont utilisés indirectement, ou permettre de creuser des questions liées mais non centrales à notre fil principal.

Les trois premières sections traitent des questions se posant particulièrement lors de l'étude des systèmes urbains ou territoriaux.

1. Un lien formel entre différents modèles stochastiques de croissance urbaine permet de poser un cadre général pour ce genre d'approche, et d'illustrer le lien implicite entre notre approche mesoscopique et notre approche macroscopique.
2. La sensibilité des lois d'échelles à la définition de la ville est étudiée analytiquement pour un modèle simple de système urbain. Cette perspective renforce la méthodologie d'analyse de sensibilité des modèles à la configuration spatiale introduite en 3.1.
3. Le contexte bibliographique et formel de la notion de données synthétiques permet également de situer celle-ci.

Nous développons ensuite des cadres méthodologiques généraux liés à l'étude des systèmes.

5. Dans le cadre de systèmes incluant des optimisation multi-attributs, une méthode d'analyse de sensibilité à la structure des données, est introduite. Elle n'est pas directement appliquée dans notre travail mais suggère des pistes pour l'application des modèles mesoscopiques de morphogenèse, puisque ceux-ci se basent sur une telle optimisation par les agents.
6. Un cadre général pour la modélisation des systèmes complexes socio-techniques, pose les premières bases d'une part d'une formalisation du *perspectivisme appliqué* mais également de la formalisation du cadre de connaissances suggérée en 8.3.

Enfin, le dernier développement concerne les méthodes d'épistémologie quantitative.

5. Les détails techniques de la méthode utilisée en 2.2 sont développés dans le cadre d'application au corpus de la revue *Cybergeo*. Les considérations sont fondamentalement méthodologiques, et doivent être également mises en perspective avec l'article thématique companion que nous adaptions en C.4.

B.1 AN UNIFIED FRAMEWORK FOR STOCHASTIC MODELS OF URBAN GROWTH

Urban growth modeling fall in the case of tentatives to find self-consistent rules reproducing dynamics of an urban system, and thus in our logic of system morphogenesis. We examine here methodological issues linked to different frameworks of urban growth.

B.1.1 *Introduction*

Various stochastic models aiming to reproduce population patterns on large temporal and spatial scales (city systems) have been discussed across various fields of the literature, from economics to geography, including models proposed by physicists. We propose here a general framework that allows to include different famous models (in particular Gibrat, Simon and Preferential Attachment model) within an unified vision. Furthermore, bridges between models lead to the possible transfer of analytical results to some models that are not directly tractable.

Seminal models of urban growth are Simon [Simon, 1955] (later generalized as e.g. [Haran and Vining, 1973]) and Gibrat models. Many examples of variants and extensions can be given across disciplines. [Benguigui and Blumenfeld-Lieberthal, 2007] give an equation-based dynamical model, whereas [Gabaix, 1999] shows that the Gibrat model produces Zipf's law in a stationary state. [Gabaix and Ioannides, 2004] reviews urban growth approaches in economics. A model adapted from evolutive urban theory is described in [Favaro and Pumain, 2011] and extends the Gibrat model by adding propagation of innovation between cities. The question of empirical scales at which it is consistent to study urban growth was also tackled in the particular case of France [Bretagnolle, Paulus, and Pumain, 2002], which shows that long time scales (more than a few decade) are appropriate to study dynamics of urban systems at a small spatial scale.

B.1.2 *Framework*

What we propose as a framework can be understood as a meta-model, i.e. a modular general modeling process within each model can be understood as a limit case or as a specific case of another model. More simply it should be a diagram of formal relations between models.

The ontological aspect is also tackled by embedding the diagram into an ontological state space (which discretization corresponds to the "bricks" of the incremental construction of [Cottineau, Chapron, and Reuillon, 2015]). It constructs a sort of model classification or model-

C : (Florent) à ce stade
on ne sait pas si tu vas
faire 1 ou N modèles,
c'est un choix qu'il te
faut défendre avant d'en
arriver là

C : (Florent) PAS UTILE
ICI JE PENSE

B.1.3 Derivations

Generalization of Preferential Attachment

[Yamasaki et al., 2006] give a generalization of the classical Preferential Attachment Network Growth model, as a birth and death model with evolving entities. More precisely, network units gain and lose population (equivalent to links connexions) at fixed probabilities, and new unit can be created at a fixed rate.

Link between Gibrat and Preferential Attachment Models

Let consider a strictly positive growth Gibrat model given by $P_i(t) = R_i(t) \cdot P_i(t-1)$ with $R_i(t) > 1$, $\mu_i(t) = \mathbb{E}[R_i(t)]$ and $\sigma_i(t) = \mathbb{E}[R_i(t)^2]$. On the other hand, we take a simple preferential attachment, with fixed attachment probability $\lambda \in [0, 1]$ and new arrivants number $m > 0$. We derive that Gibrat model can be statistically equivalent to a limit of the preferential attachment model, assuming that the moment-generating function of $R_i(t)$ exists. Classical distributions that could be used in that case, e.g. log-normal distribution, are entirely defined by two first moments, making this assumption reasonable.

Lemme 1 *The limit of a Preferential Attachment model when $\lambda \ll 1$ is a linear-growth Gibrat model, with limit parameters $\mu_i(t) = 1 + \frac{\lambda}{m \cdot (t-1)}$.*

Preuve Starting with first moment, we denote $\bar{P}_i(t) = \mathbb{E}[P_i(t)]$. Independence of Gibrat growth rate yields directly $\bar{P}_i(t) = \mathbb{E}[R_i(t)] \cdot \bar{P}_i(t-1)$. Starting for the preferential attachment model, we have $\bar{P}_i(t) = \mathbb{E}[P_i(t)] = \sum_{k=0}^{+\infty} k \mathbb{P}[P_i(t) = k]$. But

$$\{P_i(t) = k\} = \bigcup_{\delta=0}^{\infty} (\{P_i(t-1) = k - \delta\} \cap \{P_i \leftarrow P_i + 1\}^{\delta})$$

where the second event corresponds to city i being increased δ times between $t-1$ and t (note that events are empty for $\delta \geq k$). Thus, being careful on the conditional nature of preferential attachment formulation, stating that $\mathbb{P}[\{P_i \leftarrow P_i + 1\} | P_i(t-1) = p] = \lambda \cdot \frac{p}{\bar{P}(t-1)}$ (total population $P(t)$ assumed deterministic), we obtain

$$\begin{aligned} \mathbb{P}[\{P_i \leftarrow P_i + 1\}] &= \sum_p \mathbb{P}[\{P_i \leftarrow P_i + 1\} | P_i(t-1) = p] \cdot \mathbb{P}[P_i(t-1) = p] \\ &= \sum_p \lambda \cdot \frac{p}{\bar{P}(t-1)} \mathbb{P}[P_i(t-1) = p] = \lambda \cdot \frac{\bar{P}_i(t-1)}{\bar{P}(t-1)} \end{aligned}$$

It gives therefore, knowing that $P(t-1) = P_0 + m \cdot (t-1)$ and denoting $q = \lambda \cdot \frac{\bar{P}_i(t-1)}{P_0 + m \cdot (t-1)}$

$$\begin{aligned}
\bar{P}_i(t) &= \sum_{k=0}^{\infty} \sum_{\delta=0}^{\infty} k \cdot \left(\lambda \cdot \frac{\bar{P}_i(t-1)}{P_0 + m \cdot (t-1)} \right)^{\delta} \cdot \mathbb{P}[P_i(t-1) = k - \delta] \\
&= \sum_{\delta'=0}^{\infty} \sum_{k'=0}^{\infty} (k' + \delta') \cdot q^{\delta'} \cdot \mathbb{P}[P_i(t-1) = k'] \\
&= \sum_{\delta'=0}^{\infty} q^{\delta'} \cdot (\delta' + \bar{P}_i(t-1)) = \frac{q}{(1-q)^2} + \frac{\bar{P}_i(t-1)}{(1-q)} \\
&= \frac{\bar{P}_i(t-1)}{1-q} \left[1 + \frac{1}{\bar{P}_i(t-1)} \frac{q}{(1-q)} \right]
\end{aligned}$$

As it is not expected to have $\bar{P}_i(t) \ll P(t)$ (fat tail distributions), a limit can be taken only through λ . Taking $\lambda \ll 1$ yields, as $0 < \bar{P}_i(t)/P(t) < 1$, that $q = \lambda \cdot \frac{\bar{P}_i(t-1)}{P_0 + m \cdot (t-1)} \ll 1$ and thus we can expand in first order of q , what gives $\bar{P}_i(t) = \bar{P}_i(t-1) \cdot \left[1 + \left(1 + \frac{1}{\bar{P}_i(t-1)} \right) q + o(q) \right]$

$$\bar{P}_i(t) = \bar{P}_i(t-1) \cdot \left[1 + \left(1 + \frac{1}{\bar{P}_i(t-1)} \right) q + o(q) \right]$$

et donc

$$\bar{P}_i(t) \simeq \left[1 + \frac{\lambda}{P_0 + m \cdot (t-1)} \right] \cdot \bar{P}_i(t-1)$$

It means that this limit is equivalent in expectancy to a Gibrat model with $\mu_i(t) = \mu(t) = 1 + \frac{\lambda}{P_0 + m \cdot (t-1)}$.

For the second moment, we can do an analog computation. We have still

$$\mathbb{E}[P_i(t)^2] = \mathbb{E}[R_i(t)^2] \cdot \mathbb{E}[P_i(t-1)^2]$$

and

$$\mathbb{E}[P_i(t)^2] = \sum_{k=0}^{+\infty} k^2 \mathbb{P}[P_i(t) = k]$$

We obtain the same way

$$\begin{aligned}
\mathbb{E}[P_i(t)^2] &= \sum_{\delta'=0}^{\infty} \sum_{k'=0}^{\infty} (k' + \delta')^2 \cdot q^{\delta'} \cdot \mathbb{P}[P_i(t-1) = k'] \\
&= \sum_{\delta'=0}^{\infty} q^{\delta'} \cdot \left(\mathbb{E}[P_i(t-1)^2] + 2\delta' \bar{P}_i(t-1) + \delta'^2 \right) \\
&= \frac{\mathbb{E}[P_i(t-1)^2]}{1-q} + \frac{2q\bar{P}_i(t-1)}{(1-q)^2} + \frac{q(q+1)}{(1-q)^3} \\
&= \frac{\mathbb{E}[P_i(t-1)^2]}{1-q} \left[1 + \frac{q}{\mathbb{E}[P_i(t-1)^2]} \left(\frac{2\bar{P}_i(t-1)}{1-q} + \frac{(1+q)}{(1-q)^2} \right) \right]
\end{aligned}$$

We have therefore an equivalence between the Gibrat model as a continuous formulation of a Preferential Attachment (or Simon model) in the limit given before. ■

Link between Simon and Preferential Attachment

A rewriting of Simon model yields a particular case of the generalized preferential attachment, in particular by vanishing death probability.

Link between Favaro-Pumain and Gibrat

[Favaro and Pumain, 2011] generalizes Gibrat models with innovation propagation dynamics. Theoretically, a process-based model equivalent to the Favaro-Pumain should then fill the missing case in model classification at the corresponding discretization. Simpop models do not fill that case as they stay at the scale of city systems, as for Marius models [Cottineau, 2014]. These must also have their counterparts in discrete microscopic formulation.

★ ★

★

B.2 SENSITIVITY OF URBAN SCALING LAWS TO SPATIAL EXTENT

At the center of evolutive urban theory are hierarchy and associated scaling laws. We develop here a brief methodological investigation on the sensitivity of scaling laws to city definition.

Scaling laws have been shown to be universal of urban systems at many scales and for many indicators. Recent studies question however the consistence of scaling exponents determination, as their value can vary significantly depending on thresholds used to define urban entities on which quantities are integrated, even crossing the qualitative border of linear scaling, from infra-linear to supra-linear scaling. We use a simple theoretical model of spatial distribution of densities and urban functions to show analytically that such behavior can be derived as a consequence of the type of spatial distribution and the method used. Numerical simulation confirm the theoretical results and reveals that results are reasonably independent of spatial kernel used to distribute density.

Scaling laws for urban systems, starting from the well-known rank-size Zipf's law for city size distribution [Gabaix, 1999], have been shown to be a recurrent feature of urban systems, at many scales and for many types of indicators. They reside in the empirical constatation that indicators computed on elements of an urban system, that can be cities for system of cities, but also smaller entities at a smaller scale, do fit relatively well a power-law distribution as a function of entity size, i.e. that for entity i with population P_i , we have for an integrated quantity A_i , the relation $A_i \simeq A_0 \cdot \left(\frac{P_i}{P_0}\right)^\alpha$. Scaling exponent α can be smaller or greater than 1, leading to infra or supralinear effects. Various thematic interpretation of this phenomena have been proposed, typically under the form of processes analysis. The economic literature has produced abundant work on the subject (see [Gabaix and Ioannides, 2004] for a review), but that are generally weakly spatial, thus of poor interest to our approach that deals precisely with spatial organization. Simple economic rules such as energetic equilibria can lead to simple power-laws [Bettencourt, Lobo, and West, 2008] but are difficult to fit empirically. A interesting proposition by PUMAIN is that they are intrinsically due to the evolutionary character of city systems, where complex emergent interaction between cities generate such global distributions [Pumain et al., 2006]. Although a tempting parallel can be done with self-organizing biological systems, PUMAIN insists on the fact that the ergodicity assumption for such systems is not reasonable in the case of geographical systems and that the analogy can difficultly be exploited [Pumain, 2012b]. Other explanations have been proposed at other scales, such as the urban growth model at the mesoscopic scale (city scale) given in [Louf and Barthelemy, 2014b] that shows that the congestion within transportation networks may be one reason for city shapes and corresponding scaling laws.

Note that “classic” urban growth models such as Gibrat’s model do provide first order approximation of scaling systems, but that interactions between agents have to be incorporated into the model to obtain better fit on real data, such as the Favaro-Pumain model for innovation cycles propagation proposed in [Favaro and Pumain, 2011], that generalize a Gibrat model for French cities with an ontology similar to Simpop models.

However, the incautious application of scaling exponents computations was recently pointed as misleading in most cases, as [Arcaute et al., 2015] shows the variability of computed exponents to the parameters defining urban areas, such as density thresholds. [] studies empirically for France the influence of 3 parameters playing a role in city definition, that are a density threshold θ to delimitate boundaries of an urban area, a number of commuters threshold θ_c that is the proportion of commuters going to core area over which the unity is considered belonging to the area, and a cut-off parameter P_c under which entities are not taken into account for the linear regression providing the scaling exponent. Remarkable results are that exponents can significantly vary and move from infra-linear to supra-linear when threshold varies. A systematic exploration of parameter space produces phase diagrams of exponents for various quantities. One question raising immediately is how these variation can be explained by the features of spatial distribution of variables. Do they result from intrinsic mechanisms present in the system or can they be explained more simply by the fact that the system is particularly spatialized ? We prove on a toy analytical model that even simple distributions can lead to such significant variations in the exponents, along one dimension of parameters (density threshold), directing the response towards the second explanation.

Nous dérivons par la suite l’expression de la variation des exposants d’échelle dans le cas simple d’une distribution en mixture d’exponentielle.

We formalize the simple theoretical context in which we will derive the sensitivity of scaling to city definition. Let consider a polycentric city system, which spatial density distributions can be reasonably constructed as the superposition of monocentric fast-decreasing spatial kernels, such as an exponential mixture model [Anas, Arnott, and Small, 1998]. Taking a geographical space as \mathbb{R}^2 , we take for any $\vec{x} \in \mathbb{R}^2$ the density of population as

$$d(\vec{x}) = \sum_{i=1}^N d_i(\vec{x}) = \sum_{i=1}^N d_i^0 \cdot \exp\left(\frac{-\|\vec{x} - \vec{x}_i\|}{r_i}\right)$$

where r_i are spread parameters of kernels, d_i^0 densities at origins, \vec{x}_i positions of centers. We furthermore assume the following constraints :

1. To simplify, cities are monocentric, in the sense that for all $i \neq j$, we have $\|\vec{x}_i - \vec{x}_j\| \gg r_i$.
2. It allows to impose structural scaling in the urban system by the simple constraint on city populations P_i . One can compute by integration that $P_i = 2\pi d_i^0 r_i^2$, what gives by injection into the scaling hypothesis $\ln P_i = \ln P_{\max} - \alpha \ln i$, the following relation between parameters : $\ln [d_i^0 r_i^2] = K' - \alpha \ln i$.

To study scaling relations, we consider a random scalar spatial variable $a(\vec{x})$ representing one aspect of the city, that can be everything but has the dimension of a spatial density, such that the indicator $A(D) = \mathbb{E}[\iint_D a(\vec{x}) d\vec{x}]$ represents the expected quantity of a in area D . We make the assumption that $a \in \{0; 1\}$ ("counting" indicator) and that its law is given by $\mathbb{P}[a(\vec{x}) = 1] = f(d(\vec{x}))$. Following the empirical work done in [Cottineau, 2015], the integrated indicator on city i as a function of θ is given by

$$A_i(\theta) = A(D(\vec{x}_i, \theta))$$

where $D(\vec{x}_i, \theta)$ is the area centered in \vec{x}_i where $d(\vec{x}) > \theta$. Assumption 1 ensures that the area are roughly disjoint circles. We take furthermore a simple amenity such that it follows a local scaling law in the sense that $f(d) = \lambda \cdot d^\beta$. It seems a reasonable assumption since it was shown that many urban variable follow a fractal behavior at the intra-urban scale [Keersmaecker, Frankhauser, and Thomas, 2003] and that it implies a power-law distribution [Chen, 2010]. We make the additional assumption that $r_i = r_0$ does not depend on i , what is reasonable if the urban system is considered from a large scale. The estimated scaling exponent $\alpha(\theta)$ is then the result of the log-regression of $(A_i(\theta))_i$ against $(P_i(\theta))_i$ where $P_i(\theta) = \iint_{D(\vec{x}_i, \theta)} d$.

B.2.1 Analytical Derivation of Sensitivity

With above notations, let derive the expression of estimated exponent for quantity a as a function of density threshold parameter θ . The quantity computed for a given city i is, thanks to the monocentric assumption and in a spatial range and a range for θ such that $\theta \gg \sum_{j \neq i} d_j(\vec{x})$, allowing to approximate $d(\vec{x}) \simeq d_i(\vec{x})$ on $D(\vec{x}_i, \theta)$, is computed by

$$\begin{aligned} A_i(\theta) &= \lambda \cdot \iint_{D(\vec{x}_i, \theta)} d^\beta = 2\pi \lambda d_i^0 \beta \int_{r=0}^{r_0 \ln \frac{d^0}{\theta}} r \exp\left(-\frac{r\beta}{r_0}\right) dr \\ &= \frac{2\pi d_i^0 \beta r_0^2}{\beta^2} \left[1 + \beta \ln \frac{\theta}{d_i^0} \left(\frac{\theta}{d_i^0}\right)^\beta - \left(\frac{\theta}{d_i^0}\right)^\beta \right] \end{aligned}$$

We obtain in a similar way the expression of $P_i(\theta)$

$$P_i(\theta) = 2\pi d_i^0 r_0^2 \left[1 + \ln \left[\frac{\theta}{d_i^0} \right] \frac{\theta}{d_i^0} - \frac{\theta}{d_i^0} \right]$$

The Ordinary-Least-Square estimation, solving the problem $\inf_{\alpha, C} \|(\ln A_i(\theta) - C - \alpha \ln P_i(\theta))_i\|^2$, gives the value $\alpha(\theta) = \frac{\text{Cov}[(\ln A_i(\theta))_i, (\ln P_i(\theta))_i]}{\text{Var}[(\ln P_i(\theta))_i]}$. As we work on city boundaries, threshold is expected to be significantly smaller than center density, i.e. $\theta/d_i^0 \ll 1$. We can develop the expression in the first order of θ/d_i^0 and use the global scaling law for city sizes, what gives

$$\ln A_i(\theta) \simeq K_A - \alpha \ln i + (\beta - 1) \ln d_i^0 + \beta \ln \frac{\theta}{d_i^0} \left(\frac{\theta}{d_i^0} \right)^\beta$$

and

$$\ln P_i(\theta) = K_P - \alpha \ln i + \ln \left[\frac{\theta}{d_i^0} \right] \frac{\theta}{d_i^0}$$

Developing the covariance and variance gives finally an expression of the scaling exponent as a function of θ , where k_j, k_j' are constants obtained in the development :

$$\alpha(\theta) = \frac{k_0 + k_1 \theta + k_2 \theta^\beta + k_3 \theta^{\beta+1} + k_4 \theta \ln \theta + k_5 \theta^\beta \ln \theta + k_6 \theta^\beta (\ln \theta)^2 + k_7 \theta^{\beta+1} (\ln \theta)^2 + k_8 \theta^{\beta+1} \ln \theta}{k'_0 + k'_1 \ln \theta + k'_2 \theta \ln \theta + k'_3 \theta^2 + k'_4 \theta^2 \ln \theta + k'_5 \theta^2 (\ln \theta)^2}$$

This rational fraction in θ and $\ln \theta$ predicts the evolution of the scaling exponent when the threshold varies.

B.3 GENERATION OF CORRELATED SYNTHETIC DATA

Cette section correspond à l'introduction et la formalisation de [Raimbault, 2016b].



Generation of hybrid synthetic data resembling real data to some criteria is an important methodological and thematic issue in most disciplines which study complex systems. Interdependencies between constituting elements, materialized within respective relations, lead to the emergence of macroscopic patterns. Being able to control the dependance structure and level within a synthetic dataset is thus a source of knowledge on system mechanisms. We propose a methodology consisting in the generation of synthetic datasets on which correlation structure is controlled. The method is applied in a first example on financial time-series and allows to understand the role of interferences between components at different scales on performances of a predictive model. A second application on a geographical system is then proposed, in which the weak coupling between a population density model and a network morphogenesis model allows to simulate territorial configurations. The calibration on morphological objective on european data and intensive model exploration unveils a large spectrum of feasible correlations between morphological and network measures. We demonstrate therein the flexibility of our method and the variety of possible applications.

B.3.1 Context

The use of synthetic data, in the sense of statistical populations generated randomly under constraints of patterns proximity to the studied system, is a widely used methodology, and more particularly in disciplines related to complex systems such as therapeutic evaluation [Abadie, Diamond, and Hainmueller, 2010], territorial science [Moeckel, Spiekermann, and Wegener, 2003; Pritchard and Miller, 2009], machine learning [Bolón-Canedo, Sánchez-Marcano, and Alonso-Betanzos, 2013] or bio-informatics [Bulcke et al., 2006]. It can consist in data de-segregation by creation of a microscopic population with fixed macroscopic properties, or in the creation of new populations at the same scale than a given sample, with criteria of proximity to the real sample. These criteria will depend on expected applications and can for example vary from a restrictive statistical fit on given indicators, to

weaker assumptions of similarity in aggregated patterns. In the case of chaotic systems, or systems where emergence plays a strong role, a microscopic property does not directly imply given macroscopic patterns, which reproduction is indeed one aim of modeling and simulation practices in complexity science. With the rise of new computational paradigms [Arthur, 2015], data (simulated, measured or hybrid) shape our understanding of complex systems. Methodological tools for data-mining and modeling and simulation (including the generation of synthetic data) are therefore crucial to be developed.

Whereas first order (in the sense of distribution moments) is generally well used, it is not systematic nor simple to control generated data structure at second order, i.e. covariance structure between generated variables. Some specific examples can be found, such as in [Ye, 2011] where the sensitivity of discrete choices models to the distributions of inputs and to their dependance structure is examined. It is also possible to interpret complex networks generative models [Newman, 2003] as the production of an interdependence structure for a system, contained within link topology. We introduce here a generic method taking into account dependance structure for the generation of synthetic datasets, more precisely with the mean of controlled correlation matrices.

Domain-specific methods aforementioned are too broad to be summarized into a same formalism. We propose a framework as generic as possible, centered on the control of correlations structure in synthetic data.

B.3.2 Formalization

Let \vec{X}_I a multidimensional stochastic process (that can be indexed e.g. with time in the case of time-series, but also space, or discrete set abstract indexation). We assume given a real dataset $\mathbf{X} = (X_{i,j})$, interpreted as a set of realizations of the stochastic process. We propose to generate a statistical population $\tilde{\mathbf{X}} = \tilde{X}_{i,j}$ such that

1. a given criteria of proximity to data is verified, i.e. given a precision ε and an indicator f , we have $\|f(\mathbf{X}) - f(\tilde{\mathbf{X}})\| < \varepsilon$
2. level of correlation is controlled, i.e. given a matrix R fixing correlation structure (symmetric matrix with coefficients in $[-1, 1]$ and unity diagonal), we have $\text{Var}[(\tilde{X}_i)] = \Sigma R \Sigma$, where the standard deviation diagonal matrix Σ is estimated on the synthetic population.

The second requirement will generally be conditional to parameter values determining generation procedure, either generation models being simple or complex (R itself is a parameter). Formally, synthetic processes are parametric families $\tilde{X}_i[\vec{\alpha}]$. We propose to apply the

C : explicit the fact that real data may come out of different parameter values ?

methodology on very different examples, both typical of complex systems : financial high-frequency time-series and territorial systems. We illustrate the flexibility of the method, and claim to help building interdisciplinary bridges by methodology transposition and reasoning analogy. In the first case, proximity to data is the equality of signals at a fundamental frequency, to which higher frequency synthetic components with controlled correlations are superposed. It follows a logic of hybrid data for which hypothesis or model testing is done on a more realistic context than on purely synthetic data. This example that has no thematic link with the thesis, is presented in Appendix ???. In the second case, morphological calibration of a population density distribution model allows to respect real data proximity. Correlations of urban form with transportation network measures are empirically obtained by exploration of coupling with a network morphogenesis model. The control is in this case indirect as feasible space is empirically determined.



B.4 A DISCREPANCY-BASED FRAMEWORK

La multidimensionalité est un aspect fondamental du comportement des systèmes complexes, notamment dans leur processus d'optimisation. La plupart des explorations et calibrations que nous avons mené sont multi-objectif, mais les ontologies des modèles impliquent souvent des agents dont les objectifs sont multiples. Par ailleurs, la question de la sensibilité des modèles aux données a déjà été soulevée en 3.1. Nous faisons ici la jonction entre ces deux problèmes en étudiant la robustesse d'évaluations multi-objectifs à la structure des données, dans le cas particulier des évaluations multi-attributs. Ce travail ouvre des perspectives d'application aux modèles que nous avons développé, comme par exemple pour les modèles de morphogenèse mesoscopique pour lesquels les agents utilisent une fonction d'utilité multi-attribut pour l'attribution des nouvelles localisations.

★ ★

★

Cette section a été publiée en anglais comme [Raimbault, 2017a]. Elle est ici traduite et adaptée.

★ ★

★

Multi-objective evaluation is a necessary aspect when managing complex systems, as the intrinsic complexity of a system is generally closely linked to the potential number of optimization objectives. However, an evaluation makes no sense without its robustness being given (in the sense of its reliability). Statistical robustness computation methods are highly dependent of underlying statistical models. We propose a formulation of a model-independent framework in the case of integrated aggregated indicators (multi-attribute evaluation), that allows to define a relative measure of robustness taking into account data structure and indicator values. We implement and apply it to a synthetic case of urban systems based on Paris districts geography, and to real data for evaluation of income segregation for Greater Paris metropolitan area. First numerical results show the potentialities of this new method. Furthermore, its relative independence to

system type and model may position it as an alternative to classical statistical robustness methods.

B.4.1 *Introduction*

General Context

Multi-objective problems are organically linked to the complexity of underlying systems. Indeed, either in the field of *Complex Industrial Systems*, in the sense of engineered systems, where construction of Systems of Systems (SoS) by coupling and integration often leads to contradictory objectives [Marler and Arora, 2004], or in the field of *Natural Complex Systems*, in the sense of non engineered physical, biological or social systems that exhibit emergence and self-organization properties, where objectives can e.g. be the result of heterogeneous interacting agents (see [Newman, 2011] for a large survey of systems concerned by this approach), multi-objective optimization can be explicitly introduced to study or design the system but is often already implicitly ruling the internal mechanisms of the system. The case of socio-technical Complex Systems is particularly interesting as, following [Haken and Portugali, 2003], they can be seen as hybrid systems embedding social agents into “technical artifacts” (sometimes to an unexpected degree creating what PICON describes as *cyborgs* [Picon, 2013]), and thus cumulate propensity to be at the origin of multi-objective issues¹. The new notion of *eco-districts* [Souami, 2012] is a typical example where sustainability implies contradictory objectives. The example of transportation systems, which conception shifted during the second half of the 20th century from cost-benefit analysis to multi-criteria decision-making, is also typical of such systems [Bavoux et al., 2005]. Geographical system are now well studied from such a point of view in particular thanks to the integration of multi-objective frameworks within Geographical Information Systems [Carver, 1991]. As for the micro-case of eco-districts, meso and macro urban planning and design may be made sustainable through indicators evaluation [Jégou et al., 2012].

A crucial aspect of an evaluation is a certain notion of its reliability, that we call here *robustness*. Statistics naturally include this notion since the construction and estimation of statistical models give diverse indicators of the consistence of results [Launer and Wilkinson, 2014]. The first example that comes to mind is the application of the law of large numbers to obtain the *p-value* of a model fit, that can be interpreted as a confidence measure of estimates. Besides, confidence intervals and *beta-power* are other important indicators of statistical

¹ We design by *Multi-Objective Evaluation* all practices including the computation of multiple indicators of a system (it can be multi-objective optimization for system design, multi-objective evaluation of an existing system, multi-attribute evaluation ; our particular framework corresponds to the last case).

robustness. Bayesian inference provide also measures of robustness when distribution of parameters are sequentially estimated. Concerning multi-objective optimization, in particular through heuristic algorithms (for example genetic algorithms, or operational research solvers), the notion of robustness of a solution concerns more the stability of the solution on the phase space of the corresponding dynamical system. Recent progresses have been done towards unified formulation of robustness for a multi-objective optimization problem, such as [Deb and Gupta, 2006] where robust Pareto-front as defined as solutions that are insensitive to small perturbations. In [Barrico and Antunes, 2006], the notion of degree of robustness is introduced, formalized as a sort of continuity of other solutions in successive neighborhood of a solution.

However, there still lack generic methods to estimate robustness of an evaluation that would be model-independent, i.e. that would be extracted from data structure and indicators but that would not depend on the method used. Some advantages could be for example an *a priori* estimation of potential robustness of an evaluation and thus to decide if the evaluation is worth doing. We propose here a framework answering this issue in the particular case of Multi-attribute evaluations, i.e. when the problem is made unidimensional by objectives aggregation. It is data-driven and not model-driven in the sense that robustness estimation does not depend on how indicators are computed, as soon as they respect some assumptions that will be detailed in the following.

Proposed Approach

OBJECTIVES AS SPATIAL INTEGRALS We assume that objectives can be expressed as spatial integrals, so it should apply to any territorial system and our application cases are urban systems. It is not that restrictive in terms of possible indicators if one uses suitable variables and integrated kernels : in a way analog to the method of geographically weighted regression [Brunsdon, Fotheringham, and Charlton, 1998], any spatial variable can be integrated against regular kernels of variable size and the result will be a spatial aggregation which sense depends on kernel size. The example we use in the following such as conditional means or sums suit well the assumption. Even an already spatially aggregated indicator can be interpreted as a spatial indicator by using a Dirac distribution on the centroid of the corresponding area.

LINEARLY AGGREGATED OBJECTIVES A second assumption we make is that the multi-objective evaluation is done through linear aggregation of objectives, i.e. that we are tackling a multi-attribute optimization problem. If $(q_i(\vec{x}))_i$ are values of objectives functions, then weights $(w_i)_i$ are defined in order to build the aggregated decision-

making function $q(\vec{x}) = \sum_i w_i q_i(\vec{x})$, which value determines then the performance of the solution. It is analog to aggregated utility techniques in economics and is used in many fields. The subtlety lies in the choice of weights, i.e. the shape of the projection function, and various approaches have been developed to find weights depending on the nature of the problem. Recent work [Dobbie and Dail, 2013] proposed to compare robustness of different aggregation techniques through sensitivity analysis, performed by Monte-Carlo simulations on synthetic data. Distribution of biases were obtained for various techniques and some showed to perform significantly better than others. Robustness assessment still depended on models used in that work.

The rest of the paper is organized as follows : section 2 describes intuitively and mathematically the proposed framework ; section 3 then details implementation, data collection for case studies and numerical results for an artificial intra-urban case and a metropolitan real case ; section 4 finally discuss limitations and potentialities of the method.

B.4.2 Framework Description

Intuitive Description

We describe now the abstract framework allowing theoretically to compare robustnesses of evaluations of two different urban systems. Our framework is a generalization of an empirical method proposed in [**ecodistrictReport**] besides a more general benchmarking study on indicator sense and relevance in a sustainability context. Intuitively, it relies on empirical base resulting from the following axioms :

- Urban systems can be seen from the information available, i.e. raw data describing the system. As a data-driven approach, this raw data is the basis of our framework and robustness will be determined by its structure.
- From data are computed indicators (objective functions). We assume that a choice of indicators is an intention to translate particular aspects of the system, i.e. to capture a realization of an “urban fact” (*fait urbain*) in the sense of MANGIN [Mangin and Panerai, 1999] - a sort of stylized fact in terms of processes and mechanisms, having various realizations on spatially distinct systems, depending on each precise context.
- Given many systems and associated indicators, a common space can be built to compare them. In that space, data represents more or less well real systems, depending e.g. on initial scale, precision of data, missing data. We precisely propose to capture

that through the notion of point cloud discrepancy, which is a mathematical tool coming from sampling theory expressing how a dataset is distributed in the space it is embedded in [Dick and Pillichshammer, 2010].

Synthesizing these requirements, we propose a notion of *Robustness* of an evaluation that captures both, by combining data reliability with relative importance,

1. *Missing Data* : an evaluation based on more refined datasets will naturally be more robust.
2. *Indicator importance* : indicators with more relative influence will weight more on the total robustness.

Formal Description

INDICATORS Let $(S_i)_{1 \leq i \leq N}$ be a finite number of geographically disjoints territorial systems, that we assume described through raw data and intermediate indicators, yielding $S_i = (X_i, Y_i) \in \mathcal{X}_i \times \mathcal{Y}_i$ with $\mathcal{X}_i = \prod_k \mathcal{X}_{i,k}$ such that each subspace contain real matrices : $\mathcal{X}_{i,k} = \mathbb{R}^{n_{i,k}^X p_{i,k}^X}$ (the same holding for \mathcal{Y}_i). We also define an ontological index function $I_X(i, k)$ (resp. $I_Y(i, k)$) taking integer values which coincide if and only if the two variables have the same ontology in the sense of [Livet et al., 2010], i.e. they are supposed to represent the same real object. We distinguish “raw data” X_i from which indicators are computed via explicit deterministic functions, from “intermediate indicators” Y_i that are already integrated and can be e.g. outputs of elaborated models simulating some aspects of the urban system. We define the partial characteristic space of the “urban fact” by

$$(\mathcal{X}, \mathcal{Y}) \underset{\text{def}}{=} \left(\prod \tilde{\mathcal{X}}_c \right) \times \left(\prod \tilde{\mathcal{Y}}_c \right) = \left(\prod_{\mathcal{X}_{i,k} \in \mathcal{D}_X} \mathbb{R}^{p_{i,k}^X} \right) \times \left(\prod_{\mathcal{Y}_{i,k} \in \mathcal{D}_Y} \mathbb{R}^{p_{i,k}^Y} \right) \quad (25)$$

with $\mathcal{D}_X = \{\mathcal{X}_{i,k} | I(i, k) \text{ distincts, } n_{i,k}^X \text{ maximal}\}$ (the same holding for \mathcal{Y}_i). It is indeed the abstract space on which indicators are integrated. The indices c introduced as a definition here correspond to different indicators across all systems. This space is the minimal space common to all systems allowing a common definition for indicators on each.

Let $X_{i,c}$ be the data canonically projected in the corresponding subspace, well defined for all i and all c . We make the key assumption that all indicators are computed by integration against a certain kernel, i.e. that for all c , there exists H_c space of real-valued functions on $(\tilde{\mathcal{X}}_c, \tilde{\mathcal{Y}}_c)$, such that for all $h \in H_c$:

1. h is “enough” regular (tempered distributions e.g.)

2. $q_c = \int_{(\tilde{X}_c, \tilde{y}_c)} h$ is a function describing the “urban fact” (the indicator in itself)

Typical concrete example of kernels can be :

- A mean of rows of $\mathbf{X}_{i,c}$ is computed with $h(x) = x \cdot f_{i,c}(x)$ where $f_{i,c}$ is the density of the distribution of the assumed underlying variable.
- A rate of elements respecting a given condition C , $h(x) = f_{i,c}(x) \chi_{C(x)}$
- For already aggregated variables \mathbf{Y} , a Dirac distribution allows to express them also as a kernel integral.

AGGREGATION Weighting objectives in multi-attribute decision-making is indeed the crucial point of the processes, and numerous methods are available (see [Wang et al., 2009] for a review for the particular case of sustainable energy management). Let define weights for the linear aggregation. We assume the indicators normalized, i.e. $q_c \in [0, 1]$, for a more simple construction of relative weights. For i, c and $h_c \in H_c$ given, the weight $w_{i,c}$ is simply constituted by the relative importance of the indicator $w_{i,c}^L = \frac{\hat{q}_{i,c}}{\sum_c \hat{q}_{i,c}}$ where $\hat{q}_{i,c}$ is an estimator of q_c for data $\mathbf{X}_{i,c}$ (i.e. the effectively calculated value). Note that this step can be extended to any sets of weight attributions, by taking for example $\tilde{w}_{i,c} = w_{i,c} \cdot w'_{i,c}$ if \mathbf{w}' are the weights attributed by the decision-maker. We focus here on the relative influence of attributes and thus choose this simple form for weights.

ROBUSTNESS ESTIMATION The scene is now set up to be able to estimate the robustness of the evaluation done through the aggregated function. Therefore, we apply an integral approximation method similar to methods introduced in [Varet, 2010], since the integrated form of indicators indeed brings the benefits of such powerful theoretical results. Let $\mathbf{X}_{i,c} = (\vec{X}_{i,c,l})_{1 \leq l \leq n_{i,c}}$ and $D_{i,c} = \text{Disc}_{\tilde{X}_c, L^2}(\mathbf{X}_{i,c})$ the discrepancy of data points cloud² [Niederreiter, 1972]. With $h \in H_c$, we have the upper bound on the integral approximation error

$$\left\| \int h_c - \frac{1}{n_{i,c}} \sum_l h_c(\vec{X}_{i,c,l}) \right\| \leq K \cdot \|h_c\| \cdot D_{i,c}$$

where K is a constant independent of data points and objective function. It directly yields

² The discrepancy is defined as the L2-norm of local discrepancy which is for normalized data points $\mathbf{X} = (x_{ij}) \in [0, 1]^d$, a function of $t \in [0, 1]^d$ comparing the number of points falling in the corresponding hypercube with its volume, by $\text{disc}(t) = \frac{1}{n} \sum_i \mathbb{1}_{\prod_j x_{ij} < t_j} - \prod_j t_j$. It is a measure of how the point cloud covers the space.

$$\left\| \int \sum w_{i,c} h_c - \frac{1}{n_{i,c}} \sum_l w_{i,c} h_c(\vec{X}_{i,c,l}) \right\| \leq K \sum_c |w_{i,c}| \|h_c\| \cdot D_{i,c}$$

Assuming the error reasonably realized (“worst case” scenario for knowledge of the theoretical value of aggregated function), we take this upper bound as an approximation of its magnitude. Furthermore, taking normalized indicators implies $\|h_c\| = 1$. We propose then to compare error bounds between two evaluations. They depend only on data distribution (equivalent to *statistical robustness*) and on indicators chosen (sort of *ontological robustness*, i.e. do the indicators have a real sense in the chosen context and do their values make sense), and are a way to combine these two type of robustnesses into a single value.

We thus define a *robustness ratio* to compare the robustness of two evaluations by

$$R_{i,i'} = \frac{\sum_c w_{i,c} \cdot D_{i,c}}{\sum_c w_{i',c} \cdot D_{i',c}} \quad (26)$$

The intuitive sense of this definition is that one compares robustness of evaluations by comparing the highest error done in each based on data structure and relative importance.

By taking then an order relation on evaluations by comparing the position of the ratio to one, it is obvious that we obtain a complete order on all possible evaluations. This ratio should theoretically allow to compare any evaluation of an urban system. To keep an ontological sense to it, it should be used to compare disjoints sub-systems with a reasonable proportion of indicators in common, or the same sub-system with varying indicators. Note that it provides a way to test the influence of indicators on an evaluation by analyzing the sensitivity if the ratio to their removal. On the contrary, finding a “minimal” number of indicators each making the ratio strongly vary should be a way to isolate essential parameters ruling the sub-system.

B.4.3 Results

IMPLEMENTATION Preprocessing of geographical data is made through QGIS [QGis, 2011] for performance reasons. Core implementation of the framework is done in R [R Core Team, 2015] for the flexibility of data management and statistical computations. Furthermore, the package DiceDesign [Franco et al., 2009] written for numerical experiments and sampling purposes, allows an efficient and direct computation of discrepancies. Last but not least, all source code is openly available on the git repository of the project³ for reproducibility purposes [Ram, 2013].

³ at <https://github.com/JusteRaimbault/RobustnessDiscrepancy>

Implementation on Synthetic Data

We propose in a first time to illustrate the implementation with an application to synthetic data and indicators, for intra-urban quality indicators in the city of Paris.

DATA COLLECTION We base our virtual case on real geographical data, in particular for *arrondissements* of Paris. We use open data available through the OpenStreetMap project [Bennett, 2010] that provides accurate high definition data for many urban features. We use the street network and position of buildings within the city of Paris. Limits of *arrondissements*, used to overlay and extract features when working on single districts, are also extracted from the same source. We use centroids of buildings polygons, and segments of street network. Dataset overall consists of around 200k building features and 100k road segments.

VIRTUAL CASES We work on each district of Paris (from the 1st to the 20th) as an evaluated urban system. We construct random synthetic data associated to spatial features, so each district has to be evaluated many time to obtain mean statistical behavior of toy indicators and robustness ratios. The indicators chosen need to be computed on residential and street network spatial data. We implement two mean kernels and a conditional mean to show different examples, linked to environmental sustainability and quality of life, that are required to be maximized. Note that these indicators have a real meaning but no particular reason to be aggregated, they are chosen here for the convenience of the toy model and the generation of synthetic data. With $a \in \{1 \dots 20\}$ the number of the district, $A(a)$ corresponding spatial extent, $b \in B$ building coordinates and $s \in S$ street segments, we take

- Complementary of the average daily distance to work with car per individual, approximated by, with $n_{cars}(b)$ number of cars in the building (randomly generated by associated of cars to a number of building proportional to motorization rate $\alpha_m = 0.4$ in Paris), d_w distance to work of individuals (generated from the building to a uniformly generated random point in spatial extent of the dataset), and d_{max} the diameter of Paris area, $\bar{d}_w = 1 - \frac{1}{|B \in A(a)|} \cdot \sum_{b \in A(a)} n_{cars}(b) \cdot \frac{d_w}{d_{max}}$
- Complementary of average car flows within the streets in the district, approximated by, with $\varphi(s)$ relative flow in street segment s , generated through the minimum of 1 and a log-normal distribution adjusted to have 95% of mass smaller than 1 what mimics the hierarchical distribution of street use (corresponding to betweenness centrality), and $l(s)$ segment length, $\bar{\varphi} = 1 - \frac{1}{|s \in A(a)|} \cdot \sum_{s \in A(a)} \varphi(s) \cdot \frac{l(s)}{\max(l(s))}$

- Relative length of pedestrian streets \bar{p} , computed through a randomly uniformly generated dummy variable adjusted to have a fixed global proportion of segments that are pedestrian.

TABLE 27: Numerical results of simulations for each district with $N = 50$ repetitions. Each toy indicator value is given by mean on repetitions and associated standard deviation. Robustness ratio is computed relative to first district (arbitrary choice). A ratio smaller than 1 means that integral bound is smaller for upper district, i.e. that evaluation is more robust for this district. Because of the small size of first district, we expected a majority of district to give ratio smaller than 1, what is confirmed by results, even when adding standard deviations.

Arrdt	$<\bar{d}_w> \pm \sigma(\bar{d}_w)$	$<\bar{\varphi}> \pm \sigma(\bar{\varphi})$	$<\bar{p}> \pm \sigma(\bar{p})$	$R_{i,1}$
1 th	0.731655 ± 0.041099	0.917462 ± 0.026637	0.191615 ± 0.052142	1.000000 ± 0.000000
2 th	0.723225 ± 0.032539	0.844350 ± 0.036085	0.209467 ± 0.058675	1.002098 ± 0.039972
3 th	0.713716 ± 0.044789	0.797313 ± 0.057480	0.185541 ± 0.065089	0.999341 ± 0.048825
4 th	0.712394 ± 0.042897	0.861635 ± 0.030859	0.201236 ± 0.044395	0.973045 ± 0.036993
5 th	0.715557 ± 0.026328	0.894675 ± 0.020730	0.209965 ± 0.050093	0.963466 ± 0.040722
6 th	0.733249 ± 0.026890	0.875613 ± 0.029169	0.206690 ± 0.054850	0.990676 ± 0.031666
7 th	0.719775 ± 0.029072	0.891861 ± 0.026695	0.209265 ± 0.041337	0.966103 ± 0.037132
8 th	0.713602 ± 0.034423	0.931776 ± 0.015356	0.208923 ± 0.036814	0.973975 ± 0.033809
9 th	0.712441 ± 0.027587	0.910817 ± 0.015915	0.202283 ± 0.049044	0.971889 ± 0.035381
10 th	0.713072 ± 0.028918	0.881710 ± 0.021668	0.210118 ± 0.040435	0.991036 ± 0.038942
11 th	0.682905 ± 0.034225	0.875217 ± 0.019678	0.203195 ± 0.047049	0.949828 ± 0.035122
12 th	0.646328 ± 0.039668	0.920086 ± 0.019238	0.198986 ± 0.023012	0.960192 ± 0.034854
13 th	0.697512 ± 0.025461	0.890253 ± 0.022778	0.201406 ± 0.030348	0.960534 ± 0.033730
14 th	0.703224 ± 0.019900	0.902898 ± 0.019830	0.205575 ± 0.038635	0.932755 ± 0.033616
15 th	0.692050 ± 0.027536	0.891654 ± 0.018239	0.200860 ± 0.024085	0.929006 ± 0.031675
16 th	0.654609 ± 0.028141	0.928181 ± 0.013477	0.202355 ± 0.017180	0.963143 ± 0.033232
17 th	0.683020 ± 0.025644	0.890392 ± 0.023586	0.198464 ± 0.033714	0.941025 ± 0.034951
18 th	0.699170 ± 0.025487	0.911382 ± 0.027290	0.188802 ± 0.036537	0.950874 ± 0.028669
19 th	0.655108 ± 0.031857	0.884214 ± 0.027816	0.209234 ± 0.032466	0.962966 ± 0.034187
20 th	0.637446 ± 0.032562	0.873755 ± 0.036792	0.196807 ± 0.026001	0.952410 ± 0.038702

As synthetic data are stochastic, we run the computation for each district $N = 50$ times, what was a reasonable compromise between statistical convergence and time required for computation. Table 1 shows results (mean and standard deviations) of indicator values and robustness ratio computation. Obtained standard deviation confirm

that this number of repetitions give consistent results. Indicators obtained through a fixed ratio show small variability what may a limit of this toy approach. However, we obtain the interesting result that a majority of districts give more robust evaluations than 1st district, what was expected because of the size and content of this district : it is indeed a small one with large administrative buildings, what means less spatial elements and thus a less robust evaluation following our definition of the robustness.

Application to a Real Case: Metropolitan Segregation

The first example was aimed to show potentialities of the method but was purely synthetic, hence yielding no concrete conclusion nor implications for policy. We propose now to apply it to real data for the example of metropolitan segregation.

DATA We work on income data available for France at an intra-urban level (basic statistical units IRIS) for the year 2011 under the form of summary statistics (deciles if the area is populated enough to ensure anonymity), provided by INSEE⁴. Data are associated with geographical extent of statistical units, allowing computation of spatial analysis indicators.

INDICATORS We use here three indicators of segregation integrated on a geographical area. Let assume the area divided into covering units S_i for $1 \leq i \leq N$ with centroids (x_i, y_i) . Each unit has characteristics of population P_i and median income X_i . We define spatial weights used to quantify strength of geographical interactions between units i, j , with d_{ij} euclidian distance between centroids : $w_{ij} = \frac{P_i P_j}{(\sum_k P_k)^2} \cdot \frac{1}{d_{ij}}$ if $i \neq j$ and $w_{ii} = 0$. The normalized indicators are the following

- Spatial autocorrelation Moran index, defined as weighted normalized covariance of median income by $\rho = \frac{N}{\sum_{ij} w_{ij}} \cdot \frac{\sum_{ij} w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}$
- Dissimilarity index (close to Moran but integrating local dissimilarities rather than correlations), given by $d = \frac{1}{\sum_{ij} w_{ij}} \sum_{ij} w_{ij} |\tilde{X}_i - \tilde{X}_j|$ with $\tilde{X}_i = \frac{X_i - \min(X_k)}{\max(X_k) - \min(X_k)}$
- Complementary of the entropy of income distribution that is a way to capture global inequalities $\varepsilon = 1 + \frac{1}{\log(N)} \sum_i \frac{X_i}{\sum_k X_k} \cdot \log \left(\frac{X_i}{\sum_k X_k} \right)$

Numerous measures of segregation with various meanings and at different scales are available, as for example at the level of the unit

⁴ <http://www.insee.fr>

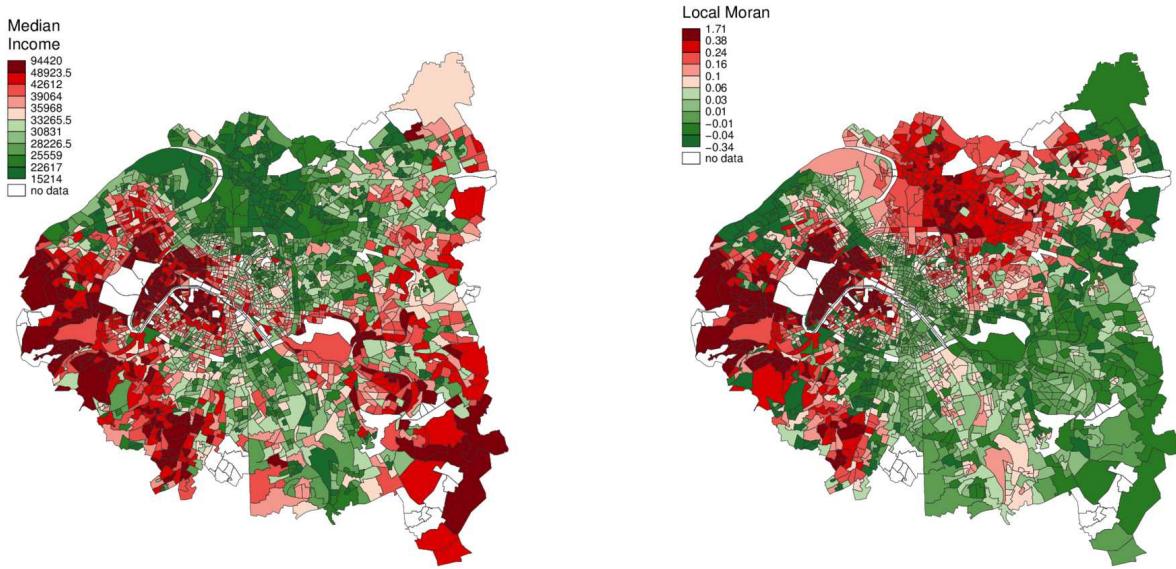


FIGURE 108: Maps of Metropolitan Segregation. Maps show yearly median income on basic statistical units (IRIS) for the three departments constituting mainly the Great Paris metropolitan area, and the corresponding local Moran spatial autocorrelation index, defined for unit i as $\rho_i = N / \sum_j w_{ij} \cdot \frac{\sum_j w_{ij} (X_j - \bar{X})(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2}$. The most segregated areas coincide with the richest and the poorest, suggesting an increase of segregation in extreme situations.

by comparison of empirical wage distribution with a theoretical null model [Louf and Barthelemy, 2016]. The choice here is arbitrary in order to illustrate our method with a reasonable number of dimensions.

RESULTS We apply our method with these indicators on the Greater Paris area, constituted of four *départements* that are intermediate administrative units. The recent creation of a new metropolitan governance system [Gilli and Offner, 2009] underlines interrogations on its consistence, and in particular on its relation to intermediate spatial inequalities. We show in Fig. ?? maps of spatial distribution of median income and corresponding local index of autocorrelation. We observe the well-known West-East opposition and district disparities inside

Paris as they were formulated in various studies, such as [Guérois and Le Goix, 2009] through the analysis of real estate transactions dynamics. We then apply our framework to answer a concrete question that has implications for urban policy : *how are the evaluation of segregation within different territories sensitive to missing data ?* To do so, we proceed to Monte Carlo simulations (75 repetitions) during which a fixed proportion of data is randomly removed, and the corresponding robustness index is evaluated with renormalized indicators. Simulations are done on each *department* separately, each time relatively to the robustness of the evaluation of full Greater Paris. Results are shown in Fig. 2. All areas present a slightly better robustness than the reference, what could be explained by local homogeneity and thus more fiable segregation values. Implications for policy that can be drawn are for example direct comparisons between areas : a loss of 30% of information on 93 area corresponds to a loss of only 25% in 92 area. The first being a deprived area, the inequality is increased by this relative lower quality of statistical information. The study of standard deviations suggest further investigations as different response regimes to data removal seem to exist.

B.4.4 Discussion

Applicability to Real situations

IMPLICATIONS FOR DECISION-MAKING The application of our method to concrete decision-making can be thought in different ways. First in the case of a comparative multi-attribute decision process, such as the determination of a transportation corridor, the identification of territories on which the evaluation may be flawed (i.e. has a poor relative robustness) could allow a more refined focus on these and a corresponding revision of datasets or an adapted revision of weights. In any case the overall decision-making process should be made more reliable. A second direction lays in the spirit of the real application we have proposed, i.e. the sensitivity of evaluation to various parameters such as missing data. If a decision appears as reliable because data have few missing points, but the evaluation is very sensitive to it, one will be more careful in the interpretation of results and taking the final decision. Further work and testing will however be needed to understand framework behavior in different contexts and be able to pilot its application in various real situations.

INTEGRATION WITHIN EXISTING FRAMEWORKS The applicability of the method on real cases will directly depend on its potential integration within existing framework. Beyond technical difficulties that will surely appear when trying to couple or integrate implementations, more theoretical obstacles could occur, such as fuzzy formulations of functions or data types, consistency issues in databases,

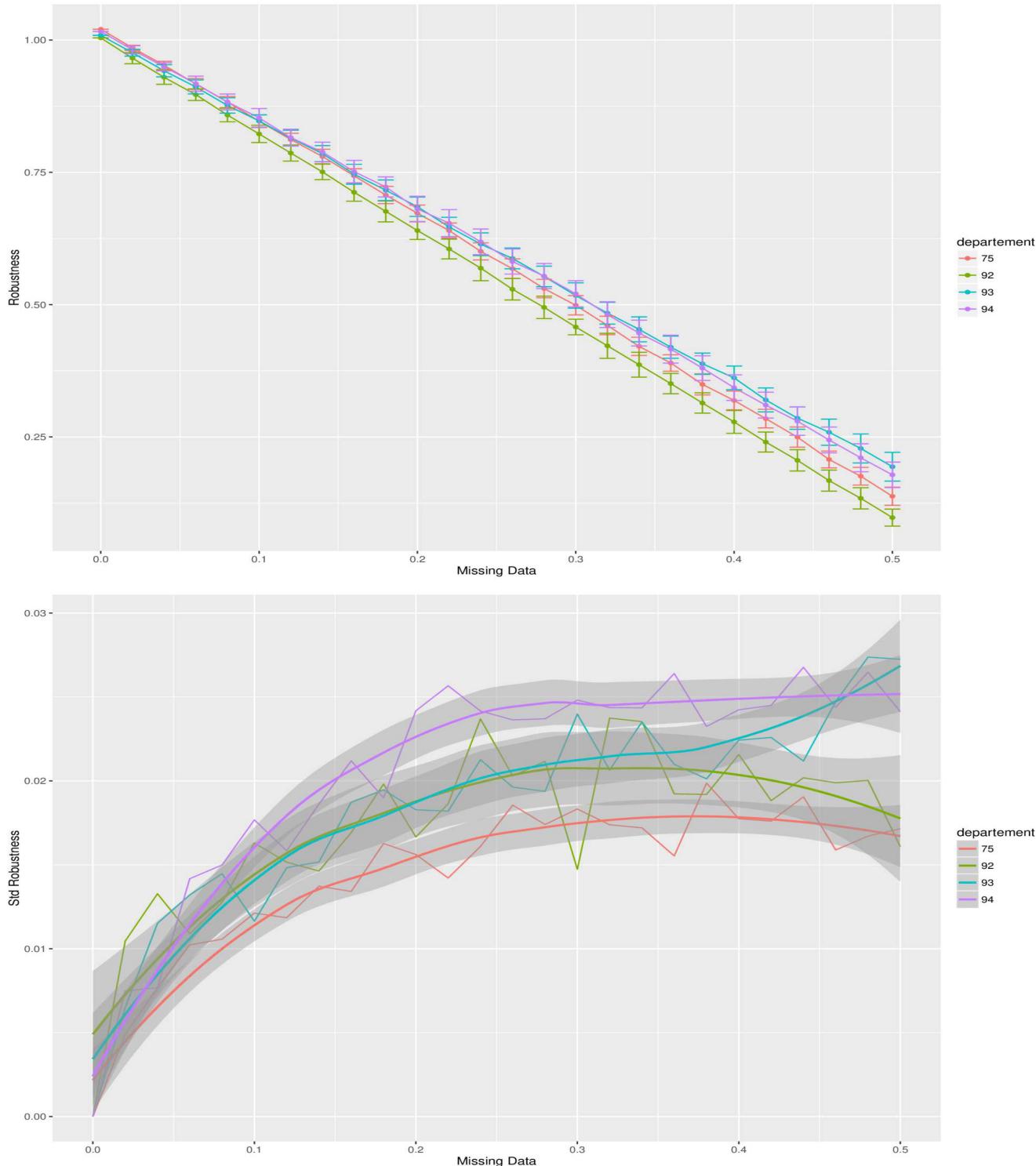


FIGURE 109: Sensitivity of robustness to missing data. *Left.* For each department, Monte Carlo simulations ($N=75$ repetitions) are used to determine the impact of missing data on robustness of segregation evaluation. Robustness ratios are all computed relatively to full metropolitan area with all available data. Quasi-linear behavior translates an approximative linear decrease of discrepancy as a function of data size. The similar trajectory of poorest departments (93,94) suggest the correction to linear behavior being driven by segregation patterns. *Right.* Corresponding standard deviations of robustness ratios. Different regimes (in particular 93 against others) unveil phase transitions at different levels of missing data, meaning that the evaluation in 94 is from this point of view more sensitive to missing data.

etc. Such multi-criteria framework are numerous. Further interesting work would be to attempt integration into an open one, such as e.g. the one described in [Tivadar et al., 2014] which calculates various indices of urban segregation, as we have already illustrated the application on metropolitan segregation indexes.

AVAILABILITY OF RAW DATA In general, sensitive data such as transportation questionnaires, or very fine granularity census data are not openly available but provided already aggregated at a certain level (for instance French Insee Data are publicly available at basic statistical unit level or larger areas depending on variables and minimal population constraints, more precise data is under restricted access). It means that applying the framework may imply complicated data research procedure, its advantage to be flexible being thus reduced through additional constraints.

Validity of Theoretical Assumptions

A possible limitation of our approach is the validity of the assumption formulating indicators as spatial integrals. Indeed, many socio-economic indicators are not necessarily depending explicitly on space, and trying to associate them with spatial coordinates may become a slippery slope (e.g. associate individual economic variables with individual residential coordinates will have a sense only if the use of the variable has a relation with space, otherwise it is a non-legitimate artifact). Even indicators which have a spatial value may derive from non-spatial variables, as [Kwan, 1998] points out concerning accessibility, when opposing integrated accessibility measures with individual-based non necessarily spatial-based (e.g. individual decisions) measures. Constraining a theoretical representation of a system to fit a framework by changing some of its ontological properties (always in the sense of real meaning of objects) can be understood as a violation of a fundamental rule of modeling and simulation in social science given in [Banos, 2013], that is that there can be an universal “language” for modeling and some can not express some systems, having for consequence misleading conclusion due to ontology breaking in the case of an over-constrained formulation.

Framework Generality

We argue that the fundamental advantage of the proposed framework is its generality and flexibility, since robustness of the evaluations are obtained only through data structure if ones relaxes constraints on the value of weight. Further work should go towards a more general formulation, suppressing for example the linear aggregation assumption. Non-linear aggregation functions would require however to present particular properties regarding integral inequalities. For

example, similar results could search in the direction of integral inequalities for Lipschitzian functions such as the one-dimensional results of [Dragomir, 1999].

Conclusion

We have proposed a model-independent framework to compare the robustness of multi-attribute evaluations between different urban systems. Based on data discrepancy, it provide a general definition of relative robustness without any assumption on model for the system, but with limiting assumptions that are the need of linear aggregation and of indicators being expressed through spatial kernel integrals. We propose a toy implementation based on real data for the city of Paris, numerical results confirming general expected behavior, and an implementation on real data for income segregation on Greater Paris metropolitan areas, giving possible insights into concrete policy questions. Further work should be oriented towards sensitivity analysis of the method, application to other real cases and theoretical assumptions relaxation, i.e. the relaxation of linear aggregation and spatial integration.

★ ★

★

B.5 SOCIO-TECHNICAL SYSTEMS

Nous développons ici un cadre formel pour la modélisation des systèmes socio-techniques. Plus précisément, celui-ci implémente l'idée de perspectivisme appliquée pour éclairer une structure possible des opérations de couplages de perspectives. Il peut par ailleurs également être compris comme un travail préliminaire pour la formalisation du cadre de connaissance suggérée en 8.3 (sans encore inclure la structure algébrique pour l'opération sur les données).

B.5.1 Context

Scientific Context

The structural misunderstandings between Social Sciences and Humanities on one side, and so-called Exact Sciences on the other side, far from being a generality, seems to have however a significant impact on the structure of scientific knowledge [Hidalgo, 2015]. In particular, the place of theory (and indeed the signification of this term itself) in the elaboration of knowledge has a totally different place, partly because of the different *perceived complexities*⁵ of studied objects: for example, mathematical constructions and by extent theoretical physics are *simple* in the sense that they are generally analytically solvable (or at least semi-analytically), whereas Social Science subjects such as humans or society (to give a *cliché* exemple) are *complex* in the sense of complex systems⁶, thus a stronger need of a constructed theoretical (generally empirically based) framework to identify and define the objects of research that are necessarily more arbitrary in the framing of their boundaries, relations and processes, because of the multitude of possible viewpoints: PUMAIN suggests indeed in [Pumain, 2005] a new approach to complexity deeply rooted in social sciences that "would be measured by the diversity of disciplines needed to elaborate a notion". These differences in backgrounds are naturally desirable in the spectrum of science, but things can get nasty when playing on overlapping terrains, typically complex systems problematics as already detailed, as the exemple of geographical urban systems has recently shown [Dupuy and Benguigui, 2015]. Complex System Science⁷ is presented by some as a "new kind

⁵ We used the term *perceived* as most of systems studied by physics might be described as simple whereas they are intrinsically complex and indeed not well understood [Laughlin, 2006].

⁶ for which no unified definition exists but of which fields of application range broadly from neuroscience to quantitative finance, including e.g. quantitative sociology, quantitative geography, integrative biology, etc. [Newman, 2011], and for which study various complementary approaches may be applied, such as Dynamical Systems, Agent-based Modeling, Random Matrix Theory

⁷ that we deliberately call that way although there is a running debate on whether it can be seen as a Science in itself or more as a different way to do Science.

of Science" [Wolfram, 2002], and would at least be a symptom of a shift in scientific practices, from analytical and "exact" approaches to computational and evidence-based approaches [Arthur, 2015], but what is sure is that it brings, together with new methodologies, new scientific fields in the sense of converging interests of various disciplines on transversal questions or of integrated approaches on a particular field [Bourgine, Chavaliarias, and al., 2009].

Objectives

Within that scientific context, the study of what we will call *Socio-technical Systems*, which we define in a rather broad way as hybrid complex systems including social agents or objects that interact with technical artifacts and/or a natural environment⁸, lies precisely between social sciences and hard sciences. The example of Urban Systems is the best example, as already before the arrival of approaches claiming to be "more exact" than soft approaches (typically by physicists, see e.g. the positioning of [Louf and Barthelemy, 2014c], but also by scientists coming from social sciences such as BATTY [Batty, 2013b]), many aspects of urban systems were already in the field of exact sciences, such as urban hydrology, urban climatology or technical aspects of transportation systems, whereas the core of their study relied in social sciences such as geography, urbanism, sociology, economy. Therefore a necessary place of theory in their study: following [Livet et al., 2010], the study of complex systems in social science is an interaction between empirical analysis, theoretical constructions, and modeling.

We propose in this section to construct a theory, or rather a theoretical framework, that would ease some aspects of the study of such systems. Many theories already exist in all fields related to this kind of problems, and also at higher levels of abstraction concerning methods such as agent-based modeling e.g., but there is to our knowledge no theoretical framework including all of the following aspects that we consider as being crucial (and that can be understood as an informal basis of our theory):

1. a precise definition and emphasis on the notion of coupling between subsystems, in particular allowing to qualify or quantify a certain degree of coupling: dependence, interdependence, etc. between components.
2. a precise definition of scale, including timescale and scales for other dimensions.

⁸ geographical systems in the sense of [Dollfus and Dastès, 1975] are the archetype of such systems, but that definition may cover other type of systems such as an extended transportation system, social systems taken with an environmental context, complicated industrial systems taken with users, etc.

3. as a consequence of the previous points, a precise definition of what is a system.
4. the inclusion of the notion of emergence in order to capture multi-scale aspects of systems.
5. a central place of ontology in the definition of systems, i.e. of the sense in the real world given to its objects⁹.
6. taking into account heterogeneous aspects of the same system, that could be heterogeneous components but also complementary intersecting views.

The rest of this section is organized as follows: we construct the theory in the following subsection, staying at an abstract level, and propose a first application to the question of co-evolving subsystems. We then discuss positioning regarding existing theories, and possible developments and concrete applications.

B.5.2 *Construction of the theory*

Perspectives and Ontologies

The starting point of the theory construction is a perspectivist epistemological approach on systems introduced by GIREE [Giere, 2010c]. To sum up, it interprets any scientific approach as a perspective, in which someone pursues some objective and uses what is called *a model* to reach it. The model is nothing more than a scientific medium. VARENNE developed [Varenne, 2010a] a functional model typology that can be interpreted as a refinement of this theory. Let for now relax this possible precision and use perspectives as proxies of the undefined objects and concepts. Indeed, different views on the same object (being complementary or diverging) have the property to share at least the object in itself, thus the proposition to define objects (and more generally systems) from a set of perspectives on them, that verify some properties that we formalize in the following.

A perspective is defined in our case as a dataflow machine M (that corresponds to the model as medium) in the sense of [Golden, Aiguier, and Krob, 2012] that gives a convenient way to represent it and to introduce timescales and data, to which is associated an ontology O in the sense of [Livet et al., 2010], i.e. a set of elements each corresponds to an entity (which can be an object, an agent, a process, etc.) of the real world. We include only two aspect (the model and the objects represented) of Giere's theory, making the assumption that purpose and producer of the perspective are indeed contained in the ontology if they make sense for studying the system.

⁹ as already explained before, this positioning along with the importance of structure may be related to Ontic Structural Realism [Frigg and Votsis, 2011] in further developments.

Définition 2 A perspective on a system is given by a dataflow machine $M = (i, o, T)$ and an associated ontology O . We assume that the ontology can be decomposed into atomic elements $O = (O_j)_{j \in I}$.

The atomic elements of the ontology can be particular elements such as agents or components of the system, but also processes, interactions, states, or concepts for example. The ontology can be seen as the exhaustive and rigorous description of the content of the perspective. The assumption of a dataflow machine implies that possible inputs and outputs can be quantified, what is not necessarily restrictive to quantitative perspectives, as most of qualitative approaches can be translated into discrete variables as soon as the set of possibles is known or assumed.

The system is then defined “reversely”, i.e. from a set of perspectives on what would constitute then the system:

Définition 3 A system is a set of perspectives on a system: $S = (M_i, O_i)_{i \in I}$, where I may be finite or not.

We denote by $\mathcal{O} = (O_{j,i})_{j \in I, i \in I}$ the set of all elements within ontologies.

Note that at this level of construction, there is not necessarily any structural consistence in what we call a system, as given our broad definition could allow for example to consider as a system a perspective on a car together with a perspective on a system of cities what makes reasonably no sense at all. Further definitions and developments will allow to be closer from classical definition of a system (interacting entities, designed artifacts, etc.). The same way, the definition of a subsystem will be given further. The introduced elements of our approach help to tackle so far points three, five and six of the requirements.

PRECISION ON THE RECURSIVE ASPECT OF THE THEORY One direct consequence of these definitions must be detailed: the fact that they can be applied recursively. Indeed, one could imagine taking as perspective a system in our sense, therefore a set of perspectives on a system, and do that at any order. If ones takes a system in any classical sense, then the first order can be understood as an epistemology of the system, i.e. the study of diverse perspectives on a system. A set of perspectives on related systems may in some conditions be a domain or a field, thus a set of perspectives on various related systems the epistemology of a field. These are more analogies to give the idea behind the recursive character of the theory. It is indeed crucial for the meaning and consistence of the theory because of the following arguments:

- The choice of perspectives in which a system consists is necessarily subjective and therefore understood as a perspective, and

a perspective on a system if we are able to build a general ontology.

- We will use relations between ontologies in the following, which construction based on emergence is also subjective and seen as perspectives.

Ontological Graph

We propose then to capture the structure of the system by linking ontologies. This approach could eventually be linked to structural realism epistemological positioning [Frigg and Votsis, 2011] as knowledge of the world is partly contained here in structure of models. Therefore, we choose to emphasize the role of emergence as we believe that it may be one practical minimalist way to capture quite well complex systems structure¹⁰. We follow on that point the approach of BEDAU on different type of emergences, in particular his definition of weak emergence given in [Bedau, 2002]. Let recall briefly definitions we will use in the following. BEDAU starts from defining emerging properties and then extends it to phenomena, entities, etc. The same way, our framework is not restricted to objects or properties and wraps thus the generalized definitions into emergence between ontologies. We will apply the notion of emergence under the two following forms¹¹:

- *Nominal emergence*: one ontology O' is included in an other O but the aspect of O that is said to be nominally emergent regarding O' does not depend on O'.
- *Weak emergence*: one part of an ontology O can be *computationally* derived by aggregation of elements and interactions between elements of an ontology O'.

As developed before, the presence of emergence, and especially weak emergence, will consist in itself in a perspective. It can be conceptual and postulated as an axiom within a thematic theory, but also experimental if clues of weak emergence are effectively measured between objects. In any case, the relation between ontologies must be encoded within an ontology, which was not necessarily introduced in the initial definition of the system.

We make therefore the following assumption for next developments:

¹⁰ what of course can not been presented as a provable claim as it depends on system definition, etc.

¹¹ the third form BEDAU recalls, *Strong emergence* will not be used, as we need only to capture dependance and autonomy, and weak emergence is more satisfying in terms of complex systems, as it does not assume “irreducible causal powers” to objects of upper scales at a given level. Nominal emergence is used to capture inclusion between ontologies.

Hypothèse 5 A system can be partially structured by extending it with an ontology that contains (not necessarily only) relations between elements of ontologies of its perspectives. We name it the coupling ontology and assume its existence in the following. We assume furthermore its atomicity, i.e. if O is in relation with O' , then any subsets of O, O' can not be in relation, what is not restrictive as a decomposition into several independent subsets ensures it if it is not the case.

It allows to exhibit emergence relations not only within a perspective itself but also between elements of different perspectives. We define then pre-order relations between subsets of ontologies:

Proposition 1 The following binary relationships are pre-orders on $\mathcal{P}(\mathcal{O})$:

- *Emergence (based on Weak Emergence):* $O' \preccurlyeq O$ if and only if O weakly emerges from O' .
- *Inclusion (based on Nominal Emergence):* $O' \Subset O$ if and only if O nominally emerges from O' .

Preuve With the convention that it can be said that an object emerges from itself, we have reflexivity (if such a convention seems absurd, we can define the relationships as O emerges from O' or $O = O'$). Transitivity is clearly contained in definitions of emergence.

Note that the inclusion relation is more general than an inclusion between sets, as it translates an inclusion “inside” the elements of the ontology.

These relations are the basis for the construction of a graph called the *ontological graph* :

Définition 4 The ontological graph is constructed by induction the following way:

1. A graph is constricted, with vertices elements of $\mathcal{P}(\mathcal{O})$ and edges of two types: $E_W = \{(O, O') | O' \preccurlyeq O\}$ and $E_N = \{(O, O') | O' \Subset O\}$
2. Nodes are reduced¹² by: if $o \in O, O'$ and $(O' \preccurlyeq O$ or $O' \Subset O)$ but not $(O \preccurlyeq O'$ or $O \Subset O')$, then $O' \leftarrow O' \setminus o$
3. Nodes with intersecting sets are merged, keeping edges linking merged nodes. This step ensures non-overlapping nodes.

¹² the reduction procedure aims to delete redundancy, keeping an entity at the higher level at which it exists.

Minimal Ontological Tree

The topological structure of the graph, that contains in a way the *structure of the system*, can be reduced into a minimal tree that captures hierarchical structure essential to the theory.

We need first to give consistence to the system:

Définition 5 *A consistent part of the ontological graph is a weakly connected component of the graph. We assume for now to work on a consistent part.*

The notion of consistent system, together with subsystem or nodes timescales that will be defined later, requires to reconstruct perspectives from ontological elements, i.e. the inverse operation of what was done in our deconstruction procedure.

Hypothèse 6 *There exists $\mathcal{O}' \subset \mathcal{P}(\mathcal{O})$ such that for any $O \subset \mathcal{O}'$, there exists a corresponding dataflow machine M such that the corresponding perspective is consistent with initial elements of the system (i.e. machines are equivalent on ontology overlaps). If $\Phi : M \mapsto O$ is the initial mapping, we denote this extended reciprocal construction by $M' = \Phi^{<-1>}(O)$.*

REMARQUE This assumption could eventually be changed into a provable proposition, assuming that the coupling ontology indeed corresponds to a coupling perspective, which dataflow machine part is consistent with coupled entities. Therein, the decomposition postulate of [Golden, Aiguier, and Krob, 2012] should allow to identify basic components corresponding to each element of the ontology, and then construct the new perspective by induction. We find however these assumptions too restrictive, as for example various ontological elements may be modeled by an irreducible machine, as a differential equations with aggregated variables. We prefer to be less restrictive and postulate the existence of the reverse mapping on some sub-ontologies, that should be in practice the ones where couplings can be effectively modeled.

Given this assumption, we can define the consistent system as the reciprocal image of the consistent part of the ontological graph. It ensures system connectivity what is a requirement for tree construction.

Proposition 2 *The tree decomposition of the ontological graph in which nodes contains strongly connected components is unique. The reduced tree, that corresponds to the ontological graph in which strongly connected components have been merged with edges kept, is called the Minimal Ontological Tree.*

Preuve (sketch of) The unicity is obtained as nodes are fixed as strongly connected components. It is trivially a tree decomposition as in a directed graph, strongly connected components do not intersect, thus the consistence of the decomposition.

Any loop $O \rightarrow O' \rightarrow \dots \rightarrow O$ in the ontological graph assumes that all its elements are equivalent in the sense of \preccurlyeq . This equivalence loops should help to define the notion of strong coupling as an application of the theory (see applications).

The Minimal Ontological Tree (MOT) is a tree in the undirected sense but a forest in the directed sense. Its topology contains a sort of system hierarchy. Consistent subsystems are defined from the set \mathcal{B} of branches of the forest, as $(\Phi^{<-1>}(\mathcal{B}), \mathcal{B})$. The timescale of a node, and by extension of a subsystem, is the union of timescales of corresponding machines. Levels of the tree are defined from root nodes, and the emergence relations between nodes implies a vertical inclusion between timescales.

Action on Data

De la même manière que les actions de groupes permettent de donner structure à l'utilisation d'un groupe sur un ensemble (généralement de données), une piste de développement puissante serait l'ajout à la théorie de l'aspect essentiel de relation à la réalité par une action des noeuds de l'arbre ontologique sur des ensembles de données. Cette opération est hors de propos pour l'instant car nous n'avons pas encore exploité la structure interne des *dataflow machines*. Une piste, que nous confirmons comme ouverture dans la section suivante 8.3, impliquerait le couplage de ce cadre avec le cadre de connaissances qui y est introduit.

Scales

Finally, we propose to define scales associated to a system. Following [Manson, 2008], an epistemological continuum of visions on scale is a consequence of differences between disciplines in the way we developed in the introduction. This proposition is indeed compatible with our framework, as the construction of scales for each level of the ontological tree results in a broad variety of scales.

Let (M, O) a subsystem and T the corresponding timescale. We propose to define the "thematic scale" (for example spatial scale) assuming a representation theorem, i.e. that an aspect (thematic aspect) of the machine can be represented as a dynamic state variable $\vec{X}(t)$. Assuming a scale operator¹³ $\|\cdot\|_S$ and that the state variable has a certain level of differentiability, the *thematic scale* if defined as $\|(d^k \vec{X}(t))_k\|_S$.

¹³ that can be of various nature: extent, probabilistic extent, spectral scales, stationarity scales, etc.

B.5.3 Application and discussion

The particular case of geographical systems

In [Dollfus and Dastès, 1975] DURAND-DASTÈS proposes a definition of geographical structure and system, structure would be the spatial container for systems viewed as complex open interacting systems (elements with attributes, relations between elements and inputs/outputs with external world). For a given system, its definition is a perspective, completed by structure to have a system in our sense. Depending on the way to define relations, it may be more or less easy to extract ontological structure.

Modularity and co-evolving subsystems

For the example of Urban Systems, urban evolutionary theory enters this framework using our previous thematic theory. The decomposition into uncorrelated subsystems yields precisely strongly coupled components as co-evolving components. The correlation between subsystems should be in a certain way positively correlated with topological distance in the tree. If we define elements of a node before merging as *strongly coupled elements*, in the case of dynamic ontologies, it provides a definition of *co-evolution* and co-evolving subsystems equivalent to the thematic definition.

Discussion

LINK WITH EXISTING FRAMEWORKS A link with the Cottineau-Chapron framework for multi-modeling [Chérel, Cottineau, and Reuillon, 2015] may be done in the case they add the bibliographical layer, which would correspond to the reconstruction of perspectives. [Reymond and Cauvin, 2013] proposes the notion of “interdisciplinary coupling” what is close to our notion of coupling perspectives. A correspondance with System of Systems approaches (see e.g. [Luzeaux, 2015] for a recent general framework englobing system modeling and system description) may be also possible as our perspectives are constructed as dataflow machines, but with the significant difference that the notion of emergence is central.

CONTRIBUTIONS TO THE STUDY OF COMPLEX SYSTEMS We do not claim to provide a theory of systems (beware of cybernetics, systemics etc. that could not model everything), but more a framework to guide research questions (e.g. in our case the direct outcomes will be quantitative epistemology that comes from system construction as perspectives ; empirical to construct robust ontologies for perspectives ; targeted thematic to unveil causal relationship/emergence for construction of ontological network ; study of coupling as possible processes containing co-evolution ; study of scales ; etc.). It may

be understood as meta-theory which application gives a theory, the thematic theory developed before being a specific implementation to territorial networked systems. We emphasize the notion of socio-technical system, crossing a social complex system approach (ontologies) with a description of technical artifacts (dataflow machines), taking the “best of both worlds”.

Reflexivity

We can draw from the construction of this theoretical framework a set of research directions, that give a general line on how trying to answer to research questions asked after the thematic theory construction.

1. The perspectivist approach implies a broad understanding of existing perspectives on a system, and of possibility of coupling between them ; thus an emphasis on applied epistemology, i.e. **Algorithmic Systematic Review** (exploration of the knowledge space), **Disciplines Mapping**(extraction of its structure) and **Datamining for Content Analysis**(refinement at the atomic level in scientific knowledge) that correspond to the three sections of chapter ??.
2. At a finer level of particularization, the knowledge of perspectives means **Knowledge of stylized facts** , i.e. empirical analysis of cases studies. These are the object of chapter ??.
3. The emphasis on coupled subsystems at different scales implies a deep understanding of coupling mechanisms, thus the need of methodological and technical developments : **Methods for Statistical Control**, **Methods for Model Exploration**, **Theoretical Study of Coupling**, **Multi-Modeling**, of which some are developed and other proposed in the methodological chapter ??.
4. Furthermore, the possibility of hidden elements within the ontology implies the test for causal relations and intermediate processes at the origin of emergence, thus e.g. the exploration of new paradigms such as role of governance within complex models as done in chapter ??.
5. Finally, the idea behind system structure contained within the ontological forest is a large set of coupled models for a given system : it means that a proper system definition (i.e. thematic problematization and exploration) and construction should yield to a structured family of models : parallel branches can be different implementations of the same process or various processes trying to explain the emerging ontology ; therefore the final objective of a family of models tackling the thematic question.

C : (Florent) TB : à ce stade peux-tu détailler lesquels vont t'intéresser ?

★ ★

★

B.6 EXPLORATION OF AN INTERDISCIPLINARY SCIENTIFIC LANDSCAPE

Les constructions méthodologiques et techniques rendant possible l’analyse épistémologique de 2.2 ont été menées dans un cadre plus large, notamment débutant avec l’analyse de corpus construits à partir de la revue *Cybergeo*. Nous détaillons ici l’aspect méthodologique des ces analyses.

* * *

*

Le contenu de cette annexe a été élaboré dans le cadre du projet commun d’analyse quantitative des publications de Cybergeo (voir C.4 pour la production commune), initié pour les 20 ans de la revue en mai 2016. Les résultats préliminaires ont été présentés comme [Raimbault, 2016c] à la conférence anniversaire, et le texte de cette annexe est extrait et traduit de [Raimbault, 2017].

* * *

*

Patterns of interdisciplinarity in science can be quantified through diverse complementary dimensions. This paper studies as a case study the scientific environment of a generalist journal in Geography, *Cybergeo*, in order to introduce a novel methodology combining citation network analysis and semantic analysis. We collect a large corpus of around 200,000 articles with their abstracts and the corresponding citation network that provides a first citation classification. Relevant keywords are extracted for each article through text-mining, allowing us to construct a semantic classification. We study the qualitative patterns of relations between endogenous disciplines within each classification, and finally show the complementarity of classifications and of their associated interdisciplinarity measures. The tools we develop accordingly are open and reusable for similar large scale studies of scientific environments.

B.6.1 *Introduction*

We develop in this paper a case study coupling citation network exploration and analysis with text-mining, aiming at mapping the scientific landscape in the neighborhood of a particular journal. We choose to study an electronic journal in Geography, named *Cybergeo*¹⁴, that publishes articles within all subfields of Geography and is in that way multidisciplinary. The choice is initially due to data availability, but ensures several constraints making it highly relevant to the context given above. First of all, the “discipline” of Geography is very broad and by essence interdisciplinary [Bracken, 2016] : the spectrum ranges from Human and Critical geography to physical geography and geomorphology, and interactions between these subfields are numerous. Secondly, bibliographical data is difficult to obtain, raising the concern of how the perception of a scientific landscape may be shaped by actors of the dissemination and thus far from objective, and making technical solutions as the ones we will consequently develop here crucial tools for an open and neutral science. Finally it makes a particularly interesting case study as the editorial policy is generalist and concerned with open science issues such as peer-review ethics transparency (Wicherts, 2016), open data and model practices, as recalled by [Pumain, 2015], and this work contributes to these by fostering the opening of reflexivity.

Our contribution is original and significant on at least two aspects :

1. we combine endogenous classifications in a network multilayer fashion, using semantic information ;
2. a large dataset is constructed from scratch to study a journal not referenced in main databases, tackling both data retrieval and large scale data processing issues.

The rest of the paper is organized as follows : we describe in the next section the dataset used and the data collection procedure. We then study properties of the citation network and describe the procedure to construct the semantic classification through text-mining. We finally study complementary measures of interdisciplinarity obtained with the different classifications.

B.6.2 *Database Construction*

Our approach imposes some requirements on the dataset used, namely:

- (i) cover a certain neighborhood of the studied journal in the citation network in order to have a consistent view on the scientific landscape;
- (ii) have at least a textual description for each node. For these to be

¹⁴ <http://cybergeo.revues.org/>

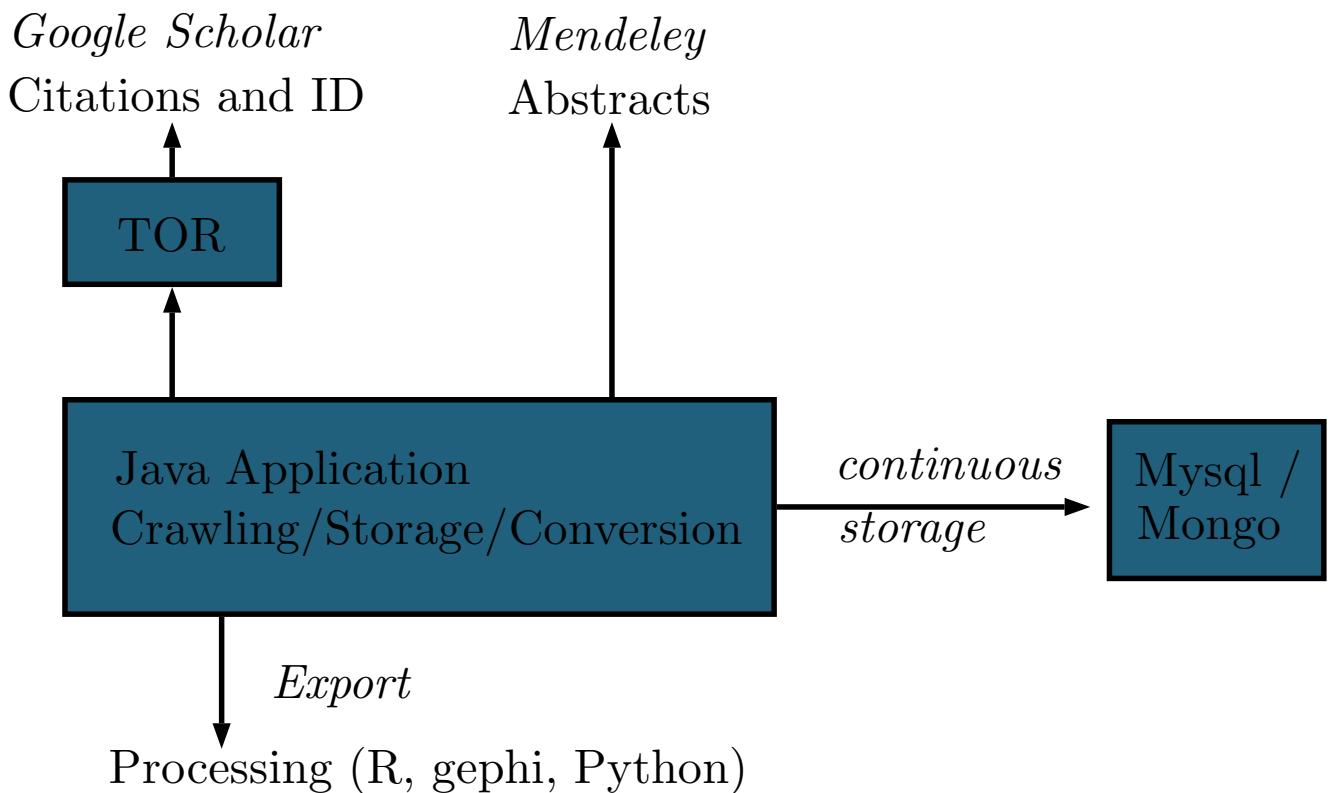


FIGURE 110: Heterogeneous Bibliographical Data Collection and processing. Architecture of the application for content (semantic data), metadata and citation data collection. The heterogeneity of tasks requires the use of multiple languages : data collection and management is done in Java, and data stored in databases (Mysql and MongoDB) ; data processing is done in python for Natural Language Processing and in R for statistical and network analyses; graph visualizations are done with Gephi software.

met, we need to gather and compile data from heterogeneous sources. We use therefore an application specifically designed, which general architecture is given in Fig. 110. Source code of the application and all scripts used in this paper are available on the open git repository of the project¹⁵. Raw and processed data are also openly available on Dataverse¹⁶. We recall that an important contribution of this paper is the construction of such an hybrid dataset from heterogeneous sources, and the development of associated tools that can be reused and further developed for similar purposes.

Initial Corpus

The production database of *Cybergeo* (snapshot taken in February 2016, provided by the editorial board), provides after pre-processing the initial database of articles, with basic information (title, abstract, publication year, authors). The processed version used is available to-

¹⁵ at <https://github.com/JusteRaimbault/HyperNetwork>

¹⁶ at <http://dx.doi.org/10.7910/DVN/VU2XKT>

gether with the full database constructed, as a mysql dump, at the address given above. This base provide also bibliographical records of articles that give all references cited by the initial base (*forward citations* for the initial corpus).

Citation Data

Citation data is collected from Google Scholar, that is the only source for incoming citations (Noruzi, 2005) in our case as the journal is poorly referenced in other databases¹⁷. We are aware of the possible biases using this single source (see e.g. [Bohannon, 2014])¹⁸, but these critics are more directed towards search results or possible targeted manipulations than the global structure of the citation network. The automatic collection requires the use of a crawling software to pipe requests, namely TorPool (Raimbault, 2016e) that provides a Java API allowing an easy integration into our application of data collection. A crawler can therethrough retrieve html pages and get backward citation data, i.e. all citing articles for a given initial article. We retrieve that way two sub-corporuses: references citing papers in *Cybergeo* and references *citing the ones cited* by *Cybergeo*. At this stage, the full corpus contains around $4 \cdot 10^5$ references.

For the sake of simplicity, we will denote by *reference* any standard scientific production that can be cited by another (journal paper, book, book chapter, conference paper, communication, etc.) and contains basic records (title, abstract, authors, publication year). We work in the following on networks of references, linked by citations.

Text Data

A textual description for all references is necessary for a complete semantic analysis. We use for this an other source of data, that is the online catalog of Mendeley reference manager software [Mendeley, 2015]. It provides a free API allowing to get various records under a structured format. Although not complete, the catalog provides a reasonable coverage in our case, around 55% of the full citation network. This yields a final corpus with full abstracts of size $2.1 \cdot 10^5$. The structure and descriptive statistics of the corresponding citation network is recalled in Fig. 111.

b.6.3 Methods and Results

Citation Network Properties

PROPERTIES As detailed above, we are able by the reconstruction of the citation network at depth ± 1 from the original 927 references

¹⁷ or was just added as in the case of *Web of Science*, indexing *Cybergeo* since May 2016 only

¹⁸ or <http://iscpif.fr/blog/2016/02/the-strange-arithmetic-of-google-scholars>

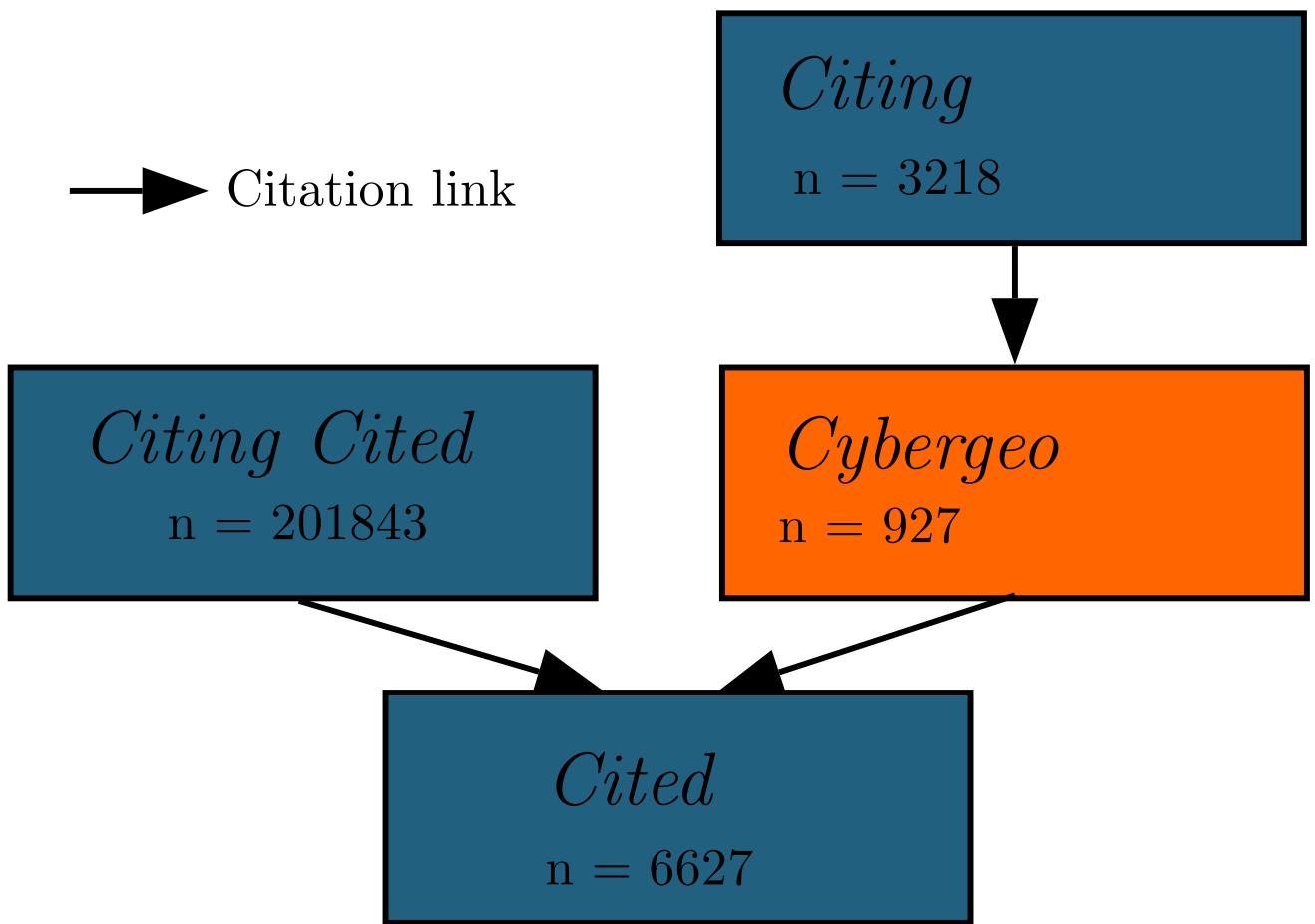


FIGURE 111: **Structure and content of the citation network.** The original corpus of *Cybergeo* is composed by 927 articles, themselves cited by a slightly larger corpus (yielding a stationary impact factor of around 3.18), cite $\simeq 6600$ references, themselves co-cited by more than $2 \cdot 10^5$ works for which we have a textual description.

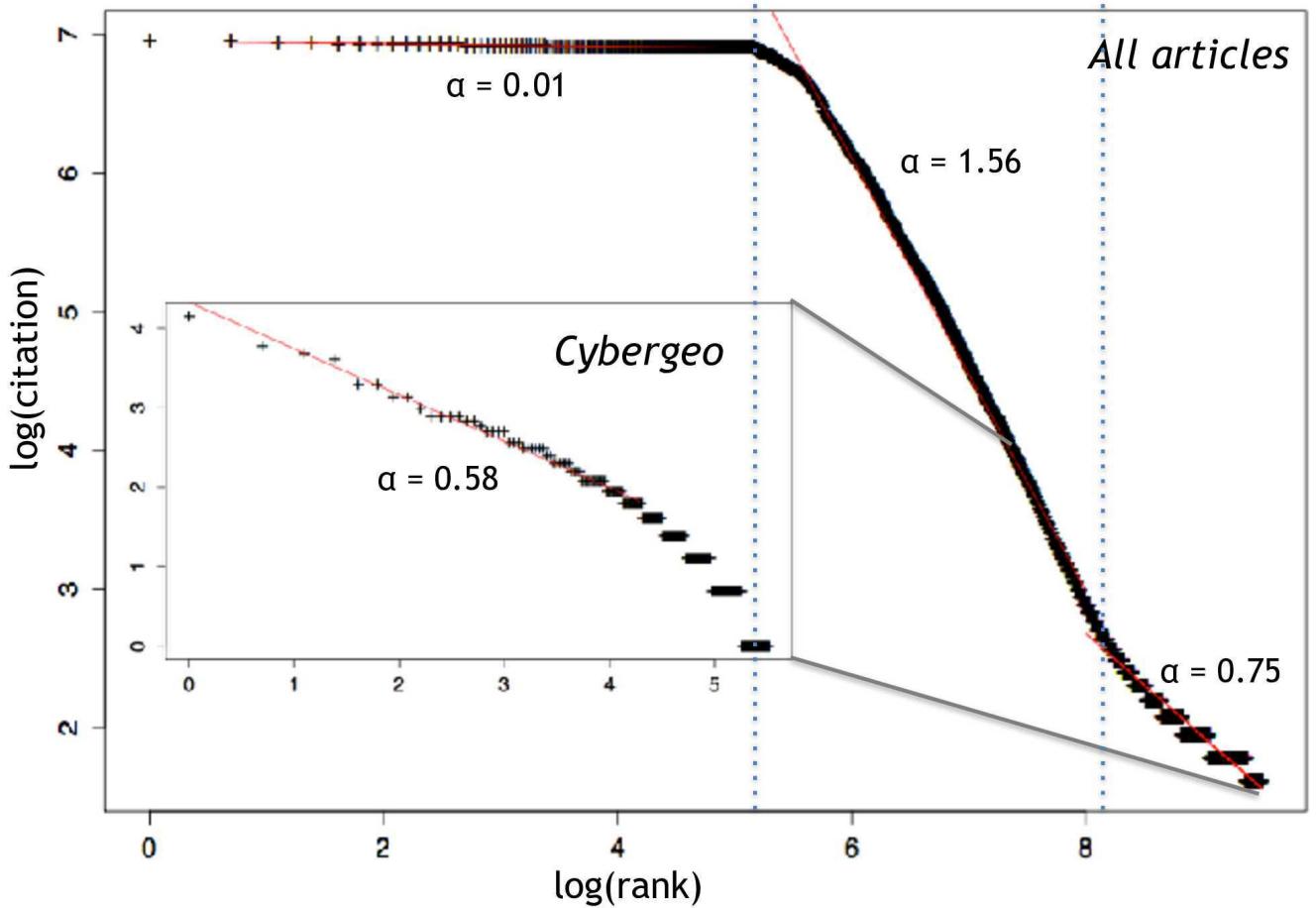


FIGURE 112: Rank-size plot of citations received. The plot unveils three superposed citations regimes, corresponding to power laws with different levels of hierarchy. The references in *Cybergeo* (inset plot) are themselves in the tail and less hierarchical.

of the journal to retrieve around $4 \cdot 10^5$ references, on which $2.1 \cdot 10^5$ have an abstract text allowing semantic analysis. A first glance on citation network properties provides useful insights. Mean in-degree (that can be interpreted as a stationary integrated impact factor) on references for which it can be defined has a value of $\bar{d} = 121.6$, whereas for articles in *Cybergeo* we have $\bar{d} = 3.18$. This difference suggests a variety for status of references, from old classical works (the most cited has 1051 incoming citations) to recent less influential works.

This diversity is confirmed by the hierarchical organisation examined in Fig. 112 that unveils three superposed regimes. More precisely, we look at the rank-size plot, given by the logarithm of the number of citations received as a function of the rank of the paper. We find, as expected (Redner, 1998), localized power-law behaviors. A first set of around 150 references shows a very low hierarchy (rank-size ex-

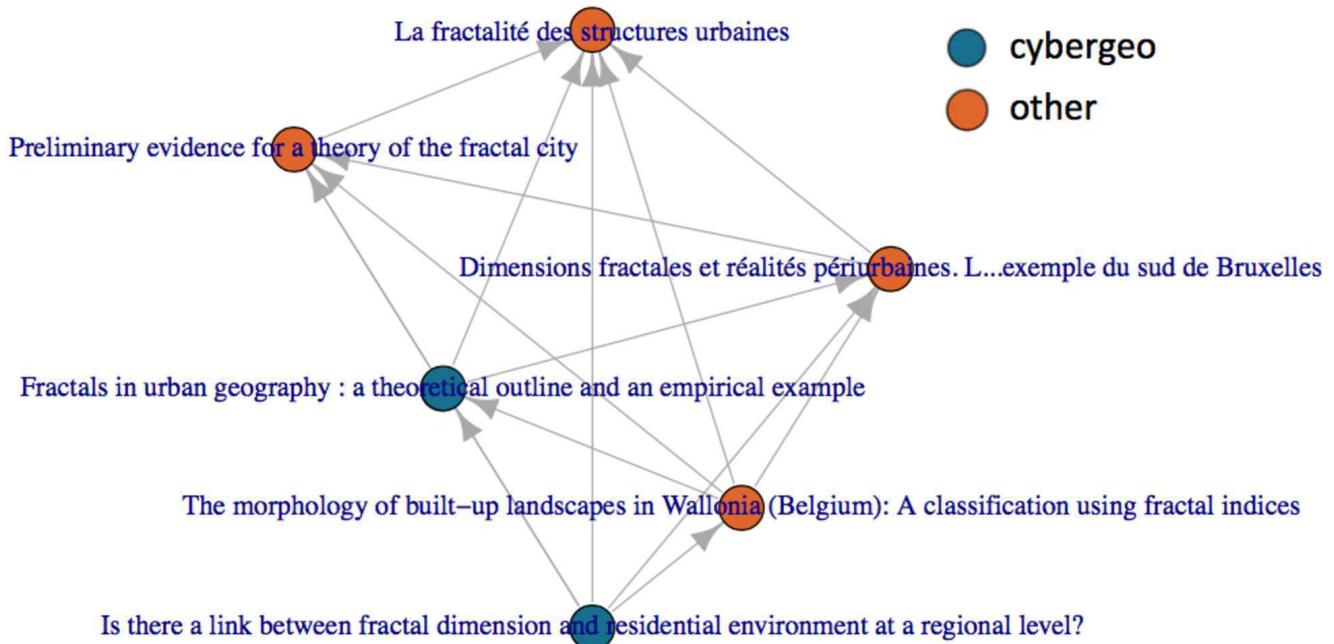


FIGURE 113: Example of a maximal clique in the citation network, paper of Cybergeo being in blue. Such topological structure reveal citation practices such as here a systematic citation of previous works in the research niche.

ponent $\alpha = 0.01$) and corresponds to classical references in different disciplines. A second regime ($\alpha = 1.56$) is much more hierarchized, followed by a last regime less hierarchical ($\alpha = 0.75$) containing more recent papers (average publication year mid-2005, against mid-1998 for the second and 1983 for the first).

Other topological properties reveal typical patterns of citation practices: for example, the existence of high-order cliques (complete sub-networks) implies citation practices which compatibility with the cumulative nature of knowledge may be questionable [Pumain, 2005], since these need always to source back the production of knowledge in the most recent works. An exemple of such a clique in shown in Fig. 113.

CITATION COMMUNITIES The citation network is a first opportunity to construct endogenous disciplines, by extracting citation communities. More precisely, this step aims at finding recurrent patterns in citations that would define a field by its citation practices. In order to be consistent with the particular data structure we have (missing incoming citations for sub-corpora at maximal depth), we filter the network by removing all nodes with degree smaller than one. This ensures that kept nodes are either at least cited by an other node (and thus there are no missing edges for these nodes) or cite at least two other nodes, what can make “bridges” between sub-

communities. The resulting network has a size of $|V| = 107164$ nodes and $|E| = 309778$ edges. It is visualized in Fig. 114.

We use a standard modularity optimization algorithm to identify communities (Blondel et al., 2008a) in this citation network. It provides 29 communities with a modularity of 0.71. In comparison, a bootstrap of 100 randomisations of links in the network gives an average modularity of $-1.0 \cdot 10^{-4} \pm 4.4 \cdot 10^{-4}$ which means that communities are highly significant.

We name the communities by inspection of the titles of most cited references in each. The 14 communities that have a size larger than 2.5% of the network are : Complex Networks, Ecology, Social Geography, Sociology, GIS, Spatial Analysis, Agent-based Modeling and Simulation (ABMS), Socio-ecology, Urban Networks, Urban Simulation, Urban Studies, Economic Geography, Accessibility/Land-use, Time Geography. These categories do not directly correspond to well-defined disciplines, as some correspond more to methods (ABMS), objects of study (Urban Studies), or paradigms (Complex Networks). Some are “specializations” of others : most papers in Urban Studies can also be classified as Critical and Social geography. This way, we construct endogenous disciplines that correspond to *scientific practices* (what is cited) more than their representation (the “official” disciplines). The relative positioning of communities in Fig. 114, obtained with a Force-Atlas algorithm, tells a lot about their respective relations : for example, social geography makes a bridge between Urban Studies and Economic Geography, whereas the connection between Socio-ecology and Urban simulations is done by GIS (what can be expected as geomatics is an interdisciplinary field). GIS also separates and connects two subfield of Ecology, on one side more thematic studies on ecological habitats, and on the other sides statistical methods. These relations already inform qualitatively patterns of interdisciplinarity, in the sense of integration measures. We will also in the following use these communities to situate the semantic classification.

Semantic Communities Construction

We now turn to the methodological details for the construction of the semantic classification. This step adapts the methodology described by [Bergeaud, Potiron, and Raimbault, 2017a], who construct a semantic classification on patent data.

RELEVANT KEYWORDS EXTRACTION We recall that our corpus with available text consists of around $2 \cdot 10^5$ abstracts of publications at a topological distance shorter than 2 from the journal *Cybergeo* in the citation network. The first important step is to extract relevant keywords from abstracts. Text processing is done with the python library `nltk` (Bird, 2006). We add a particular treatment to

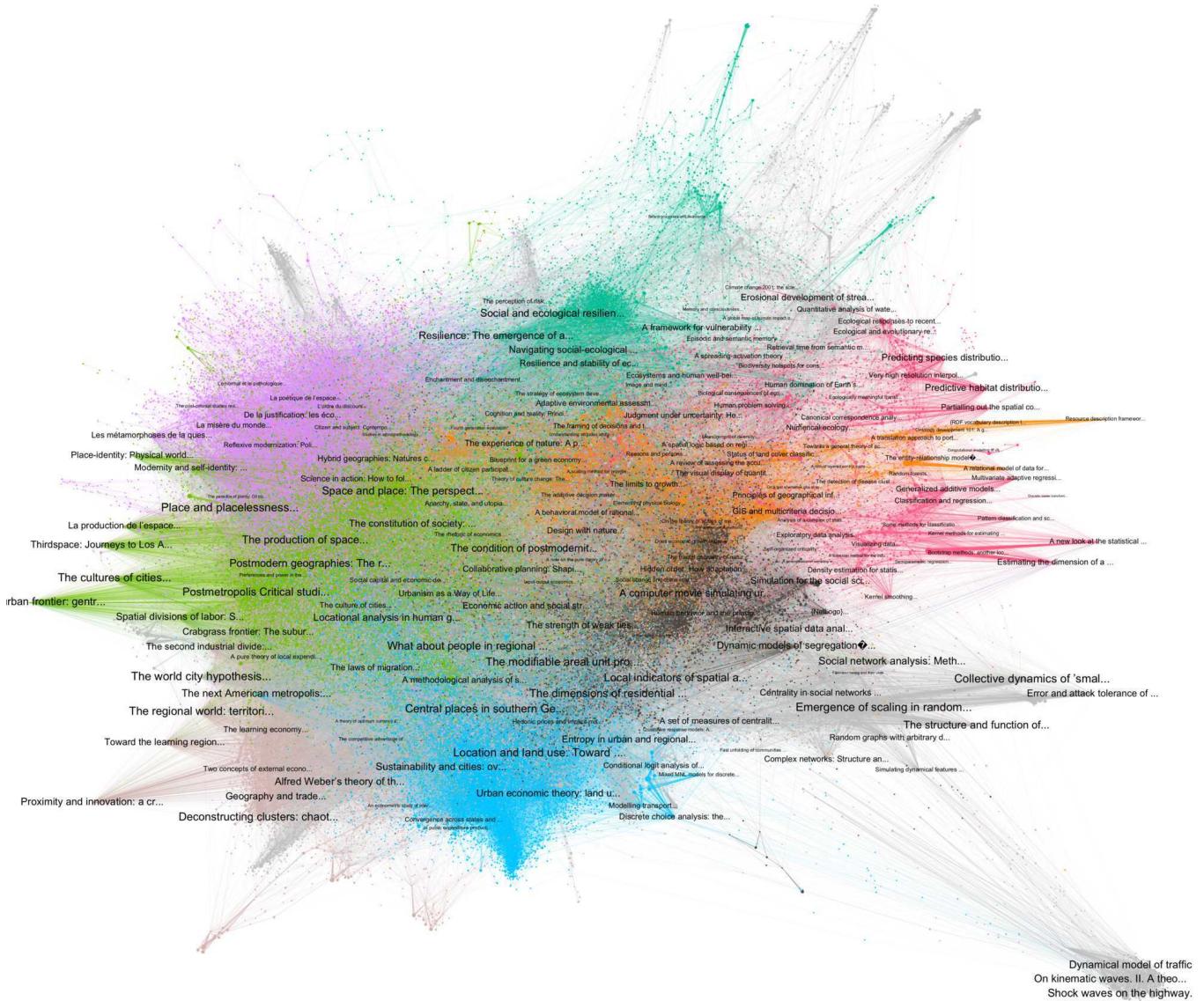


FIGURE 114: Citation Network. We show only the “core” of the citation network, composed by references with a degree larger than one ($|V| = 107164$ and $|E| = 309778$). The community detection algorithm provides 29 communities with a modularity of 0.71. Nodes and edges color gives the main communities (for example ecology in magenta, GIS in orange, Socio-ecology in turquoise, Social geography in green, Spatial analysis in blue). Node labels give shortened titles of most cited papers, size is scaled according to their in-degree. The graph is spatialized using a Force-Atlas algorithm.

the method of [Bergeaud, Potiron, and Raimbault, 2017a], as our corpus is multilingual: language detection is done with the technique of *stop-words* (Baldwin and Lui, 2010). We also use a specific tagger (the function allowing the attribution of grammatical function to words), TreeTagger (Schmid, 1994), for languages other than English.

To summarize, the keyword extraction workflow goes through the following steps :

1. Language detection is done using *stop-words*
2. Pos-tagging (detection of word functions) and stemming (extraction of the *stem*) are done differently depending on language :
 - English : `nltk` built-in pos-tagger, combined to a *Porter-Stemmer*
 - French or other : use of TreeTagger (Schmid, 1994)
3. Selection of potential *n-grams* (keywords of length n with $1 \leq n \leq 4$) following the given grammatical rules: for English $\cap\{\text{NN} \cup \text{VBG} \cup \text{JJ}\}$, and for French $\cap\{\text{NOM} \cup \text{ADJ}\}$. Other languages are a negligible proportion of the corpus and are discarded.
4. Estimation of the relevance *n-grams*, by attributing a score following the deviation of the statistical distribution of co-occurrences to a random distribution.

SEMANTIC NETWORK We keep at this stage a fixed number K_W of *n-grams*, based on their relevance score, that will be designated as the relevant keywords. We find that for large values of K_W , results are not sensitive to the total number of keywords, and take a reasonably large value for computational performance, $K_W = 50,000$. We construct the co-occurrence matrix of the relevant keywords. This co-occurrence matrix provides the semantic network as its adjacency matrix : nodes are keywords, and they are linked according to their co-occurrences.

SENSITIVITY ANALYSIS We observe the same phenomenon than in [Bergeaud, Potiron, and Raimbault, 2017a], that is the existence of nodes with large degree and not specific to a particular field : for example `model` and `space` are used in most of subfields of Geography. We also adapt the original filtering procedure, as we do not have here an exogenous information to calibrate parameters. We assume the highest degree terms do not carry specific information on particular classes and can be thus filtered given a maximal degree threshold k_{\max} . We keep the second filter on a minimal edge weight threshold θ_w . We add the supplementary constraint that keywords are also filtered on a document frequency window $[f_{\min}, f_{\max}]$ (number of references in which they appear), what is slightly different from network filtering.

A sensitivity analysis of resulting network topology to these four parameters is presented in Fig. 115. Given a filtered network, we detect communities using modularity optimization as before for the citation network. Various properties of the network can be optimized, and we look in particular at its size (number of keywords after filtering), the optimal modularity, the number of communities, and the balance between their sizes (defined as a concentration index $\sum_k s_k^2 / (\sum_k s_k)^2$). This multi-objective optimization problem does not have a unique solution as objectives are contradictory in a complex way, and a compromise point must be chosen. We take a compromise point between modularity and network size, with a high balance and a reasonable number of communities, given by $k_{\max} = 1200$, $\theta_w = 100$, $f_{\min} = 50$, $f_{\max} = 10000$. These values give a network of size 2868, with 18 communities and a modularity of 0.57.

Note that the small proportion of keywords in French is always separated from the rest of the network as they cannot co-occur with English keywords, and that with these parameter settings no French keywords are kept. All communities described in the following therefore contain only keywords in English.

SEMANTIC COMMUNITIES We obtain therein communities in the semantic network with the optimized filtering parameters. At the exception of a small proportion apparently resulting from noise (representing less than 10 keywords in 3 communities that we remove, i.e. 0.33% of keywords), communities correspond to well-defined scientific fields, domains, or approaches. Naming is also done by inspection of the most relevant keywords in each community, in order to stick here to a certain level of supervision.

Table 116 summarizes the communities, giving their names, sizes, and corresponding keywords. The most important community is related to issues in political science and critical geography, what could have been expected as several previously obtained citations communities (Social geography, Urban studies) deal with these issues. We then obtain a large cluster of terms related to biogeography, that must correspond to publications in Ecology and Socio-ecology identified before, together with a community in Environment and Climate.

In a way similar to the citation communities, but more pronounced here, we obtain endogenous “disciplines” that can correspond to real disciplines, to methodologies, to object of studies. This classification thus also unveil *effective scientific practices*, here in terms of semantic content. A class here related to complex systems can be associated to a paradigm and various approaches that were separated in the citation communities : agent-based models and complex networks for example. On the contrary, some studies that were gathered in a large domain before can be precisely differentiated in the semantic network, such as microbiology and health here that are used by stud-

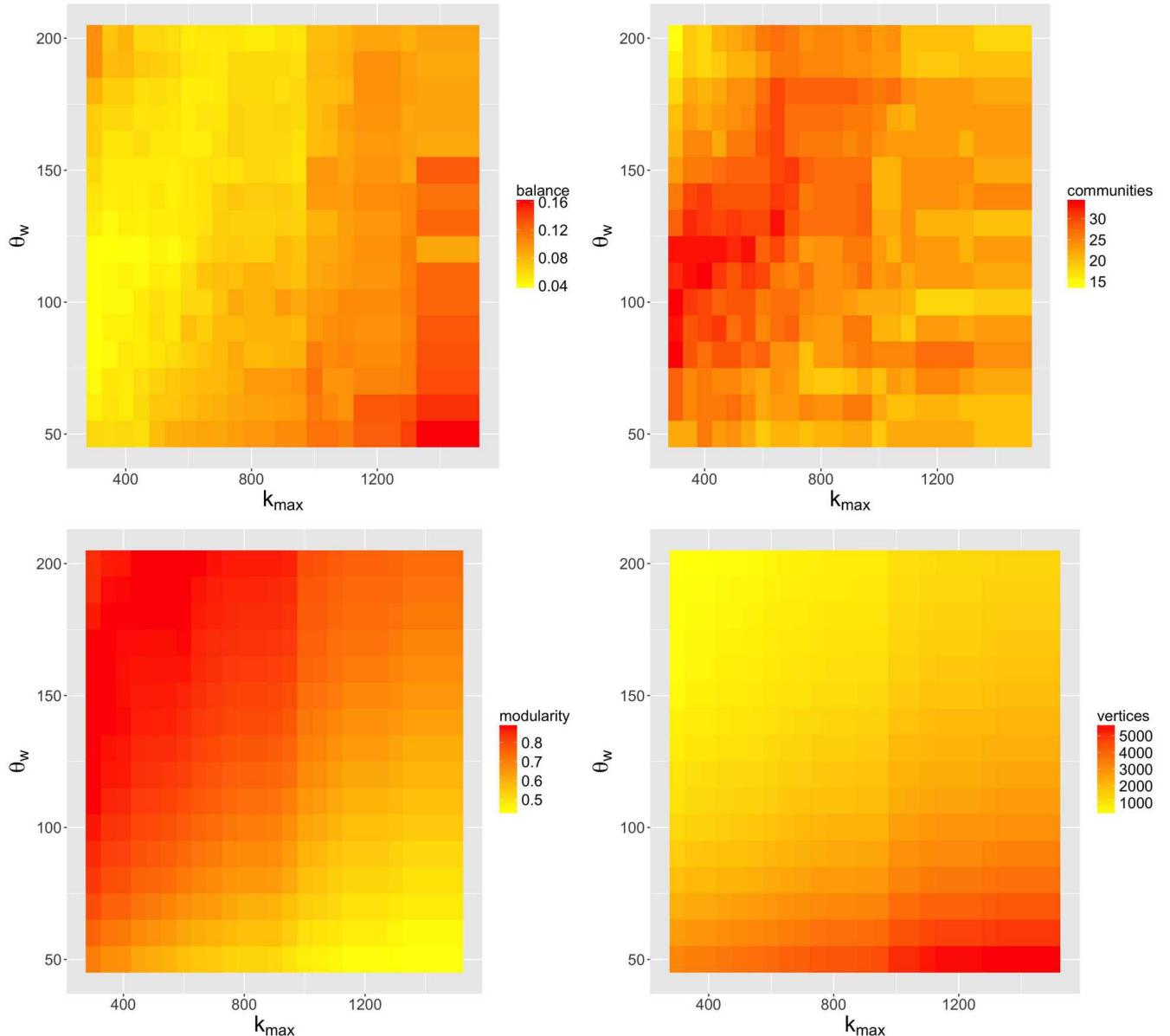


FIGURE 115: Sensitivity analysis of network indicators to filtering parameters. We show here 4 indicators (balance between community sizes, modularity of the decomposition, number of communities, number of vertices), as a function of parameters k_{\max} and θ_w , at fixed $f_{\min} = 50, f_{\max} = 10000$. Close values for these two last parameters (in a reasonable range) give similar behavior.

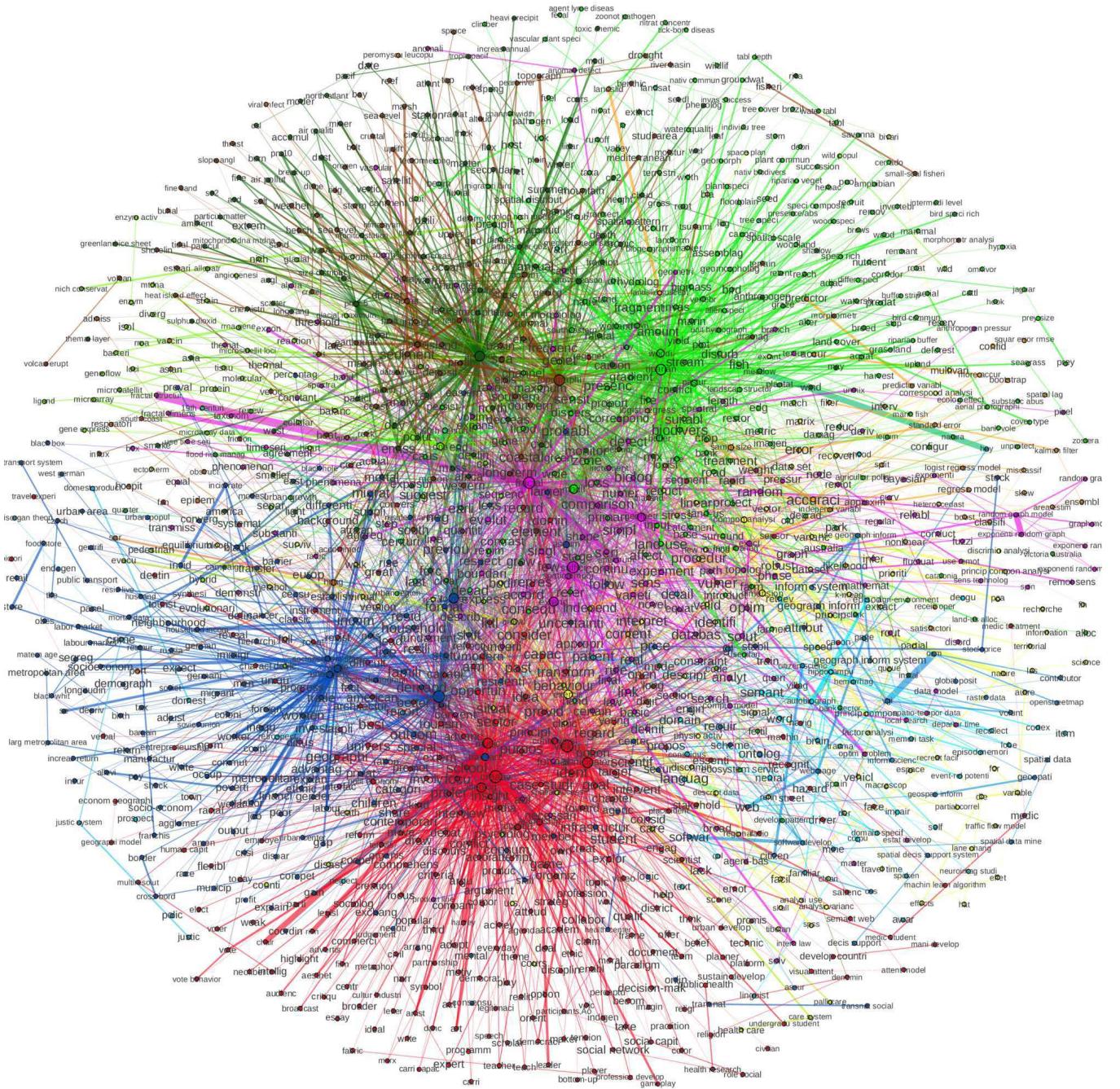


FIGURE 117: Visualization of the semantic network. Network is constructed by co-occurrences of most relevant keywords. Filtering parameters are here taken according to the multi-objective optimization done in Fig. 115, i.e. ($k_{\max} = 1200, \theta_w = 100, f_{\min} = 50, f_{\max} = 10000$). The graph spatialization algorithm (Fruchterman-Reingold), despite its stochastic and path-dependent character, unveils information on the relative positioning of communities.

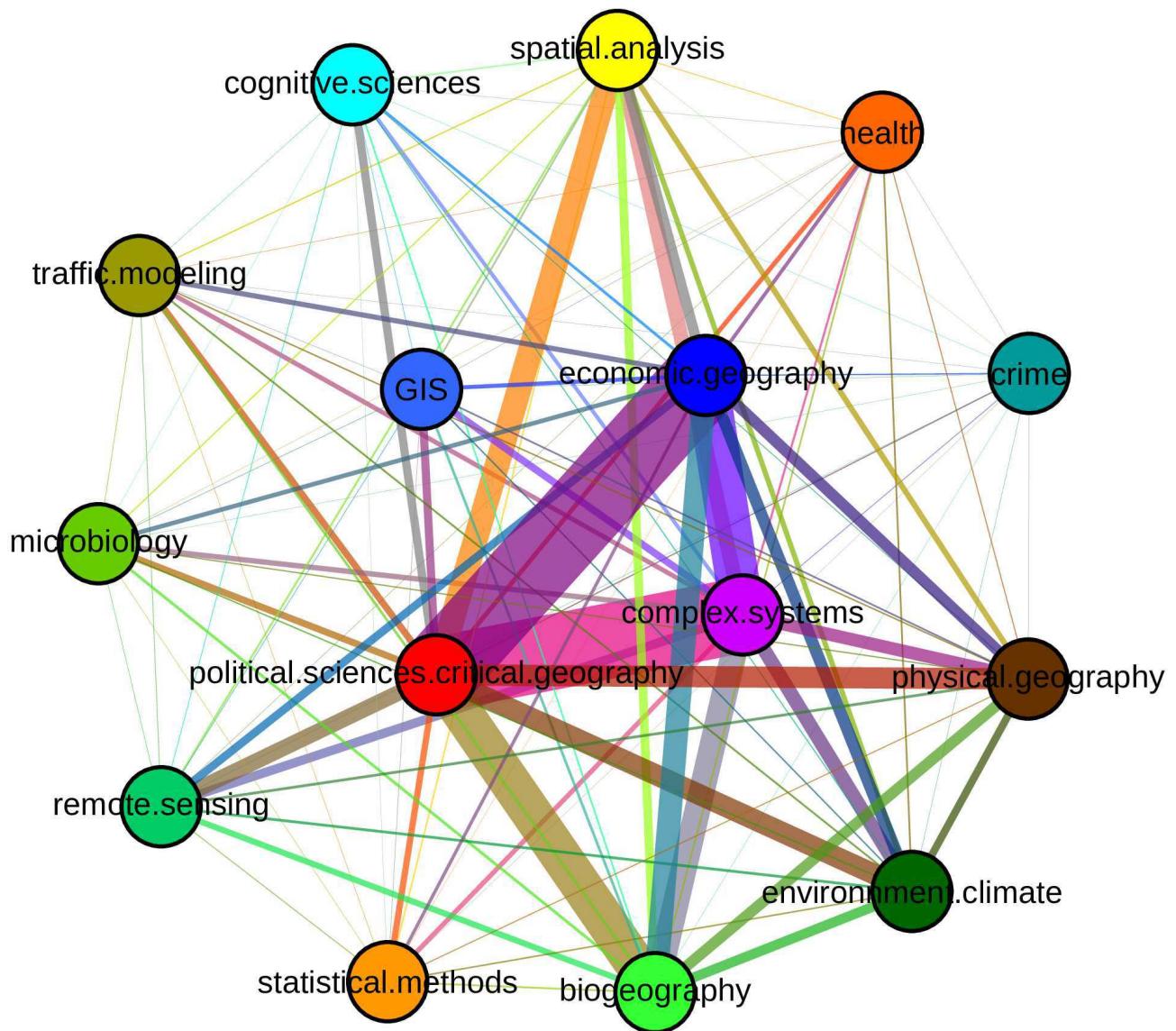


FIGURE 118: Synthesis of semantic communities and their links. Weights of links are computed as probabilities of co-occurrences of corresponding keywords within references.

FIGURE 116: Semantic communities reconstructed from community detection in the semantic network.

Name	Size	Keywords
Political sciences/critical geography	535	decision-mak, polit ideolog, democraci, stakehold
Biogeography	394	plant densiti, wood, wetland, riparian veget
Economic geography	343	popul growth, transact cost, socio-econom, household
Environment/climate	309	ice sheet, stratospher, air pollut, climat model
Complex systems	283	scale-fre, multifract, agent-bas model, self-organ
Physical geography	203	sedimentari, digit elev model, geolog, river delta
Spatial analysis	175	spatial analysi, princip compon analysi, heteroscedast
Microbiology	118	chromosom, phylogeneti, borrelia
Statistical methods	88	logist regress, classifi, kalman filter, sampl size
Cognitive sciences	81	semant memori, retrospect, neuroimag
GIS	75	geograph inform scienc, softwar design, volunt gi
Traffic modeling	63	simul model, lane chang, traffic flow, crowd behavior
Health	52	epidem, vaccin strategi, acut respiratori syndrom
Remote sensing	48	land-cov, landsat imag, lulc
Crime	17	crimin justic system, social disorgan, crime

ies related to socio-ecology or ecology in the citation network. Some very specific domains appear here as they have very few connections in their actual semantic content : for example, Geography of crime is very precise and disconnected from other communities.

We show in Fig. 117 a visualisation of the semantic network, in which the positioning of communities, induced by a Fruchterman-Reingold algorithm (that we use here to have a more precise layout in the relative positioning compared to Force Atlas (Jacomy et al., 2014)). The bridging between distant disciplines is done quite differently compared to the citation network, and reveals thus qualitatively an other dimension of interdisciplinarity, i.e. the semantics shared by disciplines. Here, the communities corresponding to Economic Geography (blue) and to Critical Geography (red) are close as in the citation network, but are linked to ecology and geomorphology (green and brown) by Complex Systems (magenta), although these were not present as a community in the citation network. Complexity methodologies such as Fractals, Scaling (West, 2017) or Networks (Newman, 2003) are indeed widely used both in social sciences and in physics or biology. The semantic analysis reveals thus that very distant disciplines, that are distant in their citation patterns, are finally close in terms of actual content.

In terms of overlaps between communities, in the sense of co-occurrences of corresponding keywords within texts of references, we show a

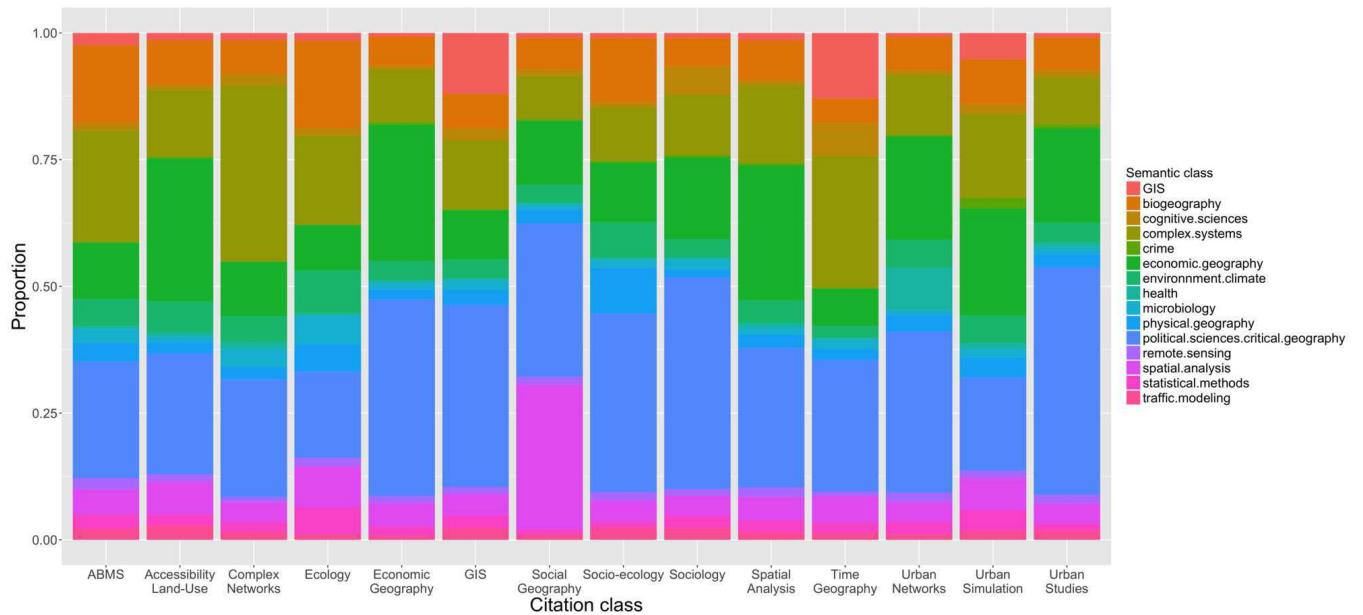


FIGURE 119: Composition of citation communities in terms of semantic content. For each citation class (horizontally), the bar is decomposed as the proportions of each semantic class (given by color).

synthesis of links between semantic communities in Fig. 118. We see that communities such as Critical Geography and Biogeography are not totally disconnected and share still a certain number of co-occurrences. More isolated communities can be spotted such as Health and Crime Geographies. Surprisingly, Statistical Methods does not share strong links with other communities, what could mean that articles dealing with methodological issues in this field are rather disconnected from the field of application, or at least do not describe it extensively. On the contrary, methods in Complex Systems are organically integrated with the thematic issues they tackle.

Semantic composition of citation communities

We can now turn to the study of the relation between classifications. First, a simple way to link them is to look at the semantic content of citation communities. Each reference has a given proportion of keywords within each semantic class, and an average composition in terms of semantic classes for each citation class can thus be computed. We show these compositions in Fig. 119. Some expected results are obtained, such as Complex Networks (citation) having the largest part in Complex Systems (semantic), or GIS (citation) the largest in GIS (semantic), and similarly for Economic Geography.

But the study of patterns that could have not been expected is very informative, and unveils practices of interdisciplinarity. For example, Time Geography (citation) uses as much GIS (semantic) as GIS (citation), what means that they should be using the corresponding

methods and tools to study the thematic question of spatio-temporal trajectories of geographical agents. The most important in terms of political science (semantic) are Urban Studies, what suggest a convergence of the City as an object of study and of the disciplines of Political Science and Critical Geography. Also interestingly, the citation communities using most biogeography are Ecology (what could have been expected) and ABMS, confirming again the role of the thematic application in complex systems methodologies.

Measuring interdisciplinarity

We had up to now a qualitative view on interdisciplinarity patterns, by looking at the relative localisation of communities within the citation and semantic classifications, and the relation between the classifications. We propose now to look at quantitative measures of interdisciplinarity, for each classification.

More precisely, for a given classification $C \in \{\text{Citation}, \text{Semantic}\}$ a reference i can be viewed as a probability vector $(p_{ij}^{(C)})_j$ on classes j that give for each class the probability to belong to it. Given this setting, we measure interdisciplinarity of one reference using Herfindhal concentration index (Porter and Rafols, 2009), that can also be called an originality index. We define originality as

$$o_i^{(C)} = 1 - \sum_j p_{ij}^{(C)}{}^2$$

For the semantic classification, probabilities are defined as the proportion of keywords of the abstract within each semantic class. With the deterministic citation classification, each reference has only one class and the originality index is always 0. Therefore in order to be able to compare the two classification, we associate a probability to each citation class for each article as the proportion of citations received from this class. The induced index is original, and measures interdisciplinarity as *how a reference is used* by different disciplines in its lifetime.

We show in Fig. 120 the statistical distribution for both indexes $o^{(\text{Semantic})}$ and $o^{(\text{Citation})}$, stratified by citation class. This allow a direct comparison between the two and also an indirect comparison by the variation of semantic distribution between citation classes. For the distribution of semantic originalities, all citation classes exhibit a similar pattern, that is a peak around large values and a smaller peak at zero. It means that either references are highly specialized and have keywords in one class only, or they use keywords from different classes in a quite even manner (for comparison, an abstract with half keywords in a class and half in an other gives an originality of 0.5). The most original, i.e. the most mixed, citation class, is Complex Networks, with a distribution clearly detached from others, what would confirm their use as a method with a lot of different problems. Social

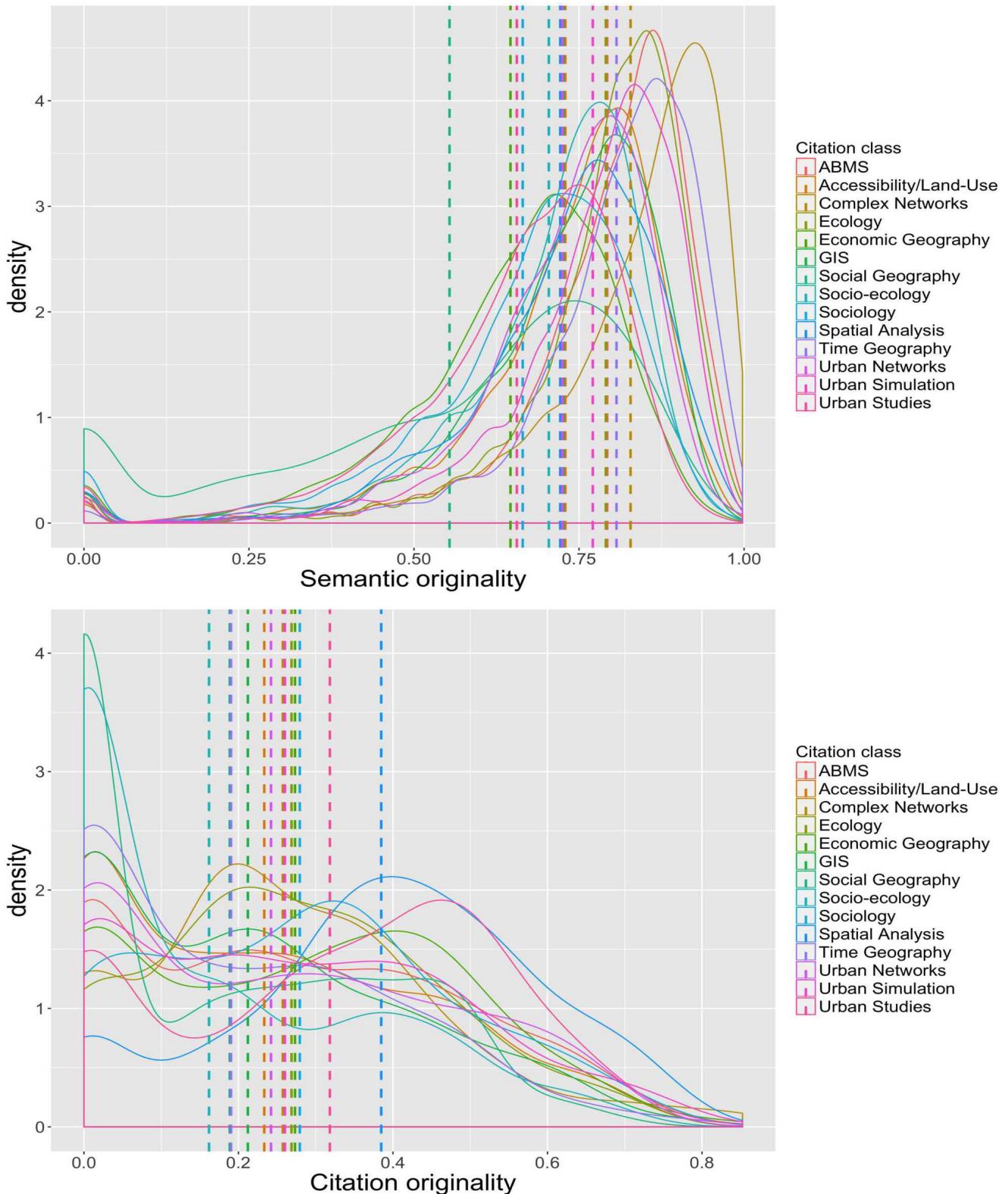


FIGURE 120: Statistical distribution of originalities. We show the smoothed probability densities of originality indexes, by citation class (given by color), for the Semantic originality $\sigma^{(\text{Semantic})}$ (top plot) and for the Citation originality $\sigma^{(\text{Citation})}$ (bottom plot). Dashed lines give the mean for each distribution, with the corresponding color.

Geography is from far the less original, with a large number of single class references, and an average far lower than other classes, what would mean an increased presence of compartmentalization within the associated disciplines.

In terms of citation originality index, the global picture is fundamentally different, as average originality indexes are all lower than 0.4 and most of distributions show their mode in 0, meaning that most references are only cited by their own citation class. Again, Social Geography is the less original, confirming a similar behavior in terms of citation practice than in terms of research content. The most original classes in average, with a peak in large values, are Spatial Analysis and Urban Simulation: this corresponds to the fact that these class feature quite generic methods that can be applied in several fields and are cited accordingly. Complex Networks do not reach the same level, but however exhibit a peak around 0.2 and no peak in 0, together with Ecology, suggesting disciplines having still significant impact in other disciplines.

To summarize, we show (i) different patterns of interdisciplinarity, depending on disciplines, in terms of scientific content (semantic) and of scientific impact (citation); and (ii) a strong qualitative difference in behavior of originalities between the two classifications, what suggests their complementarity.

Correlation between classifications

In order to strengthen the idea of a complementarity of classifications, that would each capture different dimensions of processes of knowledge production, we finally look at the correlation matrix between classifications. We use this time effective class probabilities for the citation classification, i.e. a vector of zeros except with a one at the index of the class of the reference. We compute a Pearson correlation coefficient between classes k (in semantic) and k' (in citation) as

$$\rho_{k,k'} = \frac{\text{Cov}[(p_{ik}^{(\text{Sem})})_i, (p_{ik'}^{(\text{Cit})})_i]}{\sqrt{\text{Var}[(p_{ik}^{(\text{Sem})})_i] \text{Var}[(p_{ik'}^{(\text{Sem})})_i]}}$$

where the covariance is estimated with the unbiased estimator.

The structure of the correlation matrix recalls the conclusions obtained when studying the semantic composition of citation communities, such as GIS being strongly correlated with GIS ($\rho = 0.26$), or Sociology with Political Science ($\rho = 0.16$). More importantly for our question are summary statistics of the overall matrix. It has a minimum of -0.16 (Ecology (citation) against Political Sciences (semantic)), an average of -0.002 and a maximum of 0.33 (Social geography (citation) and Spatial Analysis (semantic)). The “high” values are highly skewed, as the first decile is at -0.06 and the last at 0.09 , what

means that 80% of coefficient lie within that interval, corresponding to low correlations. In a nutshell, classifications are consistent as highest correlations are observed where one can expect them, but most of classes are uncorrelated, meaning that the classifications are quite orthogonal and therefore complementary.

b.6.4 Discussion

We have this way shown the complementarity of classifications in the qualitative patterns they unveil, but also quantitatively in terms of interdisciplinarity measures and quantitatively in terms of correlations. Our work can be extended regarding several aspects, of which we give some suggestions below.

Further Developments

A first development consists in the comparison of journals. The starting point for construction of the scientific environment, the journal *Cybergeo*, was the entry point but not the subject of our study. A development more focused on journals, trying for example to answer comparative issues, or to classify journals according to their effective level of interdisciplinarity regarding different dimensions, would be potentially interesting. The collection of precise data on the origin of references is however a first step that need to be solved first.

The performance of the semantic classification was also not quantified here. A further validation of the relevance of using complementary information contained in both classifications could be done by the analysis of modularities within the citation network, as done in [Bergeaud, Potiron, and Raimbault, 2017a]. This would however require a baseline classification to compare with, which is not available in the type of data we use. Open repository such as arXiv (for physics mainly) or Repec (for Economics) provide API to access metadata including abstracts, and could be starting points for such targeted case studies.

Applications

A first potential application of our methodology relies on the facts that both classifications unveils thematic domains (objects of study), classical disciplines, methodological communities. These different types of communities can indeed be understood as different *Knowledge Domains*. [Raimbault, 2017c] postulates co-evolving Knowledge Domains in every process of scientific knowledge production, that are Theoretical, Empirical, Modeling, Methodology, Tools and Data domains. Most of them are necessary for any process, and investigations within one conditions the advances in others. A refinement of classifications, associated with supervised classification to associate knowl-

edge domains to some communities (potentially using full texts to have more precise information on the proportion of each knowledge domains involved in each), would allow to quantify relations between domains. Furthermore, using temporal data with the date of publications, would yield an effective quantification of the *co-evolution* of domains in the sense of patterns of temporal correlations (e.g. Granger causality).

An other interesting direction is the application of our classifications to the quantification of spatial diffusion of knowledge, as [Maisonobe, 2013] does for the diffusion of a specific question in genetics. It is not clear if different dimensions of knowledge diffuse the same way: for example citation practices can be correlated to social networks and thus exhibit different patterns than effective research contents. Therefore, our work would allow to study such questions from complementary point of views.

Finally, we believe the tool we developed can contribute to an increased empowerment of authors and to the development of open science practices. Among the various visions of Open Science (Fecher and Friesike, 2014), the opening of data is always an important aspect, together with a development of reflexivity in all disciplines, beyond the sole Social Sciences to which it is classically associated. The first point is dealt with by our open tools for dataset construction, whereas the second is implied by the new knowledge of the different dimensions of the scientific environment we studied.

B.6.5 Conclusion

We have introduced a multi-dimensional approach to the understanding of interdisciplinarity, based on citation network and semantic network analysis. Starting from a generalist journal in Geography, we construct a large corpus of the citation neighborhood, from which we extract relevant keywords to elaborate a semantic classification. We then show qualitatively and quantitatively the complementarity of classifications. The methodology and associated tools are open and can be reused in similar studies for which data is difficult to access or poorly referenced in classical databases.

★ ★

★

C

DÉVELOPPEMENTS THÉMATIQUES

Cette annexe regroupe des développements thématiques, c'est-à-dire qui tombent dans les domaines empiriques, conceptuels et de modélisation. Elles peuvent être relativement éloignées à première vue de nos préoccupations principales, mais sont nécessaires pour la démonstration de points précis.

Les trois premiers développements sont importants quant à des questions empiriques, de modélisation et de méthodologie, abordées d'un point de vue thématique précis.

1. Une étude empirique de la géographie des prix du carburant aux Etats-unis, permet, sous l'hypothèse que celle-ci capture des processus à l'interface du réseau routier et des territoires, de mettre en valeur deux échelles typiques pour ces processus ainsi que la superposition de processus de gouvernance avec des effets de voisinage.
2. Un modèle multi-échelles de dynamiques de migrations résidentielles à l'échelle métropolitaine est présenté avec les premiers résultats issus de son exploration.
3. Les méthodes de données synthétiques corrélées, en lien avec 3.1 et 5.3, et présentée de dans la perspective méthodologique abstraite en B.3, est ici appliquée à une question de finance quantitative.

Les développements suivants se rapportent à des questions épistémologiques, principalement en lien avec l'interdisciplinarité.

4. Le concept de perspectivisme appliqué est introduit dans la présentation de l'application *CybergeoNetworks*, qui permet l'analyse de corpus scientifiques par la combinaison de différentes approches. Celle-ci est également cruciale quant aux questions de Science Ouverte.
5. La méthode d'analyse sémantique utilisée en 2.2 et déjà présentée en B.6 est appliquée à un corpus de brevets, ce qui nous permet de la déployer sur données massives, et également de développer la question de l'innovation, aspect thématique crucial pour la théorie évolutive.
6. Un compte rendu de la session spéciale Economie et Géographie à l'ECTQG 2017 permet d'une part d'explorer le rôle des modèles dans les démarches interdisciplinaires, et d'autre part d'illustrer la démarche du perspectivisme appliqué.

7. La question des outils de la médiation scientifique est abordée directement par la présentation d'un projet d'exploration d'outils basés sur les jeux dans le cas des questions environnementales liées aux écosystèmes d'eau douce.

* * *

*

Les publications ou communications correspondant au contenu de ces annexes sont détaillées pour chacune, avec le détail des contributions des différents collaborateurs.

C.1 ROAD NETWORK AND PRICES DRIVERS

Les interactions entre réseaux et territoires peuvent se manifester indirectement au sein de propriétés économiques locales de territoires : ainsi, le prix de l'énergie conditionne fortement l'impédance d'un réseau routier, et donc son impact sur les territoires, et réciproquement ce prix est localement produit par des sous-marchés qui sont partie intégrante des territoires. La géographie des prix du carburant est donc un marqueur indirect des interactions. Par exemple, [Orfeuil and Wiel, 2012] (p. 307) suggère un lien entre prix de l'essence et crise immobilière en région parisienne.

* * *

*

Cette annexe a été réalisée en collaboration avec l'économiste DR. A. BERGEAUD (Banque de France), dans le cadre d'une convergence des problématiques entre marchés de l'énergie et observation indirecte des interactions entre réseaux et territoires. Elle a été présenté à la conférence EWGT 2017 comme [Rimbault and Bergeaud, 2017].

* * *

*

The geography of fuel prices has many various implications, from its significant impact on accessibility to being an indicator of territorial equity and transportation policy. In this paper, we study the spatio-temporal patterns of fuel price in the US at a very high resolution using a newly constructed dataset collecting daily oil prices for two months, on a significant proportion of US gas facilities. These data have been collected using a specifically-designed large scale data crawling technology that we describe.

We study the influence of socio-economic variables, by using complementary methods: Geographically Weighted Regression to take into account spatial non-stationarity, and linear econometric modeling to condition at the state and test county level characteristics. The former yields an optimal spatial range roughly corresponding to stationarity scale, and significant influence of variables such as median income or wage per job, with a non-simple spatial behavior that confirms the importance of geographical particularities. On the other

hand, multi-level modeling reveals a strong state fixed effect, while county specific characteristics still have significant impact. Through the combination of such methods, we unveil the superposition of a governance process with a local socio-economical spatial process. We discuss one important application that is the elaboration of locally parametrized car-regulation policies.

c.1.1 *Context*

What drives the price of fuel? Using a new database on oil price at a gas station level collected during two months, we explore its variability across time and space. Variation in the cost of fuel can have many causes, from the crude oil price to local tax policy and geographical features, all having heterogeneous effect in space and time. If the evolution of the average fuel price in time is an indicator that is carefully followed and analyzed by many financial institution, its variability across space remain a rather unexplored topic in the literature. Yet, such differences can reflect variation in more indirect socio-economic indicators such as territorial inequalities and geographical singularities or consumer preferences.

There exists to our knowledge no systematic mapping in space and time of retail fuel prices for a country. The main reason is probably that the availability of data have been a significant obstacle. It is also likely that the nature of the problem may also have influence, as it lies at the crossroad of several disciplines. While economists study price elasticity and measurement in different markets, transportation geography with method such as transportation prices in spatial models, puts more emphasis on spatial distribution than on precise market mechanisms. Nevertheless, examples of somehow related works can be found. For example, [Rietveld, Bruinsma, and Van Vuuren, 2001] studies the impact of cross-border differences in fuel price and the implications for gradual spatial taxation in Netherlands. At the country-level, [Rietveld and Woudenberg, 2005] provides statistical models to explain fuel price variability across European countries. [Macharis et al., 2010] models the impact of spatial fuel price variation on patterns of inter-modality, implying that the spatial heterogeneity of fuel prices has a strong impact on user behavior. With a similar view on the geography of transportation, [Gregg et al., 2009] studies spatial distribution of gas emission at the US-state level. The geography of fuel prices also have important implications on effective costs, as shows [Combes and Lafourcade, 2005] by determining accurate transportation costs across urban areas for France. More closely related to our work, and using very similar daily open data for France, [Gautier and Saout, 2015] investigate dynamics of transmission from crude oil prices to fuel retail prices. However, they do not introduce

an explicit spatial model of prices diffusion and do not study spatio-temporal dynamics.

In this paper we adopt a different approach by proceeding to exploratory spatial analysis on US fuel prices. We show that most of the variation occurs between counties and not across time, although crude oil price was not constant during the period considered. We therefore turn to a spatial analysis of the distribution of fuel prices. Our main findings are twofold: first we show that there are significant spatial pattern in some large US regions, second we show that even if most of the observed variation can be explained by state level policies, and especially the level of tax, some county level characteristics are still significant.

Dataset

Our dataset contain daily information on fuel price at the gas station level for the whole US mainland territory. These information have been constructed from self-reported fuel price and span almost the entire universe of gas station in the US. We start by describing data collection and then give some statistics about this new dataset.

Collecting large scale heterogeneous data

The availability of new type of data has induced consequent changes in various disciplines from social science (e.g. online social network analysis ([Tan et al., 2013])) to geography (e.g. new insights into urban mobility or perspectives on “smarter” cities ([Batty, 2013a])) including economics where the availability of exhaustive individual or firm level data is seen as a revolution of the field. Most studies involving these new data are at the interface of implied disciplines, what is both an advantage but also a source of difficulties. For example misunderstandings between physics and urban sciences described in [Dupuy and Benguigui, 2015] are in particular caused by different attitudes towards unconventional data or divergent interpretations and ontologies of it. Collection and use of new data has therefore become a crucial stack in social-science. The construction of such datasets is however far from straightforward because of the incomplete and noisy nature of data. Specific technical tools have to be implemented but have often been designed to overcome one specific problem and are difficult to generalize. We develop such a tool that fills the following constraints that are typical of large scale data collection: (i) reasonable level of flexibility and generality; (ii) optimized performance through parallel collection jobs; (iii) anonymity of collection jobs to avoid any possible bias in the behavior of the data source. The architecture, at a high level, has the following structure:

- An independent pool of tasks runs continuously socket proxies to pipe requests through tor.

- A manager monitors current collection tasks, split collection between subtasks and launches new ones when necessary.
- Subtasks can be any callable application taken as argument destination urls, they proceed to the crawling, parsing and storage of collected data.

The application is open and its modules are reusable: source code is available on the repository of the project.¹ We constructed our dataset by using the tool continuously in time during two months to collect crowdsourced data available from various online sources.

Dataset

Our dataset comprises around $41 \cdot 10^6$ unique observations of retail fuel prices at the station level, spanning the period starting the 10th of January 2017 and ending the 19th of March 2017 and corresponds to 118,573 unique retail stations. For each of these stations, we associate a precise geographical location (city resolution). On average we have 377 price information by station. Prices correspond to a unique purchase mode (credit card, other modes such as cash being less than 10% in test datasets, they were discarded in the final dataset) and four possible fuel types: Diesel (18% of observations), Regular (34%), Midgrade (24%) and Premium (24%). The best coverage of stations is for Regular fuel type with on average 4,629 price information by county. We therefore choose to focus the study to this type of fuel, keeping in mind that further developments with the dataset may include comparative analysis on fuel types.

Our final dataset thus contains 14,192,352 observations from 117,155 gas station, followed during 68 days. We further aggregate these data by day, taking the average of the observed price per gallon, to obtain a panel of 5,204,398 gas station - day observations.² Table ?? gives some basic descriptive statistics of on price data showing that the distribution of oil price is highly concentrated with a small skewness (the ratio of the 99th to the 1st percentile is 1.6). Finally, in the spatial analysis, we will also use socio-economic data at the county level, available from the US Census Bureau. We shall use the latest available, which most of the time implies relying to the 2010 Census.

c.1.2 *Results*

Spatio-temporal Patterns of Prices

Before moving to a more systematic study of the variation of fuel price, we propose a first exploratory introduction to give insight about

¹ at <https://github.com/JusteRaimbault/EnergyPrice>

² The panel is not balanced as prices are not reported every day in every station. The average gas station has information on price for 44 days (over 68).

FIGURE 121: Descriptive statistics on Fuel Price (\$ per gallon)

Moyenne	Dev. Std.	p10	p25	p50	p75	p90
2.28	0.27	2.02	2.09	2.21	2.39	2.65

its spatio-temporal structure. This exercise is a crucial stage to guide further analyses, but also to understand their implications in a geographical context. To explore the data, we built a simple web application which allow to map the data in space and time. This application is available on this page.

We also show one example of mapping the data at the county level in Figure ?? where we used average price over the whole period. We clearly see regional patterns with the Southcentral and Southeast regions having the lowest prices and the Pacific cost and Northeast the highest prices. Of course, plotting aggregated data over the whole period does not bring much information about the time variation of the data. As we will show more in detail below most of the variation of fuel price occurs across space. A variance decomposition of fuel price yields only 11% of the total variance is explained by within gas station variations. Similarly, the Spearman's rank correlation coefficient between the gas station price of regular fuel in the first day of dataset and in the last day is 0.867, and the null hypothesis that these two information are independent is strongly rejected.

Since most of the variation in oil price is between gas station, we now focus mainly on spatial correlations. We will conduct the analysis at the county level for various reasons. First it appears that a variance decomposition of fuel price between and within county shows that more than 85% of the variance is between-county, second because the localization of gas station is not reliable enough to allow for a smaller granularity and third because we have many socio-economic information at this level. We therefore study the spatial autocorrelation of prices at the county level. Spatial autocorrelation can be seen as an indicator of spatial heterogeneity which we measure using the Moran index ([Tsai, 2005]), with spatial weights of the form $\exp(-d_{ij}/d_0)$ with d_{ij} being the distance between spatial entities i and j , and d_0 a decay parameter giving the spatial range of interactions accounted for in the computation. We show in Fig. ?? its variations for each day and also as a function of the decay parameter. The fluctuations in time of the daily Moran index for low and medium spatial range, confirms geographical specificities in the sense of locally changing correlation regimes. These are logically smoothed for long ranges, as price correlations drop down with distance. The behavior of spatial autocorrelation with decay distance is particularly interesting: we observe a first regime change around 10km (from constant to piecewise linear regime), and a second important one

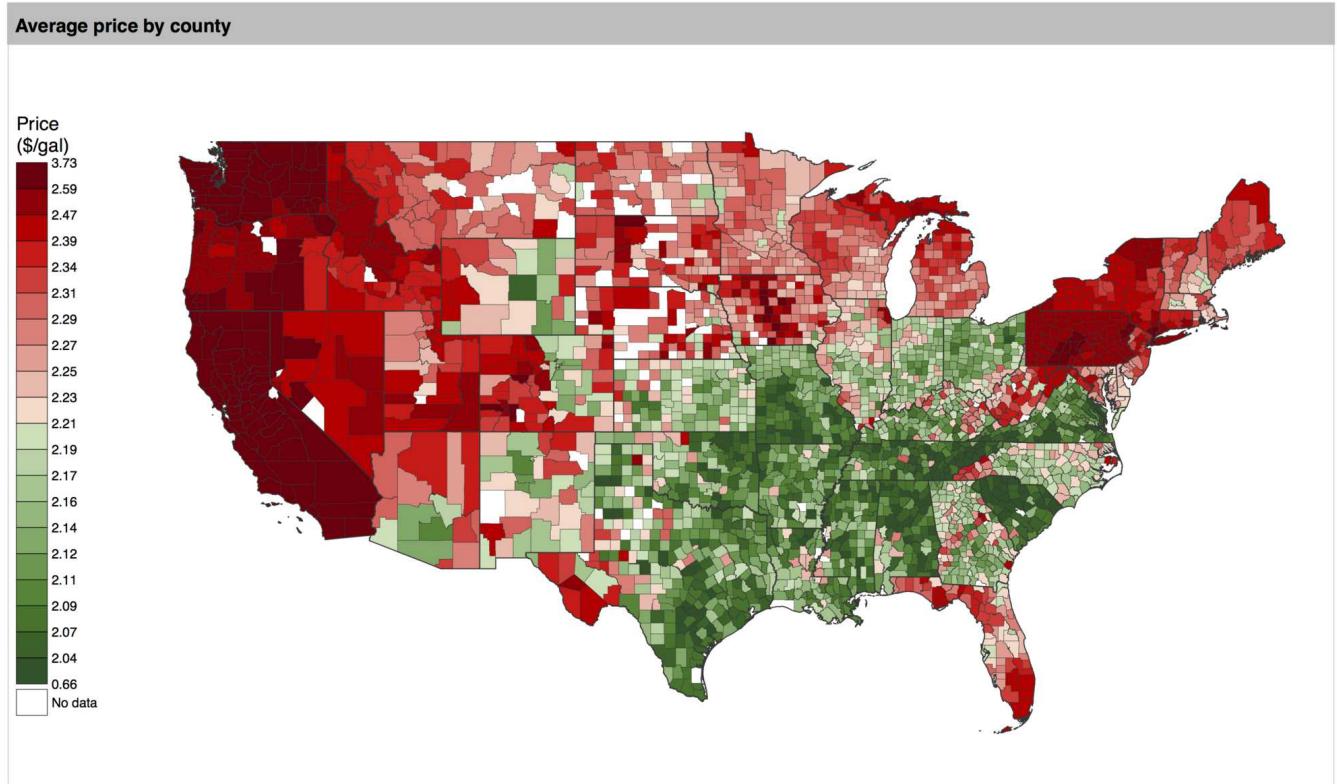


FIGURE 122: Map of mean price for counties, regular fuel, averaged over the whole period.

around 1000km, both consistent across weekly time windows. We postulate that these correspond to typical spatial scales of the involved processes: the low regime would be local specificities and the middle one the state level processes. This behavior confirms that prices are non-stationary in space, and that therefore appropriate statistical techniques must be used to study potential drivers at different level. The two next subsections follow this idea and investigate potential explicative variables of local fuel prices, using two different techniques corresponding to two complementary paradigms: geographically weighted regression that puts the emphasis on neighborhood effects, and multi-level regression taking into account administrative boundaries.

Geographically Weighted Regression

The issue of spatial non-stationarity of geographical processes has always been a source of biased aggregated analyses or misinterpretations when applying general conclusions to local cases. To take it into account into statistical models, numerous techniques have been developed, among which the simple but very elegant Geographically Weighted Regression (GWR), that estimates non-stationary regressions by weighting observations in space similarly to kernel estimation methods. This was introduced in a seminal paper by [Brunsdon, Fotheringham, and Charlton, 1996] and has been subsequently used and matured since then. The significant advantage of this technique is that an optimal spatial range in the sense of model performance can be inferred to derive a model that yields the effect of variables varying in space, thus revealing local effects that can occur at different spatial scales or across boundaries.

We proceed to multi-modeling to find the best model and associated kernel and spatial range. More specifically, we do the following: (i) we generate all possible linear models from the five potential variables (income, population, wage per job, jobs per capita, jobs); (ii) for each model and each candidate kernel shape (exponential, gaussian, bisquare, step), we determine the optimal bandwidth in the sense of both cross-validation and corrected Akaike Information Criterion (AICc) which quantifies information included in the model; (iii) we fit the models with this bandwidth. We choose the model with the best overall AICc, namely $\text{price} = \beta \cdot (\text{income}, \text{wage}, \text{percapjobs})$ for a bandwidth of 22 neighbors and a gaussian kernel,³ with an AICc of 2,900. The median AICc difference with all other models tested is 122. The global R-squared is 0.27, what is relatively good also compared to the best R-squared of 0.29 (obtained for the model with all variables, which clearly overfits with an AICc of 3010; furthermore,

³ note that the kernel shape does not have much influence as soon as gradually decaying functions are used

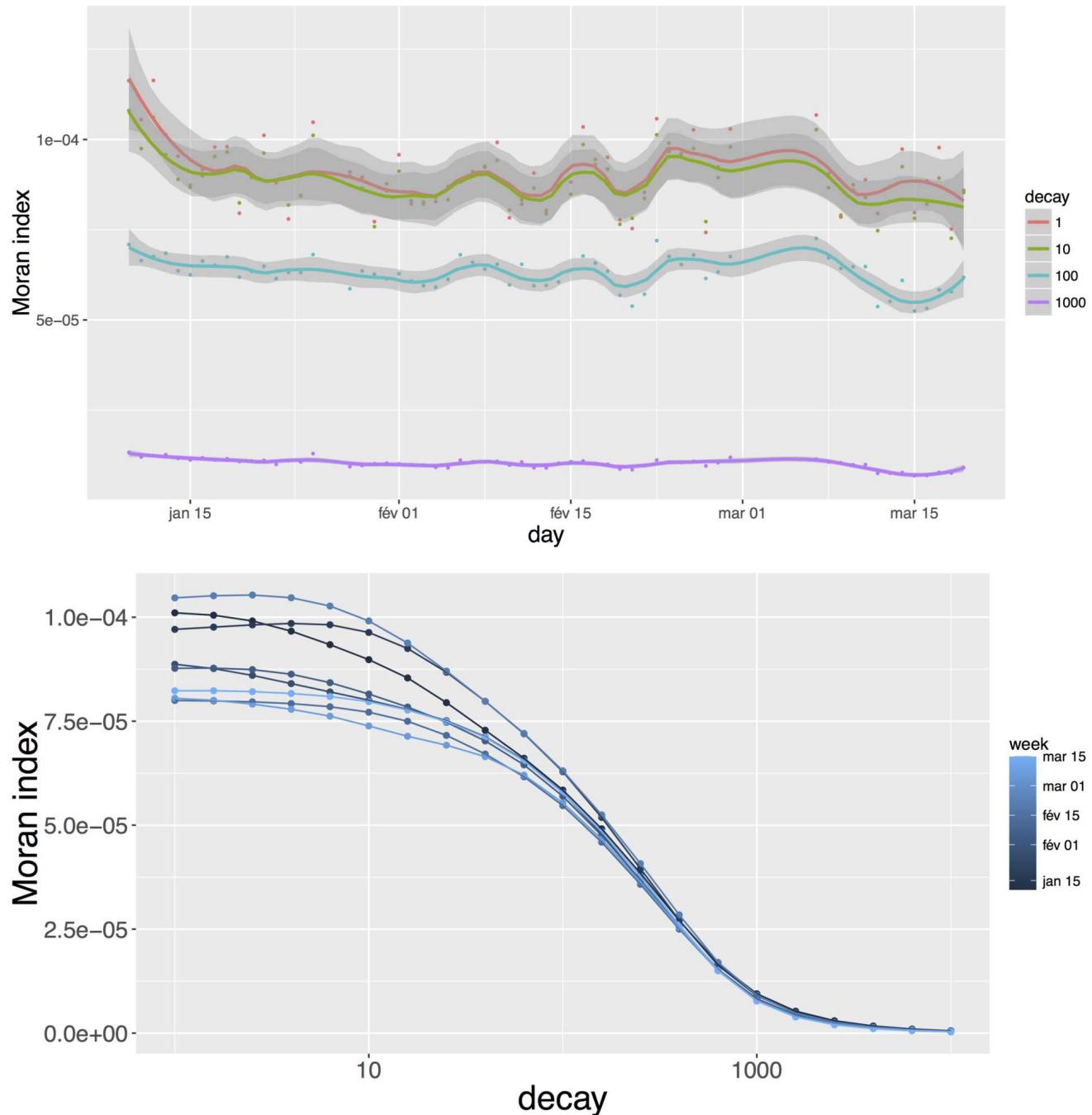


FIGURE 123: Behavior of Moran spatial-autocorrelation index. (Left) Evolution in time of Moran index computed on daily time windows, for different decay parameter values. (Right) Moran index as a function of decay parameter, computed on weekly time windows.

effective dimension is less than 5 as 90% of variance is explained by the three first principal components for the normalized variables).

The coefficients and local R-squared for the best model are shown in Fig. ?? . The spatial distribution of residuals (not shown here) seems globally randomly distributed, which confirms in a way the consistency of the approach. Indeed, if a distinguishable geographical structure had been found in the residuals, it would have meant that the geographical model or the variable considered had failed to translate spatial structure. Let now turn to an interpretation of the spatial structures we obtain. First of all, the spatial distribution of the model performance reveals that regions where these simple socio-economic factors explain do a good job in explaining prices are mostly located on the west coast, the south border, the north-east region from lakes to the east coast, and a stripe from Chicago to the south of Texas. The corresponding coefficients have different behaviors across the areas, suggesting different regimes.⁴ For example, the influence of income in each region seems to be inverted when the distance to the coast increases (from north to south-east in the west, south to north in Texas, east to west in the east), what may be a fingerprint of different economic specializations. On the contrary, the regime shifts for wage show a clear cut between west (except around Seattle) and middle/east, that does not correspond to state-policies only as Texas splits in two. The same way, jobs per capita show an opposition between east and west, what could be due for example to cultural differences. These results are difficult to interpret directly, and must be understood as a confirmation that geographical particularities matters, as regions differ in regimes of role for each of the simple socio-economic-variables. Further precise knowledge could be obtained through targeted geographical studies including qualitative field studies and quantitative analyses, that are beyond the scope of this exploratory paper and left for further research.

Finally, we extract the spatial scale of the studied processes, that is, by computing the distribution of distance to nearest neighbors with the optimal bandwidth. It yields roughly a log-normal distribution, of median 77km and interquartile 30km. We interpret this scale as the spatial stationarity scale of price processes in relation with economic agents, which can also be understood as a range of coherent market competition between gas stations.

Multi-level Regression

Since our initial database enables to look at the level of variable $x_{i,s,c,t}$, the fuel price in day t, in gas station i, in state s and in county c, we start by running high dimensional fixed effect regressions following the model:

⁴ We comment their behavior in areas where the model has a minimal performance, that we fix arbitrarily as a local R-squared of 0.5

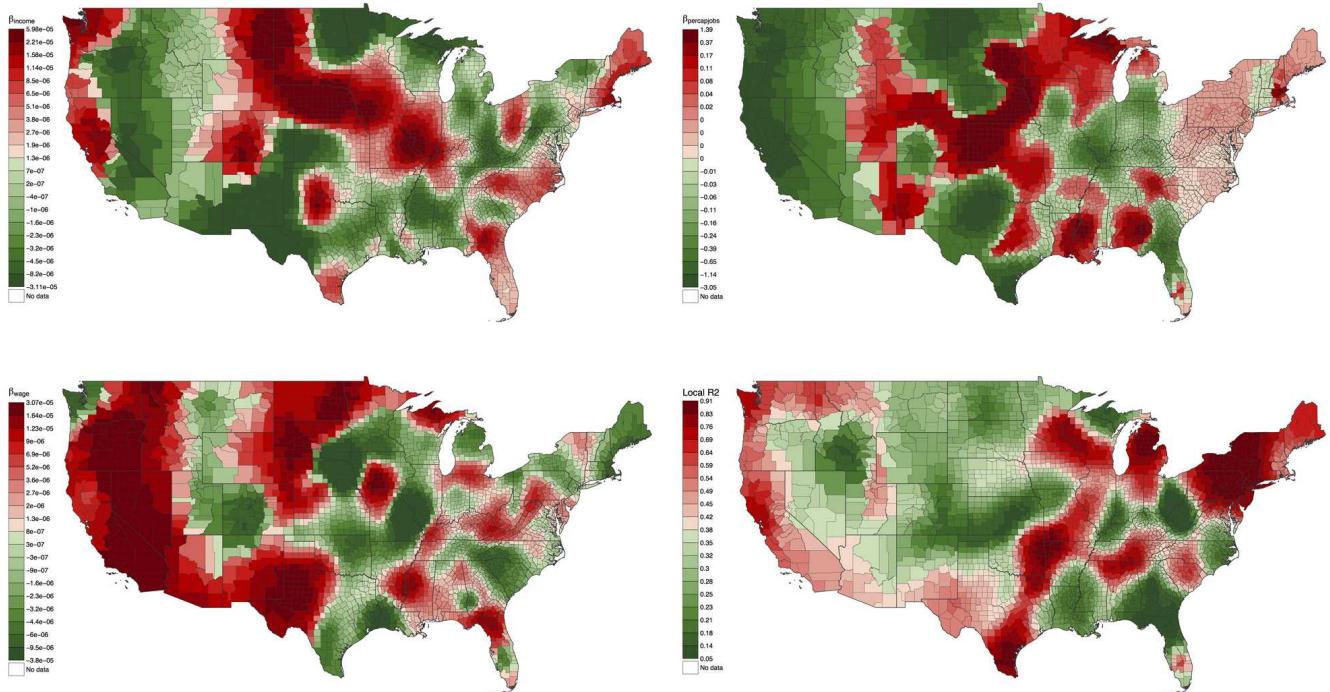


FIGURE 124: Results of GWR analyses. For the best model in the sense of AICc, we map the spatial distribution of fitted coefficient, in order from left to right and top to bottom, β_{income} , $\beta_{percapijobs}$, β_{wage} , and finally the local r-squared values.

$$x_{i,s,c,t} = \beta_s + \varepsilon_{i,s,c,t} \quad (27)$$

$$x_{i,s,c,t} = \beta_c + \varepsilon_{i,s,c,t} \quad (28)$$

$$x_{i,s,c,t} = \beta_i + \varepsilon_{i,s,c,t} \quad (29)$$

Where $\varepsilon_{i,s,c,t}$ contains an idiosyncratic error and a day fixed effect. This first analysis confirm that most of the variance can be explained by a state fixed effect and that integrating more accurate levels has only small effect on the fit of our model as measured by the R-squared.

We now turn to a different analysis, aiming at capturing the explanatory variables that account for spatial price variation of fuel. We consider the following linear model:

$$\log(x_i) = \beta_0 + X_i\beta_1 + \beta_{s(i)} + \varepsilon_i, \quad (30)$$

where x_i denotes average measured fuel price in county i aggregated across all days, X_i is a set of county specific variables and $s(i)$ is the state to which the county belongs so that $\beta_{s(i)}$ capture all state specific variation. Finally ε_i is an error term satisfying $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ if $s(i) \neq s(j)$. This clustering of standard error at the state level is motivated by finding of the previous section, showing that spatial autocorrelation of fuel price at the state level is still potentially strong. This specification aims at capturing the effect of various socio-economic variable at the county level after a state fixed effect has been removed.

We present our results in Table ???. Column (1) shows that regressing the log of price on a state fixed-effect is already enough to explain 74% of the variance. This is mostly due to tax on fuel which are set at the state level in the US. In fact, when we regress the log of oil price on the level of state tax, we find a R-squared of 0.33%. The remaining explanatory variables show that dense urban counties have higher fuel price, but this price decreases with population. This result seems sensible, desert areas have on average higher oil price. Fuel price increases with total income, decreases with poverty and decrease with the extent to which a county has voted for a Republican candidate. This last finding suggests a circular link: counties that use car the most tend to vote to politician that promote pro car policies. Adding these explanatory variables slightly increase the R-squared, suggesting that even after having removed a state fixed-effect, the price of fuel can be explained by local socio-economic features.

TABLE 28: Regressions at the county level

	(1)	(2)	(3)	(4)	(5)
Density	0.016*** (0.002)	0.016*** (0.001)	0.016*** (0.001)	0.015*** (0.001)	
Population (log)	-0.007*** (0.001)	-0.040*** (0.011)	-0.041*** (0.011)	-0.039*** (0.010)	
Total Income (log)		0.031*** (0.010)	0.031*** (0.010)	0.027*** (0.009)	
Unemployment		0.001 (0.001)	0.000 (0.001)	0.000 (0.001)	
Poverty		-0.028** (0.011)	-0.030*** (0.011)	-0.029** (0.011)	
Percentage Black			0.000*** (0.000)	-0.000 (0.000)	
Vote GOP				-0.072*** (0.015)	
R-squared	0.743	0.767	0.774	0.776	0.781
N	3,066	3,011	3,011	3,011	3,011

Notes: This table plots results from an Ordinary Least Square regression of model presented in equation (30). Density is measured as the number of inhabitant by square miles and total income is given in dollars. Poverty is measured as the number of people below the poverty threshold per inhabitant. Percentage black is the percentage of black people living in the county and vote GOP is the share of people having voted for Donald Trump in the 2016 elections. Regression includes a state fixed effect. Robust standard errors clustered at the state level are reported in parenthesis. ***, ** and * respectively indicate 0.01, 0.05 and 0.1 levels of significance.

C.1.3 Discussion

On the complementarity of Econometric and Spatial Analysis methods

One important aspect of our contribution is methodological. We show that to explore a new panel dataset, geographers and economists have different approach, leading to similar generic conclusion but with different path. Some studies have already combined GWR and multi-level regressions ([Chen and Truong, 2012]), or compared them in terms of model fit or robustness ([Lee, Kang, and Kim, 2009]). We take here a multi-disciplinary point of view and combines approaches answering to different questions, GWR aiming at finding precise explicative variables and to measure the extent of spatial correlation, whereas econometric models explain with more accuracy the effect of factors at different levels (state, county) but take these geographical characteristics as exogenous. We claim that both are necessary to understand all dimensions of the studied phenomenon.

Designing localized car-regulation policies

Another application of such analysis is to help better designing car-regulation policies. Environmental and health issues nowadays require a reasoned use of cars, in cities with the problem air pollution but also overall to reduce carbon emissions. [Fullerton and West, 2002] showed that a taxation of fuel and cars can be equivalent to a taxation on emissions. [Brand, Anable, and Tran, 2013] highlight the role of incentives for the transition towards a low carbon transportation. However, such measures can't be uniform across states or even counties for obvious reasons of territorial equity: areas with different socio-economic characteristics or with different amenities shall contribute regarding their capabilities and preferences. Knowing local prices dynamics and their drivers, in which our study is a preliminary step, may be a path to localized policies taking into account the socio-economic configuration and include an equity criterion.

Conclusion

We have described a first exploratory study of US fuel prices in space and time, using a new database at the gas station level spanning two months. Our first result is to show the high spatial heterogeneity of price processes, using interactive data exploration and autocorrelation analyses. We proceed with two complementary studies of potential drivers: GWR unveils spatial structures and geographical particularities, and yields a characteristic scale of processes around 75km; multi-level regressions show that even though most of the variation can be explained by state level characteristics, and mostly by the level of the tax on fuel that is set by the state, there are still socio-

economic specificities at the county level that can explain spatial variation of fuel price.

Perspective

★ ★

★

C.2 MULTI-SCALAR MODELING OF RESIDENTIAL DYNAMICS

Nous avons effleuré dans le chapitre introductif les questions de mobilité (quotidienne et résidentielle) comme processus voisins de ceux qui nous ont occupé tout au long de ce travail, à une autre échelle et avec d'autres ontologies. Nous avons d'autre part suggéré l'ouverture vers des modèles multi-échelles comme un développement privilégié et une application relativement immédiate des briques préliminaires que nous avons forgé ici. Cet annexe présente brièvement un travail développant précisément ces deux points, dans le cas des dynamiques résidentielles des migrants ruraux dans le Delta de la rivière des Perles en Chine.

* * *

*

Ce travail est le fruit d'une collaboration interdisciplinaire avec la sociologue et sinologue CINZIA LOSAVIO (UMR CNRS 8504 Géographie-cités), dans le cadre du projet MEDIUM. Le texte produit en collaboration est ici adapté et traduit. Ces résultats ont été présenté à la conférence internationale Urban China 2017 comme [Losavio and Rimbault, 2017].

* * *

*

This paper introduces an agent-based model of regional migration dynamics, applied to the Mega-city Region of Pearl River Delta. It focuses on residential dynamics of migrants workers and aims at taking into account the variety of migrant's profiles, based on qualitative fieldwork observations. The extensive exploration of the model, both on synthetic and real-world configurations, yield several stylized facts of migration processes and specific effects of the regional geography, which can be used to inform migration policies. We postulate that such integrated modeling approaches will be more and more appropriate to study cities in China.

C.2.1 *Introduction*

CONTEXT Over the last three decades, rural-to-urban migrant-workers have been a driving force for China's economy, raising attention on

associated socio-economical issues. However, the importance of their economic diversity and social mobility has been poorly considered in the analysis of urban development strategy.

We use an agent-based model to simulate residential dynamics of migrants in Pearl River Delta (PRD) mega city region, taking into account the full range of migrants' socio-economical status and their evolution. Mega-city regions have become a new scale of Chinese State regulation, and PRD represent the most prosperous and dynamic one in term of migration waves, standing as an ideal unit of analysis.

MEGA-CITY REGIONS Mega-city regions (MCRs) as defined by Florida, Gulden and Mellander are "integrated sets of cities and their surrounding suburban hinterlands across which labour and capital can be reallocated at very low cost" [Florida, Gulden, and Mellander, 2008]. This urban configuration recalls what Gottmann defined as *megalopolis* [Gottmann, 1961] in reference to the north-east coast of the United States. Despite this affinity in their spatial and functional configuration, MCRs perform on a different scale than megalopolis: they operate at a regional as well as at a global scale. Indeed, one of the main characteristics of MCRs is their "connectivity": spatially, they branch out into nearby rural and metropolitan areas, and economically they grow beyond their physical border, becoming international. These densely populated regions do not have a single barycenter but merging into one another they turn into highly networked spaces connected through multiple nodes. The high density of connections and the polycentrism characterizing these new economic units facilitate migrations flows and encourage regional integration.

In China, the development of mega-city regions has started right after the implementation of the Open Door Policy in 1978. But it is the gradual decentralization of the State power - which occurred in the beginning of 1990 – that promote cities and more recently mega-city regions as a new scale of Chinese State regulation [Wu, 2016]. The process of rapid economic growth and urban development molds new densely populated and industrially dynamic mega-city regions, of which the Pearl River Delta (PRD)⁵ is the most obvious example. The area was designed in 1988 as a "comprehensive economic reform area", and was granted many "one step ahead" policies to attract foreign capital. Evolving into the most important exporter center since the economic reform, the Pearl River Delta represents the most dynamic MCR in terms of migration waves [Xu and Li, 1990].

⁵ The PRD Mega City Region consists of nine cities: the core cities are Guangzhou and Shenzhen, surrounded by Dongguan, Foshan, Zhongshan, Zhuhai, Huizhou, Jiangmen, and Zhaoqing. The model does not include Hong Kong and Macau, which are part of the PRD Mega Urban Region.

MIGRANT WORKERS Taking the PRD as the spatial unit of the model, we aim to reproduce migrant workers' residential patterns taking into account the full range of their socio-economical status. Migration patterns and key related issues have extensively been studied from very diverse perspectives, ranging for example from racialization issues [Han, 2010] to big data analysis of their spatio-temporal behavior [Yang et al., 2017]. However, migrant workers are generally considered and treated as a uniform category, which stand at the bottom of the urban society, carrying the stigma of the rural household registration system. The rural-urban dual structure has been for years the only approach to define and understand migrant-workers, but the process of rapid economic growth China have been experiencing accelerated social transformation. We postulate that studying migrant workers, by merely considering their *hukou* status and place of registration is not sufficient anymore to apprehend such a complex and diversified social category. Others aspects such as migrant workers economical, cultural and human capital should be taken into account.

Especially three dimensions can help differentiate number of migrant workers sub-categories: (i) the professional dimension, which not only determines migrants' economical situation but also influences their trajectory and the duration of their staying in the city as well as their residential choice; (ii) the residential dimension which impacts all aspects of migrants' urban lives – patterns of urban settlement, housing choices, residential conditions, relation with the city, neighborhood activities etc; (iii) the generational dimension.⁶

All these sub-categories have different mobility patterns, that we simulate in the model. Considering this diversity and translating it in qualitative stylized facts that correspond to precise patterns of synthetic data, this model aims at establishing a new perspective for understanding China's urban and regional mobility employing a more qualitative approach, specifying the mechanisms through which Party-State shape the parameters of migrants' choices.

C.2.2 Model

MODELING RURAL-URBAN MIGRATIONS IN CHINA Existing works in rural-urban migration modeling in China are mainly econometric studies, relying on census or on survey data. [Zhang and Zhao, 2013] estimate discrete choice models to study the trade-off between migration distance and earning difference. [Fan, 2005] shows that gravity-based models can explain well inter-provincial migratory patterns, implying an underlying strong dominant aggregation processes. The positive association between wage gap and migration rates was ob-

⁶ The generational dimension is not taken into account in the model, as simulated dynamics correspond to rather short time scales, between 10 and 20 years.

tained from time-series analysis in [Zhang and Shunfeng, 2003]. An empirical study of intra-urban migrants residential dynamics is done by [Wu, 2006]. [Xie, Batty, and Zhao, 2007] uses an agent-based model to simulate the emergence of Urban Villages. To the best of our knowledge, there was no previous attempt in the literature to model regional migrations in China from an agent-based perspective.

MODEL The model is designed to include targeted stylized facts and experiments, in particular the role of the socio-economic structure of migrant population. More precisely, a recent shift in socio-economic structure of migrating population was observed, including a rise of middle-income migrants and a relativisation of the role of *Hukou* in migration dynamics. The core of the model is thus centered on the exploration of the impact of a varying population economic structure for migrants on system dynamics, and the influence of government migration policies.

The region is represented in the model by N patches, characterized by their population $P_i(t)$ and an economic structure $E_i^{(c)}(t)$ giving a potential number of jobs for socio-economic classes c . The associated effective number of workers is denoted by $W_i^{(c)}(t)$. For the sake of simplicity, we assume a discrete number of classes. At initial time, the variables are initialized either following a synthetic data generation process (see below), or from real geographical data (abstracted and simplified to fit our context).

Urban Centers are characterized by aggregated population $\tilde{P}_k(t)$ and corresponding economic variables $\tilde{E}_k^c(t)$. An agent is a household of migrants, with location for residence and job. Socio-economic structure of the population is captured by the distribution of wealth $g(w)$, which are then stratified into categories. At a given time, the utility difference between not moving and moving to cell j from cell i , for a category c is given by

$$\Delta U_{i,j}^{(c)}(t) = \frac{Z_j^{(c)} - Z_i^{(c)}}{Z_0} + \gamma \cdot \frac{C_i^{(c)} - C_j^{(c)}}{C_0} - u_i^{(c)} - h_j^{(c)}$$

where $Z_i^{(c)}$ is generalized accessibility given by

$$Z_i^{(c)} = P_i \cdot \sum_k \left[E_k^{(c)} - W_k^{(c)} \right] \cdot \exp \left(\frac{-d_{ij}}{d_0} \right)$$

with d_{ij} effective travel distance⁷ and d_0 commuting characteristic distance ; the parameter γ is the ratio giving the relative importance

⁷ as the model does not focus on the role of transportation, we take euclidian distance, and d_0 captures typical commuting distance in both public transportation or car. A more complicated model could include an explicit transportation network and modal choice depending on socio-economic category.

of life cost compared to accessibility in the migration decisions ; $C_i^{(c)}$ is the cost of life which is a function of cell and city variables, that will be taken as $C_i^{(c)} \propto P_i^{\alpha_0} \cdot \tilde{P}_i^{\alpha_1}$; $u_i^{(c)}$ a baseline aversion to move and $h_j^{(c)}$ an exogenous variable corresponding to regulation policies; Z_0 and C_0 dimensioning parameters.

At each time step, the system evolves sequentially according to the following rules :

1. cities-level variables are updated and distributed across patches variables (in our first experiments, we will assume short time scale and skip this step)
2. new migrants enter the region and lean on social network to settle
3. migration occur within the region, randomly drawn from discrete choice probabilities obtained with the above utility difference between two patches
4. Migrants update their wealth and eventually economic category, according to an abstract “quality of place” that we associate to per-capita GDP which follows a scaling law of population.

C.2.3 Results

The model is implemented in NetLogo, the open source implementation being available with results at [https://www.github.com/JusteRaimbault/](https://www.github.com/JusteRaimbault/MigrationDynamics) MigrationDynamics. We explore the model on synthetic city systems first, to isolate results due to processes from results due to geographical configuration. With such a random model where many parameters cannot be given directly a real-world value, it is necessary to explore intensively the parameter space to obtain robust conclusions. Using the software OpenMole [Reuillon, Leclaire, and Rey-Coyrehourcq, 2013], we proceed to 1,599,495 simulations of the model on computation grid, achieving 15 years of equivalent CPU in around 2 days. We validate the model internally by checking the statistical convergence of indicators.

From the baseline experiments we learn that : (i) when migrants have a high propensity to move, the spatial repartition of jobs becomes suboptimal in intermediate regimes of stochasticity, corresponding to a regime where congestion dominates; (ii) the congestion regime corresponds to a linear decrease of job distance with randomness, meaning that social determinism creates spatial inequalities; (iii) changing the relative importance of accessibility does not affect much the aggregated dynamics: an increased gain in mobility produced by policies such as individual transportation subsidies will have no effect on migrations patterns. (iv) configurations with an intermediate value of move aversion (in which real configurations fall) yield a negative

feedback effect of time, witnessing a progressive saturation. In a “U-shape” manner, very mobile or very fixed configurations yield positive feedback of time. We then turn to specific experiments.

Adding categorization does not change the qualitative behavior of the model. The lower category appears more vulnerable to spatial inequalities created by social determinism. Concerning the influence of economic parameters, namely income inequality and income growth, we find that : (i) larger income inequalities yield stronger spatial inequities in job accessibility; (ii) larger enrichments when migrating induces a suboptimal regime for the upper category.

The application of the model on the real population and economic configuration of Pearl River Delta slightly changes conclusions: we witness for example the emergence of optimal behavior ranges for the commuting distance indicators. It means that incentives for migrations have to be specifically tuned depending on the region configuration. Other conclusions mainly hold and are therefore process-specific.

DISCUSSION A last application we are currently developing is testing the impact of localized regulation policies, i.e. having the term $h_j^{(c)}$ varying across cities and across categories, what corresponds to policies effectively observed in practice. This various stylized facts listed above may furthermore inform more general policies, such as the impact of mobility or the existence of optimal regimes for intermediate values of randomness. Further work may consist in a calibration of the model on migration trajectories with appropriate datasets, but also in a feedback of simulation results on qualitative fieldwork, trying to compare to concrete real situations.

Our modeling enterprise is aimed at being integrated, as the model is initially built with taking in consideration qualitative observations from fieldwork, and as its outputs shall in return inform qualitative research. We believe that such integrated modeling approaches will be important tools in the future of Urban China research, in particular because of the emergence of new urban regimes in Chinese cities that were never observed somewhere else before, making difficult the use of some of previous empirical knowledge on cities.

* * *

*

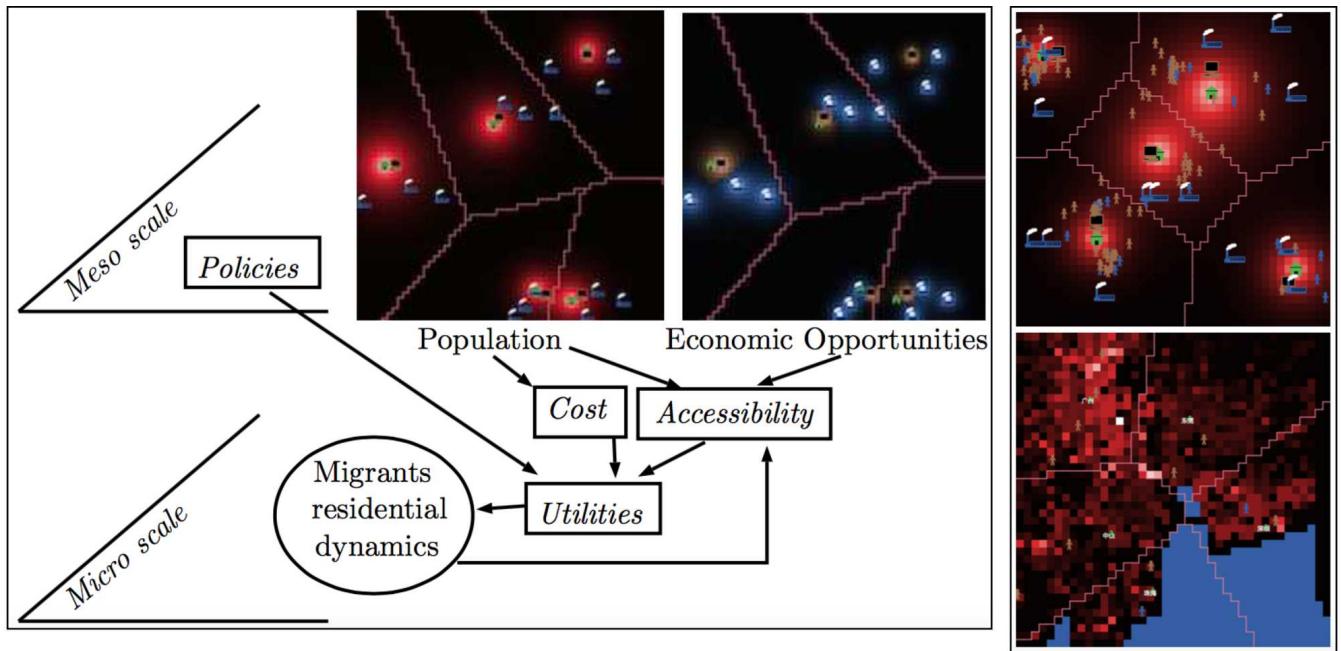


FIGURE 125: (Left) Multi-scale schema of processes included in the model. (Right) Examples of regional population configuration, for a synthetic city system (top) and Pearl River Delta (bottom).

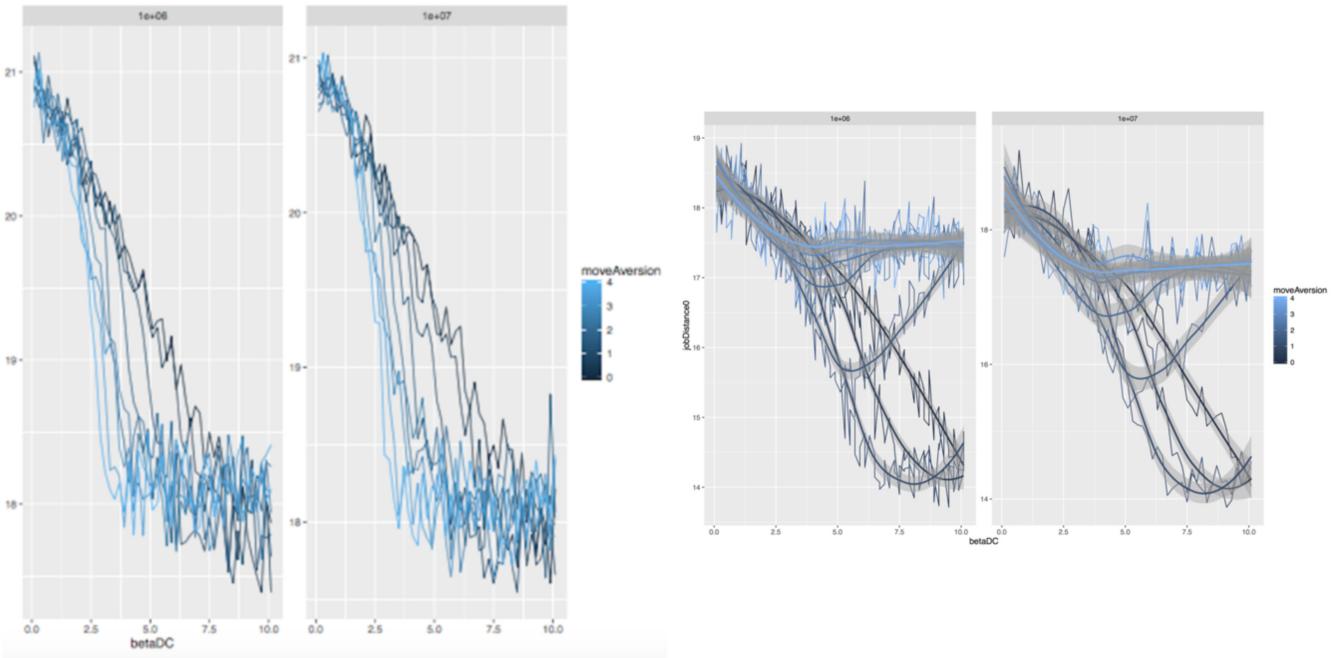


FIGURE 126: Comparison of average distance to jobs for the lowest economic category, as a function of the randomness parameter β , between synthetic city systems (two left plots) and real configuration (two right plots). The color gives the constant move aversion $u_i^{(c)}$ and plots are given for two values of cost-accessibility ratio γ . We witness the apparition of optimal values of β in the real situation, probably caused by the geography.

C.3 GENERATION OF CORRELATED SYNTHETIC DATA

Context

Our first field of application is that of financial complex systems, of which captured signals, financial time-series, are heterogeneous, multi-scalar and highly non-stationary [Mantegna and Stanley, 1999]. Correlations have already been the object of a broad bunch of related literature. For example, Random Matrix Theory allows to undress signal of noise, or at least to estimate the proportion of information undistinguishable from noise, for a correlation matrix computed for a large number of asset with low-frequency signals (daily returns mostly) [Bouchaud and Potters, 2009]. Similarly, Complex Network Analysis on networks constructed from correlations, by methods such as Minimal Spanning Tree [Bonanno, Lillo, and Mantegna, 2001] or more refined extensions developed for this purpose [Tumminello et al., 2005], yielded promising results such as the reconstruction of economic sectors structure. At high frequency, the precise estimation of interdependence parameters in the framed of fixed assumptions on asset dynamics, has been extensively studied from a theoretical point of view aimed at refinement of models and estimators [Barndorff-Nielsen et al., 2011]. Theoretical results must be tested on synthetic datasets as they ensure a control of most parameters in order to check that a predicted effect is indeed observable *all things equal otherwise*. For example, [Potiron and Mykland, 2015] obtains a bias correction for the *Hayashi-Yoshida* estimator (used to estimate integrated covariation between two brownian at high frequency in the case of asynchronous observation times) by deriving a central limit theorem for a general model that endogeneize observation times. Empirical confirmation of estimator improvement is obtained on a synthetic dataset at a fixed correlation level.

Formalization

Framework

We consider a network of assets $(X_i(t))_{1 \leq i \leq N}$ sampled at high-frequency (typically 1s). We use a multi-scalar framework (used e.g. in wavelet analysis approaches [Ramsey, 2002] or in multi-fractal signal processing [Bouchaud, Potters, and Meyer, 2000]) to interpret observed signals as the superposition of components at different time scales : $X_i = \sum_{\omega} X_i^{\omega}$. We denote by $T_i^{\omega} = \sum_{\omega' \leq \omega} X_i^{\omega'}$ the filtered signal at a given frequency ω . A recurrent problem in the study of complex systems is the prediction of a trend at a given scale. It can be viewed as the identification of regularities and their distinction from compo-

nents considered as random⁸. For the sake of simplicity, we represent such a process as a trend prediction model at a given temporal scale ω_1 , formally an estimator $M_{\omega_1} : (T_i^{\omega_1}(t'))_{t' < t} \mapsto \hat{T}_i^{\omega_1}(t)$ which aims to minimize error on the real trend $\|T_i^{\omega_1} - \hat{T}_i^{\omega_1}\|$. In the case of autoregressive multivariate estimators, the performance will depend among other parameters on respective correlations between assets. It is thus interesting to apply the method to the evaluation of performance as a function of correlation at different scales. We assume a Black-Scholes dynamic for assets [Jarrow, 1999], i.e. $dX = \sigma \cdot dW$, with W Wiener process. Such a dynamic model allows an easy modulation of correlation levels.

Data generation

We can straightforward generate \tilde{X}_i such that $\text{Var}[\tilde{X}_i^{\omega_1}] = \Sigma R \Sigma$ (with Σ estimated standard deviations and R fixed correlation matrix) and verifying $X_i^{\omega \leq \omega_0} = \tilde{X}_i^{\omega \leq \omega_0}$ (data proximity indicator : components at a lower frequency than a fundamental frequency $\omega_0 < \omega_1$ are identical). We use therefore the simulation of Wiener processes with fixed correlation. Indeed, if $dW_1 \perp\!\!\!\perp dW_1^{\perp\!\!\!\perp}$ (and $\sigma_1 < \sigma_2$ indicatively, assets being interchangeable), then

$$W_2 = \rho_{12} W_1 + \sqrt{1 - \frac{\sigma_1^2}{\sigma_2^2} \cdot \rho_{12}^2} \cdot W_1^{\perp\!\!\!\perp}$$

is such that $\rho(dW_1, dW_2) = \rho_{12}$. Next signals are constructed the same way by Gram orthonormalization. We isolate the component at the desired frequency ω_1 by filtering the signal, i.e. $\tilde{X}_i^{\omega_1} = W_i - \mathcal{F}_{\omega_0}[W_i]$ (with \mathcal{F}_{ω_0} low-pass filter with cut-off frequency ω_0). We reconstruct then the hybrid synthetic signals by

$$\tilde{X}_i = T_i^{\omega_0} + \tilde{X}_i^{\omega_1} \quad (31)$$

Results

Methodology

The method is tested on an example with two assets from foreign exchange market (EUR/USD and EUR/GBP), in a six month period from June 2015 to November 2015. Data⁹ cleaning, starting from original series sampled at a frequency around 1s, consists in a first step to the determination of the minimal common temporal range (missing

⁸ see [Gell-Mann, 1995] for an extended discussion on the construction of *schema* to study complex adaptive systems (by complex adaptive systems).

⁹ obtained from <http://www.histdata.com/>, without specified licence. For the respect of copyright, only cleaned and filtered at ω_m data are made openly available.

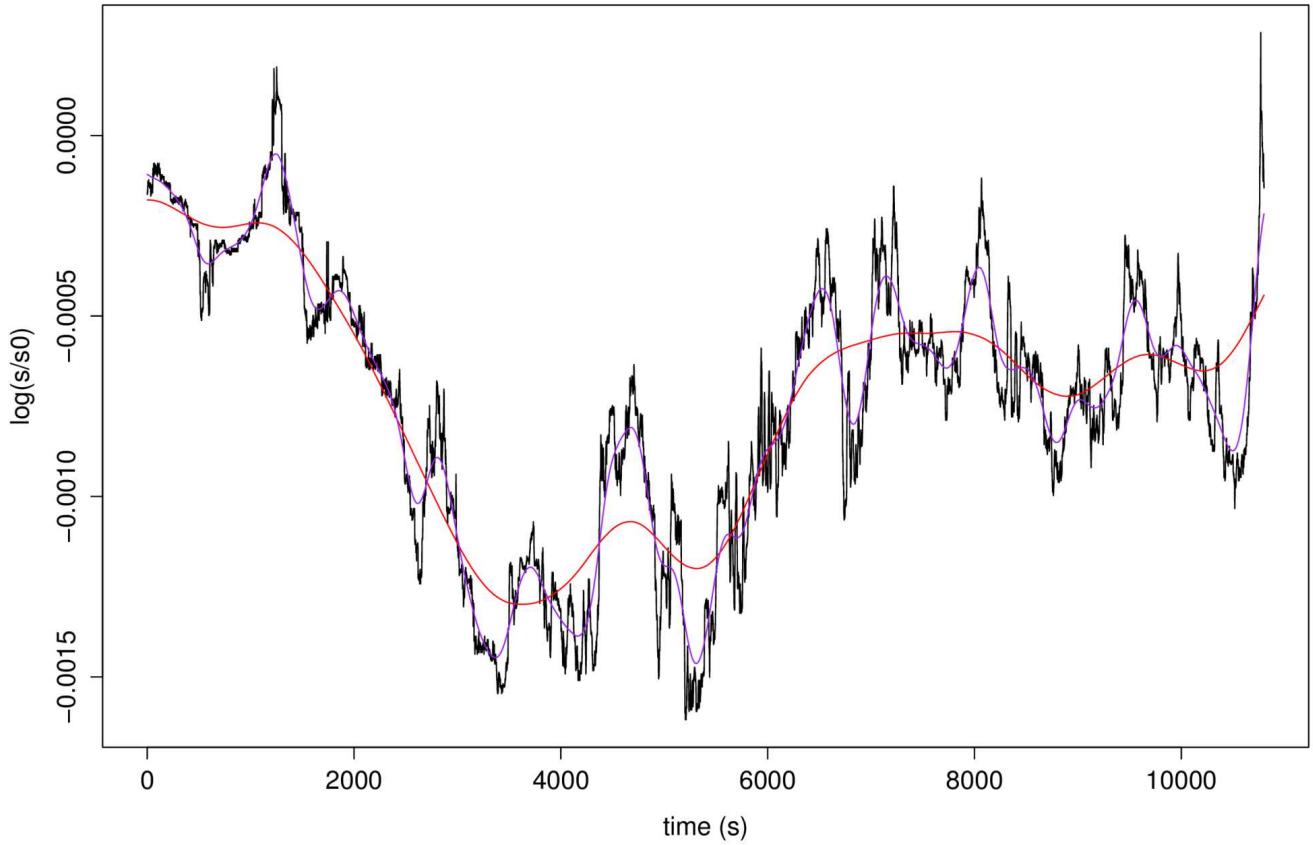
EUR/USD, 1st November 2015

FIGURE 127: Example of the multi-scalar structure of the signal, basis of the construction of synthetic signals | Log-prices are represented on a time window of around 3h for November 1st 2015 for asset EUR/USD, with 10min (purple) and 30min trends.

sequences being ignored, by vertical translation of series, i.e. $S(t) := S(t) \cdot \frac{S(t_n)}{S(t_{n-1})}$ when t_{n-1}, t_n are extremities of the “hole” and $S(t)$ value of the asset, what is equivalent to keep the constraint to have returns at similar temporal steps between assets). We study then *log-prices* and *log-returns*, defined by $X(t) := \log \frac{S(t)}{S_0}$ and $\Delta X(t) = X(t) - X(t-1)$. Raw data are filtered at a maximal frequency $\omega_m = 10\text{min}$ (which will be the maximal frequency for following treatments) for concerns of computational efficiency¹⁰. We use a non-causal gaussian filter of total width ω . We fix the fundamental frequency $\omega_0 = 24\text{h}$ and we propose to construct synthetic data at frequencies $\omega_1 = 30\text{min}, 1\text{h}, 2\text{h}$. See Fig. ?? for an example of signal structure at these different scales.

¹⁰ as time-series are then sampled at $3 \cdot \omega_m$ to avoid aliasing, a day of size 86400 for 1s sampling is reduced to a much smaller size of 432.

It is crucial to consider the interference between ω_0 and ω_1 frequencies in the reconstructed signal : correlation indeed estimated is

$$\rho_e = \rho [\Delta\tilde{X}_1, \Delta\tilde{X}_2] = \rho [\Delta T_1^{\omega_0} + \Delta\tilde{X}_1^\omega, \Delta T_2^{\omega_0} + \Delta\tilde{X}_2^\omega]$$

what yields in the reasonable limit $\sigma_1 \gg \sigma_0$ (fundamental frequency small enough), when $\text{Cov}[\Delta\tilde{X}_i^{\omega_1}, \Delta X_j^\omega] = 0$ for all $i, j, \omega_1 > \omega$ and returns centered at any scale, the correction on effective correlation due to interferences : we have at first order the expression of effective correlation

$$\rho_e = [\varepsilon_1 \varepsilon_2 \rho_0 + \rho] \cdot \left[1 - \frac{1}{2} (\varepsilon_1^2 + \varepsilon_2^2) \right] \quad (32)$$

what gives the correlation that we can effectively simulate in synthetic data.

Correlation is estimated by Pearson method, with estimator for covariance corrected for bias, i.e.

$$\hat{\rho}[X1, X2] = \frac{\hat{C}[X1, X2]}{\sqrt{\hat{\text{Var}}[X1]\hat{\text{Var}}[X2]}}$$

, where $\hat{C}[X1, X2] = \frac{1}{(T-1)} \sum_t X_1(t)X_2(t) - \frac{1}{T(T-1)} \sum_t X_1(t) \sum_t X_2(t)$ and $\hat{\text{Var}}[X] = \frac{1}{T} \sum_t X^2(t) - \left(\frac{1}{T} \sum_t X(t)\right)^2$.

The tested predictive model M_{ω_1} is a simple ARMA for which parameters $p = 2, q = 0$ are fixed (as we do not create lagged correlation, we do not expect large orders of auto-regression as these kind of processes have short memory for real data ; furthermore smoothing is not necessary as data are already filtered). It is however applied in an adaptive way¹¹. More precisely, given a time window T_W , we estimate for any t the model on $[t - T_W + 1, t]$ in order to predict signals at $t + 1$.

IMPLEMENTATION Experiments are implemented in R language, using in particular the MTS [Tsay, 2015] library for time-series models. Cleaned data and source code are openly available on the git repository of the project¹².

¹¹ adaptation level staying low, as parameters T_W, p, q and model type do not vary. We are positioned within the framework of [Potiron, 2016] which assumes a locally parametric dynamic but for which meta-parameters are fixed. We could imagine a variable T_W which would adapt for the best local fit, the same way parameters are estimated in bayesian signal processing by augmentation of the state with parameters.

¹² at <https://github.com/JusteRaimbault/SynthAsset>

RESULTS Figure ?? gives effective correlations computed on synthetic data. For standard parameter values (for example $\omega_0 = 24\text{h}$, $\omega_1 = 2\text{h}$ and $\rho = -0.5$), we find $\rho_0 \simeq 0.71$ et $\varepsilon_i \simeq 0.3$ what yields $|\rho_e - \rho| \simeq 0.05$. We observe a good agreement between observed ρ_e and values predicted by 32 in the interval $\rho \in [-0.5, 0.5]$. On the contrary, for larger absolute values, a deviation increasing with $|\rho|$ and as ω_1 decreases : it confirms the intuition that when frequency decreases and becomes closer to ω_0 , interferences between the two components are not negligible anymore and invalidate independence assumptions for example.

We apply then the predictive model described above to synthetic data, in order to study its mean performance as a function of correlation between signals. Results for $\omega_1 = 1\text{h}, 1\text{h}30, 2\text{h}$ are shown in Fig. ???. The a priori counter-intuitive result of a maximal performance at vanishing correlation for one of the assets confirms the role of synthetic data to better understand system mechanisms : the study of lagged correlations shows an asymmetry in the real data that we can understand at a daily scale as an increased influence of EUR/GBP on EUR/USD with a rough two hours lag. The existence of this *lag* allows a “good” prediction of EUR/USD thanks to fundamental component. This predictive power is perturbed by added noises in a way that increases with their correlation. The more noises correlated are, the more the model will take them into account and will make false predictions because of the markovian character of simulated brownian¹³.

This case study stays a *toy-model* and has no direct practical application, but demonstrates however the relevance of using simulated synthetic data. Further developments can be directed towards the simulation of more realistic data (presence of consistent *lagged correlation* patterns, more realistic models than Black-Scholes) and apply it on more operational predictive models.

★ ★

★

¹³ the model used has theoretically no predictive power at all on pure brownian

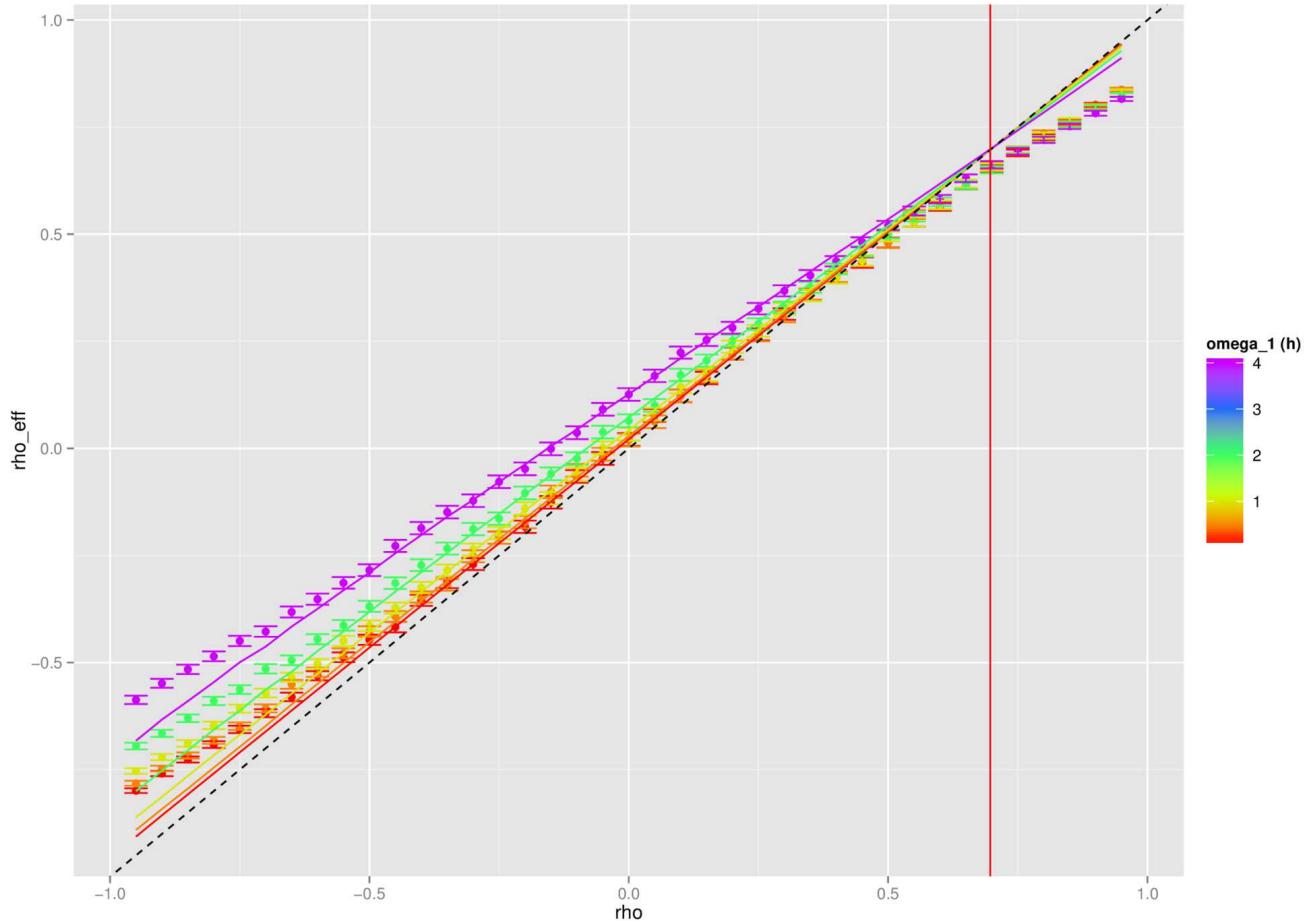


FIGURE 128: Effective correlations obtained on synthetic data. Dots represent estimated correlations on a synthetic dataset corresponding to 6 months between June and November 2015 (error-bars give 95% confidence intervals obtained with standard Fisher method) ; scale color gives filtering frequency $\omega_1 = 10\text{min}, 30\text{min}, 1\text{h}, 2\text{h}, 4\text{h}$; solid lines give theoretical values for ρ_e obtained by ?? with estimated volatilities (dotted-line diagonal for reference) ; vertical red line position is the theoretical value such that $\rho = \rho_e$ with mean values for ε_i on all points. We observe for high absolute correlations values a deviation from corrected values, what should be caused by non-verified independence and centered returns assumptions. Asymmetry is caused by the high value of $\rho_0 \simeq 0.71$.

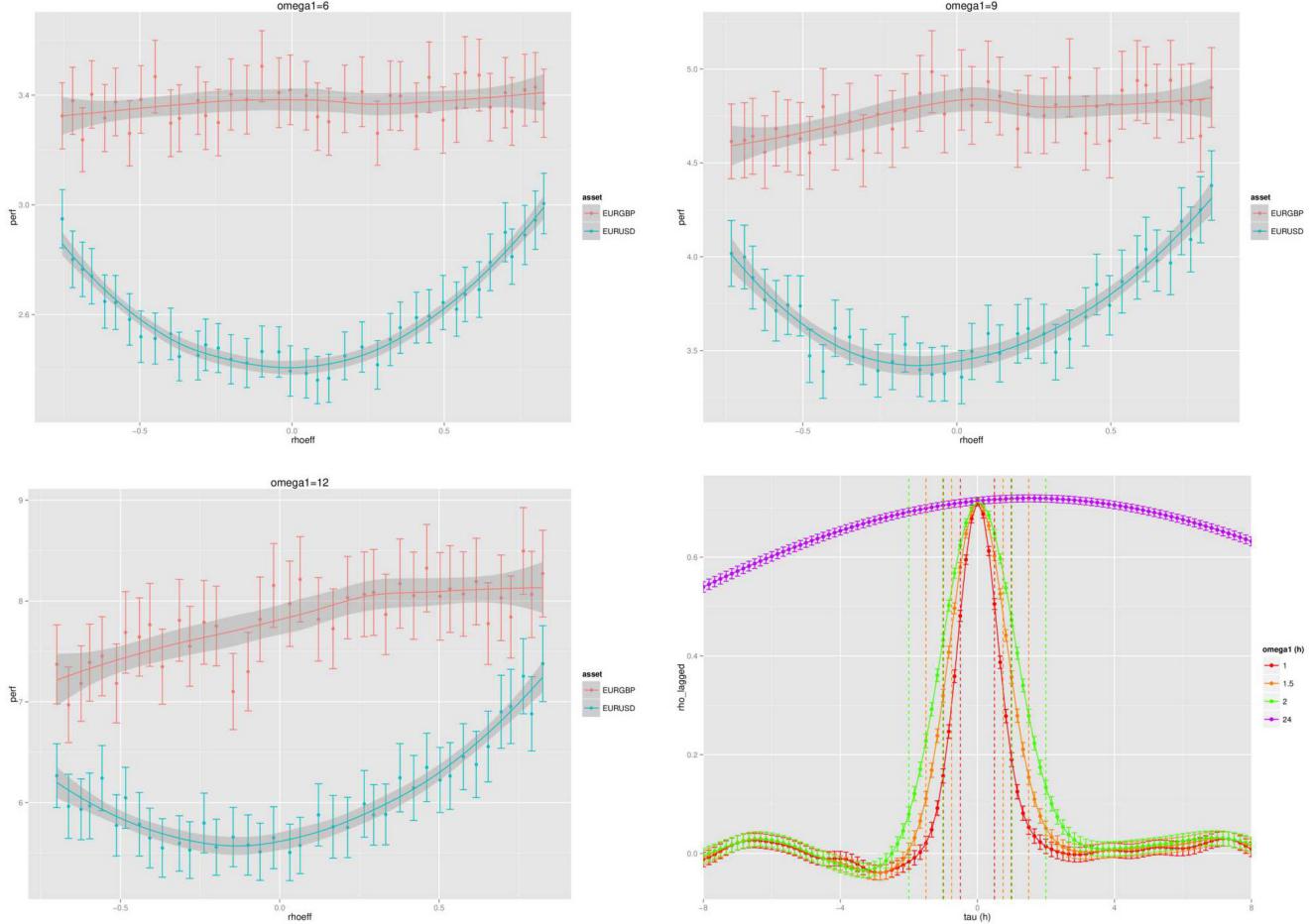


FIGURE 129: Performance of a predictive model as a function of simulated correlations. From left to right and top to bottom, three first graphs show for each asset the normalized performance of an ARMA model ($p = 2, q = 0$), defined as $\pi = \left(\frac{1}{T} \sum_t (\hat{X}_i(t) - M_{\omega_1}[\hat{X}_i](t))^2 \right) / \sigma [\hat{X}_i]^2$ (95% confidence intervals computed by $\pi = \bar{\pi} \pm (1.96 \cdot \sigma[\pi]) / \sqrt{T}$, local polynomial smoothing to ease reading). It is interesting to note the U-shape for EUR/USD, due to interference between components at different scales. Correlation between simulated noises deteriorates predictive power. The study of *lagged correlations* (here $\rho[\Delta X_{EURUSD}(t), \Delta X_{EURGBP}(t - \tau)]$) on real data clarifies this phenomenon : fourth graph show an asymmetry in curves at any scale compared to zero lag ($\tau = 0$) what leads fundamental components to increase predictive power for the dollar, amelioration then perturbed by correlations between simulated components. Dashed lines show time steps (in equivalent τ units) used by the ARMA at each scale, what allows to read the corresponding lagged correlation on fundamental component.

C.4 CYBERGEO NETWORKS : A MULTI-DIMENSIONAL AND SPATIALIZED BIBLIOMETRIC

L'analyse du corpus de *Cybergeo* a également été occasion de réflexivité et de creuser l'idée de perspectivisme appliquée par la combinaison d'approches méthodologiques. Cette annexe montre leur complémentarité et les connaissances nouvelles qui peuvent être produites par leur couplage, par en particulier ici leur spatialisation.

* * *

*

Cette annexe est le fruit d'une collaboration dans le cadre des 20 ans de la revue Cybergeo : initiée par D. PUMAIN (Université Paris 1) et A. BANOS (Université Paris 1), une équipe interdisciplinaire composée de C. COTINEAU (University College London), P.-O. CHASSET (LISER), H. COMMENGES (Université Paris 1), a mené une analyse par méthodes multiples et complémentaires du corpus de la revue Cybergeo. L'article correspondant (soumis à Journal of Informetrics) est ici traduit et adapté.

* * *

*

Bibliometrics have become commonplace and widely used by authors and journal to monitor, evaluate and identify their readership in an ever-increasing publishing scientific world. With this contribution, we aim to move from the near-real time counts to investigate the semantic proximities and evolution of the papers published in the online journal *Cybergeo* since its creation in 1996. We compare three strategies for building semantic networks, using keywords (self-declared themes), citations (areas of research using the papers published in *Cybergeo*) and full-texts (themes derived from the words used in writing). We interpret these networks and semantic proximities with respect to their temporal evolution as well as spatial expressions, by considering the countries studied in the papers under inquiry. Finally, we compare the three methods and conclude that their complementarity can help go beyond simple statistics to better understand the epistemological evolution of a scientific community and the readership target of the journal.

C.4.1 *Introduction*

Since the seminal work of Thomas Kuhn in the early 1960s the development of science studies has been based on three disciplinary pillars: history of science, philosophy of science, sociology of science. In the 1980s political science grew in importance studying the links between knowledge production and knowledge utilisation. This “political turn” began with the creation of the journal *Knowledge* in 1979. Since late 1990s, science studies has been affected by a “spatial turn” and eventually emerged a geography of science (Livingstone, 1995; Livingstone, 2003; Livingstone and Withers, 2005; Withers, 2009). Our work follows this trend: we propose in this paper a spatialised bibliometrics approach.

Faced with the increasing number of articles, journals and channels of publication used by researchers in an open access and digital world, journals need tools to identify their readership and authors need this information to better reach their target audience, using the right keywords, vocabulary and citations. This paper provides a set of complementary digital tools which meet three requirements: 1) to go beyond the usual citation metrics and give semantic and network analytics directly from the scientific contents of the papers; 2) to situate the position of sets of papers according to the semantic fields of their topics; 3) to identify the significant variations in research topics that may be linked with the geographical origin of authors or to the country they choose to analyse. This last point is especially interesting for our first case of study which is a journal of geography.

The 20-year anniversary of the first journal exclusively digital in social science – Cybergeo –, was the occasion to analyse a consistent corpus of over 700 articles published in 7 languages, with respect to the geography of its authorship and readership. We performed a quantitative epistemology analysis of the scientific papers published since 1996 to measure their similarities according three types of textual indicators: their keywords (the way authors advertise their research), their citation network (the way the paper is used by other fields and disciplines), or their full-text (the vocabulary used to write the paper and present the research).

These analyses are complementary and shows the evolution of a journal towards emergent themes of research. It also highlights the need for Cybergeo to keep extending its authorship base beyond the French-speaking community, in order to match its ambition of a European Journal of Geography. Our contribution consists in these specific epistemological conclusions, but also in a broader methodological and technical input on handling interactively large-scale heterogeneous scientific corpus. We show how the coupling of complementary views can create a second order knowledge: the spatial embedding of the three classification methods unveils unexpected patterns. Further-

more, the dedicated tool that we designed is available as an open source software, that can be used by journals for a collective scientific reflexivity, but also by institutions and individual scientists for a bottom-up empowerment of Open Science.

The rest of the paper is organized as follows: we first review similar initiatives tackling heterogeneous or multidimensional approaches to bibliometrics, and describe the case study we work on. We then develop technical details of the different methods used, and how these are coupled through interactive spatial data exploration ; describe results at the first order (each method) and at the second order (achieved through coupling) ; and finally discuss broader implication for quantitative epistemology and reflexivity for Open Science.

Bibliographic Context

Studies in bibliometrics having as a main focus the complementarity of different approaches are rather sparse. [Omodei, De Domenico, and Arenas, 2017] shows that taking into account citation and discipline data into a multilayer network is useful to understand patterns of interdisciplinarity. [Cronin and Sugimoto, 2014] is an attempt of an overview of the complex nature of measuring scientific publications and the intrinsic multidimensional nature of knowledge production. It provides both recent technical contributions with critical approaches. It insists on the “Janus-faced nature of metrics”, confirming that reducing knowledge production to a few dimensions is not only wrong but also dangerous for science. The geographical dimension of science has been studied by numerous targeted studies, such as [Maisonneuve, 2013] that investigates the diffusion of specific questions and practices in molecular biology across the world.

Cybergeo as a case study

Cybergeo was founded in 1996 as a digital-only European journal of geography. Since then, 737 scientific articles have been published (until May 2016) by 1351 authors from 51 countries. These articles have generated 2710 citations altogether over the last twenty years, which corresponds to half the number of other articles cited in Cybergeo (5545).

Most contributions come from a French institution (561), although French-speaking countries (35 papers from an author affiliated in Canada, 21 in Switzerland) and neighbouring countries (UK: 23 contributions, Italy: 18) are well represented too (fig. 130). The geographical subjects of the articles themselves show a larger diversity, as the world is almost fully covered (fig. 130). However, France and neighbouring countries such as Spain and Germany are the main focus of the majority of articles, although the United States are the 5th most

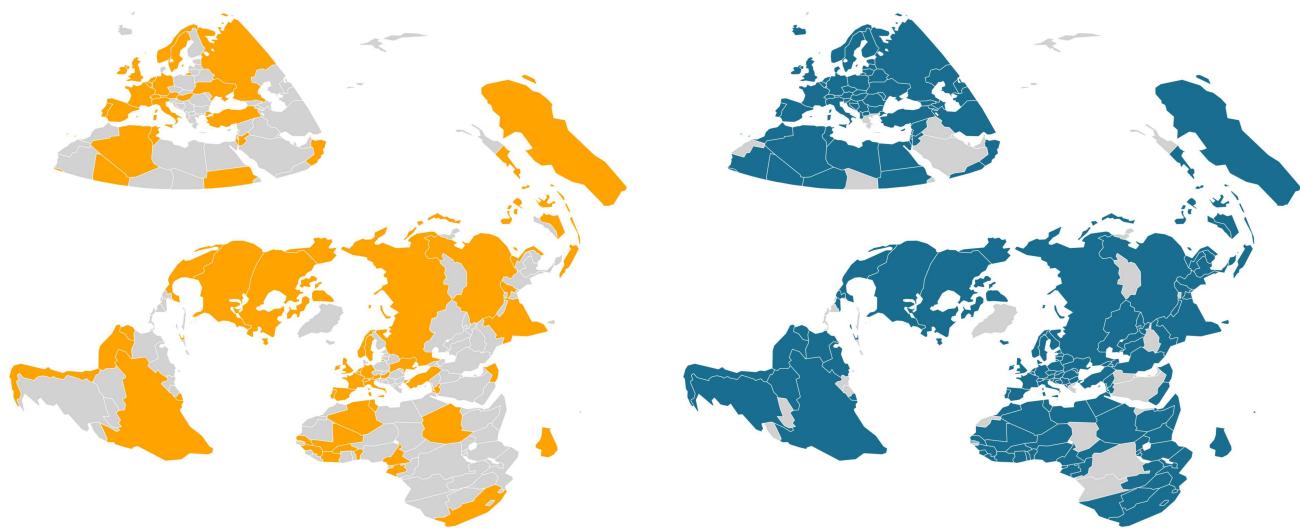


FIGURE 130: Countries with at least one author | 1996-2015 (Left), Countries studied at least once | 1996-2015 (Right)

studied single country. By linking authors to their geographical subject (fig. 131), we find different patterns:

- European and North American countries studying each other through Cybergeo articles;
- American countries being studied by authors affiliated in Europe and North-America;
- African and Asian countries being studied mainly by Europeans and marginally by Americans and themselves;
- Russia and Australia being studied by Western authors and studying their own hinterland.

Finally, the temporal evolution shows an accelerated growth of the number of authors – although the number of articles by 5-year period remains stable –, a spread of geographical coverage – with more articles published about emerging countries and extra-European territories –, along with a growing connexion in citation networks. There is a reinforcing bias towards a French-speaking authorship, revealed by the origin of authors as well as by the share of papers published in French.

c.4.2 Methods

One main aspect of our contribution is the complementary combination of different methodologies, each having its potentialities and pitfalls, but also specific questions and objects of study. We detail in



FIGURE 131: Geographical origins and destinations of papers, 1996-2015

this section the different methods and how they are coupled to produce new knowledge.

Internal semantic network

The first exploration method is based on the set of keywords declared by the articles' authors in the Cybergeo journal. We consider articles and keywords as a bipartite network. This network can be decomposed in two simple networks: a network of articles (vertices) linked by common keywords (edges); a network of keywords (vertices) linked by common articles (edges) (Roth and Cointet, 2010). We consider the second one as a semantic network.

The vertices (keywords) are described by two variables: frequency and degree. The frequency is the number of articles citing the keyword. The degree is the total degree of the vertices in the network, that is, the number of edges linking a given keyword to the others (there is no distinction between in- and out-degree as the network is undirected). Both variables are distinct but correlated. The edges are described by three variables: observed weight, expected weight and modal weight. For two given keywords the observed weight is the number of articles citing both keywords. The expected weight is the probability that the edge exists considering only the vertices' degree:

$$P_{i \rightarrow j} = \frac{w_i w_j}{w(w - w_i)} \quad P_{j \rightarrow i} = \frac{w_i w_j}{w(w - w_j)}$$

$$P_{i \leftrightarrow j} = P_{i \rightarrow j} \cup P_{j \rightarrow i}$$

$$w_{i \leftrightarrow j}^e = \frac{w}{2} P_{i \leftrightarrow j}$$

The probability of a link between i and j ($P_{i \rightarrow j}$) is defined as the cross-product of the marginal sums (w_i and w_j) divided by the total weight (w). This can be seen as a quasi-modularity measure or a quasi chi-squared distance. The only difference is the null diagonal that creates asymmetric probabilities. The expected weight ($w_{i \leftrightarrow j}^e$) is the product of the probability and the mid-sum of weights. Eventually the *modal weight* is computed as a ratio between the observed weight and the squared-root expected weight of the edge (such as a Pearson residual in a chi-square analysis of a contingency table). This modal weight can be used as a preferential attachment measure.

Based on this preferential attachment measure two kinds of visualisations are proposed: semantic fields and communities. The semantic field shows a any given keyword at the centre of the plot and all its neighbours at a distance inversely proportional to the modal weight. The communities are computed with the Louvain algorithm (Blondel

et al., 2008b). This community detection method is chosen among others because it is based on the modularity such as the modal weight above defined.

External semantic network

The second methodological development focus on the combination of citation network exploration with network semantic analysis. The method applied for this development is described in details by [Rimbault, 2017]. Citation Networks have been widely used in studies of science, for example as a predictive tool for the success of a paper (Newman, 2014), or to unveil emerging research fronts (Shibata et al., 2008). Indeed, the bibliography of a paper contains a sort of scientific positioning, as a heritage to which it aims to contribute and which fields it is based on. The other way, reverse citations, i.e. contributions citing a given paper, up to a given level, shows how the knowledge produced was understood, interpreted and used, and in particular by which field (on this point the interesting example of [Jacobs, 2016], heavily cited today by most of quantitative studies of the city by physicists, shows how unexpected the type of the audience can be).

We define the citation neighborhood of our corpus as all the articles citing articles published in *Cybergeo*, all the articles citing the ones cited by *Cybergeo*, and all the articles citing these ones. (having thus a network at depth 2). The citation data is collected using automatic data collection similarly to [Rimbault, 2017].

Having constructed this citation neighborhood, we introduce a method to analyze its content through text mining. More precisely, we focus on the *relevant* keywords of abstracts, in a precise sense, which was introduced by [Chavalarias and Cointet, 2013] to study the evolution of scientific fields, and later refined and scaled to big data on a Patent database by [Bergeaud, Potiron, and Rimbault, 2017a]. Using co-occurrences of n-grams (keywords with multiple components, obtained after a first text cleaning and filtering), the deviation from an uniform distribution across texts using a chi-squared test gives a measure of keyword relevance, on which a fixed number is filtered. The weighted co-occurrence network between relevant keywords captures their second order relationship and we assume that its topology contains information on the structure of disciplines that are present in the citation network. A sensitivity analysis of community structure to network filtering parameters (minimal edge weight, minimal and maximal document occurrence for nodes) yield a robust network with optimal community structure, what allows to associate to each paper a list of keywords and corresponding disciplines. These are complementary to the declared keywords and the full text themes presented in the next subsection, as they reveal how authors position

in the semantic landscape associated to the citation neighborhood, or what are their “cultural backgrounds”.

Topics allocation using full text documents

The third and last exploration method details the allocation of topics in full text documents, and is thus complementary to the previous ones that used declared keywords and relevant keywords within abstracts of the citation neighborhood. Topic classification of texts documents is an intense field of research, that have developed several algorithms. In this field, a topic is considered as a set of words frequently used together in the same document, and a text document as a mixture of topics. Following a long standing development in natural language processing from the weighting scheme of words called Term frequency-inverse document frequency (tfidf) introduced by [Salton and McGill, 1986] to first generative probabilistic model of [Hofmann, 1999], [Blei, Ng, and Jordan, 2003] have lastly proposed an evolution with the Latent Dirichlet Allocation model (LDA).

The LDA method consider texts in a destructured way, i.e. words proximity or words presence in a same sentence are irrelevant. Articles become thus bags of words. To alleviate the disadvantage of de-structuring the text, different methods can be used. The probabilistic tagging method proposed by [Schmid, 1994] and used in this article aims at categorizing each word by its function in the sentence, allowing us to filter only nouns, articles and verbs. The tagging includes also the transformation from plural forms into singular, and from conjugated verb form into infinitive. Each word is then associated with a frequency in a document, which can be weighted using the tfidf weights. We use in this article one of the many forms of tfidf given by

$$\text{tfidf}_{t,d,D} = \log(1 + f_{t,d}) \cdot \log \left(\frac{N}{f_{t,D}} \right)$$

where $f_{t,d}$ is the frequency of the term t in the document d , N is the total number of documents in the corpus and $f_{t,D}$ is the number of documents D containing the term t . This way, after having destructured text documents, filtered only nouns, articles and verbs, and finally weighted each word, we produce the matrix of weights per document and word in terms of topics, using the LDA model. LDA is a Bayesian hierarchical model (fig. ??). We give details of its structure in the following. This model considers three levels: corpus, document and word. Each level is defined by a set of probabilistic distributions and their parameters. Then, at the corpus level, the model has parameters α and β . α is a vector of positive reals (once per topic). β is a matrix describing the probability of each word included in the dictionary (columns) for each topic (rows). Afterwards, at the document level, the generative process begins to draw the number of

words from a Poisson distribution with parameter ϵ and a vector θ from a Dirichlet distribution with parameter α . Finally, at the word level, the topic z of the word is drawn from a multinomial distribution with parameter θ and a word is drawn from a multinomial distribution with parameter the vector probability line of the topic z from the parameter matrix β . All these parameters are estimated here using a Markov chain Monte Carlo algorithm, called Gibbs sampler from [Geman and Geman, 1984]. This algorithm generates a vast amount of draws of each parameters needed by the generative model, and produces a distribution of the value of each parameter. The main product is the β parameter, which is the probability of a word per topic, that can be then analyzed in order to understand the topic found in the corpus. During this process, the number of topics is a fixed parameter. In order to select an optimal number, [Blei, Ng, and Jordan, 2003] proposed a graphical method aiming to identify the number of topics with a minimal perplexity and a maximal entropy.

Geographical aggregation of semantic profiles

In order to produce the maps in figures 130 to 131 and analyses at the country level, articles have been geo-tagged in two ways. Firstly, the country of affiliation of the author(s) has been coded following the 2-letter identifiers of the International Organization for Standardization. Secondly, the articles were read one by one to extract the major geographical subjects. Articles were tagged with a country if this country or a sub-region of it constituted the focus of the study. In the case of European countries, different sets of countries were associated with the publication, depending on the perimeter of the subject (for instance: EU15, EU25, Schengen area, EuroMed, etc.). Given a semantic characterisation of articles (using keywords, citations or full-texts), it is then possible to determine two semantic profiles of countries: one using countries as authoring origins and one using countries as subject 'destination'. This semantic profile of a country X is made of the mean share of themes Y present in articles authoring from or studying country X. All in all, given the three semantic characterisations of articles (using keywords, citations and full-texts) and the two geographical allocation of articles (authoring or studied), each country has a maximum of six distinct semantic profiles. We use these semantic profiles to cluster countries. The clustering method applied is an ascending hierarchical clustering algorithm using the Ward criterion of distance maximisation. When analysing authoring clusters, we consider groups of countries from which a certain geography is made and written. This option is interesting in a reflexive aspect but practically more hazardous because of the high concentration of emission and the consequently low number of emitting countries. Therefore, in the results section, we base our clustering on studied countries. When analysing clusters of studied countries, we consider how cer-

tain groups of territories are studied, what words authors use to talk about them and in which research areas the papers about them are used.

Open data + interactivity = reproducibility & transparency

Last but not least, our methodological contribution is also closely linked to issues of reflexivity, transparency and reproducibility in the process of knowledge production. It is now a well sustained idea that all these aspects are closely linked and that their strong coupling participates to a virtuous circle enhancing and accelerating knowledge production, as seen in the various approaches of Open Science (Fecher and Friesike, 2014). For example, open peer review is progressively emerging as an alternative way to the rigid and slow classical canons of scientific communication: [Ross-Hellauer, 2017] proceeds to a systematic review on the notion to give an unified definition and understand its potential benefits and pitfalls. In the domain of computational science, tools are numerous to ensure reproducibility and transparency but require a strict discipline of use and are not easily accessible (Wilson et al., 2017). Open Science suggest transparency of the knowledge production process itself, but also of the knowledge communication patterns: on this point we claim that interactive exploration of quantitative epistemological patterns are necessary. We build therefore an interactive application to allow the exploration of heterogeneous scientific corpuses.

The web application is available online at <http://shiny.parisgeo.cnrs.fr/CybergeoNetworks/>. Source code and data, both for analyses and the web application, are available on the open git repository of the project at <https://github.com/AnonymousAuthor3/cybergeo20>.

c.4.3 Results

Internal semantic network

COMMUNITIES AND SEMANTIC FIELDS The community detection algorithm finds a modularity optimum with 10 clusters: mobility and transportation; imagery and GIS; climate and environment; history and epistemology; sustainability, risk, planning; Economic geography; Territory and population; urban dynamics; statistics and modelling; emotional geography. Some clusters concentrate a large number of keywords and articles, such as "imagery and GIS" or "statistics and modelling". This result was expected because of the original aim and scope of the journal. Beside the main clusters and a set of medium-sized clusters, two small and totally unexpected clusters emerged: "emotional geography" and "climate and environment". The CybergeoNetworks application proposes a set of visualisation parameters to draw the communities (see Figure 132) such as setting the size

of vertices and edges according to different variables (degree, number of articles, modal weight).

The above explained modal weight metrics can be used to draw semantic fields. The CybergeoNetworks application proposes the full list of keywords. The user chooses one keyword from that list, the word is placed in the centre of the plot and all its neighbours are arranged at a distance inversely proportional to the preferential attachment (modal weight). The application proposes visualisation parameters such as setting the character size according to the weight of the keywords (number of articles or degree in the network) (see Figure 132). Some proximities are expected ("urban" is closely linked to "city"), some are expected knowing the original scope of the journal in the field of theoretical and quantitative geography ("model" or "spatial statistics" are linked to "city"). Some proximities are totally unexpected: for "city" the preferential attachment of keywords like "movie", "web", "virtual".

Spatialised communities Using the keywords distributions to draw the semantic profile of the 128 countries studied in a Cybergeo article, we obtain a clustering in 4 groups representing 16.5% of the initial inertia. Its geographical distribution is shown in figure 133 with the average profile of each group.

Countries are differentiated firstly by whether or not the articles studying them also declare keywords related to transport and mobility, history and epistemology, urban systems and/or emotional geography. Indeed, the first group of 83 countries (in blue, Fig. 133) is defined by these themes. The corresponding countries are the most developed and rich territories of the world, including emergent countries such as the BRICS. The keywords used to advertise the articles about them follow the fashions of geography, with mentions of emotions and mobility for instance.

The countries of the other groups over-represent the keywords related to:

- methods (in orange) such as statistics and modelling. The countries associated with these keywords are all located in central and southern Africa, with the exception of Lao. These countries are studied by a small number of articles which focus on methodological approaches. For example, the only article studying Rwanda (Querriau et al., 2004) relates to an optimal location problem whereas Vallée (2009) uses 'multilevel modelling' as a keyword for the only article about Lao.
- sustainability and risks (in yellow). This is the case of articles about Indonesia for example, which all relate to aleas and vulnerability: to tsunamis (Ozer and De Longueville, 2005), to volcanoes (Bélizal, Lavigne, and Grancher, 2011) and to water scarcity (Putra and Baier, 2009).

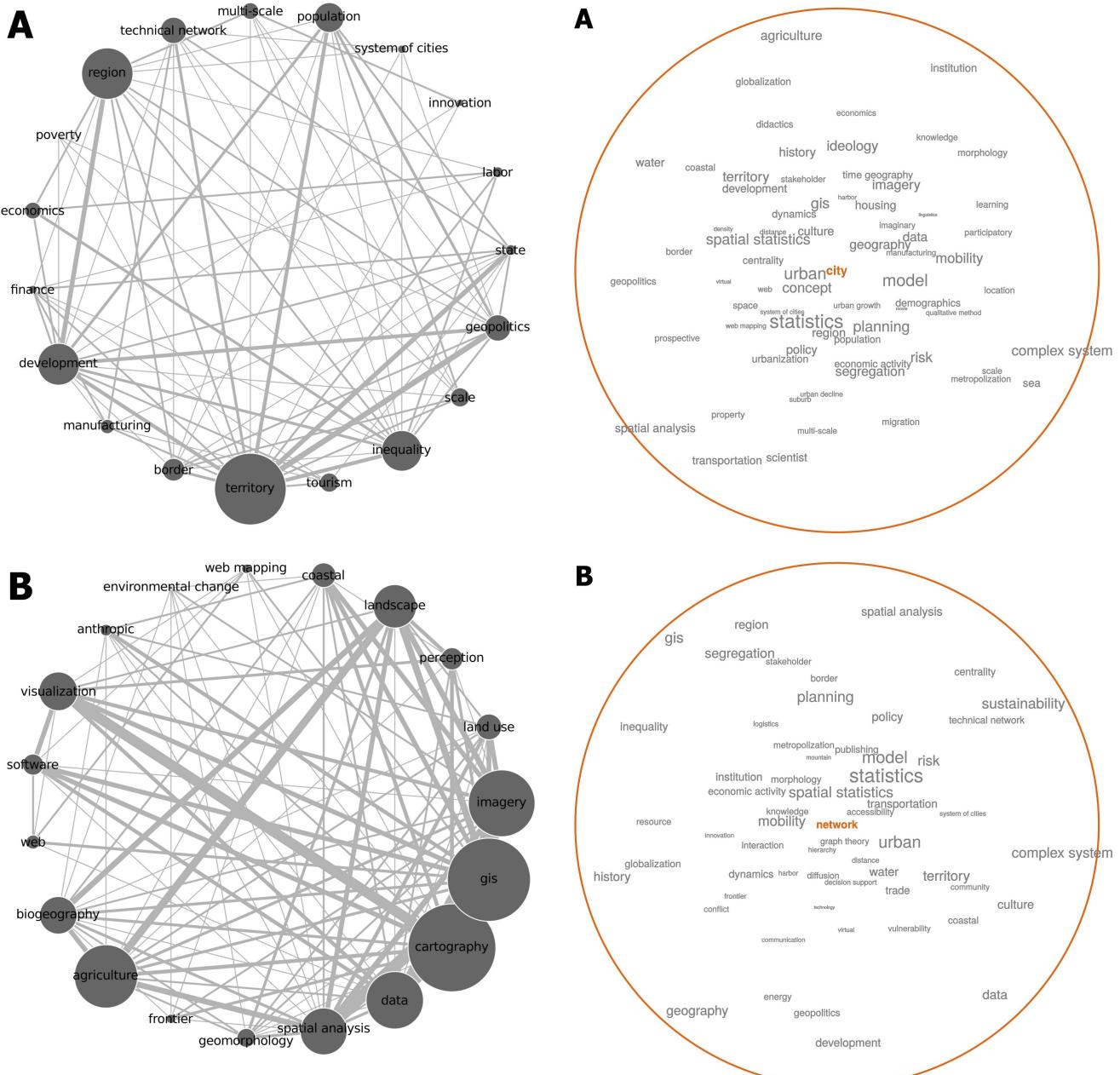


FIGURE 132: Community structure of the internal semantic network. (A-Territory; B-Imagery & GIS). Semantic fields. (A-City; B-Network)

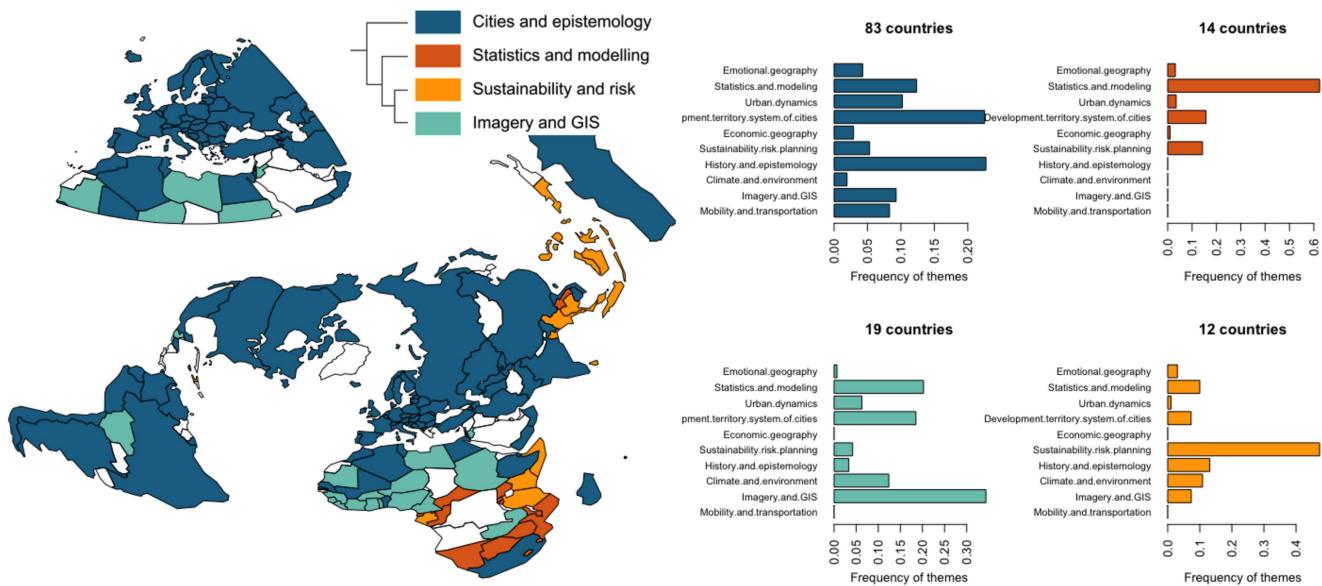


FIGURE 133: Geographical communities of declared interest (Left); Corresponding semantic profile of groups (Right)

- Finally, 19 countries are associated with keywords related to imagery and GIS (teal colour). They are located primarily in Saharan Africa. In many cases, this happens because the articles present a methodology which uses aerial and satellite images to substitute missing socioeconomic data (Ackermann, Mering, and Quensiere, 2003; Devaux, Fotsing, and Chéry, 2007).

Thus, drawing communities of declared interest, we find an interesting dichotomy between rich countries on the one hand, which are studied extensively in the literature and for which authors use trendy keywords to singularise themselves from past and concurrent work; and developing countries on the other hand, which are associated with more technical keywords reflecting a narrower spectrum of domains and specific data challenges.

External semantic network

The application allows to explore the citation neighborhood of chosen articles, in terms of semantic contents (the visualisation of full networks are technically not feasible as the full corpus contains around 200,000 articles). Wordclouds give the content of the article and the content of the articles in the neighborhood, with each word being associated to the semantic communities. The user can therefore situate a work within a semantic context, and we expect that unanticipated connexions can be made with this tools, as authors may not be aware of similar works in alien disciplines.

COMMUNITIES STRUCTURES As explained before, the raw semantic network is optimized for modularity and size, taking a compromise between these two opposite objectives, when edge and node filtering parameter vary. This provides 12 communities, that can correspond to existing disciplines, to methodological issues, or to very precise thematic subjects. The communities are, in order of importance in terms of proportion of total keywords : Political Science/Communication; Biogeography; Social and Economic Geography; Climate; Physical Geography; Commerce; Spatial Analysis; Microbiology; Neuroscience; GIS; Agriculture; Health. This method has the characteristic of grouping keywords by co-occurrences, revealing thus the actual structure of abstracts contents: it is both an advantage when revealing links as for the large field of Social and Economic Geography, but can also blur information by grouping more detailed communities. Very precise small communities such as Health Geography appear as they are strongly isolated from the rest of the communities. This structure is particular, and shows a dimension of knowledge that for example classical citation analysis do not reveal.

SPATIALISED COMMUNITIES Using the previous networks to draw the semantic profile of the 128 countries studied in a Cybergeo article, we obtain a clustering in 5 groups representing 19.3% of the initial inertia. Its geographical distribution is shown in figure 134 with the average profile of each group.

The largest group of countries largely overlaps with the largest cluster of keywords communities (cf. previous section). Indeed, rich and emergent countries are studied in articles used in similar ways in citation networks. There are further divides among this group. A first subgroup (in blue) of countries is studied by Cybergeo articles cited preferentially in the fields of commerce, socio-economic and politics analysis. These correspond to articles mostly in Economics and Social Sciences.

The nearest subgroup of countries (in orange) comprises Australia, Azerbaijan, Iran, Lao, the Philippines and Iceland. It corresponds to countries treated by articles cited preferentially in methodological fields (spatial analysis and GIS). Indeed, the only article about Iran presents a collaborative decision support system (Jelokhani-Niaraki and Malczewski, 2012) while the only article about Australia reviews online cartographic products (Escobar, Polley, and Williamson, 2000). This kind of articles then tends to stay in the citation clique of geomatics.

The third subgroup refers to countries of South-East Asia, Western Africa, Yemen and Chile. The articles studying them are cited preferentially in the fields of biogeography and socioeconomic studies, although they match the average profile.

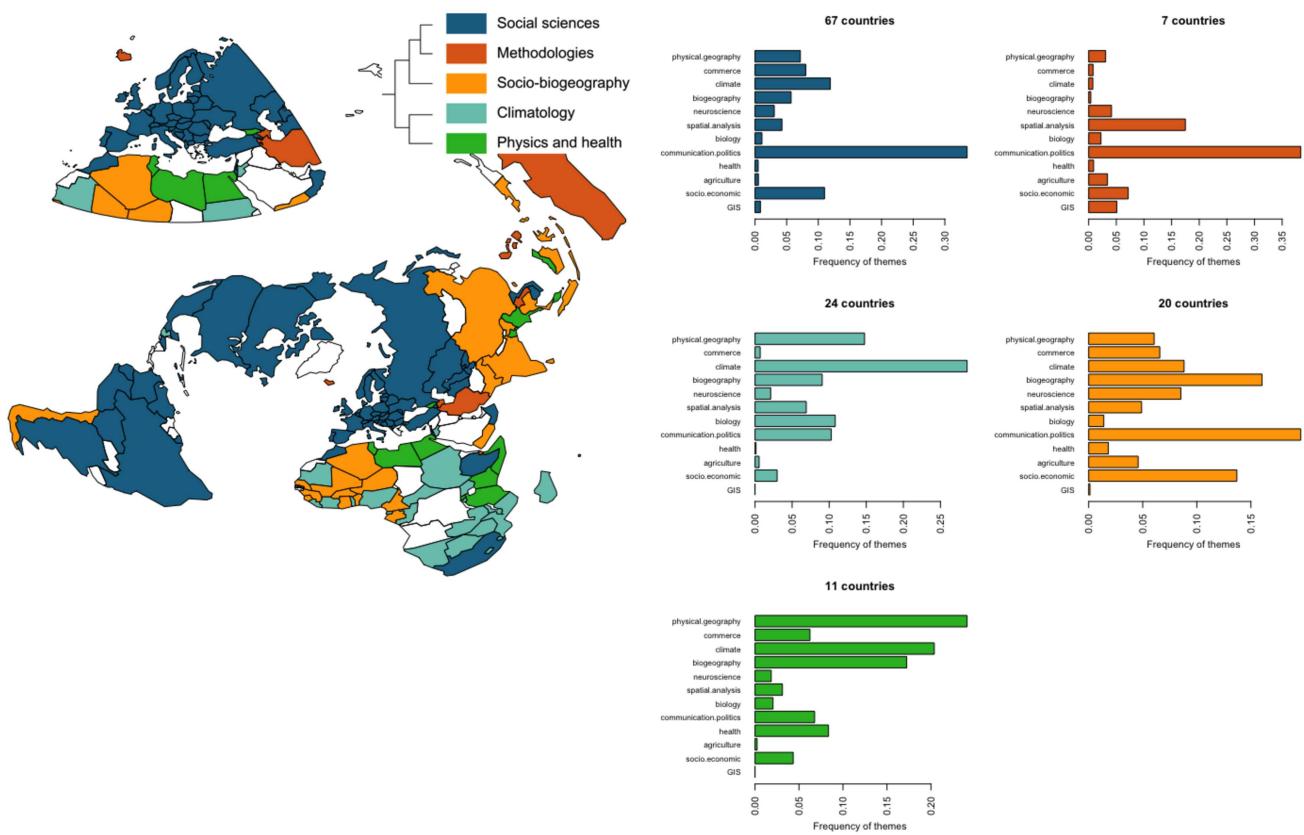


FIGURE 134: Geographical communities of bibliographical use (*Left*); Corresponding semantic profile of groups (*Right*).

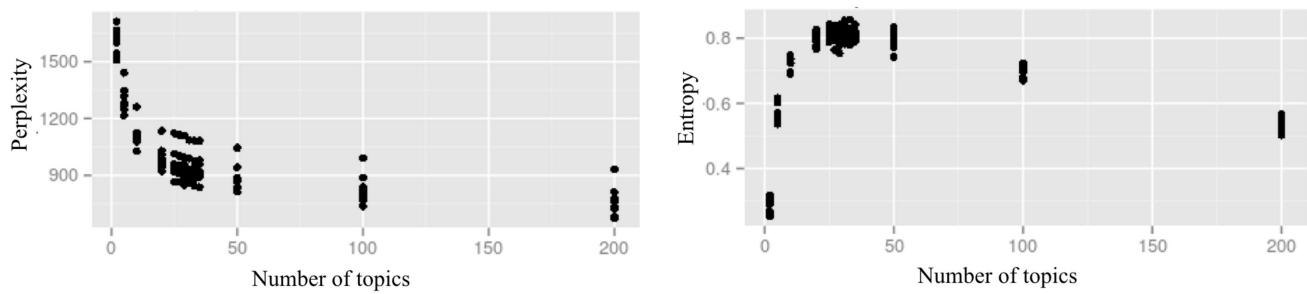


FIGURE 135: Perplexity and Entropy of the LDA model per number of topics.

In the second group of countries, we find a first subgroup of sub-Saharan countries (in teal) associated with papers cited in the climatology citation community. The second subgroup is composed of East African, North African and South-East Asian countries (in green) associated with papers cited in the fields of physical geography and health.

Thus, drawing communities of bibliographical use, we find an interesting dichotomy between rich countries on the one hand, which are associated with papers cited in broad communities, including topical and methodological fields; and poor and developing countries on the other hand, which are associated with papers cited mainly in relation to natural hazards, health and risks in the literature.

Topics allocation using full text documents

EVOLUTION OF THE TOPICS ADDRESSED IN THE CORPUS After deconstructing text documents and filtering nouns, articles and verbs, our corpus counts no less than four millions words, which leads to a dictionary of 137 224 unique words. Then, the optimal number of topics was chosen by estimating the LDA parameters with different numbers of topics: 2, 5, 10, 20, 50, 100 and 200. Similar results were obtained using a different scheme being more precise between 20 and 40. We test the robustness of results to stochasticity by iterating ten times each parameter estimation for a specific number of topics. We then consider that 20 topics is an optimal according to perplexity (fig. ??) and entropy (fig. ??) indicators.

The final result of the topic allocation model is the parameter matrix β . It describes the probability p of each word of the dictionary to belong to a topic. We present below a part of this matrix describing for each topic index the translated (except for the index 7) words in decreasing order of their probability to belong to this topic (tab. ??). It is then interesting to observe how many documents addressed a topic per year, i.e. the topics evolution in the Cybergeo corpus (fig. 136). We can distinguish several evolution profile: decreasing, punctual, regular and increasing topics. Articles about cartography (11)

Table 29: List of words and its probability to belong to the topic

Index	Words (probability) list
1	district (8), city (7), housing (5), household (3)
2	move (6), mobility (5), density (4), accessibility (3), indicator (3), simulation (3), modeling (2), scenario (2)
3	image (8), soil (7), occupation (4), surface (4), vegetation (2), map (2), resolution (2), pixel (2), landscape (2), fire (2)
4	planning (7), governance (4), urban planning (2), sustainability (2), document (2), participation (2)
5	risk (9), vulnerability (5), hazard (3), flood (2), city (2), water (2), disaster (2), management (2)
6	firm (7), healthcare (3), care (2)
7	the (16), and (8), The (2), for (2), are (2)
8	index (4), agent (3), graph (2), vertex (2), mountain (2)
9	city (5)
10	water (8), exploitation (4), management (3), farmer (2), agriculture (2), parcel (2)
11	map (10), cartography (3), journal (3), http (2), image (2), atlas (2)
12	geography (8), geographer (3), author (3), document (3), science (2)
13	island (5), student (3), geography (3), education (2), identity (2), image (2), university (2)
14	city (23), agglomeration (4), metropolis (3), area (2), urbanization (2)
15	Border (3), China (2), State (2), States (2), Brazil (2), Asia (2), Nation (2)
16	vote (5), map (3), party (2)
17	village (10), Pole (3), Departement (3), Area (2), Map (2), University (2), Student (2)
18	village (5), season (2), rain (2), resort (2), valley (2), precipitation (2), speed (2)
19	landscape (9), heritage (2), image (2), tourist (2)
20	port (4), sea (3), wind (2), station (2), breeze (2), Tunis (2), temperature (2)

tends to decrease. Articles about remote sensing (3) was mainly produced in 2000, just like articles about water management (10) in 2004 and 2011. Articles about agglomeration (14) are regularly produced. Topics such as district (1) and mobility (2) tends to increase.

SPATIALISED FULL-TEXT COMMUNITIES Using the full texts to draw the semantic profile of the 128 countries studied in a Cybergeo article, we obtain a clustering in 4 groups representing 13.4% of the initial inertia. Its geographical distribution is shown in figure 137 with the average profile of each group.

In this clustering analysis, we do not find the dichotomy of countries based on their wealth and economic development levels. The link between semantic and geographical proximity is also less obvious at the world level, although one region is strikingly revealed: the institutional boundaries of Europe. The group of countries included in the EU27 plus the USA, Brazil and Chile (in yellow) appear strongly similar in terms of vocabulary used to talk about them. In particular, themes related to issues of administrative boundaries ("communes") and regional planning ("amenagement") describe these countries well (for example: (Le Néchet, 2011a; Lusso, 2009; Santa-maria, 2009)). Two subgroups neighbour this cluster in the clustering tree. The first one includes countries studied by papers written in English. The second subgroup includes countries from all continents and corresponds to papers written preferentially with words such as "eau" (water) and "entreprise" (enterprise). Finally, 59 countries are distant from these groups in that words used to write about them refer to villages and borders ("frontiere"), in contexts as diverse as Canada, Ecuador, Malaysia or Zimbabwe. The communities of vocabulary and writing practice thus appear less straightforward and less linked to geographical proximity. The main result lays in the fact that there is a specific set of words used to write about the European Union, a sort of EU27 Novlang made of words like "Eurovision", "subsidiarity" and "Spatial Development Perspectives".

c.4.4 Discussion

Why three classifications? Evaluating the complementarity of approaches

This section backs up the previous qualitative comparison of approaches through their spatialization by quantitative measures of their complementarity. Although we have seen that the communities obtained from the three different methods are semantically and geographically distinct, we do not know precisely how they complement each other. The overlapping analysis is complicated by the fact that articles belong simultaneously to several clusters for each classification. Therefore, we compare the methods 2 by 2 by computing the share of articles classified simultaneously in each possible pair of clusters

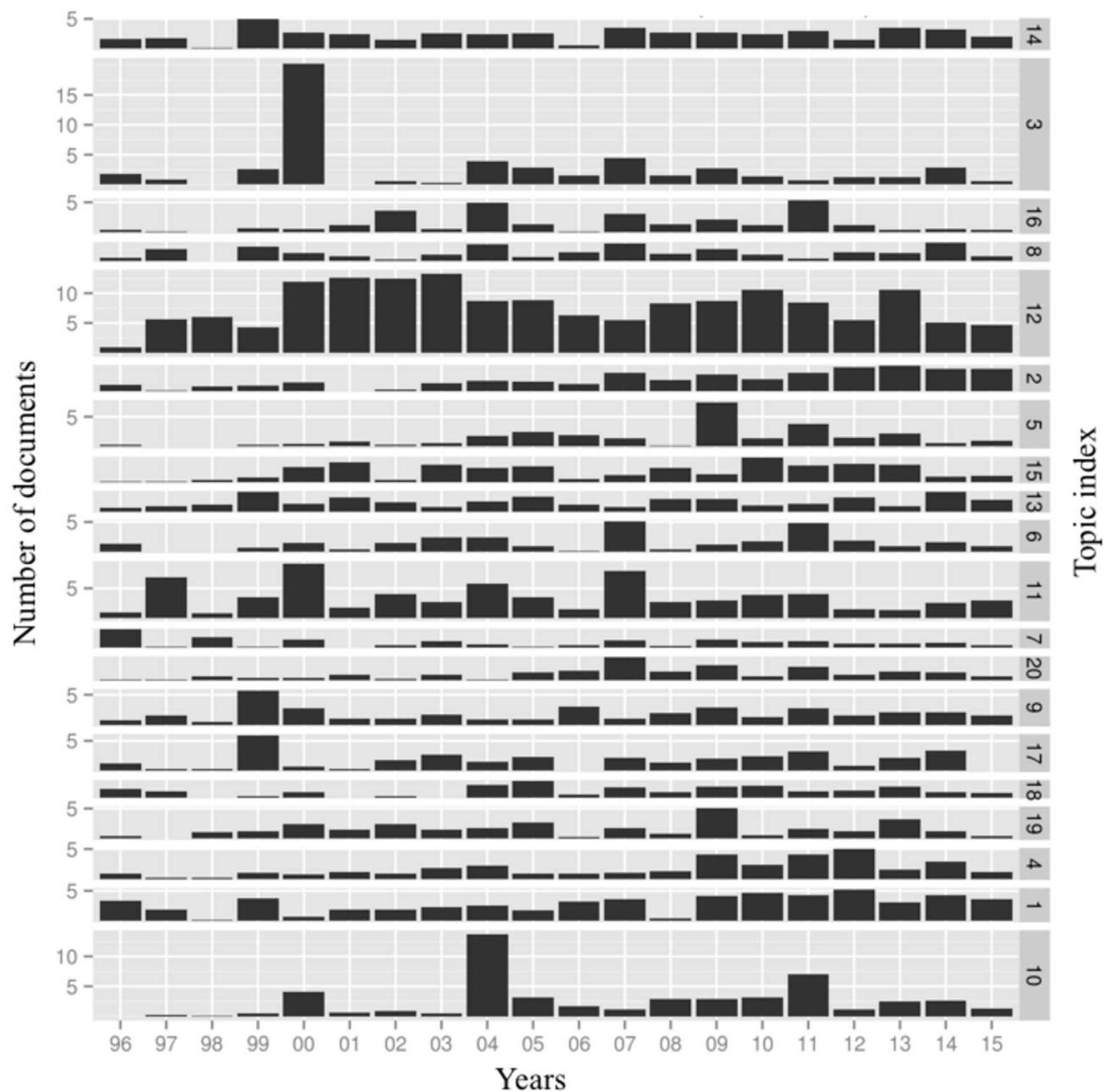


FIGURE 136: Number of documents addressing a topic per year, between 1996 and 2015.

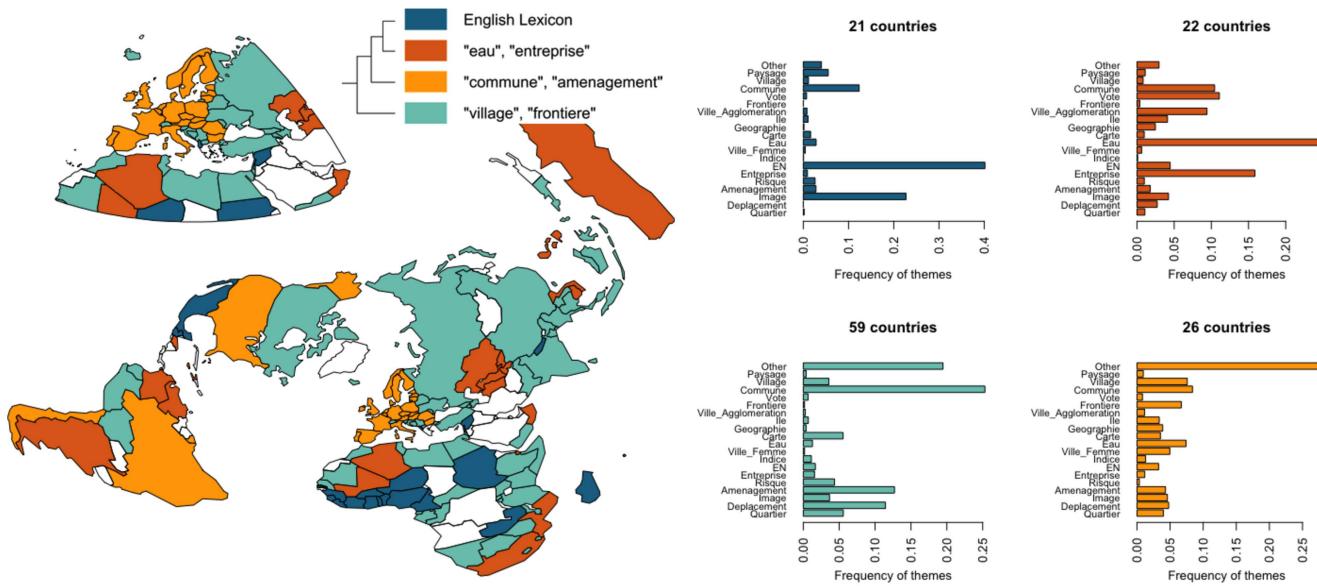


FIGURE 137: Geographical communities of writing practice (Left) ; Corresponding semantic profile of groups (Right)

from the two methods. In other words, if a method M_1 (for ex. based on citation communities) is composed of n categories and a method M_2 (for ex. based on keywords communities) is composed of m categories, we compute for each article $n \cdot m$ products of co-occurrences and then sum these products into flows for the whole Cybergeo corpus. If the methods were equivalent ways of describing and clustering articles, we would expect all the flows between communities to be $1 : 1$, $n : 1$ or $1 : n$, given that the methods do not give the same number of clusters. If the methods were completely orthogonal, we should find that each flow is proportional to the size of the origin cluster and of the size of the destination clusters. The fact that we find $n : n$ flows and that they are not determined entirely by the size of the clusters at origin and destination means that our three methods of semantic clustering are not equivalent nor orthogonal (Figure 138). On the contrary, they shed different lights on the journal corpus.

For instance, there are clear preferential positive and negative relations between some citations communities and keywords communities (figure 138). On the one hand, 35% of the Cybergeo articles cited by papers in the GIS cluster are characterized by keywords identified as "Imagery and GIS". On the other hand, there is no article cited by papers in the "crime" cluster which have keywords of the "Climate and environment" community. These relationships make sense, because the way a paper is advertised by its keywords is one of the first elements indicating the potential reader that the paper is relevant or not. Interestingly, the "complex systems" citation community is characterized by a variety of keywords communities (27% of the

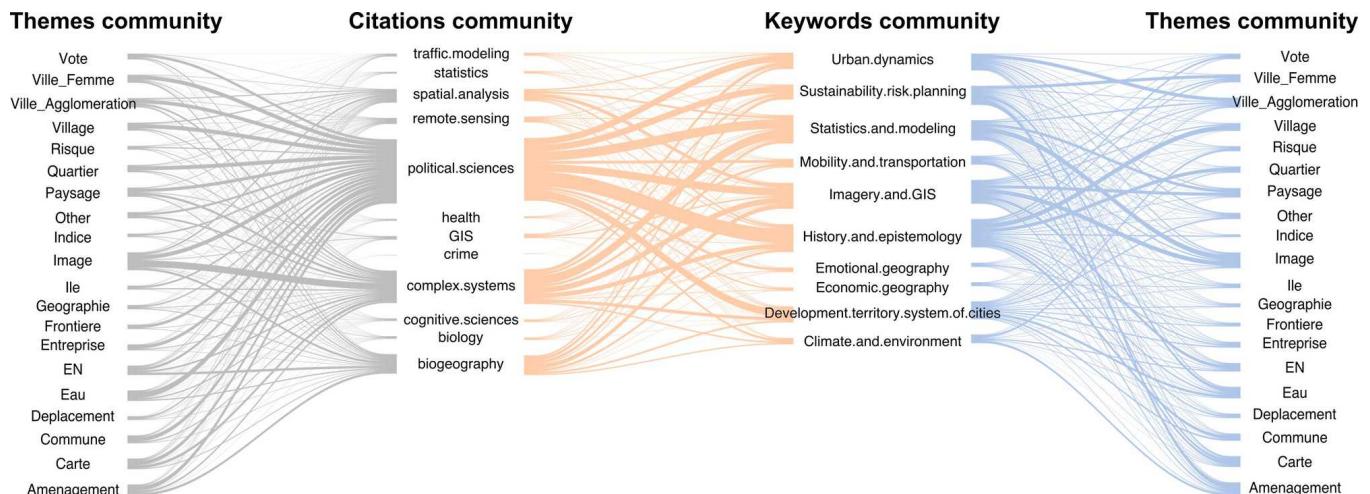


FIGURE 138: Overlap between semantic communities.

articles cited by this community are tagged in the "statistics and modelling" cluster, 17% in "Imagery and GIS" cluster, 13% in "history and epistemology", 11% in "urban dynamics"). This suggests that the field of complex systems, being unified by methods rather than objects of inquiry, are more open to diverse topics than other citation communities. It could also mean that within *Cybergeo*, authors of articles relevant to the complex systems community advertise their paper with keywords from the discipline of geography rather than methods only in order to attract topical reader as well.

Looking at the relations between keywords communities and themes communities, we find that some topics require specific words to write about them. For example, "Imagery and GIS"-tagged articles use more words from the "EN" theme category, which corresponds to English words (rather than French). Urban studies are distinguished between its quantitative side (advertised by keywords around "urban dynamics" and using words such as "agglomeration") and its qualitative side (advertised by keywords around "sustainability, risk and planning" and using words such as "*femme*": woman). Interestingly, the words like "risk" (*risque*) are used themselves more in articles tagged around "Climate and environment" than around "sustainability, risk and planning". Finally, the flows between themes communities and citations communities appear roughly proportional to the size of clusters at origin and destination, suggesting that citations are rather independent of the vocabulary used in the articles. This is reflected in the quantitative analysis below (figure ??), this pair having the smallest mean absolute correlation. In short, the words that count in a citation strategy are much more the keywords than the actual content of the paper.

We synthesize the flow relations between classifications by looking at their covariance structure in an aggregated way. More precisely,

given the probability matrices $(p_{ki}) = (P_i)$ and $(p_{kj}) = (P_j)$ summarizing two classifications, where articles are indexed by rows, we estimate the correlation matrix between their columns $\rho_{ij} = \hat{\rho}[P_i, P_j]$ using a standard Pearson correlation estimator. We look then at aggregated measures, namely minimal correlation, maximal correlation and mean absolute correlation. In order to have a reference to interpret the values of these correlations, we compare them to two null models obtained by bootstrapping random corpuses. The estimate for the lower null model (ρ_0) is expected to minimize correlation and is obtained by shuffling all rows of one of the two matrices, which is done successively on both to ensure symmetry. The upper null model (ρ_+) is constructed by computing correlations between one matrix and the same where a fixed proportion of rows have been shuffled. We set this proportion to 50%, which is a rather high level of similarity, and compute the model for both matrices each time. Average and standard deviations are computed for null models on $b = 10000$ bootstrap repetitions. Table ?? summarizes the results. We find that the maximal correlation for the Cybergeo corpus, which can be interpreted as a maximum overlap between approaches of semantic clustering, is always significantly smaller (around $5 \cdot \sigma$) than for the upper null model. This confirms that our three classifications are highly independent of one another in their main components. It is interesting to note that for Keywords/Themes, the mean absolute correlation is within the standard error range of the mean absolute correlation of the upper null model, suggesting that these two must be rather close on small overlaps. They are actually closer than with Citations for all indicators. We also confirm that Themes/Citations has the lowest mean absolute overlap.

To make these conclusions more robust, we complement the analysis with a network modularity analysis, which is a widely applied method to evaluate the relevance of a classification within a network. To be able to compare two classifications, since the citation network is too sparse for any analysis as mentioned, we evaluate the modularity of a classification within the network induced by the other. More precisely, given a distance threshold θ and two documents given by their probabilities within a classification $\vec{p}_i^{(c)}, \vec{p}_j^{(c)}$, we consider the network with documents as nodes linked if and only if $d(\vec{p}_i^{(c)}, \vec{p}_j^{(c)}) < \theta$ with d euclidian distance. We can then compute the multi-class modularity of the other classification in the sense of [Nicosia et al., 2009]. We show in Fig. 140, for different thresholds, the modularities normalized by the modularity of the network classification within its own network. The closest the measure is to 1, the closer are the classifications. Most of couple have low values for large ranges of θ , confirming the previous conclusions of orthogonality. Furthermore, the different behavior as a function of θ (increasing or decreasing) suggests differ-

FIGURE 139: Correlations between classifications.

	$\min \rho$	$\min \rho_0$	$\min \rho_+$	$\max \rho$	$\max \rho_0$	$\max \rho_+$	$\langle \rho \rangle$	$\langle \rho_0 \rangle$	$\langle \rho_+ \rangle$
Themes/Citations	-0.30 ±0.019	-0.12 ±0.071	-0.17 ±0.071	0.36 ±0.042	0.21 ±0.042	0.69 ±0.070	0.059	0.043 ±0.0021	0.073 ±0.012
Citations/Keywords	-0.26 ±0.015	-0.096 ±0.047	-0.20 ±0.047	0.30 ±0.027	0.13 ±0.027	0.64 ±0.068	0.070	0.034 ±0.0026	0.092 ±0.0081
Keywords/Themes	-0.20 ±0.013	-0.11 ±0.030	-0.13 ±0.030	0.51 ±0.032	0.17 ±0.032	0.66 ±0.075	0.091	0.040 ±0.0022	0.080 ±0.020

Notes: For each pair of classification and measure, we also give average and standard deviation for lower (ρ_0) and upper (ρ_+) null models, obtained by bootstrapping $b = 10000$ random corpuses.

ent *internal structures* of classification, what is consistent with the fact that they rely on different processes to classify data.

Together with the visual diagrams, these analyses show the complementarity of classifications in the exploration of semantic diversity of publication in a 20 year old journal.

Applied Perspectivism

Our approach can be understood as an “applied perspectivism”, which we believe is a way to enhance second order knowledge creation and to ensure reflexivity. Perspectivism is an epistemological position defended by [Giere, 2010c], that aims at going beyond the constructivism-reductivism debates. Focusing on scientific agents as carriers of knowledge creation, any scientific enterprise is a certain *perspective* on the world, taken by the agent for a given purpose and through a given medium that is considered as the *model*. Perspectives are necessary complementary as they result from different approaches to the same objects, even if the definition of objects and research questions will not necessarily be the same. Coupling perspectives should thus be a typical feature of interdisciplinarity. We position our work as a deliberate attempt to couple complementary perspectives on the same corpus. [Varenne, 2017] recalls that one of the various function of models is to foster coupling between theories through coupling of models themselves, allowing the creation of novel knowledge within the virtuous spiral between disciplinarity and interdisciplinarity coined

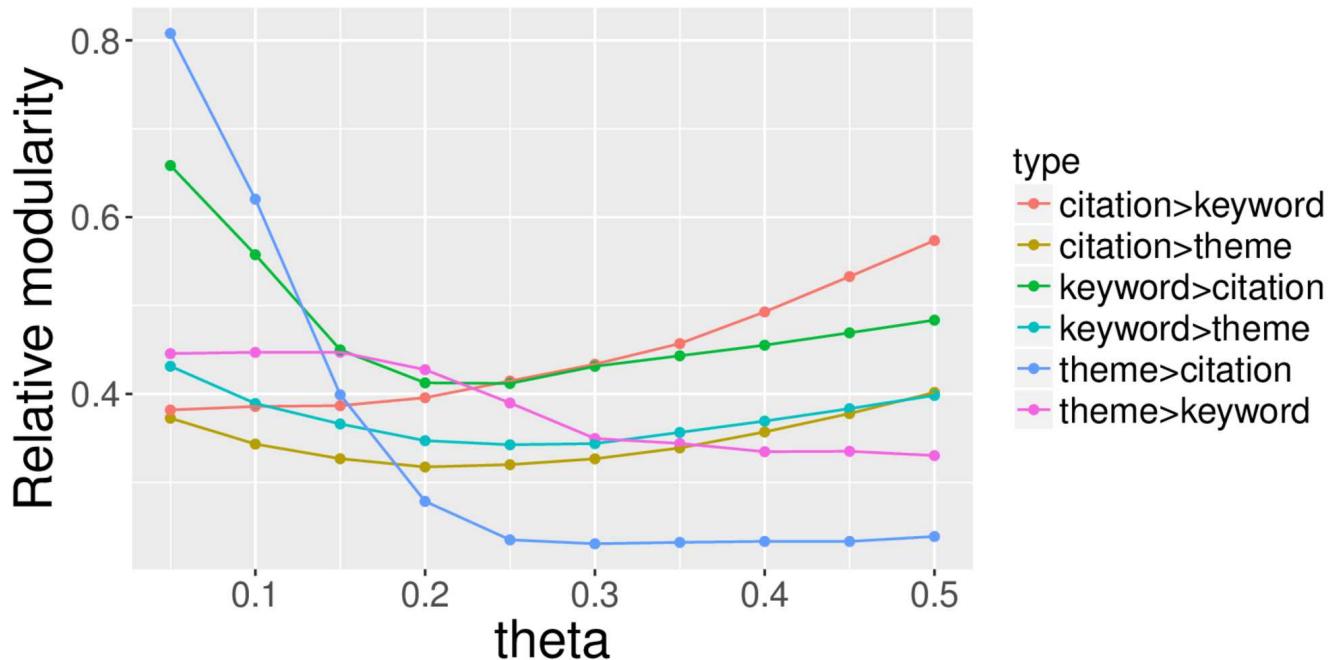


FIGURE 140: Evaluating the complementarity of classification through network modularities. The plot gives the relative modularity of the first classification in the network induced by the second with the threshold θ (see text), for each couple of classifications (color).

by [Banos, 2013]. Our work aims precisely at accelerating and improving such processes.

Fostering Open Science and Reflexivity

The open tools and software we provide participate to a larger effort of reflexivity tools in the context of Open Science. It is aimed at being complementary to existing platforms, like the Community Explorer for the community of Complex Systems developed by IS-CPIF¹⁴ that provides an interactive visualisation of social research networks combined to semantic networks based on self-declared keywords provided by researchers. An other example closer to what we developed is Gargantext¹⁵ that provides corpus exploration functionalities. Linkage¹⁶ is a similar tool with different methods, using latent topic allocation for networks with textual annotations (Bouveyron, Latouche, and Zreik, 2016). We differentiate from these by exploring simultaneously multiple dimensions of semantic classification and more importantly by adding the geographical aspect. Furthermore, in comparison to various tools that private publishers are beginning to introduce, the open and collaborative nature of our work is crucial. For example, [Bohannon, 2014] suggests that one must stay careful

¹⁴ available at <https://communityexplorer.org>

¹⁵ <https://gargantext.org/>

¹⁶ <https://linkage.fr/>

when using search results from a popular academic search engine, as the mechanisms of the ranking algorithm and thus the multiple biases are unknown. The comparison is similar with text-mining paying services provided by private companies, as we suggest that a subtle synergy between knowledge content and knowledge production processes (that is allowed by open tools only) can be more beneficial to both.

We have studied a scientific corpus of a journal in Geography, combining multiple points of view through their embedding in the geographical space. This work is therefore in itself reflexive, illustrating the kind of new approach to science it aims at promoting. We believe that the open tools we develop in this context will contribute to the empowerment of authors within Open Science.

★ ★

★

C.5 CLASSIFYING PATENTS BASED ON THEIR SEMANTIC CONTENT

Les classifications par hyper-réseau obtenues au Chapitre 2 ont été permises par diverses contributions techniques complémentaires. La construction du réseau sémantique, incluant l'extraction des mots-clés et la quantification de leur pertinence, puis son analyse, ont été initialement lancés dans le cadre de l'analyse de la revue *Cybergeo* (voir B.6). L'application à des corpus massifs et la méthode d'extraction de sous-réseau optimal par optimisation de Pareto ont été développé dans le cadre de l'application qui est présentée ici, au corpus de brevets déposés aux Etats-unis de 1976 à 2013.

Cette annexe a ainsi un caractère essentiel du point de vue des méthodes et des outils (même si nous la classons dans les développements thématiques de par son caractère thématique autonome), mais également pour son contenu propre au vu des futurs développements possibles, comme nous avons suggéré pour une caractérisation quantitative de la diffusion de l'innovation dans le cadre d'une investigation empirique des hypothèses de la Théorie Evolutive des Villes.

* * *

*

Cette annexe résulte d'une collaboration avec DR A. BERGEAUD (Paris School of Economics) et DR Y. POTIRON (Keyo University), dans le cadre d'une convergence des problématiques respectives d'analyse sémantique de corpus massifs et de caractérisation endogène de l'innovation. Elle a été publiée comme [Bergeaud, Potiron, and Raimbault, 2017b]. Elle est ici traduite et adaptée.

* * *

*

In this paper, we extend some usual techniques of classification resulting from a large-scale data-mining and network approach. This new technology, which in particular is designed to be suitable to big data, is used to construct an open consolidated database from raw data on 4 million patents taken from the US patent office from 1976 onward. To build the pattern network, not only do we look at each patent title, but we also examine their full abstract and extract

the relevant keywords accordingly. We refer to this classification as *semantic approach* in contrast with the more common *technological approach* which consists in taking the topology when considering US Patent office technological classes. Moreover, we document that both approaches have highly different topological measures and strong statistical evidence that they feature a different model. This suggests that our method is a useful tool to extract endogenous information.

Introduction

Innovation and technological change have been described by many scholars as the main drivers of economic growth as in [Aghion and Howitt, 1992] and [Romer, 1990]. [Griliches, 1990] advertised the use of patents as an economic indicator and as a good proxy for innovation. Subsequently, the easier availability of comprehensive databases on patent details and the increasing number of studies allowing a more efficient use of these data (e.g. [Hall, Jaffe, and Trajtenberg, 2001]) have opened the way to a very wide range of analysis. Most of the statistics derived from the patent databases relied on a few key features: the identity of the inventor, the type and identity of the rights owner, the citations made by the patent to prior art and the technological classes assigned by the patent office post patent's content review. Combining this information is particularly relevant when trying to capture the diffusion of knowledge and the interaction between technological fields as studied in [Youn et al., 2015]. With methods such as citation dynamics modeling discussed in [Newman, 2014] or co-authorship networks analysis in [Sarigöl et al., 2014], a large body of the literature such as [Sorenson, Rivkin, and Fleming, 2006] or [Kay et al., 2014] has studied patents citation network to understand processes driving technological innovation, diffusion and the birth of technological clusters. Finally, [Bruck et al., 2016] look at the dynamics of citations from different classes to show that the laser/ink-jet printer technology resulted from the recombination of two different existing technologies.

Consequently, technological classification combined with other features of patents can be a valuable tool for researchers interested in studying technologies throughout history and to predict future innovations by looking at past knowledge and interaction across sectors and technologies. But it is also crucial for firms that face an ever changing demand structure and need to anticipate future technological trends and convergence (see, e.g., [Curran and Leker, 2011]) to adapt to the resulting increase in competition discussed in [Katz, 1996] and to maintain market share. Curiously, and in spite of the large number of studies that analyze interactions across technologies [Furman and Stern, 2011], little is known about the underlying “innovation network” (e.g. [Acemoglu, Akcigit, and Kerr, 2016]).

In this monograph, we propose an alternative classification based on semantic network analysis from patent abstracts and explore the new information emerging from it. In contrast with the regular technological classification which results from the choice of the patent reviewer, semantic classification is carried automatically based on the content of the patent abstract. Although patent officers are experts in their fields, the relevance of the existing classification is limited by the fact that it is based on the state of technology at the time the patent was granted and cannot anticipate the birth of new fields. To correct for this, the USPTO regularly make changes in its classification in order to adapt to technological change (for example, the “nanotechnology” class (977) was established in 2004 and retroactively to all relevant previously granted patents). In contrast we don’t face this issue with the semantic approach. The semantic links can be clues of one technology taking inspiration from another and good predictors of future technology convergence (e.g. [Preschitschek et al., 2013] study semantic similarities from the whole text of 326 US-patents on *phytosterols* and show that semantic analysis have a good predicting power of future technology convergence). One can for instance consider the case of the word *optic*. Until more recently, this word was often associated with technologies such as photography or eye surgery, while it is now almost exclusively used in a context of semi-transistor design and electro-optic. This semantic shift did not happen by chance but contains information on the fact that modern electronic extensively uses technologies that were initially developed in optic.

Previous research has already proposed to use semantic networks to study technological domains and detect novelty. [Yoon and Park, 2004] was one of the first to enhance this approach with the idea of visualizing keywords network illustrated on a small technological domain. The same approach can be used to help companies identifying the state of the art in their field and avoid patent infringement as in [Park and Yoon, 2014] and [Yoon and Kim, 2011]. More closely related to our methodology, [Gerken and Moehrle, 2012] develop a method based on patent semantic analysis of patent to vindicate the view that this approach outperform others in the monitoring of technology and in the identification of novelty innovation. Semantic analysis has already proven its efficiency in various fields, such as in technology studies (e.g. [Choi and Hwang, 2014] and [Fattori, Pedrazzi, and Turra, 2003]) and in political science (e.g. [Gurciullo et al., 2015]).

Building on such previous research, we make several contributions by fulfilling some shortcomings of existing studies, such as for example the use of frequency-selected single keywords. First of all, we develop and implement a novel fully-automatized methodology to classify patents according to their semantic abstract content, which is to the best of our knowledge the first of its type. This includes the following refinements for which details can be found in Section

??: (i) use of multi-stems as potential keywords; (ii) filtering of keywords based on a second-order (co-occurrences) relevance measure and on an external independent measure (technological dispersion); (iii) multi-objective optimization of semantic network modularity and size. The use of all this techniques in the context of semantic classification is new and essential from a practical perspective.

Furthermore, most of the existing studies rely on a subsample of patent data, whereas we implement it on the full US Patent database from 1976 to 2013. This way, a general structure of technological innovation can be studied. We draw from this application promising qualitative stylized facts, such as a qualitative regime shift around the end of the 1990s, and a significant improvement of citation modularity for the semantic classification when comparing to the technological classification. These thematic conclusions validate our method as a useful tool to extract endogenous information, in a complementary way to the technological classification.

On the account of this information, we believe that patent officers could benefit very much from looking at the semantic network when considering potential citation candidates of a patent in review.

The paper is organized as follows. Section presents the patent data, the existing classification and provide details about the data collection process. Section ?? explains the construction of the semantic classes. Section ?? tests their relevance by providing exploratory results. Finally, section ?? discusses potential further developments and conclude.

Background

In our analysis, we will consider all utility patents granted in the United States Patent and Trademark Office (USPTO) from 1976 to 2013. A clearer definition of utility patent is given in ?? of [Bergeaud, Potiron, and Raimbault, 2017b]. Also, additional information on how to correctly exploit patent data can be found in [Hall, Jaffe, and Trajtenberg, 2001] and [Lerner and Seru, 2015].

An existing classification: the USPC system

Each USPTO patent is associated with a non-empty set of technological classes and subclasses. There are currently around 440 classes and over 150,000 subclasses constituting the United State Patent Classification (USPC) system. While a technological class corresponds to the technological field covered by the patent, a subclass stands for a specific technology or method used in this invention. A patent can have multiple technological classes, on average in our data a patent has 1.8 different classes and 3.9 pairs of class/subclass. At this stage, two features of this system are worth mentioning: (i) classes and subclasses are not chosen by the inventors of the patent but by the examiner dur-

ing the granting process based on the content of the patent; (ii) the classification has evolved in time and continues to change in order to adapt to new technologies by creating or editing classes. When a change occurs, the USPTO reviews all the previous patents so as to create a consistent classification.

A bibliographical network between patents: citations

As with scientific publications, patents must give reference to all the previous patents which correspond to related prior art. They therefore indicate the past knowledge which relates to the patented invention. Yet, contrary to scientific citations, they also have an important legal role as they are used to delimit the scope of the property rights awarded by the patent. One can consult [OECD, 2009] for more details about this. Failing to refer to prior art can lead to the invalidation of the patent (e.g. [Dechezleprat, Martin, and Mohnen, 2014]). Another crucial difference is that the majority of the citations are actually chosen by the examiners and not by the inventors themselves. From the USPTO, we gather information of all citations made by each patent (backward citations) and all citations received by each patent as of the end of 2013 (forward citations). We can thus build a complete network of citations that we will use later on in the analysis.

Turning to the structure of the lag between the citing and the cited patent in terms of application date, we see that the mean of this lag is 8.5 years and the median is 7 years. This distribution is highly skewed, the 95th percentile is 21 years. We also report 164,000 citations with a negative time lag. This is due to the fact that some citations can be added during the examination process and some patents require more time to be granted than others.

In what follows, we choose to restrict attention to pairs of citations with a lag no larger than 5 years. We impose this restriction for two reasons. First, the number of citations received peaks 4-5 years after application. Second, the structure of the citation lag is necessarily biased by the truncation of our sample: the more recent patents mechanically receive less citations than the older ones. As we are restricting to citations received no later than 5 years after the application date, this effect will only affect patents with an application date after 2007.

Data collection and basic description

Each patent contains an abstract and a core text which describe the invention. To see what a patent looks like in practice, one can refer to the USPTO patent full-text database <http://patft.uspto.gov/netahtml/PTO/index.html> or to Google patent which publishes USPTO patents in pdf format at <https://patents.google.com>. Although including the full core texts would be natural and probably very useful in a systematic text-mining approach as done in [Tseng, Lin, and Lin,

2007], they are too long to be included and thus we consider only the abstracts for the analysis. Indeed, the semantic analysis counts more than 4 million patents, with corresponding abstracts with an average length of 120.8 words (and a standard deviation of 62.4), a size that is already challenging in terms of computational burden and data size. In addition, abstracts are aimed at synthesizing purpose and content of patents and must therefore be a relevant object of study (see [Adams, 2010]). The USPTO defines a guidance stating that an abstract should be “a summary of the disclosure as contained in the description, the claims, and any drawings; the summary shall indicate the technical field to which the invention pertains and shall be drafted in a way which allows the clear understanding of the technical problem, the gist of the solution of that problem through the invention, and the principal use or uses of the invention” (PCT Rule 8).

We construct from raw data a unified database. Data is collected from USPTO patent redbook bulk downloads, that provides as raw data (specific dat or xml formats) full patent information, starting from 1976. Detailed procedure of data collection, parsing and consolidation are available in ?? . The latest dump of the database in Mongodb format is available at <http://dx.doi.org/10.7910/DVN/BW3ACK>. Collection and homogenization of the database into a directly usable database with basic information and abstracts was an important task as USPTO raw data formats are involved and change frequently.

We count 4,666,365 utility patents with an abstract granted from 1976 to 2013. A very small number of patents have a missing abstract, these are patents that have been withdrawn and we do not consider them in the analysis. The number of patents granted each year increases from around 70,000 in 1976 to about 278,000 in 2013. When distributed by the year of application, the picture is slightly different. The number of patents steadily increase from 1976 to 2000 and remains constant around 200,000 per year from 2000 to 2007. Restricting our sample to patent with application date ranging from 1976 to 2007, we are left with 3,949,615 patents. These patents cite 38,756,292 other patents with the empirical lag distribution that has been extensively analyzed in [Hall, Jaffe, and Trajtenberg, 2001]. Conditioned on being cited at least once, a patent receives on average 13.5 citations within a five-year window. 270,877 patents receive no citation during the next five years following application, 10% of patents receive only one citation and 1% of them receive more than 100 citations. A within class citation is defined as a citation between two patents sharing at least one common technological class. Following this definition, 84% of the citations are within class citations. 14% of the citations are between two patents that share the exact same set of technological classes.

Towards a Complementary Classification

Potentialities of text-mining techniques as an alternative way to analyze and classify patents are documented in [Tseng, Lin, and Lin, 2007]. The author's main argument, in support of an automatic classification tool for patent, is to reduce the considerable amount of human effort needed to classify all the applications. The work conducted in the field of natural language processing and/or text analysis has been developed in order to improve search performance in patent databases, build technology map or investigate the potential infringement risks prior to developing a new technology (see [Abbas, Zhang, and Khan, 2014] for a review). Text-mining of patent documents is also widely used as a tool to build networks which carry additional information to the simplistic bibliographic connections model as argued in [Yoon and Park, 2004]. As far as the authors know, the use of text-mining as a way to build a global classification of patents remains however largely unexplored. One notable exception can be found in [Preschitschek et al., 2013] where semantic-based classification is shown to outperform the standard classification in predicting the convergence of technologies even in small samples. Semantic analysis reveals itself to be more flexible and more quickly adaptable to the apparition of new clusters of technologies. Indeed, as argued in [Preschitschek et al., 2013], before two distinct technologies start to clearly converge, one should expect similar words to be used in patents from both technologies.

Finally, a semantic classification where patents are gathered based on the fact that they share similar significant keywords has the advantage of including a network feature that cannot be found in the USPC case, namely that each patent is associated with a vector of probability to belong to each of the semantic classes (more details on this feature can be found in Section ??). Using co-occurrence of keywords, it is then possible to construct a network of patents and to study the influence of some key topological features. As reviewed previously, the use of co-occurrences is the usual way to construct a semantic network. Other hybrid technique such as bipartite semantic/authors networks, do not have the nice feature of relying solely on endogenous semantic information contained in data.

Semantic Classification Construction

In this section, we describe methods and empirical analysis leading to the construction of semantic network and the corresponding classification.

Keywords extraction

Let \mathcal{P} be the set of patents, we first assign to a patent $p \in \mathcal{P}$ a set of potentially significant keywords $K(p)$ from its text $A(p)$ (that corresponds to the concatenation of its own title and abstract). $K(p)$ are extracted through a similar procedure as the one detailed in [Chavalarrias and Cointet, 2013]:

1. Text parsing and Tokenization: we transform raw texts into a set of words and sentences, reading it (parsing) and splitting it into elementary entities (words organized in sentences).
2. Part-of-speech tagging: attribution of a grammatical function to each of the tokens defined previously.
3. Stem extraction: families of words are generally derived from a unique root called stem (for example `compute`, `computer`, `computation` all yield the same stem `comput`) that we extract from tokens. At this point the abstract text is reduced to a set of stems and their grammatical functions.
4. Multi-stems construction: these are the basic semantic units used in further analysis. They are constructed as groups of successive stems in a sentence which satisfies a simple grammatical function rule. The length of the group is between 1 and 3 and its elements are either nouns, attributive verbs or adjectives. We choose to extract the semantics from such nominal groups in view of the technical nature of texts, which is not likely to contain subtle nuances in combinations of verbs and nominal groups.

Text processing operations are implemented in python in order to use built-in functions `nltk` library [NLTK, 2015] for most of above operations. This library supports most of state-of-the-art natural language processing operations. Source code is openly available on the repository of the project at <https://github.com/JusteRaimbault/PatentsMining>.

Keywords relevance estimation

RELEVANCE DEFINITION Following the heuristic in [Chavalarrias and Cointet, 2013], we estimate relevance score in order to filter multi-stem. The choice of the total number of keywords to be extracted, which we shall denote K_w , is important, too small a value would yield similar network structures but including less information whereas very large values tend to include too many irrelevant keywords. We choose to set this parameter to $K_w = 100,000$. We first consider the filtration of $k \cdot K_w$ (with $k = 4$) to keep a large set of potential keywords but still have a reasonable number of co-occurrences to be computed.

This step has only very marginal effects on the nature of the final keywords but is necessary for computational purposes. The filtration is done on the *unithood* u_i , defined for keyword i as $u_i = f_i \cdot \log(1 + l_i)$ where f_i is the multi-stem's number of apparitions over the whole corpus and l_i its length in words. A second filtration of K_w keywords is done on the *termhood* t_i , where the formal definition can be found in (33). It is computed as a chi-squared score on the distribution of the stem's co-occurrences and then compared to a uniform distribution within the whole corpus. Intuitively, uniformly distributed terms will be identified as plain language and they are thus not relevant for the classification. More precisely, we compute the co-occurrence matrix (M_{ij}), where M_{ij} is defined as the number of patents where stems i and j appear together. The *termhood* score t_i is defined as

$$t_i = \sum_{j \neq i} \frac{(M_{ij} - \sum_k M_{ik} \sum_k M_{jk})^2}{\sum_k M_{ik} \sum_k M_{jk}}. \quad (33)$$

MOVING WINDOW ESTIMATION The previous scores are estimated on a moving window with fixed time length following the idea that the present relevance is given by the most recent context and thus that the influence vanishes when going further into the past. Consequently, the co-occurrence matrix is chosen to be constructed at year t restricting to patent which applied during the time window $[t - T_0; t]$. Note that the causal property of the window is crucial as the future cannot play any role in the current state of keywords and patents. This way, we will obtain semantic classes which are exploitable on a T_0 time span. For example, this enables us to compute the modularity of classes in the citation network as in section ???. In the following, we take $T_0 = 4$ (which corresponds to a five year window) consistently with the choice of maximum time lag for citations made in Section ???. Accordingly, the sensitivity analysis for $T_0 = 2$ can be found in Appendix ??.

Construction of the semantic network

We keep the set of most relevant keywords \mathcal{K}_W and obtain their co-occurrence matrix as defined in Section ???. This matrix can be directly interpreted as the weighted adjacency matrix of the semantic network. At this stage, the topology of raw networks does not allow the extraction of clear communities. This is partly due to the presence of hubs that correspond to frequent terms common to many fields (e.g. *method*, *apparat*) which are wrongly filtered as relevant. We therefore introduce an additional measure to correct the network topology: the concentration of keywords across technological classes, defined as:

$$c_{\text{tech}}(s) = \sum_{j=1}^{N^{(\text{tec})}} \frac{k_j(s)^2}{(\sum_i k_i(s))^2},$$

where $k_j(s)$ is the number of occurrences of the s th keyword in each of the j th technological class taken from one of the $N^{(\text{tec})}$ USPC classes. The higher c_{tech} , the more specific to a technological class the node is. For example, the terms **semiconductor** is widely used in electronics and does not contain any significant information in this field. We use a threshold parameter, defined as θ_c , and keep nodes with $c_{\text{tech}}(s) > \theta_c$. Likewise, edges with low weights correspond to rare co-occurrences and are considered to be noise. To account for this we define the threshold parameter for edges θ_w , and we filter edges with a weight below θ_w , following the rationale that two keywords are not linked “by chance” if they appear simultaneously a minimal number of time. To control for size effect, we normalize by taking $\theta_w = \theta_w^{(0)} \cdot N_P$ where N_P is the number of patents in the corpus ($N_P = |\mathcal{P}|$). $\theta_w^{(0)}$ is thus a varying parameter interpreted as a noise threshold *per patent*. Communities are then extracted using a standard modularity maximization procedure as described in [Clauset, Newman, and Moore, 2004] to which we add the two constraints captured by θ_w and θ_c , namely that edges must have a weight greater than θ_w and nodes a concentration greater than θ_c . At this stage, both parameters θ_c and $\theta_w^{(0)}$ are unconstrained and their choice is not straightforward. Indeed, many optimization objectives are possible, such as the modularity, network size or number of communities. We find that modularity is maximized at a roughly stable value of θ_w across different θ_c for each year, corresponding to a stable $\theta_w^{(0)}$ across years, which leads us to choose $\theta_w^{(0)} = 4.1 \cdot 10^{-5}$. Then for the choice of θ_c , different candidates points lie on a Pareto front for the bi-objective optimization on number of communities and network size. There is a priori no reason to choose any specific point among the different optima. Consequently, we have tried the analysis with all the candidate values for θ_c and found that the results are the most reasonable when taking $\theta_c = 0.06$ (see Fig. 141). We show in Fig. 142 an example of semantic network visualization.

Characteristics of Semantic Classes

For each year t , we define as $N_t^{(\text{sem})}$ the number of semantic classes which have been computed by clustering keywords from patents appeared during the period $[t - T_0, t]$ (we recall that we have chosen $T_0 = 4$). Each semantic class $k = 1, \dots, N_t^{(\text{sem})}$ is characterized by a set of keywords $K(k, t)$ which is a subset of \mathcal{K}_W selected as described in previous sections. The cardinal of $K(k, t)$ distribution across each semantic class k is highly skewed with a few semantic classes contain-

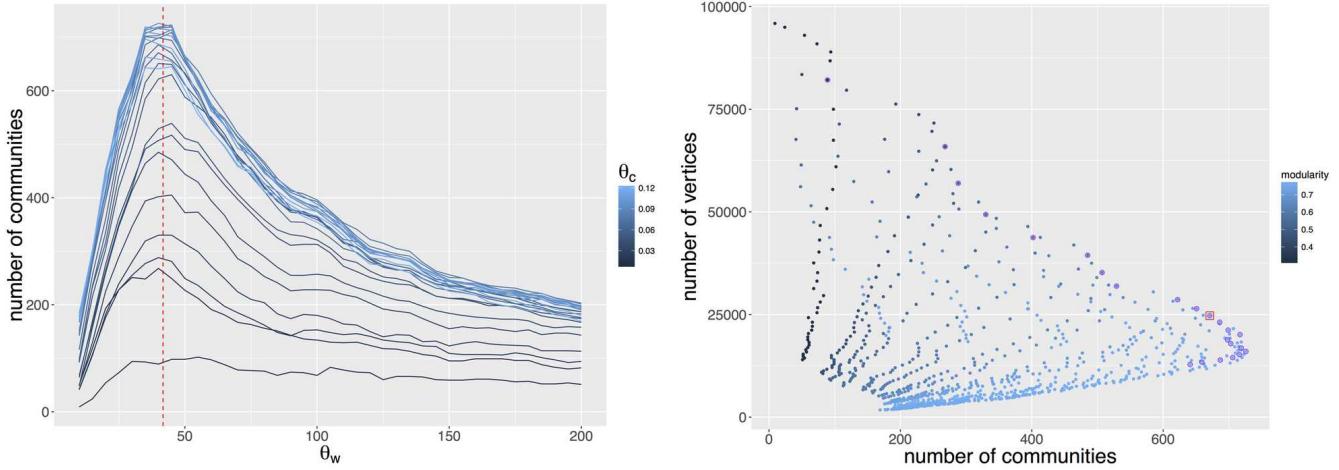


FIGURE 141: Sensitivity analysis of network community structure to filtering parameters. We consider a specific window 2000-2004 and the obtained plots are typical. (*Left panel*) We plot the number of communities as a function of the edge threshold parameter θ_w for different values of the node threshold parameter θ_c . The maximum is roughly stable across θ_c (dashed red line). (*Right panel*) To choose θ_c , we do a Pareto optimization on communities and network size: the compromise point (red overline) on the Pareto front (purple overline: possible choices after having fixed $\theta_w^{(0)}$; blue level gives modularity) corresponds to $\theta_c = 0.06$.

ing over 1,000 keywords, most of them with roughly the same number of keywords. In contrast, there are also many semantic classes with only two keywords. There are around 30 keywords by semantic class on average and the median is 2 for any t . Fig. 143 shows that the average number of keywords is relatively stable from 1976 to 1992 and then picks around 1996 prior to going down.

TITLE OF SEMANTIC CLASSES USPC technological classes are defined by a title and a highly accurate definition which help retrieve patents easily. The title can be a single word (e.g.: class 101: “Printing”) or more complex (e.g.: class 218: “High-voltage switches with arc preventing or extinguishing devices”). As our goal is to release a comprehensive database in which each patent is associated with a set of semantic classes, it is necessary to give an insight on what these classes represent by associating a short description or a title as in [Tseng, Lin, and Lin, 2007]. In our case, such description is taken as a subset of keywords taken from $K(k, t)$. For the vast majority of semantic classes that have less than 5 keywords, we decide to keep all of these keywords as a description. For the remaining classes which feature around 50 keywords on average, we rely on the topological properties of the semantic network. [Yang et al., 2000] suggest to retain only the most frequently used terms in $K(k, t)$. Another possibility is to select 5 keywords based on their network centrality with the idea that very central keywords are the best candidates to describe

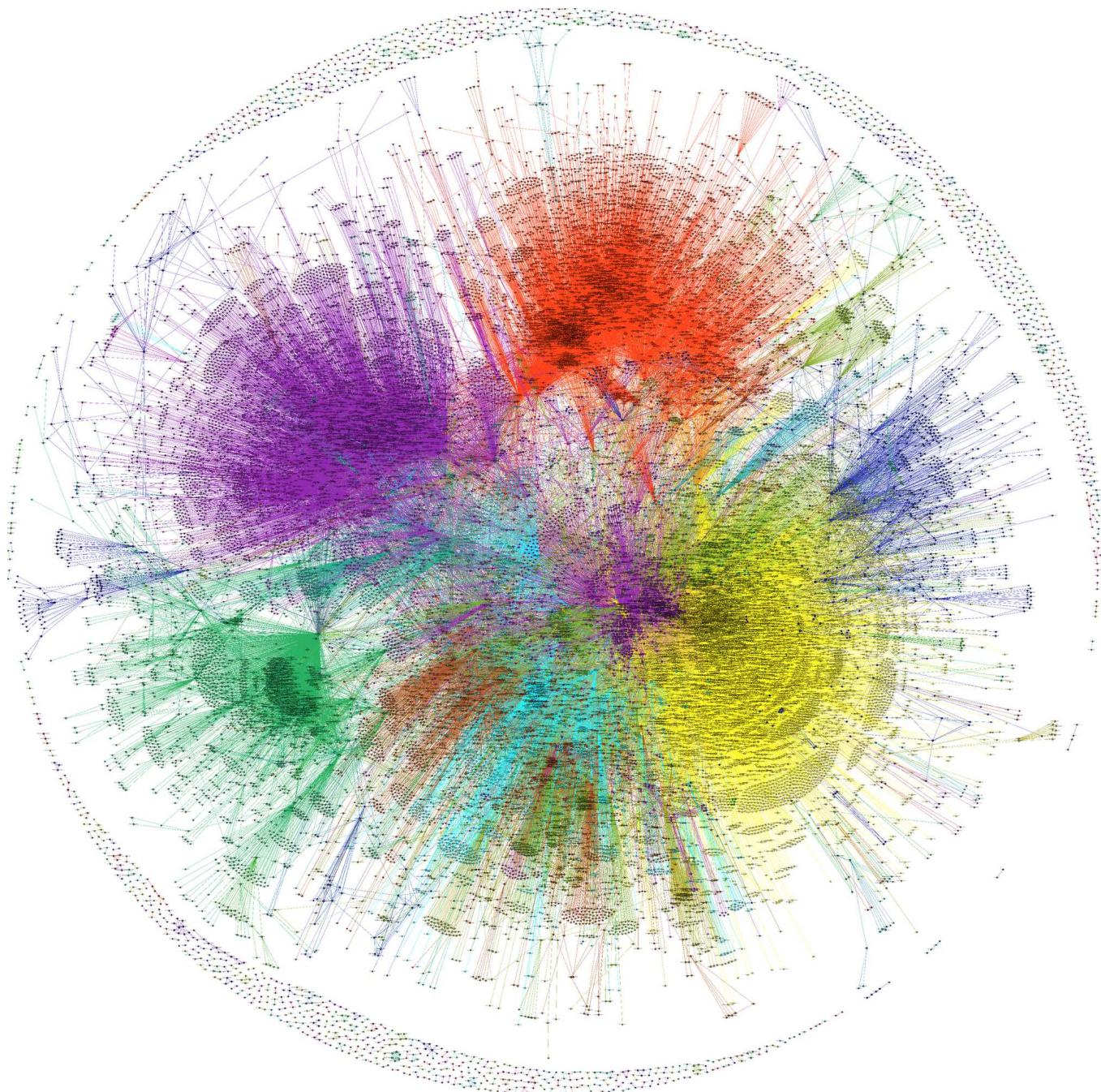


FIGURE 142: An example of semantic network visualization. We show the network obtained for the window 2000-2004, with parameters $\theta_c = 0.06$ and $\theta_w = \theta_w^{(0)} \cdot N_p = 4.5e^{-5} \cdot 9.1e^5$. The corresponding file in a vector format (.svg), that can be zoomed and explored, is available as ??.

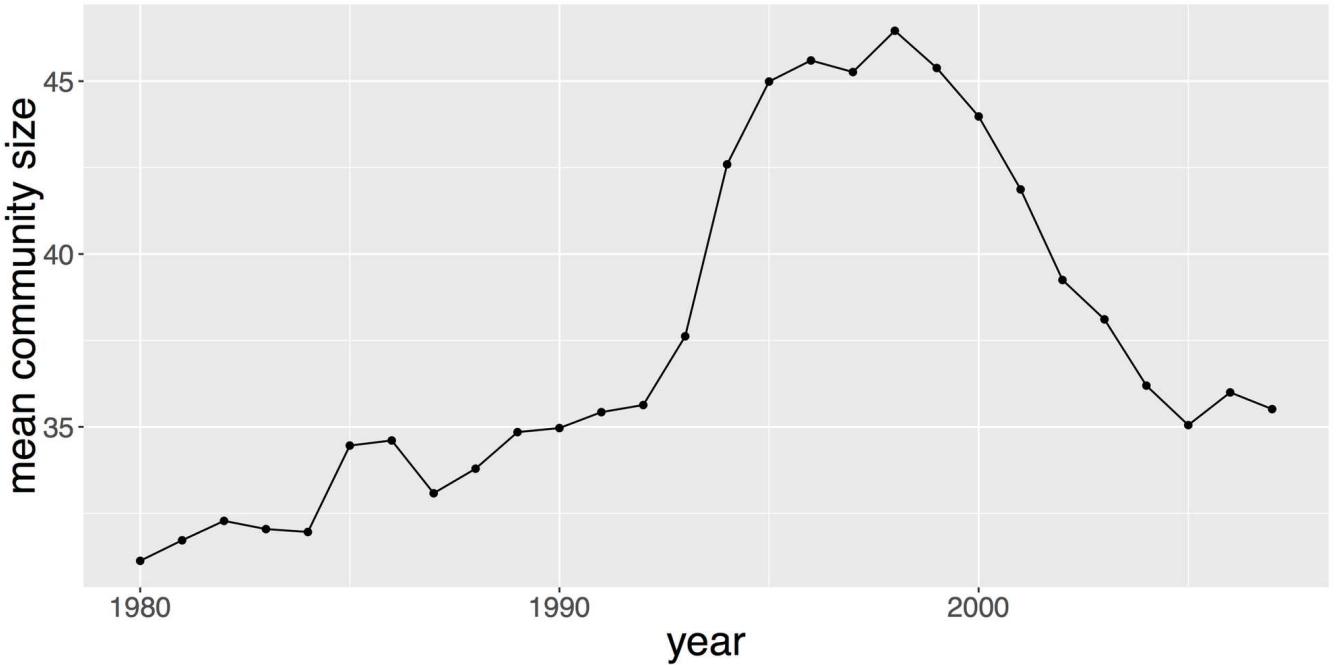


FIGURE 143: This figure plots the average number of keywords by semantic class for each time window $[t-4; t]$ from $t = 1980$ to $t = 2007$.

the overall idea captured by a community. For example, the largest semantic class in 2003-2007 is characterized by the keywords: Support Packet; Tree Network; Network Wide; Voic Stream; Code Symbol Reader.

SIZE OF TECHNOLOGICAL AND SEMANTIC CLASSES We consider a specific window of observations (for example 2000-2004), and we define Z the number of patents which appeared during that time window. For each patent $i = 1, \dots, Z$ we associate a vector of probability where each component $p_{ij}^{(sem)} \in [0, 1]$, with $j = 1, \dots, N(sem)$ and where

$$\sum_{j=1}^{N^{(sem)}} p_{ij}^{(sem)} = 1$$

(when there is no room for confusion, we drop the subscript t in $N_t^{(sem)}$). On average across all time windows, a patent is associated to 1.8 semantic classes with a positive probability. Next we define the size of a semantic class as

$$S_j^{(sem)} = \sum_{i=1}^Z p_{ij}^{(sem)}.$$

Correspondingly, we aim to provide a consistent definition for technological classes. For that purpose, we follow the so-called “fractional count” method, which was introduced by the USPTO and consists

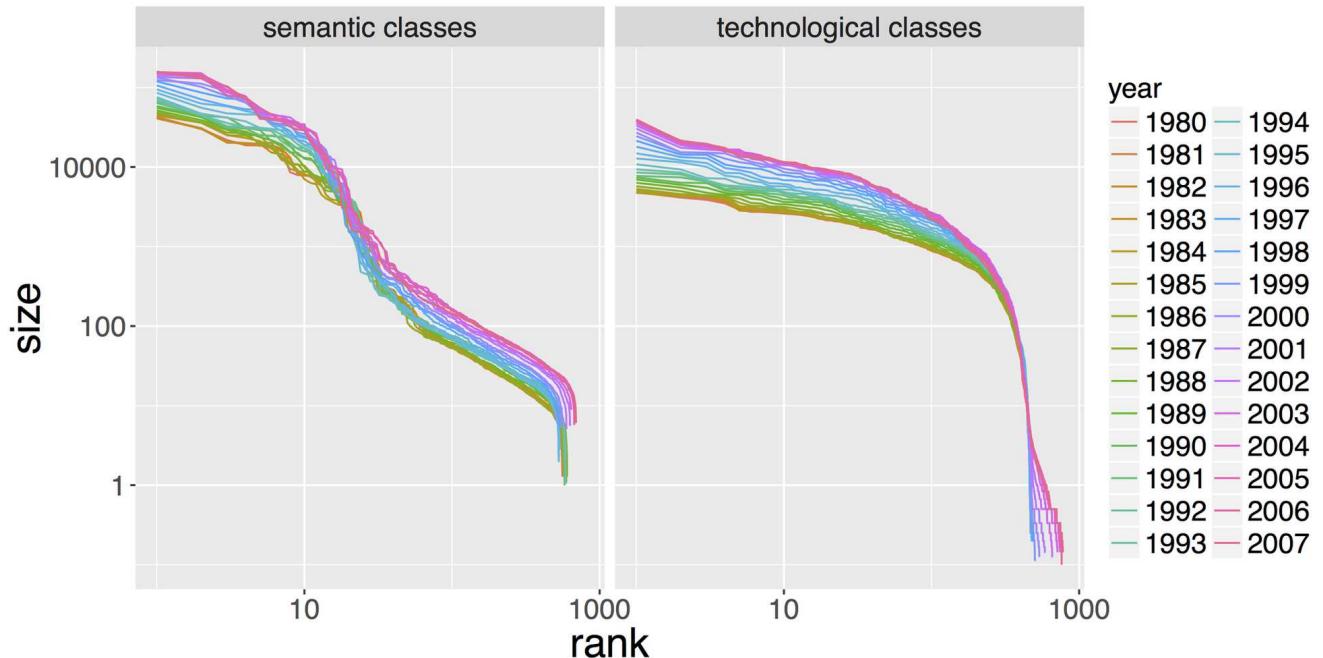


FIGURE 144: Sizes of classes. Yearly from $t = 1980$ to $t = 2007$, we plot the size of semantic classes (left-side) and technological classes (right-side) for the corresponding time window $[t - 4, t]$, from the biggest to the smallest. The formal definition of size can be found in Section ???. Each color corresponds to one specific year. Yearly semantic classes and technological classes present a similar hierarchical structure which confirms the comparability of the two classifications. Over time, curves are translated and levels of hierarchy stays roughly constant.

in dividing equally the patents between all the classes they belong to. Formally, we define the number of technological classes as $N^{(tec)}$ (which is not time dependent contrary to the semantic case) and for $j = 1, \dots, N^{(tec)}$ the corresponding matrix of probability is defined as

$$p_{ij}^{(tec)} = \frac{B_{ij}}{\sum_{k=1}^{N^{(tec)}} B_{ik}},$$

where B_{ij} equals 1 if the i th patent belongs to the j th technological class and 0 if not. When there is no room for confusion, we will drop the exponent part and write only p_{ij} when referring to either the technological or semantic matrix. Empirically, we find that both classes exhibit a similar hierarchical structure in the sense of a power-law type of distribution of class sizes as shown in Fig. ???. This feature is important, it suggests that a classification based on the text content of patents has some separating power in the sense that it does not divide up all the patents in one or two communities.

Potential Refinements of the Method

Our semantic classification method could be refined by combining it with other techniques such as Latent Dirichlet Allocation which is a widely used topic detection method (e.g. [Blei, Ng, and Jordan, 2003]), already used on patent data as in [Kaplan and Vakili, 2015] where it provides a measure of idea novelty and the counter-intuitive stylized facts that breakthrough invention are likely to come out of local search in a field rather than distant technological recombination. Using this approach should first help further evaluate the robustness of our qualitative conclusions (external validation). Also, depending on the level of orthogonality with our classification, it can potentially bring an additional feature to characterize patents, in the spirit of multi-modeling techniques where neighbor models are combined to take advantage of each point of view on a system.

Our use of network analysis can also be extended using newly developed techniques of hyper-network analysis. Indeed, patents and keywords can for example be nodes of a bipartite network, or patents be links of an hyper-network, in the sense of multiple layers with different classification links and citation links. The combination of citation network modeling by Stochastic Block Modeling with topic modeling was studied for scientific papers by [Zhu et al., 2013b], outperforming previous link prediction algorithms. [Iacovacci, Wu, and Bianconi, 2015] provide a method to compare macroscopic structures of the different layers in a multilayer network that could be applied as a refinement of the overlap, modularity and statistical modeling studied in this paper. Furthermore, it has recently been shown that measures of multilayer network projections induce a significant loss of information compared to the generalized corresponding measure [De Domenico et al., 2015], which confirms the relevance of such development that we left for further research.

An other potential research development would be to further exploit the temporal structure of our dataset. Indeed, large progress have recently been made in complex network analysis of time-series data (see [Gao, Small, and Kurths, 2017] for a review). For example, [Gao et al., 2015] develops a method to construct multiscale network from time series, which could in our case be a solution to identify structures in patents trajectories at different levels, and be an alternative to the single scale modularity analysis we use.

Results

In this section, we present some key features of our resulting semantic classification showing both complementary and differences with the technological classification. We first present several measures derived from this semantic classification at the patent level: Diversity, Originality, Generality (Section ??) and Overlapping (Section ??). We then

show that the two classifications show highly different topological measures.

Patent Level Measures

Given a classification system (technological or semantic classes), and the associated probabilities p_{ij} for each patent i to belong to class j (that were defined in Section ??), one can define a patent-level diversity measure as one minus the Herfindhal concentration index on p_{ij} by

$$D_i^{(z)} = 1 - \sum_{j=1}^{N^{(z)}} p_{ij}^2, \text{ with } z \in \{\text{tec, sem}\}.$$

We show in Fig. 145 the distribution over time of semantic and technological diversity with the corresponding mean time-series. This is carried with two different settings, namely including/not including patents with zero diversity (i.e. single class patents). We call other patents “complicated patents” in the following. First of all, the presence of mass in small probabilities for semantic but not technological diversity confirms that the semantic classification contains patent spread over a larger number of classes. More interestingly, a general decrease of diversity for complicated patents, both for semantic and technological classification systems, can be interpreted as an increase in invention specialization. This is a well-known stylized fact as documented in [Archibugi and Pianta, 1992]. Furthermore, a qualitative regime shift on semantic classification occurs around 1996. This can be seen whether or not we include patents with zero diversity. The diversity of complicated patents stabilizes after a constant decrease, and the overall diversity begins to strongly decrease. This means that on the one hand the number of single class patents begins to increase and on the other hand complicated patents do not change in diversity. It can be interpreted as a change in the regime of specialization, the new regime being caused by more single-class patents.

More commonly used in the literature are the measures of originality and generality. These measures follow the same idea than the above-defined diversity in quantifying the diversity of classes (whether technological or semantic) associated with a patent. But instead of looking at the patent’s classes, they consider the classes of the patents that are cited or citing. Formally, the originality O_i and the generality G_i of a patent i are defined as

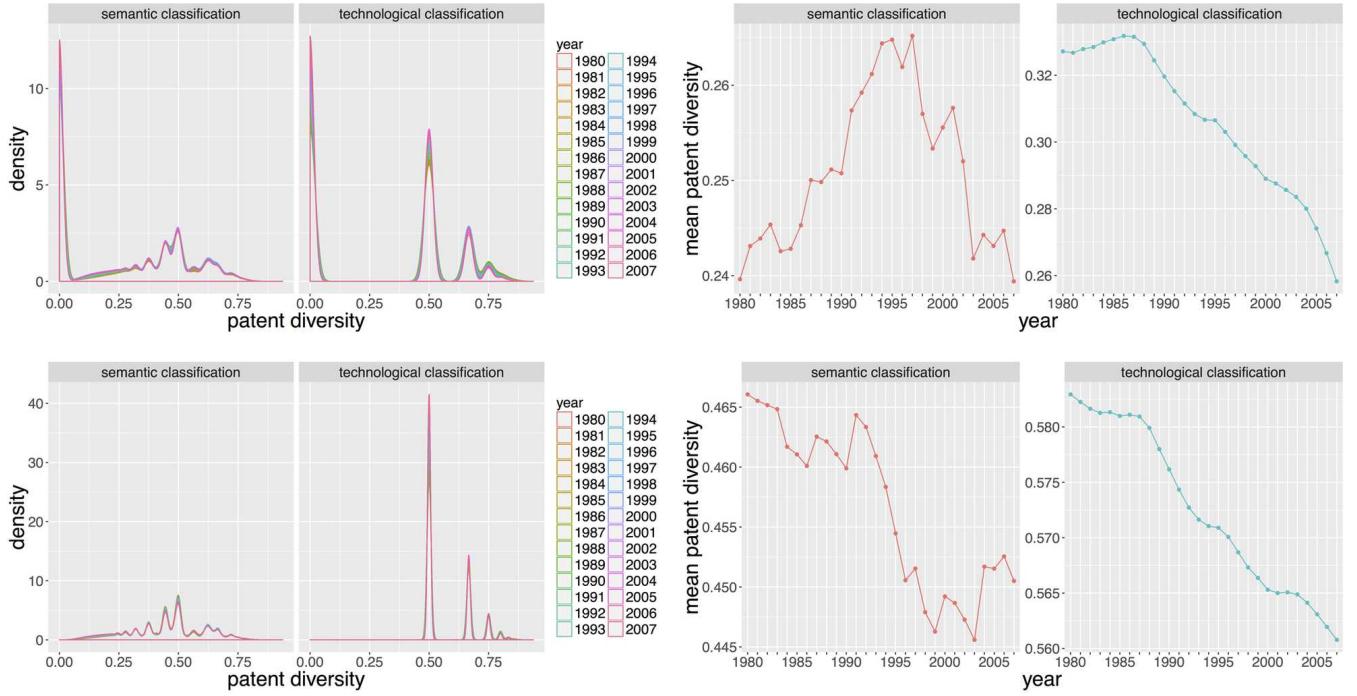


FIGURE 145: Patent level diversities. Distributions of diversities (Left column) and corresponding mean time-series (Right column) for $t = 1980$ to $t = 2007$ (with the corresponding time window $[t - 4, t]$). The first row includes all classified patents, whereas the second row includes only patents with more than one class (i.e. patents with diversity greater than 0).

$$O_i^{(z)} = 1 - \sum_{j=1}^{N^{(z)}} \left(\frac{\sum_{i' \in I_i} p_{i'j}}{\sum_{k=1}^{N^{(z)}} \sum_{i' \in I_i} p_{i'k}} \right)^2 \quad \text{and} \quad G_i^{(z)} = 1 - \sum_{j=1}^{N^{(z)}} \left(\frac{\sum_{i' \in \tilde{I}_i} p_{i'j}}{\sum_{k=1}^{N^{(z)}} \sum_{i' \in \tilde{I}_i} p_{i'k}} \right)^2,$$

where $z \in \{tec, sem\}$, I_i denotes the set of patents that are cited by the i th patent within a five year window (i.e. if the i th patent appears at year t , then we consider patents on $[t - T_0, t]$) when considering the originality and \tilde{I}_i the set of patents that cite patent i after less than five years (i.e. we consider patents on $[t, t + T_0]$) in the case of generality. Note that the measure of generality is forward looking in the sense that $G_i^{(z)}$ used information that will only be available 5 years after patent applications. Both measures are lower on average based on semantic classification than on technological classification. Fig. 146 plots the mean value of $O_i^{(sem)}$, $O_i^{(tec)}$, $G_i^{(sem)}$ and $G_i^{(tec)}$.

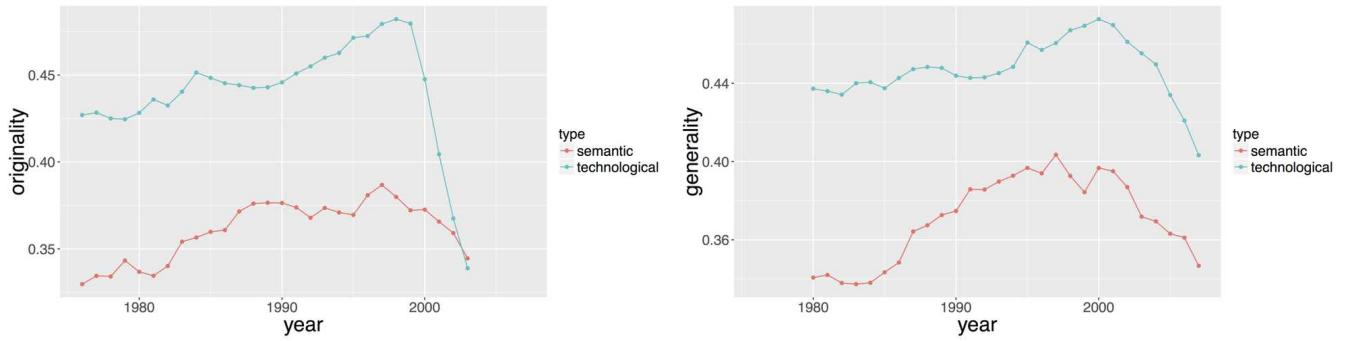


FIGURE 146: Patent level originality (left hand side) and generality (right hand side) for $t = 1980$ to $t = 2007$ (with the corresponding time window $[t - 4, t]$) as defined in subsection ??.

Classes overlaps

A proximity measure between two classes can be defined by their overlap in terms of patents. Such measures could for example be used to construct a metrics between semantic classes. Intuitively, highly overlapping classes are very close in terms of technological content and one can use them to measure distance between two firms in terms of technology as done in [Bloom, Schankerman, and Reenen, 2013]. Formally, recalling the definition of (p_{ij}) as the probability for the i th patent to belong to the j th class and N_P as the number of patents it writes

$$\text{Overlap}_{jk} = \frac{1}{N_P} \cdot \sum_{i=1}^{N_P} p_{ij} p_{ik}. \quad (34)$$

The overlap is normalized by patent count to account for the effect of corpus size: by convention, we assume the overlap to be maximal when there is only one class in the corpus. A corresponding relative overlap is computed as a set similarity measure in the number of patents common to two classes A and B , given by $\sigma(A, B) = 2 \cdot \frac{|A \cap B|}{|A| + |B|}$.

INTRA-CLASSIFICATION OVERLAPS The study of distributions of overlaps inside each classification, i.e. between technological classes and between semantic classes separately, reveals the structural difference between the two classification methods, suggesting their complementary nature. Their evolution in time can furthermore give insights into trends of specialization. We show in Fig. 147 distributions and mean time-series of overlaps for the two classifications. The technological classification globally always follow a decreasing trend, corresponding to more and more isolated classes, i.e. specialized inventions, confirming the stylized fact obtained in previous subsec-

tion. For semantic classes, the dynamic is somehow more intriguing and supports the story of a qualitative regime shift suggested before. Although globally decreasing as technological overlap, normalized (resp. relative) mean overlap exhibits a peak (clearer for normalized overlap) culminating in 1996 (resp. 1999). Looking at normalized overlaps, classification structure was somewhat stable until 1990, then strongly increased to peak in 1996 and then decrease at a similar pace up to now. Technologies began to share more and more until a breakpoint when increasing isolation became the rule again. An evolutionary perspective on technological innovation [Ziman, 2003] could shed light on possible interpretations of this regime shift: as species evolve, the fitness landscape first would have been locally favorable to cross-insemination, until each fitness reaches a threshold above which auto-specialization becomes the optimal path. It is very comparable to the establishment of an ecological niche [Holland, 2012], the strong interdependency originating here during the mutual insemination resulting in a highly path-dependent final situation.

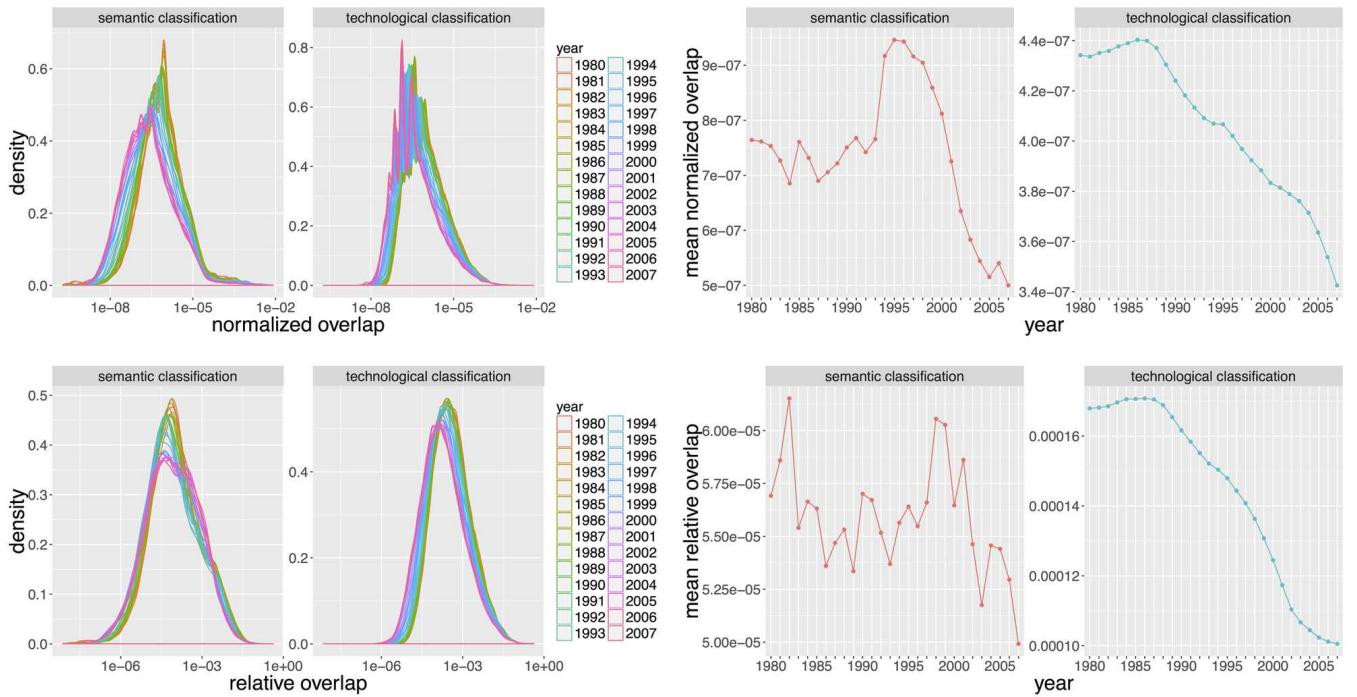


FIGURE 147: Intra-Classification overlaps. (Left column) Distribution of overlaps O_{ij} for all $i \neq j$ (zero values are removed because of the log-scale). Right column) Corresponding mean time-series. (First row) Normalized overlaps. (Second row) Relative overlaps.

INTER-CLASSIFICATION OVERLAPS Overlaps between classifications are defined as in (), but with j standing for the j th technological class and k for the k th semantic class: p_{ij} are technological probabilities and p_{ik} semantic probabilities. They describe the relative correspon-

dence between the two classifications and are a good indicator to spot relative changes, as shown in Fig. 148. Mean inter-classification overlap clearly exhibits two linear trends, the first one being constant from 1980 to 1996, followed by a constant decrease. Although difficult to interpret directly, this stylized fact clearly unveils a change in the *nature* of inventions, or at least in the relation between content of inventions and technological classification. As the tipping point is at the same time as the ones observed in the previous section and since the two statistics are different, it is unlikely that this is a mere coincidence. Thus, these observations could be markers of a hidden underlying structural changes in processes.

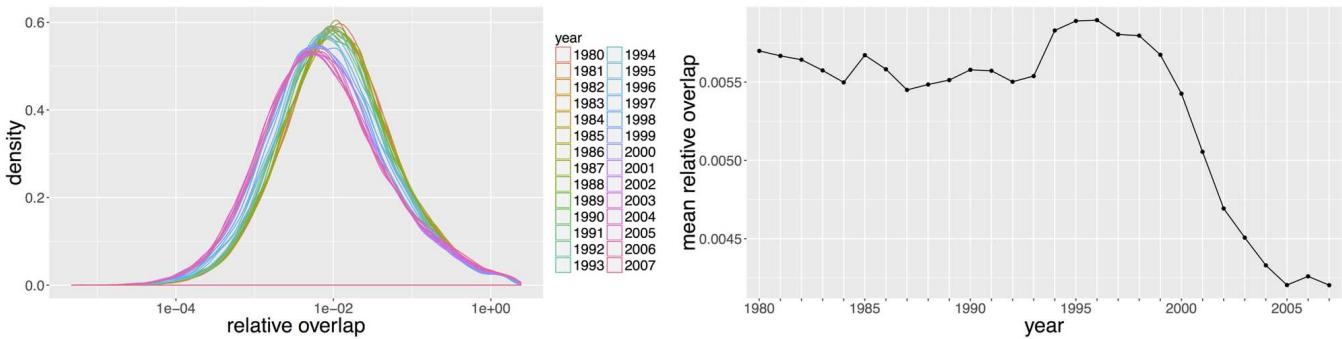


FIGURE 148: **Distribution of relative overlaps between classifications.** (Left) Distribution of overlaps at all time steps; (Right) Corresponding mean time-series. The decreasing trend starting around 1996 confirms a qualitative regime shift in that period.

Citation Modularity

An exogenous source of information on relevance of classifications is the citation network described in Section ???. The correspondence between citation links and classes should provide a measure of accuracy of classifications, in the sense of an external validation since it is well-known that citation homophily is expected to be quite high (see, e.g, [Acemoglu, Akcigit, and Kerr, 2016]). This section studies empirically modularities of the citation network regarding the different classifications. Modularity is a simple measure of how communities in a network are well clustered (see [Clauset, Newman, and Moore, 2004] for the accurate definition). Although initially designed for single-class classifications, this measure can be extended to the case where nodes can belong to several classes at the same time, in our case with different probabilities as introduced in [Nicosia et al., 2009]. The simple directed modularity is given in our case by

$$Q_d^{(z)} = \frac{1}{N_p} \sum_{1 \leq i, j \leq N_p} \left[A_{ij} - \frac{k_i^{in} k_j^{out}}{N_p} \right] \delta(c_i, c_j),$$

with A_{ij} the citation adjacency matrix (i.e. $A_{ij} = 1$ if there is a citation from the i th patent to the j th patent, and $A_{ij} = 0$ if not), $k_i^{in} = |I_i|$ (resp. $k_i^{out} = |\tilde{I}_i|$) in-degree (resp. out-degree) of patents (i.e. the number of citations made by the i th patent to others and the number of citations received by the i th patent). Q_d can be defined for each of the two classification systems: $z \in \{\text{tec}, \text{sem}\}$. If $z = \text{tec}$, c_i is defined as the main patent class, which is taken as the first class whereas if $z = \text{sem}$, c_i is the class with the largest probability.

Multi-class modularity in turns is given by

$$Q_{ov}^{(z)} = \frac{1}{N_P} \sum_{c=1}^{N^{(z)}} \sum_{1 \leq i, j \leq N_P} \left[F(p_{ic}, p_{jc}) A_{ij} - \frac{\beta_{i,c}^{out} k_i^{out} \beta_{j,c}^{in} k_j^{in}}{N_P} \right],$$

where

$$\beta_{i,c}^{out} = \frac{1}{N_P} \sum_j F(p_{ic}, p_{jc}) \text{ and } \beta_{j,c}^{in} = \frac{1}{N_P} \sum_i F(p_{ic}, p_{jc}).$$

We take $F(p_{ic}, p_{jc}) = p_{ic} \cdot p_{jc}$ as suggested in [Nicosia et al., 2009]. Modularity is an aggregated measure of how the network deviates from a null model where links would be randomly made according to node degree. In other words it captures the propensity for links to be inside the classes. Overlapping modularity naturally extends simple modularity by taking into account the fact that nodes can belong simultaneously to many classes.

We document in Fig. ?? both simple and multi-class modularities over time. For simple modularity, $Q_d^{(\text{tec})}$ is low and stable across the years whereas $Q_d^{(\text{sem})}$ is slightly greater and increasing. These values are however low and suggest that single classes are not sufficient to capture citation homophily. Multi-class modularities tell a different story. First of all, both classification modularities have a clear increasing trend, meaning that they become more and more adequate with citation network. The specializations revealed by both patent level diversities and classes overlap is a candidate explanation for this growing modularities. Secondly, semantic modularity dominates technological modularity by an order of magnitude (e.g. 0.0094 for technological against 0.0853 for semantic in 2007) at each time. This discrepancy has a strong qualitative significance. Our semantic classification fits better the citation network when using multiple classes. As technologies can be seen as a combination of different components as shown by [Youn et al., 2015], this heterogeneous nature is most likely better taken into account by our multi-class semantic classification.

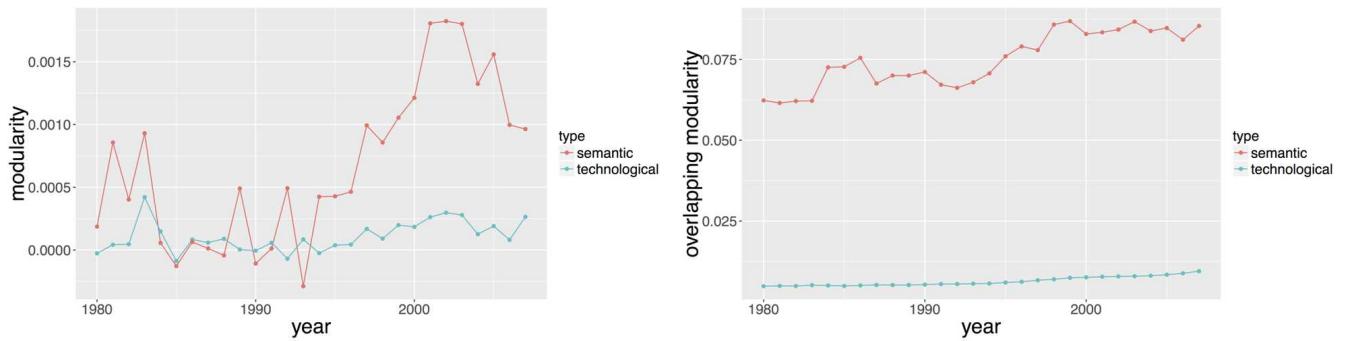


FIGURE 149: Temporal evolution of semantic and technological modularities of the citation network. (Left) Simple directed modularity, computed with patent main classes (main technological class and semantic class with larger probability). (Right) Multi-class modularity, computed following [Nicosia et al., 2009].

Perspectives

The main contribution of this study was twofold. First we have defined how we built a network of patents based on a classification that uses semantic information from abstracts. We have shown that this classification share some similarities with the traditional technological classification, but also have distinct features. Second, we provide researchers with materials resulting from our analysis, which includes: (i) a database linking each patent with its set of semantic classes and the associated probabilities; (ii) a list of these semantic classes with a description based on the most relevant keywords; (iii) a list of patent with their topological properties in the semantic network (centrality, frequency, degree, etc.). The availability of this data suggests new avenues for further research. Linking our dataset with existing open ones can lead to various powerful developments. For example, using it together with the disambiguated inventor database provided by [Li et al., 2014] could be a way to study semantic profiles of inventors, or of cities as inventor addresses are provided. The investigation of spatial diffusion of innovation between cities, which is a key component of Pumain's Evolutive Urban Theory [Pumain, 2010], would be made possible.

A first potential application is to use the patents' topological measures inherited from their relevant keywords. The fact that these measures are backward-looking and immediately available after the publication of the patent information is an important asset. It would for example be very interesting to test their predicting power to assess the quality of an innovation, using the number of forward citations received by a patent, and subsequently the future effect on the firm's market value.

Regarding firm innovative strategy, a second extension could be to study trajectories of firms in the two networks: technological and semantic. Merging these information with data on the market value of firms can give a lot of insight about the more efficient innovative strategies, about the importance of technology convergence or about acquisition of small innovative firms. It will also allow to observe innovation pattern over a firm life cycle and how this differ across technology field.

A third extension would be to use dig further into the history of innovation. USPTO patent data have been digitized from the first patent in July 1790. However, not all of them contain a text that is directly exploitable. We consider that the quality of patent's images is good enough to rely on Optical Character Recognition techniques to retrieve plain text from at least 1920. With such data, we would be able to extend our analysis further back in time and to study how technological progress occurs and combines in time. [Akcigit, Kerr, and Nicholas, 2013] conduct a similar work by looking at recombination and apparition of technological subclasses. Using the fact that communities are constructed yearly, one can construct a measure of proximity between two successive classes. This could give clear view on how technologies converged over the year and when others became obsolete and replaced by new methods.

★ ★

★

C.6 BRIDGES BETWEEN ECONOMICS AND GEOGRAPHY

Cette section rend compte d'une première expérience en "perspectivisme appliqué", c'est-à-dire la tentative de couplage de perspectives sur des objets communs pour créer des ponts entre disciplines. Dans cet esprit, une session spéciale a été organisée, conjointement avec B. CARANTINO (Paris School of Economics) à l'*European Colloquium in Theoretical and Quantitative Geography* (York, septembre 2017) pour questionner les liens entre Géographie et Economie. La question de ponts au sein des modèles, c'est-à-dire de la façon dont les modèles permettent d'utiliser des concepts économiques en géographie ou réciproquement, a été particulièrement étudiée. L'encadré 15 ci-dessous présente l'appel à communication. La session a rassemblé 11 contributions¹⁷, dont une à l'initiative d'économistes et deux autres en collaboration avec des économistes : l'effort pour intéresser des économistes à un congrès de géographie a difficilement porté ses fruits.

As Krugman points out, space is for Economic Geography the final frontier, whereas Geographical analyses are somehow far from an advanced integration of economical concepts. What are the existing and potential links? Is there unsurmountable epistemic divergences making bridging approaches irrelevant? For example, the assumptions regarding equilibrium, but also the concepts of equilibrium itself in each discipline may be irreconcilable. This session aims at giving element of answers from a modeling perspective. It is open to case studies of models at the interface and from both disciplines, integrating both elements of spatial analysis and geosimulation together with concepts and methods from economics. It is also open to theoretical or conceptual contributions, in order to bring a broader point of view. An alternative way to study the question is through quantitative epistemology studies, in order to extract empirical endogenous information on the modeling practices themselves. The diversity of views will shed light on potential enrichments on both sides, but also on recurrent difficulties and epistemological divergences, as should illustrate the study of the same objects from totally different perspectives.

FRAME 15: ECTQG 2017 Special Session : bridges between economics and geography.

Synthesis of contributions

Les contributions à la session ont permis d'apporter des éclairages sur la question à différents niveaux et selon différents domaines de connaissance. Des études de modélisation ont permis de montrer le compromis qu'il faut toujours faire entre spatialisation du modèle et pertinence des mécanismes économiques, que ce soit dans le cas d'un modèle stylisé (contribution de M. BIDA et al.) ou dans le cas de mod-

¹⁷ Le programme est disponible à <http://www.geog.leeds.ac.uk/ectqg17/programme.html>.

èles opérationnels d'évolution de l'usage du sol (contribution de E. KOOMEN et D. VASCO). Ce compromis se retrouve au niveau théorique, mais est compliqué également par des divergences épistémologiques, par exemple sur le rôle à donner aux dynamiques évolutionnaires (contribution de D. PUMAIN) ou au déséquilibre (contribution de R. WHITE et al.), qui se retrouvent dans les relations effectives entre disciplines, comme observé par une analyse bibliométrique (contribution de J. RAIMBAULT).

Un exemple concret d'objet étudié selon divers point de vue illustre ces considérations : les trajectoires de firmes. Du point de vue purement économique, des facteurs internes et les caractéristiques des locaux induisent les déménagements des entreprises (contribution de A. BERGEAUD et S. RAY), tandis que les dynamiques spatiales de celles-ci peuvent être appréhendées par leur relations spatiales et des effets d'agrégation (contribution de C. COTTINEAU et al.). A une plus petite échelle, la spatialisation de l'activité économique des firmes transnationales permet de tirer des conclusions sur la structure du système géographique (contribution d'O. FINANCE).

Enfin, les études empiriques présentée montrent comment croiser données économiques, comme usage du sol (contribution de J. DELLOYE et al.), transactions en ligne (contribution de J. BECKERS et al.) ou locations de logements (contribution de Z. SHABRINA et al.), et modèles spatialisés comme modèle d'accessibilité ou de distribution de densité.

Les discussions finales ont fait ressortir les points suivants : (i) les divergences épistémologiques ne sont pas nécessairement fondamentales si elles sont contextualisées ; (ii) les différences de comportement face au modèles des différentes disciplines sont aussi liées à la demande qui est faite à ces disciplines, comme des recommandations d'action publique pour l'économie, et relaxer les standards disciplinaires pourrait aider à la communication ; (iii) le cloisonnement bibliographique, combiné à des difficultés d'intelligibilité, est un point crucial sur lequel des progrès considérables sont possibles, notamment par l'utilisation des nouvelles données et méthodes en analyse textuelle et datamining.

Ainsi, les ponts potentiels sont bien présents, et les outils et méthodes permettant de faciliter leur concrétisation ne demandent qu'à être développés. Un exemple d'application favorisant la réflexivité et donc le dialogue interdisciplinaire est donné en C.4.

* * *

*

C.7 GAMED-BASED TOOLS AS MEDIA TO TRANSMIT FRESHWATER ECOLOGY CONCEPTS

The issue of scientific communication, in particular between agents producing knowledge, has been a recurrent theme in our work. It also plays a role in the interface with the public for scientific mediation, and the development of a mediation can in return inform interdisciplinary enterprises. We develop here two models as games, with a similar objective to transmit freshwater ecology concepts. This reinforces the idea of the model as a crucial instrument of scientific mediation.

* * *

*

This section is the output of an interdisciplinary collaboration with the ecotoxicologist DR. HÉLÈNE SERRA (Université de Bordeaux and Ineris) and was presented at the SETAC 2016 conference as [Serra and Raimbault, 2016].

* * *

*

C.7.1 *Introduction*

There is an increasing expectation on people to be aware and to get involved in the environmental issues that our world is facing. However, expert knowledge is often required to understand most of these issues. One of the challenges in science today lies in explaining complex issues in a simple and understandable way to an unspecialized audience. Games can turn out to be a good medium for scientific vulgarization. Indeed, the first form of learning we all experienced was by playing. Games are very popular, and from an educational point of view, they present many advantages. They are dynamic and interactive. Therefore, the player engagement increases, as well as its knowledge retention. In addition, the player is immersed into a new world and discovers a virtual environment where he needs to develop strategies and to identify crucial processes. Those characteristics can be wisely used to spread scientific topics, and gamification has al-

ready been proposed as a tool for an easier propagation of scientific thinking [Morris et al., 2013] such as in pharmacology [Cain and Pi-ascik, 2015] or geosciences [Reynard et al., 2015]. In this context, our project aims at developing game-based tools to transmit the basic concepts of freshwater ecology. We choose to focus on a classical board game and on a computer based game because they are complementary in the targeted audience (groups versus online gamers) and the possibilities offered, in particular regarding the interactions between players and the system dynamics.

c.7.2 *Material and Methods*

The general methodology is divided in five steps: (1) selection of species; (2) definition of the instructions (object, game board, rules); (3) incorporation of environmental stressors (biotic and abiotic), (4) design and construction of interfaces (board and computer model); (5) test with players. All steps are necessarily interdependent and are tackled in parallel during the development of the games.

While the board game is inspired by past experiences of player, the computer game is based on a model of simulation of the ecosystem. In order to introduce notions of equilibrium and its perturbations that occur at a larger time scale than on the board game, we propose to implement an agent-based model (ABM) and to couple its dynamics with gaming actions. ABM have already been widely used in ecology [Grimm et al., 2005]. Therefore, we selected a trophic chain dynamic model (extended prey-predator model) that can capture fish behavioral rules and spatially heterogeneous environment. It is particularly suitable for the game implementation: fish behaviors are influenced by players whereas the ecosystem is disturbed by external events.

c.7.3 *Results*

Both games are based on the same general rules, even if slight modifications have to be expected according to the type of game. The virtual ecosystem is presented from a fish perspective. The object of the game is to reach a given number of adults and juveniles that will guarantee the stability of the population in the lake. The external perturbations are illustrated by “events” that are supposed to reflect abiotic (e.g. water temperature, light, water scarcity) and biotic (e.g. chemicals, parasites, fisherman) stressors. The rationale behind lies in maximizing interactions between players (predation and competition, see fig. 1) and to illustrate feeding and reproduction strategies from different perspectives (from a big solitary fish to a shoal fish, including a invasive fish species).

The board game

To maintain the populations in the board game, each player has to find resources accordingly to his fish species. The resources are converted into "units" that can be used thereafter by the player for different purposes, such as reproduction, juvenile growth, to escape a predator or to attack a pray.

The current version of the game includes four players, each of them being a different species, namely the roach (*Rutilus rutilus*), the pumpkinseed (*Lepomis gibbosus*), the zander (*Sander lucioperca*), and the bleak (*Alburnus alburnus*).

The board is basically composed of boxes. Each of them represents a type of resource (e.g. crustacean, plants, insects), and some boxes are combined with an "event" to include the external perturbations in the game. The player has 2 token on the board (one male and one female) and is moving them by throwing dice. The ecological characteristics of each species are kept on a record paper by each player. It describes the species-specific rules (feeding preferences, time and resources needed to reproduce, how to escape/attack etc). A first prototype is currently being tested to determine and adjust the board game design, the ecological characteristics of each species and the characterization of events, in particular their impacts on players. The design of the board is under progress and will figure the edge of a lake.

Computer-based game

The player controls an ecosystem with preys (the roach) and predators (the pumpkinseed). The objective of the game is to maintain the stability of the ecosystem. The concepts illustrated are population dynamic and ecosystem resilience.

An agent-based model2 (ABM) for a simple prey-predator system is proposed as a basis of the computer game . ABM simulate the behavior and interactions between agents (fish) to reconstruct the population dynamic (bottom-up approach). Stochasticity is included with spatialized interactions (smoothed brownian motions), illustrating the randomisation of prey-predator interactions .Discrete dynamics consist in the following steps : (a) wandering of species in their preferred zone of the lake; (b) trophic interactions (fish-fish and fish-ressources) ; (c) renewing of fish (reproduction) and of ressources. The model parameters include reproduction rates, movement parameters, etc.

(Netlogo software). Large-scale model exploration and calibration are currently running (Using OpenMole model exploration platform [Reuillon, Leclaire, and Rey-Coyrehourcq, 2013] to explore the NetLogo implementation which source code is openly available on the repository of the project at <https://github.com/JusteRaimbault/MediationEcotox>) in order to find parameter ranges at which ecosystem is in equilib-

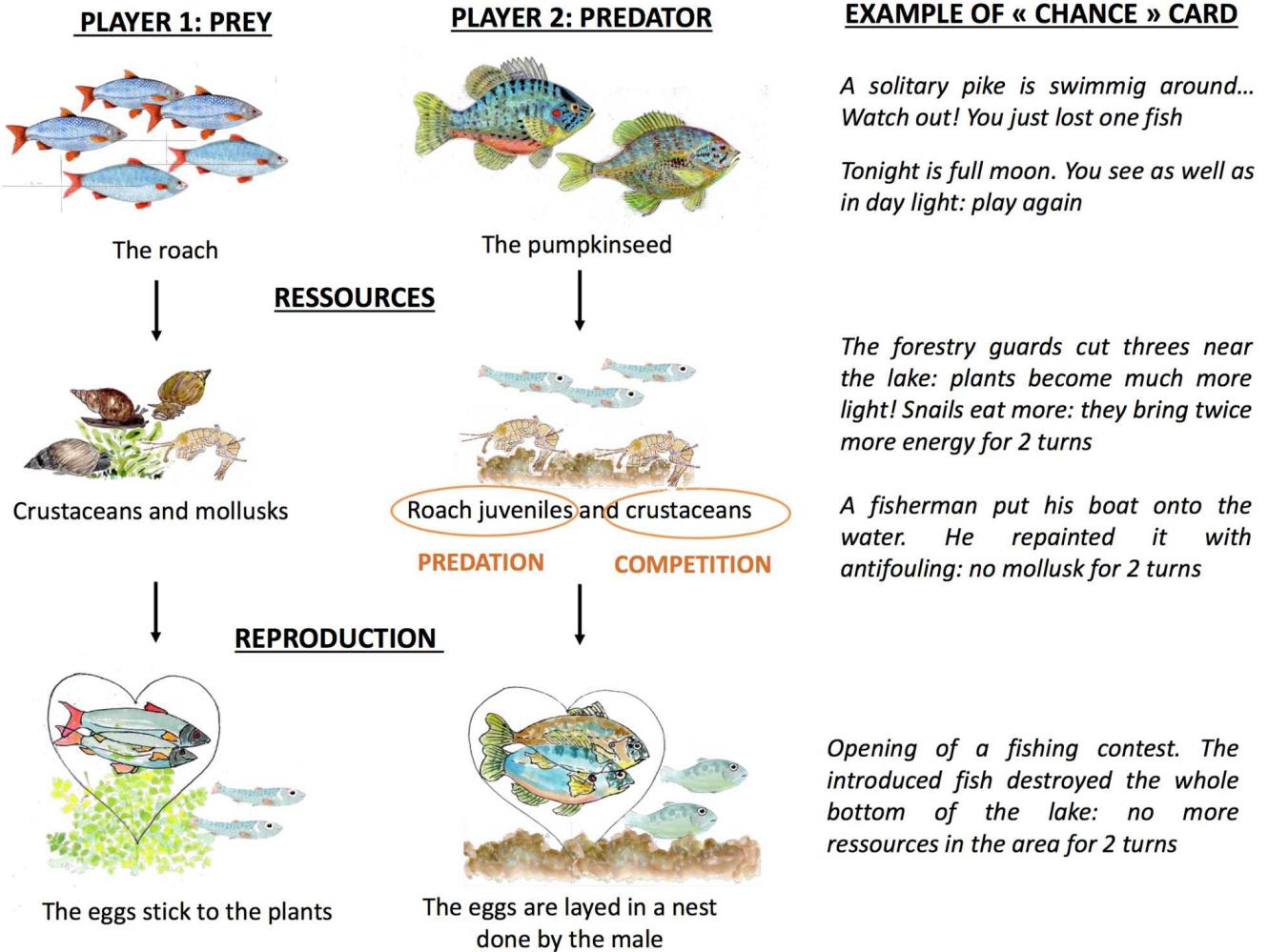


FIGURE 150: Principles of the board game.

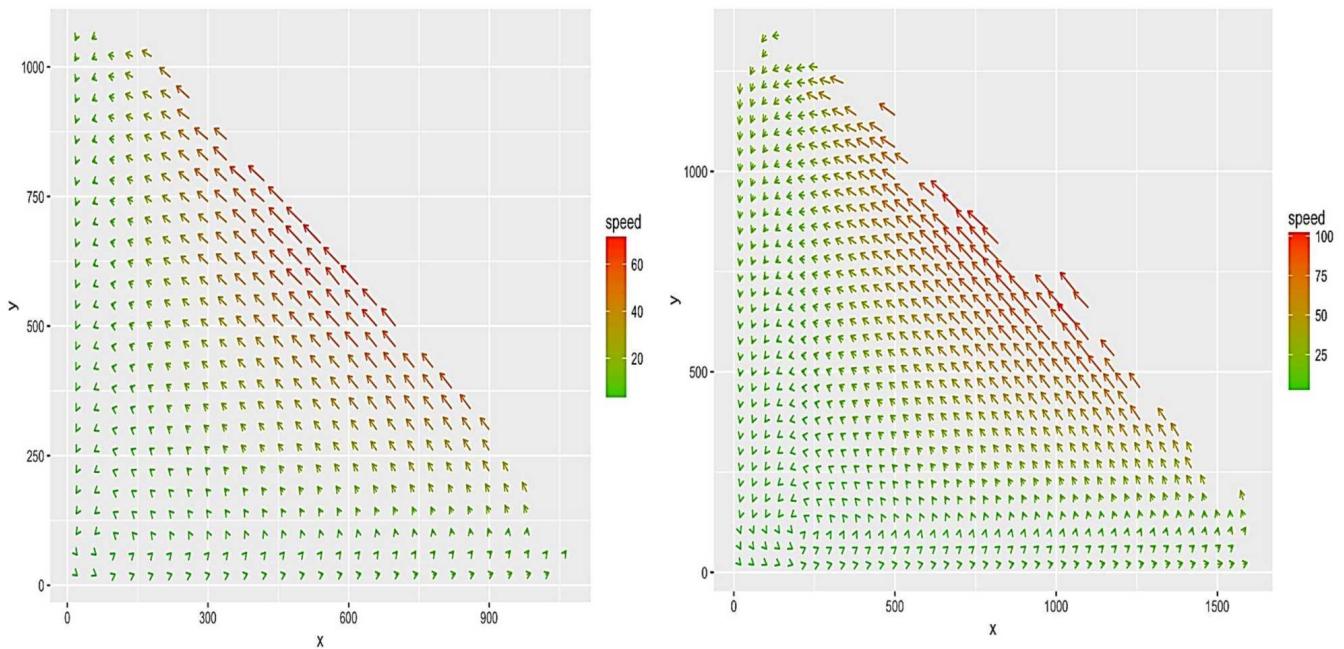


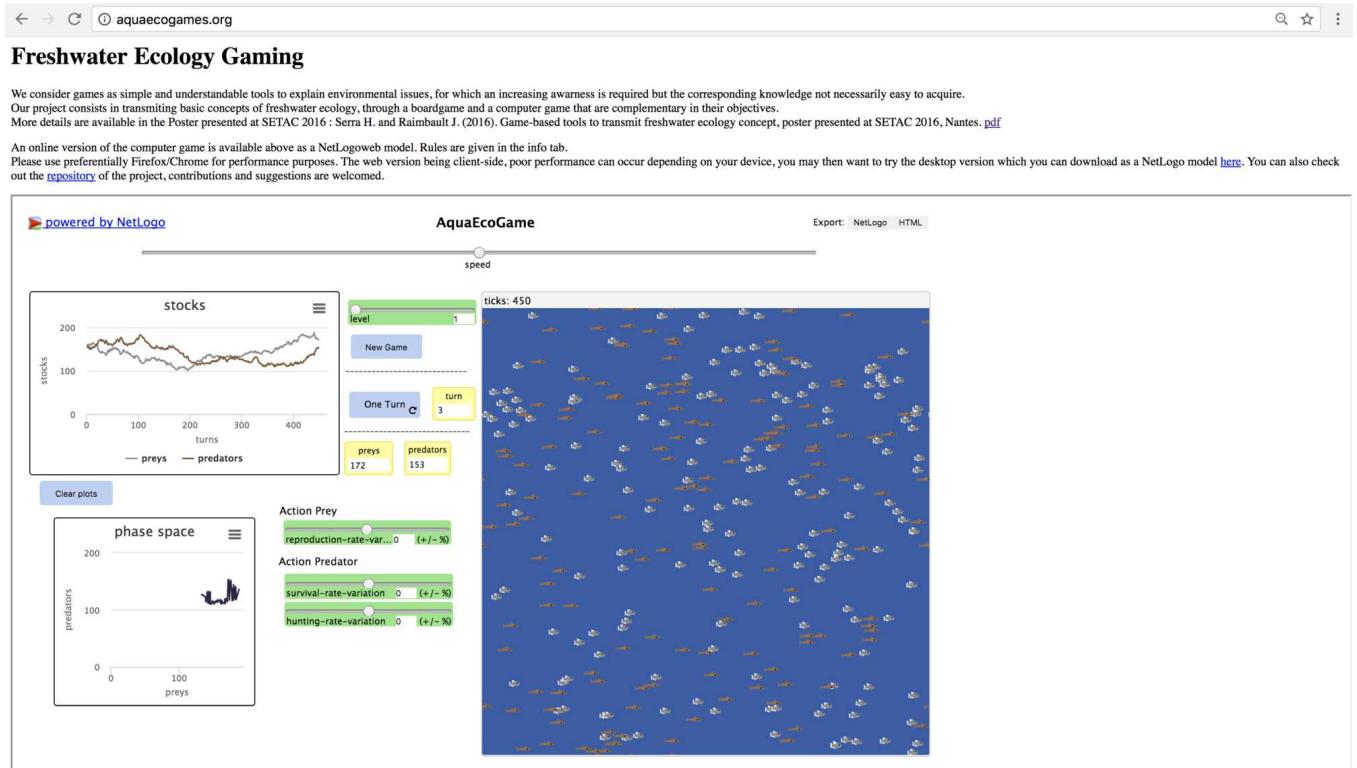
FIGURE 151: Examples of phase diagrams of the prey-predator model.

rium1. Systematic exploration of parameter space using OpenMole software³, to verify theoretical average trajectories in phase space, which allows analytical and numerical determination of initial position (attractor) and justify the use of this system for the game.

The player starts the game with a stable ecosystem: the initial position is the attractor. The button « one turn » makes the ecosystem evolve during 50 time steps. The player sees the changes in the fish populations simultaneously on the screen. The trajectory can be corrected towards the attractor in the phase space by changing the parameters of the model (predator survival, prey reproduction and hunting behavior). External events randomly perturbate the ecosystem. The game includes 5 levels of difficulty based on event strength available at <http://aquaecogames.org/>

C.7.4 Discussion

FIGURE 152: A prototype of each game is currently available for testing and refinements are expected while experiencing the games. In a short term, next versions of the games will be developed after player feedback and will include the aesthetic design of the games and refined processes parameters. Mid-term and long-term objectives are oriented towards an online version of the computer game as described before, and the use of crowdfunding platforms to offer and diffuse the board game.



The very first objective of our games remains to be entertaining, keeping in mind that the ludic rather than pedagogical aspects are central in the success of such game-based media. If players forget that the game is about ecology, our precise objective is reached. It would mean that the underlying scientific concepts are clearly understood.

★ ★

★

D

DONNÉES

This appendix lists and describes the different open datasets created and used in the thesis.

D.1 GRAND PARIS TRAFFIC DATA

D.1.1 *Description*

Ce jeu de données, utilisé sur deux mois pour l'analyse de 3.2, s'étend finalement sur deux ans de février 2016 à février 2018. Il est constitué des temps de parcours sur les axes autoroutiers principaux de la métropole parisienne, à une granularité temporelle de 2 minutes.

D.1.2 *Specification*

CITATION Rimbault J., 2018, Replication Data for: Investigating the empirical existence of static user equilibrium, doi:10.7910/DVN/X22ODA, Harvard Dataverse, V1

TYPE AND FORMAT Liste des liens routiers, avec temps effectif et temps théorique de parcours, et le moment d'observation (timestamps) ; au format sqlite3.

LICENSE Domaine public CCo.

AVAILABILITY La base est disponible sur le Harvard Dataverse à <http://dx.doi.org/10.7910/DVN/X22ODA>.

D.2 TOPOLOGICAL ROAD NETWORK

D.2.1 *Description*

La simplification des réseaux routiers, opérée à grande échelle pour l'Europe et la Chine sur les données d'OpenStreetMap, produit les graphes topologiques correspondants comme décrit en 4.1 et en A.4.

L'intérêt de ce jeu de données est la possibilité d'utilisation directe pour l'étude de mesures de graphes des réseaux routiers, sur une étendue spatiale quelconque. En effet, la création du réseau topologique à l'échelle considérée a requis un effort computationnel considérable, pas forcément accessible au plus grand nombre.

D.2.2 *Specification*

CITATION Raimbault, Juste, 2018, "Simplified road networks, Europe and China", doi:10.7910/DVN/RKDZMV, Harvard Dataverse, V1

TYPE AND FORMAT Les données sont sous forme de liste des liens, au format extraction compressée de postgresql (dump).

LICENSE Domaine public CCo.

AVAILABILITY La base est disponible sur le Harvard Dataverse à <http://dx.doi.org/10.7910/DVN/RKDZMV>.

D.3 INTERVIEWS

Un matériau de recherche qui serait plus “qualitatif” au sens classique, n'a pas de raison d'être moins ouvert que des bases de données “quantitatives”. Dans le cas d'entretiens, l'ouverture des transcriptions est essentielle pour la reproductibilité puisqu'il s'agit du dernier (et du premier) stade avant la traduction non reproductible en interprétations. Nous pensons également qu'elle est cruciale pour exploiter l'ensemble de leur potentiel, l'ouverture permettant leur réutilisation et donc possiblement réactions ou débats. Des initiatives dans cette direction commencent à émerger, comme le *Qualitative Data Repository*¹ qui permet d'archiver et de présenter de manière cohérente un corpus qualitatif, souvent décrit de façon parcellaire et conjointement aux analyses dans les articles [Elman and Kapiszewski, 2018].

D.3.1 *Description*

Interview with Denise Pumain, 2017/03/31

Cet entretien est intervenu dans le contexte d'une collecte de matériau empirique pour la rédaction de [Raimbault, 2017c], qui a permis entre autre la construction du cadre de connaissances développé en 8.3. L'entretien est principalement centré sur la genèse de la Théorie Evolutionniste des Villes.

Interview with Romain Reuillon, 2017/04/11

Cet entretien intervient dans le même contexte que le précédent, en cherchant à apporter un éclairage du point de vue des méthodes et outils. Il retrace en particulier la genèse d'OpenMole.

¹ <https://qdr.syr.edu/>

Interview with Clémentine Cottineau, 2017/05/05

Cet entretien vise à comprendre le point de vue d'une géographe à l'interface interdisciplinaire (participation à l'ERC Geodiversity) sur la théorie évolutive des villes et son élaboration en termes de domaines de connaissance.

Interview with Denise Pumain, 2017/12/15

Ce deuxième entretien avec D. PUMAIN se concentre plus particulièrement sur les effets structurants des infrastructures de transport et coévolution, du point de vue de la géographie.

Interview with Alain Bonnafous, 2018/01/09

Cet entretien s'intéresse aux effets structurants des infrastructures de transport, du point de vue de l'économie des transports, ainsi qu'au positionnement interdisciplinaire de l'économie des transports.

D.3.2 Specification

CITATION Raimbault J., 2017. JusteRaimbault/Entretiens vo.2 (Version vo.2). Zenodo. <http://doi.org/10.5281/zenodo.556331>

TYPE AND FORMAT Transcription des entretiens au format texte.

LICENSE Creative commons CC-BY-NC.

AVAILABILITY Les entretiens sont disponibles sur le dépôt git dédié à <https://github.com/JusteRaimbault/Entretiens>, et les versions successives sont accessibles à <https://doi.org/10.5281/zenodo.596954>.

D.4 SYNTHETIC DATA AND SIMULATION RESULTS

Les résultats de calculs ou de simulations utilisés pour l'ensemble des résultats présentés sont disponibles de manière ouverte, soit sur le dépôt git soit sur un dépôt dataverse dédié dans le cas d'articles autonomes ou de fichiers massifs. Les liens sont les suivants pour les dépôts particuliers :

- Résultats de l'exploration du corpus Cybergeo <http://dx.doi.org/10.7910/DVN/VU2XKT> ; Epistémologie quantitative et modélographie <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/QuantEpistemo/HyperNetwork/data>
- Indicateurs morphologiques et topologiques pour l'Europe et la Chine <http://dx.doi.org/10.7910/DVN/RHLM5Q>

- Simulation de données synthétiques par le modèle RBD pour l'identification de régimes de causalité spatio-temporelle <http://dx.doi.org/10.7910/DVN/KGHZZB>
- Calibration du modèle macroscopique d'interactions <https://github.com/JusteRaimbault/CityNetwork/tree/master/Results/NetworkNecessity/InteractionGibrat/calibration>
- Simulation et calibration du modèle de morphogenèse pour la densité <http://dx.doi.org/10.7910/DVN/WSUSBA>
- Simulation du couplage faible des modèles de densité et de croissance de réseau <http://dx.doi.org/10.7910/DVN/UIHBC7>
- Simulation du modèle SimpopNet <http://dx.doi.org/10.7910/DVN/RW8S36>
- Simulations du modèle de co-évolution macroscopique <http://dx.doi.org/10.7910/DVN/TYBNFQ> et <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/MacroCoevol/MacroCoevol/calibres> pour la calibration
- Simulations du modèle de co-évolution mesoscopique <http://dx.doi.org/10.7910/DVN/0BQ4CS>
- Simulations du modèle Lutecia <http://dx.doi.org/10.7910/DVN/V3KI2N>

★ ★

★

E

OUTILS

Cet annexe rend compte des outils développés et utilisés pour l'ensemble des analyses. Comme nous l'avons précisé en section 8.3, les outils correspondent généralement à l'implémentation de méthodes (étant alors des *proto-méthodes*, du moins dans notre cas où nous n'utilisons pas d'appareillage physique de mesure), mais correspondent bien à un domaine de connaissance à part entière et avec une certaine indépendance.

Nous distinguons et décrivons ici :

- les *packages* ou *logiciel* développés dans le cadre de ce travail, mais qui peuvent remplir des fonctions bien plus larges et peuvent être distribués de manière autonome ;
- l'implémentation des modèles de simulation et des algorithmes de traitement des données ;
- des outils ou pratiques facilitant particulièrement une science ouverte et fluide.

★ ★

★

E.1 SOFTWARES AND PACKAGES

This appendix lists and describes the different open datasets created and used in the thesis.

E.1.1 *largeNetwoRk: Network Import and simplification for R*

DESCRIPTION Simplification of european road network, Package LargeNetwoRk

CHARACTERISTICS Le package est intégralement écrit en R, et requiert une connexion avec une base PostgreSQL (avec extension Post-Gis installée). Le code source est disponible à <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/TransportationNetwork/NetworkSimplification> avec la documentation.

FUNCTIONS Les principales fonctions suivantes sont implémentées :

- `constructLocalGraph` : construit un graphe topologique à partir des lignes spatiales issues de la base postgis (dans une étendue spatiale précisée)
- `graphFromSpdf` : construit un graphe topologique à partir d'une structure de données spatiale (permet d'importer depuis un fichier shp par exemple)
- `mergeGraphs` : fusion de deux graphes voisins dans l'espace
- `simplifyGraph` : simplification d'un graphe (voir algorithme en A.4)
- `connexify` : donne un graphe connecté à partir d'un graphe quelconque, par l'ajout de connecteurs
- `exportGraph` : exporte un graphe topologique dans la base de données

Un script complet permet par ailleurs l'exécution de l'algorithme *split and merge* décrit en A.4 pour la simplification de grandes étendues spatiales.

PARTICULARITIES Handling of large size databases imposes sequential processing ; use of external program osmosis for conversion from osm data to pgsql.

E.1.2 *Transportation networks and accessibility in R*

DESCRIPTION Le package tRansport pour le langage R rend transparent les calculs d'indicateurs pour les réseaux de transport en commun et les calculs d'accessibilité associés. A partir de jeux de données

comprenant lignes et stations pour différents modes de transports en commun, il permet de construire un réseau topologique multimodal et de calculer différentes mesures étant donné des variables géographiques.

CHARACTERISTICS Le package est écrit en langage R et produit des graphes selon la structure du package `igraph`. Le code source et la documentation sont disponibles à <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/TransportationNetwork/NetworkAnalysis>.

FUNCTIONS Les principales fonctions suivantes sont disponibles :

- `addTransportationLayer` : construit un graphe à partir d'une couche du réseau, ou ajoute une couche à un réseau existant, à partir d'une description shapefile des liens et des noeuds (gares) du réseau.
- `addPointsLayer` : ajoute une couche de points, qui peuvent être alors origine ou destination d'itinéraires. Ils sont reliés à la gare la plus proche par des connecteurs dont la vitesse est spécifiée.
- `addAdministrativeLayer` : fonction similaire, qui connecte les centroïdes d'une couche de polygones représentant typiquement des zones administratives, conservant leur attributs comme attributs de noeuds.
- `computeAccess` : calcule l'accessibilité entre des points du réseau de transport, selon plusieurs spécifications : temps de trajet, pondération à l'origine et/ou à la destination par des données spécifiées.

E.1.3 *The morphology NetLogo extension to measure Urban Form*

DESCRIPTION L'extension `morphology` pour NetLogo5 permet de calculer de manière efficiente et transparente les indicateurs morphologiques introduits en 4.1 (indice de Moran, entropie, distance moyenne, hiérarchie), pour la distribution spatiale d'une variable de patch quelconque.

CHARACTERISTICS L'extension est écrite en `scala` et est compatible avec les versions 5 de NetLogo. Elle est disponible à <https://github.com/JusteRaimbault/nl-spatialmorphology>.

PARTICULARITIES Les indicateurs impliquant une convolution (indice de Moran, distance moyenne) sont implementés par transformée de Fourier rapide, permettant de faire passer la complexité d'un $O(N^4)$ à un $O(N^2 \cdot \log^2 N)$ si N est la taille d'un côté de la grille.

E.1.4 *TorPool*

DESCRIPTION TorPool is a java based Tor wrapper available with an api (currently only java, R version projected) at <https://github.com/JusteRaimbault/TorPool>. It allows among other purposes tricky data retrieval.

FUNCTIONS Le logiciel se lance sous forme d'executable jar, et ouvre une plage spécifiée de ports en proxy socks5 local vers le réseau Tor.

La bibliothèque java associée permet (i) d'établir une connexion avec les proxys (ii) de demander un renouvellement des instances, permettant un changement de circuit dans le réseau.

E.1.5 *Scientific Corpus Mining*

DESCRIPTION Les outils développés dans le cadre du Chapitre 2, et des Annexes B.6 et C.5 permettent de manière générale la fouille de corpus scientifiques, du point de vue du réseau de citation et du réseau sémantique.

CHARACTERISTICS Comme rappelé en B.6, les tâches requises sont assez hétérogènes, et différents langages sont alors mobilisés : Java pour la collection des données, python pour l'analyse textuelle, R pour les analyses de réseau. La version des différents scripts utilisés pour le Chapitre 2 est disponible à <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/QuantEpistemo/HyperNetwork>.

FUNCTIONS Les fonctions suivantes sont assurées : (i) collecte du réseau de citation à partir d'un corpus initial, collecte des résumés d'un corpus ; (ii) extraction des mots-clés sous forme de n-grams, estimation de la pertinence des mots clés ; (iii) construction des réseaux sémantiques et de citation.

E.2 ARCHITECTURE AND SOURCES FOR ALGORITHMS AND MODELS OF SIMULATION

And yet it is. It makes no sense to put code listings in the core of the text if there is no particular algorithmic detail that requires attention. As soon as implementation biases are avoided, architecture and source for a computational model should be independent from its formal description (but provided along model description with source code as already mentioned before).

We give in this appendix architectural details on main models of simulation or algorithms we used. Langage and size (in code lines) are provided, along with architectural remarkable features. See <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models> for all models, empirical analysis and small experiments.

E.2.1 Algorithmic Systematic Review

OBJECTIVE Implement systematic literature review algorithm.

LOCATION <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/QuantEpistemo/AlgoSR/AlgoSRJavaApp>

CHARACTERISTICS

- Langage : Java
- Taille : 7116

PARTICULARITIES

- HashConsing used for unique bibliography object, specific hash-code switching if id available or only titles (proceed to lexical distance comparison in that latest case).
- API to context currently being replaced by Python scripts.

ARCHITECTURE Classical object oriented, see code.

ADDITIONAL SCRIPTS R for result exploration and visualization.

E.2.2 Indirect Bibliometrics

OBJECTIVE Multi-layer network analysis of scientific corpuses : cybergeo journal, corpus in 2.2

LOCATION <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/QuantEpistemo/HyperNetwork>

CHARACTERISTICS

- Langage : Python, R and Java.
- Taille : 2210

PARTICULARITIES Utilise des bases de données sqlite, sql ou MongoDB selon les opérations.

ARCHITECTURE See schema chapter 3.

E.2.3 Static correlations**OBJECTIVE**

LOCATION <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/StaticCorrelations>

CHARACTERISTICS

- Langage : R
- Taille : 1862

E.2.4 Spatio-temporal causalities**OBJECTIVE**

LOCATION <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/SpatioTempCausalities>

CHARACTERISTICS

- Langage : R
- Taille : 8627

E.2.5 Macroscopic model of interactions

OBJECTIVE Interaction macroscopic model

LOCATION <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/InteractionGibrat>

CHARACTERISTICS

- Langage : NetLogo, scala, R
- Taille : 5918

PARTICULARITIES Morphological indicators in scala implemented with Fast Fourier transform ; with R communication in NetLogo.

E.2.6 *Density Urban Growth*

OBJECTIVE Density-based urban morphogenesis model

LOCATION <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Synthetic/Density>

CHARACTERISTICS

- Langage : NetLogo, scala, R
- Taille : 5065

E.2.7 *Correlated data generation*

OBJECTIVE Weak coupling of density generation and network generation.

LOCATION <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Synthetic/Network>

CHARACTERISTICS

- Langage : NetLogo (réseau) and scala (densité)
- Taille : 3188

PARTICULARITIES Network heuristic easier to implement and explore in netlogo

ARCHITECTURE OpenMole allows coupling between modules through exploration script.

E.2.8 *Macroscopic co-evolution*

OBJECTIVE Implementation of macro-coevolution model

LOCATION <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/MacroCoevol>

CHARACTERISTICS

- Langage : NetLogo
- Taille : 4950

PARTICULARITIES

DATA USED Population des aires urbaines Françaises 1830-1999

ADDITIONAL SCRIPTS Exploration et calibration (oms), exploration des résultats (R)

E.2.9 Morphogenesis co-evolution

OBJECTIVE Implementation of meso-coevolution model

LOCATION <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/MesoCoevol>

CHARACTERISTICS

- Langage : NetLogo
- Taille : 5386

ADDITIONAL SCRIPTS Exploration et calibration (oms), exploration des résultats (R)

E.2.10 Lutecia Model

OBJECTIVE Implementation of Lutecia model, chapter 7.3.

LOCATION <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Governance/Lutecia/Lutecia>

CHARACTERISTICS

- Langage : NetLogo
- Taille : 8866

PARTICULARITIES Shortest path dynamical programming using matrices.

ADDITIONAL SCRIPTS oms for model exploration, R for result exploration

E.2.11 Static User Equilibrium

OBJECTIVE

LOCATION <https://github.com/JusteRaimbault/TransportationEquilibrium/tree/master/Models>

CHARACTERISTICS

- Langage : python, R
- Taille : $\simeq 300$

E.2.12 Geography of fuel prices**OBJECTIVE**

LOCATION <https://github.com/JusteRaimbault/EnergyPrice/tree/master/Models>

CHARACTERISTICS

- Langage : python, R
- Taille : 1469

PARTICULARITIES

★ ★

★

E.3 TOOLS AND WORKFLOW FOR AN OPEN REPRODUCIBLE RESEARCH

We briefly evoke here tools or workflows currently under development or testing, aimed at easing an open reproducible research and making it more transparent.

E.3.1 *NetLogo documentation generator*

Documentation generation is central for reproducibility as it can automatize implementation description. NetLogo does not provide a documentation generator and we experimented a Doxygen wrapper for NetLogo code, that basically consists in transforming NetLogo code into Java code and parsing documentation comment blocks. An experimental version is available at <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Doc>.

E.3.2 *git as a reproducibility tool*

The use if git as a reproducibility and transparency tool was emphasized in [Ram, 2013], for various reasons such as exact history tracing, easy cloning, past commit branching.

It furthermore can help individual workflow for advantages such as automatic backup, organisation, experiments tracking.

E.3.3 *Open Review*

Le processus de revue de ce manuscrit a expérimentalement testé la revue ouverte, par l'utilisation du dépôt git et de commandes L^AT_EXspécifiques. La commande basique \comment permet aux relecteurs d'insérer leur commentaires à l'endroit approprié (et se place alors en annotation de marge dans le manuscrit) et permet une discussion jusqu'à 5 réponses successives par des arguments optionnels. Une *pull request* depuis la branche du relecteur permet d'intégrer les retours. D'autres commandes permettent par exemple de marquer les changements ou d'insérer des listes de tâches.

L'un des intérêts de cette démarche est qu'il est possible a posteriori de reconstruire le processus de revue, et que celui-ci est entièrement ouvert (pour une éventuelle revue de la revue). L'automatisation par parcours du réseau de l'historique du dépôt git est même facilement envisageable.

E.3.4 Towards a git-compatible figures metadata handler

The issue of meta-data for figures is a crucial issue, as it is often difficult to keep a trace of all parameter values that have generated it, along with the corresponding code. Tricks may furthermore happen in script environments such as R or python when variables are accidentally modified without code modification.

Keeping an exhaustive trace of the exact dataset, code and history that has generated a precise figure is a necessary condition for exact reproducibility. We are elaborating a git-compatible tool that would automatically handle these metadata, for example by branching and associating the unique commit hash to the figure. The final idea would be to have under each figure a unique identifier linking to the associated reproducing environment.

★ ★

★

REFLEXIVE ANALYSIS

We have throughout all this work the crucial role of a reflexivity in the research process. It has played a role for the definition of objects or questions asked, in a concrete way in works directly based on fieldwork, or in the elaboration of theories with a recursive aspect. Without pretending having exhaustively constructed a “meta-viewpoint” as recommended by [Morin, 1991] for the construction of a complex thinking, we suggest to have brought preliminary elements of answer.

We propose here, as a “meta-conclusion”, to proceed to a quantitative analysis for reflexivity, by applying the methods we developed to our work itself. We do in a first part the analysis of the scientific landscape from the corpus of our bibliography, and then analyze in a second part the evolution of the knowledge produced in terms of projects and of knowledge domains.

This approach is particularly original since to the best of our knowledge there exist no monograph explicitly including its own analysis using quantitative tools. We defend a more systematic use of such approaches, to foster the development of a knowledge at the second order.

F.1 HYPERNETWORK ANALYSIS

We apply here the methodology using citation and semantic networks developed in Chapter 2. The initial corpus is constituted by all our bibliography¹ which includes 834 references.

F.1.1 Citation Network

We reconstruct the citation network at depth two from this initial corpus, and obtain a consequent network ($|V| = 177428$, $|E| = 203317$), of average degree 2.29 (average in-degree 1.15). The core of the network, constituted by vertices with a degree larger than or equal to 2 for the largest connected component (which covers 98% of the network), has a size of $|V| = 19714$ and $|E| = 47348$.

A detection of communities using the Louvain algorithm gives a directed modularity of 0.74 for 19 communities with a size larger than 10. We interpret the communities by the tags given in Table 30. We recover domains that are directly covered and used in our work

¹ Fixed at the 27/11/2017, and available at https://github.com/JusteRaimbault/CityNetwork/raw/master/Models/Reflexivity/data/CityNetwork_20171127.bib.

TABLE 30: **Citation communities.** The size of communities is given as a proportion of the size of the core of the network.

Community	Size
Economic Geography	12.4 %
Power Laws	9.1 %
Networks	7.92 %
Spatial Urban Growth Models	7.67 %
Physics of Cities	7.43 %
ABM	7.37 %
Complexity	7.19 %
LUTI	7.16 %
Urban Systems	5.15 %
Spatial Statistics	5.13 %
Evolutionary Economic Geography	5.03 %
Spatio-temporal data	3.18 %
Datamining	2.81 %
Quantitative Epistemology	2.43 %
Space Syntax/Procedural modeling	2.43 %
Fractals	2.02 %
VGI	1.8 %
Biological Networks	1.33 %
Chaos	0.624 %

(Urban Systems, Spatial Models of Urban Growth), and other neighbors mentioned but not directly used (Fractals, Economic Geography, Space Syntax).

The citation network is visualized in Fig. 153. The position of communities is very instructive to situate our work, which forms bridges between different domains depending on the viewpoint chosen. If we take the point of view of urban systems, the corresponding community (in red) makes a bridge between LUTI models (turquoise) and urban growth models (black) on one side, and economic geography on the other side (in green). If we take the viewpoint of simulation models (ABM community, purple), the link is established between Power Laws (light blue) and networks and spatial networks (magenta and orange). Auxiliary communities are attached at the periphery: Quantitative Epistemology (yellow) is close to network analysis, whereas the analysis of spatio-temporal processes (dark green) is relatively independent. Communities within which our models can be thematically classified (growth models and urban systems) are located at the core of the compact part of the network: this confirms that the direction explored are not auxiliary, and that all principal auxiliary

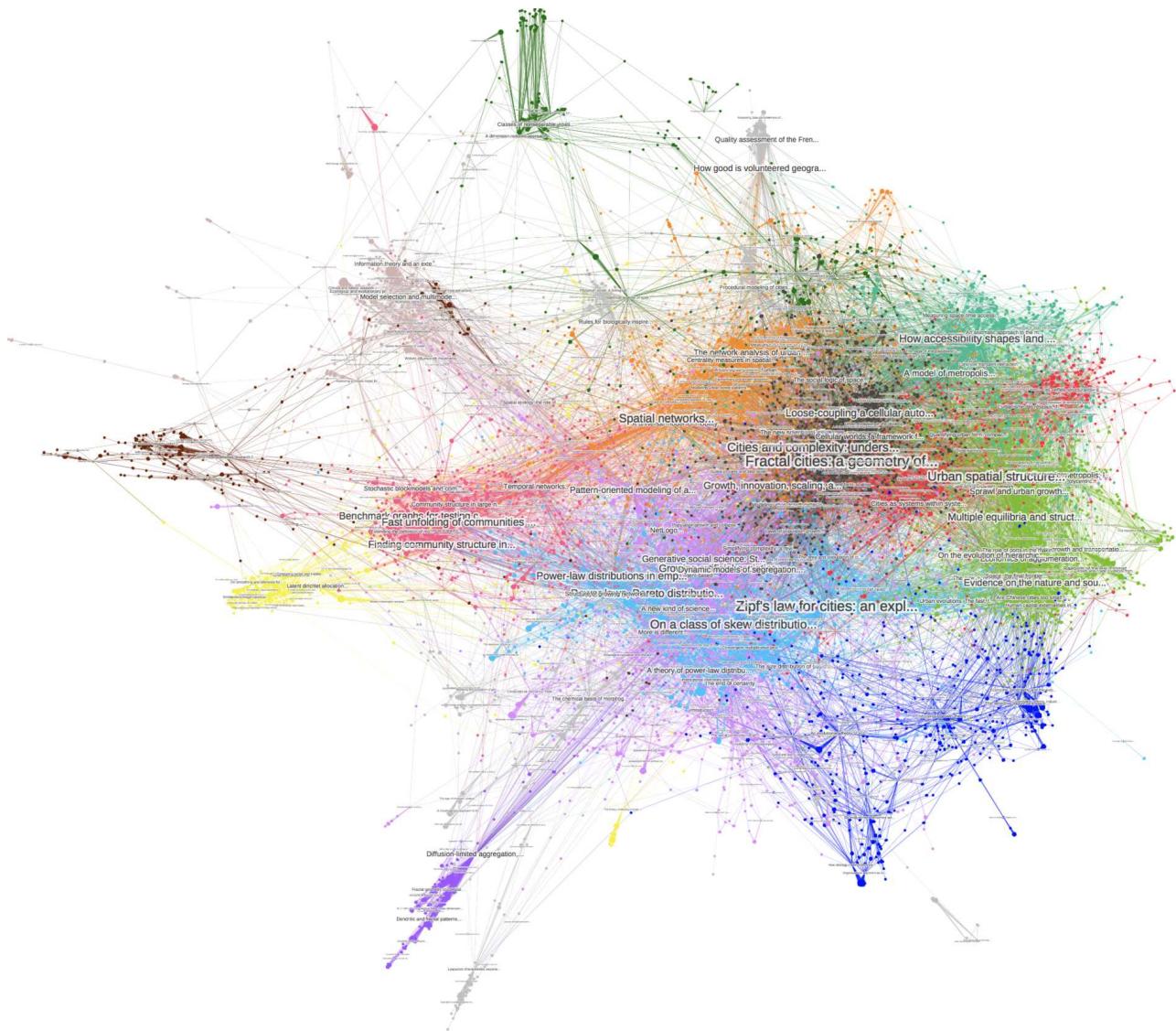


FIGURE 153: Citation network. We visualize only the core of the network, constituted here by nodes with a degree larger than or equal to 2. The network is spatialized with the algorithm Force Atlas 2. The size of labels is proportional to the degree of nodes, and the color gives the community.

domains evoked where “necessary” in the sense of a strong connection between communities here.

F.1.2 Semantic network

After collecting the abstracts, we obtain 91412 references on which it is possible to proceed to the semantic analysis. The construction of the raw co-occurrences network, after filtering links with a weight smaller than 5, and for a number of keywords $K_W = 50000$, yields a semantic network with $|E| \simeq 16 \cdot 10^6$. The sensitivity analysis to filtering parameters suggests to choose $k_{\min} = 0$, $k_{\max} = 500$, $f_{\max} = 10000$, $\theta_w = 5$, what produces a semantic network of size $|V| = 37482$ and $|E| = 218926$, with 26 communities and a modularity of 0.78. Main communities can be labeled as: toxicology, chemistry, political sciences, theoretical ecology, urban systems, sustainability, innovation economics, spatial analysis, physiology, physics, networks, bioanthropology, health, statistics, microbiology, transportation, biological networks, health geography, botany, evolution, ecology, genetics.

It is less evident to use this typology to understand our work, in comparison with the citation network, since remote domains (toxicology, chemistry, physiology, botany) can be found in relatively small amount in our citation corpus (coming from common citations on morphogenesis or ecology for example) but form then communities that are particularly isolated in the semantic network. We give in Fig. 154 the distribution of semantic interdisciplinaries for each citation community. At the exception of evolutionary economic geography which is relatively flat (and thus rather closed) and voluntary geographical information (VGI) which exhibits a peak at 0 (which is expected for such a specific domain), citation communities have fundamentally the same interdisciplinarity profile.

F.2 INTERACTION BETWEEN PROJECTS

We propose here to quantify the evolution of the different projects and of their interactions, and also of associated knowledge domains. A table of the time spent on each project, with a precision of half an hour, has been held between the 16/02/2015 and the 16/02/2018². A project is defined as a minimal consistent entity, either by its thematic (for example: morphogenesis model of 5.2) either by its content (case studies, geographical theory). These have been defined progressively in time, and some overlap or are the precursors of others: we have thus built a classification a posteriori under the form of “macro-projects” which globally correspond to the final articulation. We also

² It is available at <https://github.com/JusteRaimbault/CityNetwork/raw/master/Docs/Organisation/Projects.ods>. For the analyses here, we use the version frozen at the 02/12/2017.

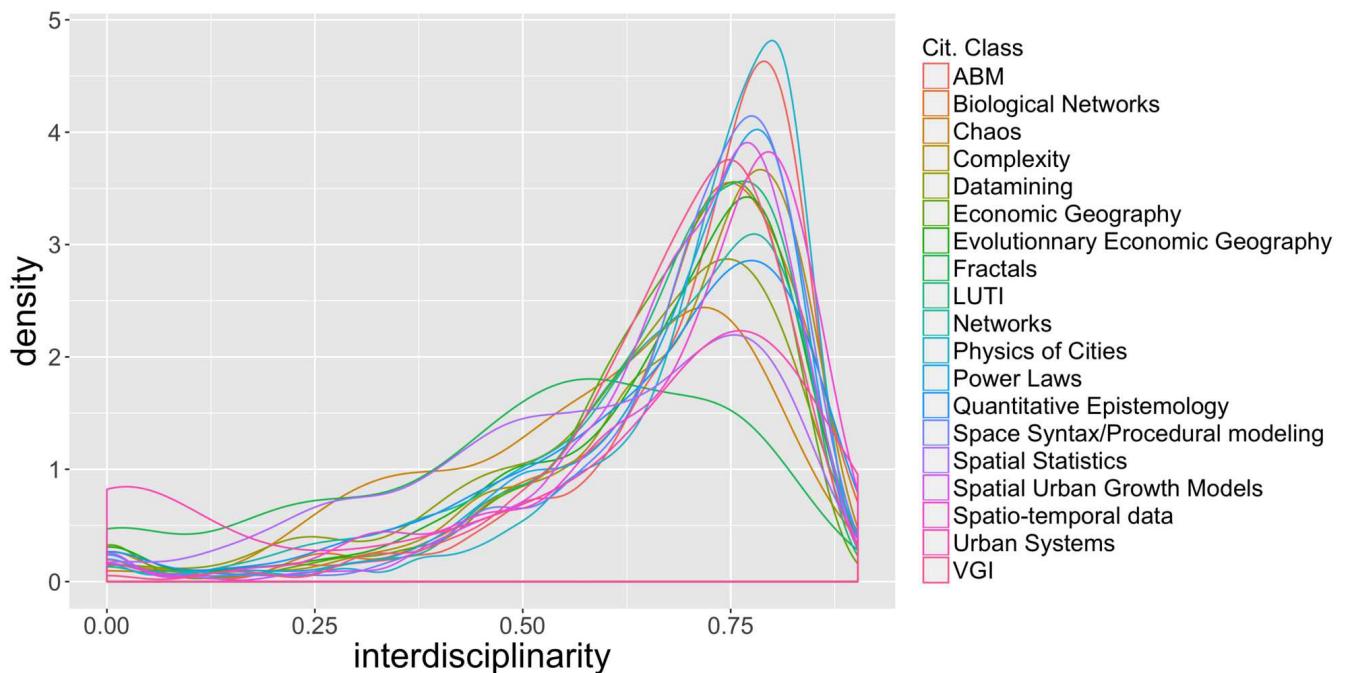


FIGURE 154: Distribution of interdisciplinaries for each citation community.

associate to them a main knowledge domain³ and the main section of this memoir to which they are attached.

The list of projects is given in Table 31, with the macro-project, the knowledge domain and the cumulated time. The Fig. 155 gives the temporal distribution according to these different modalities, in time. We confirm a non-linear organization, most of projects and chapters being treated in parallel. For example, the chapter 7 has been the object of a first preliminary exploration in the first months, and a resurgence when converging as the intellectual maturity had been acquired. Methods regularly punctuate the distribution, but culminate just before the first half. Modeling projects, similarly to empirical studies, are also regularly distributed, whereas the conceptual takes more time in the end, what confirms that it necessitates the other domains and a thorough reflection.

It is then possible to construct interaction graphs between macro-projects or knowledge domains, assuming simplifying hypotheses.

A first index of simultaneous interaction is based on an apparition at the same time. We denote $T_{i,t}$ the time for the entity i (macro-project or knowledge domain) on the temporal unit t (that we take as the week). The probability of simultaneous occurrence between i

³ Knowing that there is a non-negligible bias in the fact of attributing a unique domain to a project, since domains are generally intimately linked at the core of knowledge production itself. The constraint of data collection however leads to this segmentation which is relatively reductionist.

TABLE 31: Description of projects. The section links to the part of the memoir where the project is mainly used. Global generic tasks are not taken into account in the chapter cumulated count (Memoire: writing of this memoir; Academic: academic life; Bibliography: general readings).

Project	Macro-project	Section	Domain	Time (h)
CaseStudies	Thematic	1.2	Empirical	5.5
Modelography	QuantEpistemo	2.2	Empirical	20
QuantEpistemology	QuantEpistemo	2.2	Empirical	32
MacroCoEvol	MacroCoEvol	6.2	Modeling	72
SpatioTempCausality	CausalityRegimes	4.2	Methods	37.5
Entretiens	Thematic	D.3	Data	13
MesoCoEvol	MesoCoEvol	7.2	Modeling	60.5
Fieldwork	Thematic	1.3	Empirical	27.5
EnergyPrice	Empirical	C.1	Empirical	72.5
Morphogenesis	Morphogenesis	5.1	Conceptual	24.5
NetworkNecessity	InteractionGibrat	4.3	Modeling	158
Memoire	Memoire	-	Conceptual	489.5
SpatialStatistics	CausalityRegimes	4.2	Methods	44
BPCaseStudy	CausalityRegimes	1.2	Empirical	12
Perspectivism	Epistemology	8.3	Conceptual	8.5
RealEstate	CausalityRegimes	1.2	Empirical	18
Theory	Thematic	1.1, 8.2	Conceptual	136
CorrelatedSyntheticData	Methods	5.3	Methods	128
MediationEcotox	Methods	C.7	Methods	59
DensityGeneration	DensityGeneration	5.2	Modeling	84.5
PatentsMining	Methods	C.5	Methods	349.5
CyberGeo	Methods	B.6, C.4	Methods	332
SpaceMatters	Methods	3.1	Methods	100.5
NetworkDensityStatistics	Empirical	4.1	Empirical	176.5
NetLogoUtils	Tools	-	Tools	10
StochasticUrbanGrowth	Methods	B.1	Methods	13
TransportationEquilibrium	Empirical	3.1	Empirical	56.5
BiologicalNetwork	MesoCoEvol	7.1	Modeling	5
Discrepancy	Methods	B.4	Methods	54
Governance	Governance	7.3	Modeling	228
SyntheticData	Methods	5.2	Methods	99
Reproduction	MacroCoEvol	6.1	Modeling	46
AlgorithmicReview	QuantEpistemo	2.2	Empirical	75.5
Tools	Tools	-	Tools	137
Academic	Acad	-	NA	1388
Bibliography	Biblio	-	Conceptual	312

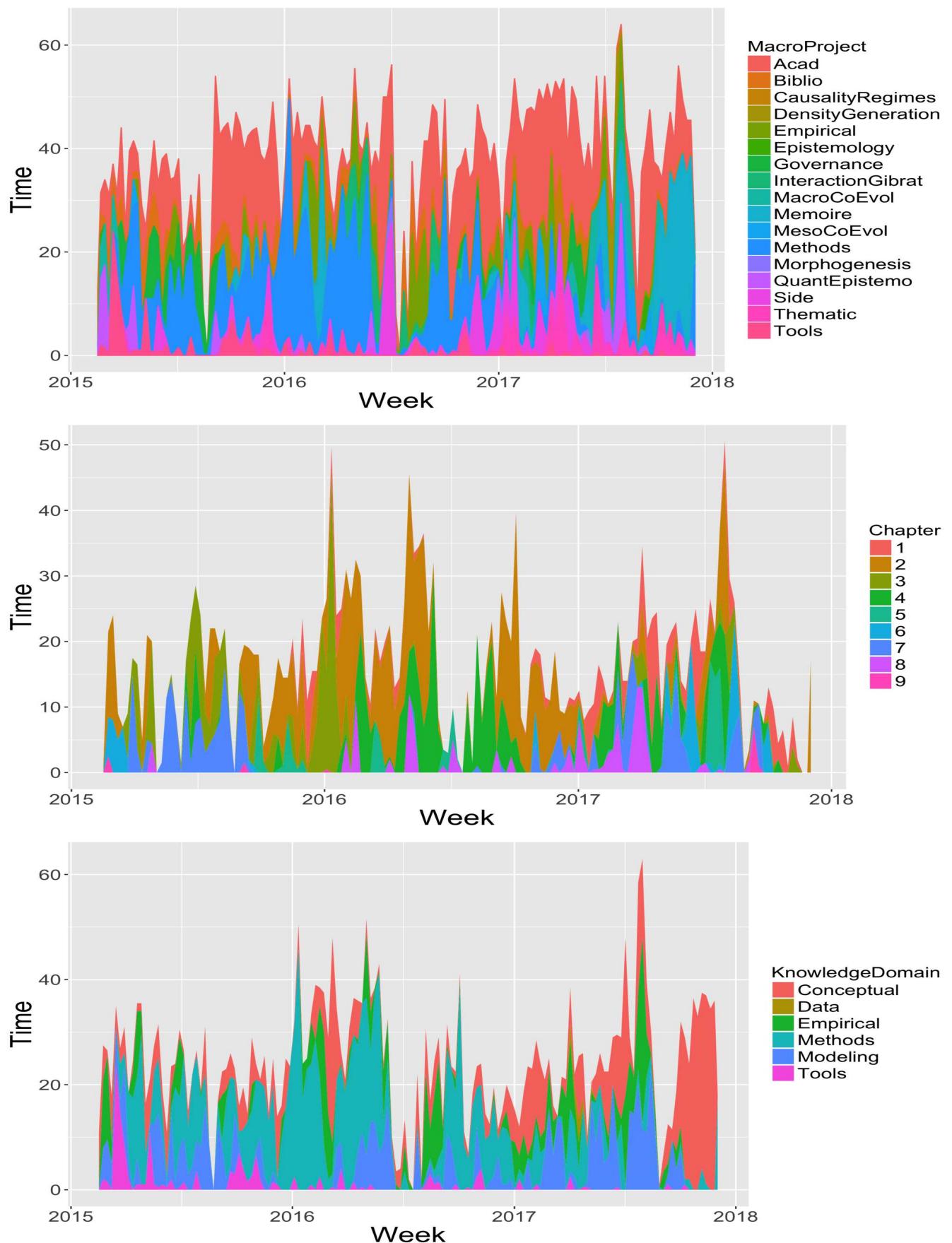


FIGURE 155: **Temporal distribution.** Times are aggregated at the level of the week and areas in color give the temporal distribution for macro-projects (first row), chapters (second row) and knowledge domains (third row).

and j is at time t given by $\frac{T_{i,t}T_{j,t}}{(\sum_i T_{i,t})^2}$, and we can sum them in time to obtain an index of interaction between entities:

$$I_{i,j} = \sum_t \frac{T_{i,t}T_{j,t}}{(\sum_i T_{i,t})^2}$$

The matrix $(I_{i,j})$ allows then to construct a network. A similar index based uniquely on co-occurrence is given by

$$C_{i,j} = \sum_t \mathbb{1}_{T_{i,t}>0} \mathbb{1}_{T_{j,t}>0}$$

We also look at lagged interactions, under the assumption that an entity at time t can trigger the one at time $t+1$, the non-symmetrical index being then

$$\tilde{I}_{i \rightarrow j} = \sum_t \frac{T_{i,t}T_{j,t+1}}{\sum_i T_{i,t} \sum_j T_{j,t+1}}$$

and the same index of lagged co-occurrence

$$\tilde{C}_{i \rightarrow j} = \sum_t \mathbb{1}_{T_{i,t}>0} \mathbb{1}_{T_{j,t+1}>0}$$

Nous montrons les graphes correspondants pour les macro-projets en Fig. 156. Concernant les macro-projets, il apparaît que le cœur du réseau de co-occurrence simultanée est constitué par la bibliographie, la vie académique et les méthodes : ces éléments sont quasi-médiatisés à chaque instant et structurent le reste de la recherche. Ensuite, les différents projets thématiques peuvent être menés relativement indépendamment, et gravitent à la périphérie du réseau. Avec le graphe dirigé des interactions retardées, il n'est guère possible de tirer une information supplémentaire, les flux étant quasi-symétriques : soit l'agrégation à la semaine n'est pas pertinente soit le délai peut être différent, soit il y a effectivement réciprocité, cette dernière hypothèse étant crédible vu l'intrication des projets.

Les graphes pour les domaines de connaissance sont donnés en Fig. 157. Outre que les données sont relativement périphériques, ce qui est attendu vu leur faible importance et leur intégration au sein d'autres projets, nous n'observons pas de motifs particuliers dans ces graphes : l'ensemble des domaines est mobilisé à la plupart des instants. Il y a également l'ensemble des relations réciproques dans le graphe dirigé, ce qui suggère éventuellement une co-évolution entre les domaines de connaissance, ce que nous allons vérifier par la suite.

Nous estimons pour chaque couple de domaine de connaissance i, j (hors du domaine données qui ne cumule que 13h au total donc trop peu de variations pour estimer une corrélation) les corrélations retardées entre les différences $\rho[\Delta T_{i,t-\tau}, T_{j,t}]$ pour $-4 \leq \tau \leq 4$ (délai

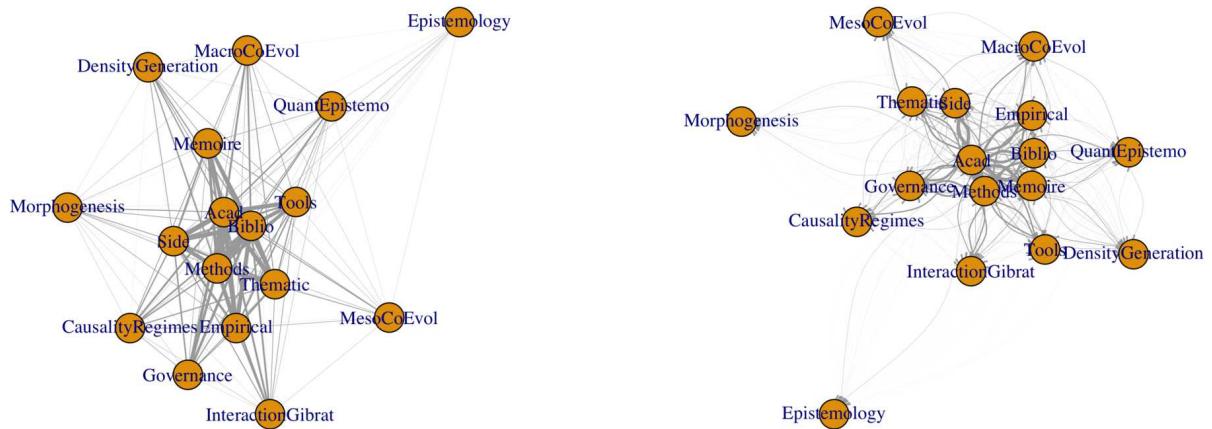


FIGURE 156: Interaction networks between projects

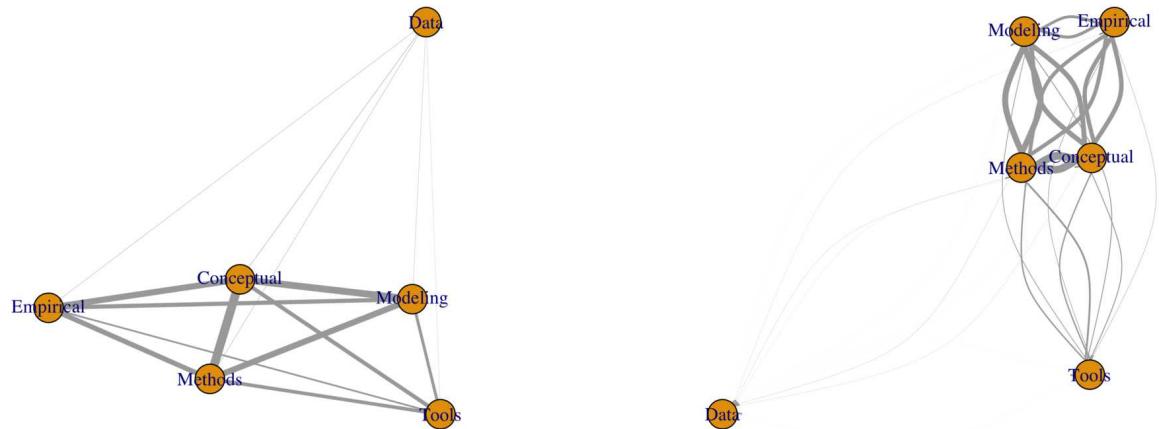


FIGURE 157: Interaction networks between knowledge domains.

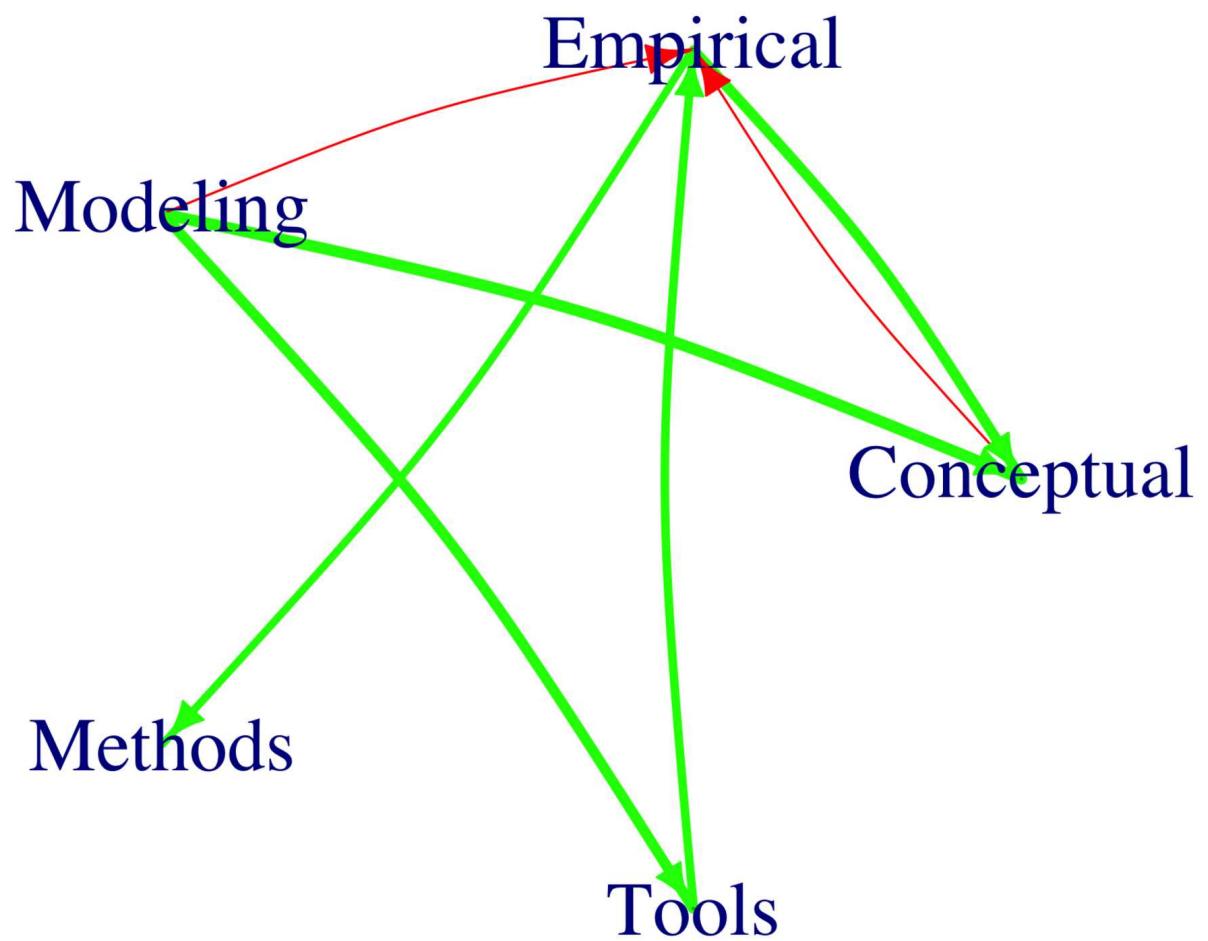
maximal d'un mois). Nous conservons les corrélations si $p < 0.05$ et sélectionnons la corrélation absolue maximale pour chaque couple de variable si elle existe.

Le graphe des corrélations retardées est donné en Fig. 158. Il existe un certain nombre de liens significatifs, et même une relation circulaire entre empirique et conceptuel, c'est-à-dire une co-évolution à proprement parler entre ces domaines. La modélisation et l'empirique induisent des travaux dans le domaine conceptuel, ce qui peut s'interpréter comme une induction des théories. Par contre, le conceptuel diminue l'empirique, ce qui pourrait être symptôme d'une trop grande déconnection avec le concret parfois.

Ainsi, même si ces résultats sont bien sûr à prendre avec précaution vu les biais intrinsèques aux données (difficulté de donner un label, réduction au sein de projets, etc.), nous suggérons une intrication des domaines de connaissance et une co-évolution pour certains. Cela peut être mis en écho avec l'hypothèse fondamentale du cadre de connaissance développé en 8.3, qui implique qu'une connaissance complexe nécessite co-évolution des domaines. Enfin, l'application des propres outils de notre travail à lui-même suggère une dimension holographique [Morin, 1986], rappelant le lien entre complexité et production de connaissance suggéré en 3.3.

★ ★

*



Caractérisation et modélisation de la co-évolution des réseaux de transport et des territoires

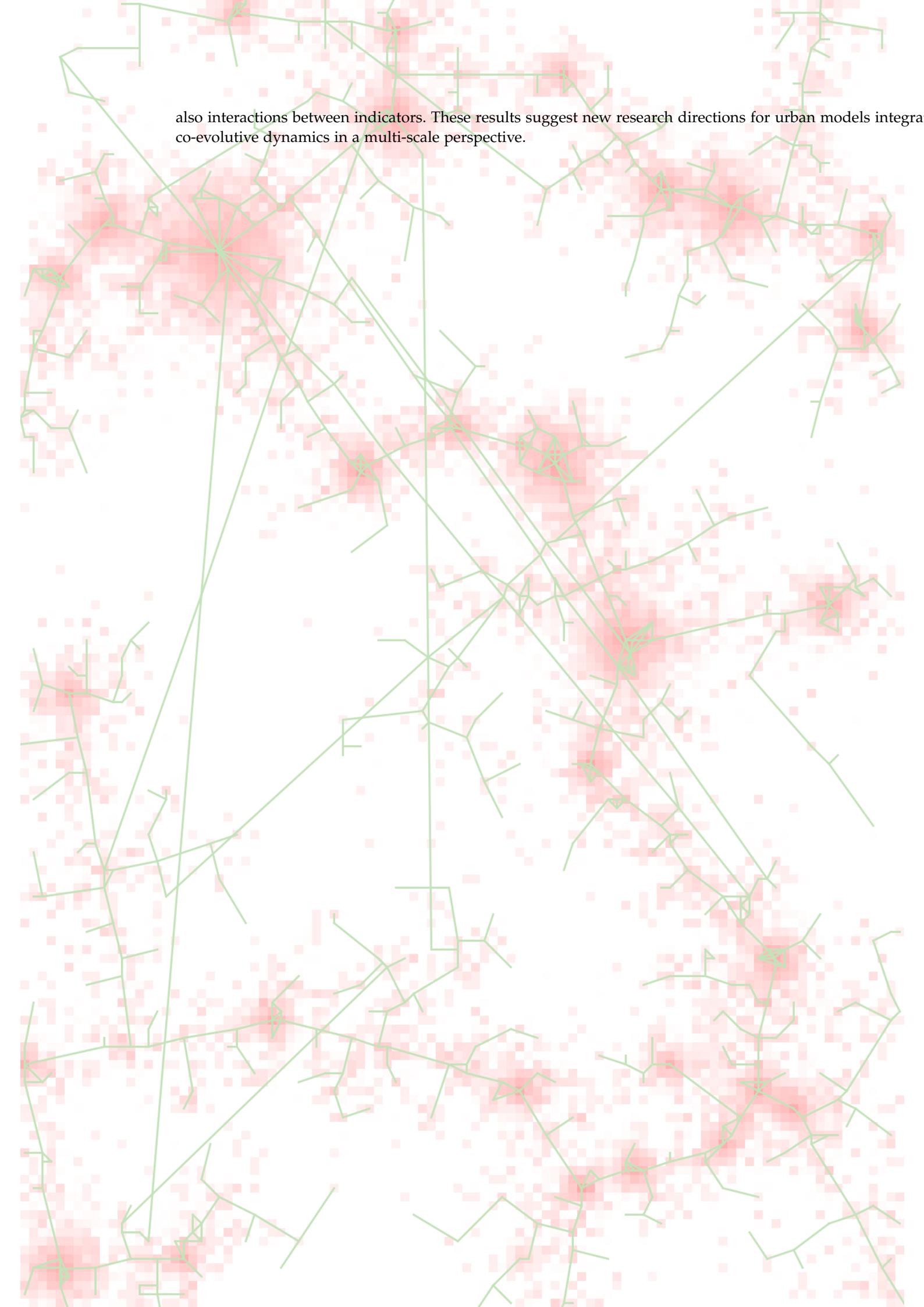
Mots-clés : Territoires ; Réseaux de Transport ; Co-évolution ; Morphogenèse ; Théorie Évolutive des Villes ; Épistémologie Quantitative ; Systèmes de Villes ; Morphologie Urbaine ; Grand Paris ; Delta de la Rivière des Perles

L'identification d'effets structurants des infrastructures de transports sur la dynamique des territoires reste un défi scientifique ouvert. Cette question est une des facettes de recherches sur la complexité des dynamiques territoriales, au sein desquelles territoires et réseaux de transport seraient en co-évolution. L'objectif de cette thèse est de mettre à l'épreuve cette vision des interactions entre réseaux et territoires, autant sur le plan conceptuel que sur le plan empirique, en les intégrant au sein de modèles de simulation des systèmes territoriaux. La nature intrinsèquement pluri-disciplinaire de la question nous conduit à mener un travail d'épistémologie quantitative, qui permet de dresser une carte du paysage scientifique et une description des éléments communs et des spécificités des modèles traitant la co-évolution entre réseaux et territoires dans chaque discipline. Nous proposons ensuite une définition de la co-évolution, ainsi qu'une méthode de caractérisation empirique, basée sur une analyse de corrélations spatio-temporelles. Deux pistes complémentaires de modélisation, correspondant à des ontologies et des échelles différentes sont alors explorées. A l'échelle macroscopique, nous construisons une famille de modèles dans la lignée des modèles d'interaction au sein des systèmes de villes développés par la Théorie Evolutive des Villes (Pumain, 1997). Leur exploration montre qu'ils capturent effectivement des dynamiques de co-évolution, et leur calibration sur des données démographiques pour le système de villes français (1830-1999) quantifie l'évolution des processus d'interaction comme l'effet tunnel ou le rôle de la centralité. A l'échelle mésoscopique, un modèle de morphogenèse capture la co-évolution de la forme urbaine et de la topologie du réseau. Il est calibré sur les indicateurs correspondants pour la forme et la topologie locales calculés pour l'ensemble de l'Europe. De multiples processus d'évolution du réseau s'avèrent être complémentaires pour reproduire la grande variété des configurations observées, au niveau des indicateurs ainsi que des interactions entre indicateurs. Ces résultats suggèrent de nouvelles pistes d'exploration des modèles urbains intégrant les dynamiques co-évolutives dans une perspective multi-échelles.

Characterizing and modeling the co-evolution of transportation networks and territories

Keywords: Territories; Transportation Networks; Co-evolution; Morphogenesis; Evolutive Urban Theory; Quantitative Epistemology; Systems of Cities; Urban Morphology; Greater Paris; Pearl River Delta

The identification of structuring effects of transportation infrastructure on territorial dynamics remains an open research problem. This issue is one of the aspects of approaches on complexity of territorial dynamics, within which territories and networks would be co-evolving. The aim of this thesis is to challenge this view on interactions between networks and territories, both at the conceptual and empirical level, by integrating them in simulation models of territorial systems. The intrinsically multidisciplinary nature of the question requires first to proceed to a quantitative epistemology analysis, that allow us to draw a map of the scientific landscape and to give a description of common features and specificities of models studying the co-evolution between network and territories within each discipline. We propose consequently a definition of co-evolution and an empirical method for its characterization, based on spatio-temporal correlation analysis. Two complementary modeling approaches, that correspond to different scales and ontologies, are then explored. At the macroscopic scale, we build a family of models inheriting from interaction models within system of cities, developed by the Evolutive Urban Theory (Pumain, 1997). Their exploration shows that they effectively capture co-evolutionary dynamics, and their calibration on demographic data for the French system of cities (1830-1999) quantifies the evolution of interaction processes such as the tunnel effect or the role of centrality. At the mesoscopic scale, a morphogenesis model captures the co-evolution of the urban form and of network topology. It is calibrated on corresponding indicators for local form and topology, computed for all Europe. Multiple network evolution processes are shown complementary to reproduce the large variety of observed configurations, at the level of indicators but



also interactions between indicators. These results suggest new research directions for urban models integrating co-evolutive dynamics in a multi-scale perspective.