

THÈSE DE DOCTORAT
pour obtenir le grade de
DOCTEUR DE L'UNIVERSITÉ PARIS VII - DENIS DIDEROT
en
GÉOGRAPHIE

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

Présentée par

JUSTE RAIMBAULT

Sous la direction de ARNAUD BANOS et FLORENT LE NÉCHET

UMR CNRS 8504 Géographie-cités
et UMR-T IFSTTAR 9403 LVMT

18 Janvier 2018 – version 3.4.2

Composition du Jury :

Arnaud Banos
Florent Le Néchet
Didier Josselin
Catherine Morency

Directeur de Recherche, CNRS (Directeur)
Maître de Conférence, Université Paris-Est (Directeur)
Directeur de Recherche, CNRS (Rapporteur)
Professeure, Ecole Polytechnique de Montréal (Rapporteuse)

[22 janvier 2018 at 12:17 – Thesis version 3.4.2]

JUSTE RAIMBAULT : *Caractérisation et modélisation de la co-évolution des réseaux de transport et des territoires*, Mémoire de Thèse de Doctorat, © 18 Janvier 2018

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 (Pu-main, 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.

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.

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。

NOTES DE LECTURE

Cette thèse devait initialement être rédigée en anglais pour sa version originale. Un premier tiers et la majorité des articles l'ont été, pour être repris et traduits par la suite, afin de répondre à une contrainte administrative d'un autre âge. Elle avait également été conçue comme une "thèse à articles", mais les fortes recommandations du CNU ont vite eu vent de cette ambition. Ainsi, la version courante est passée par maintes transformations et "lissages", afin de lui donner une forme, un fond et une identité "classiques". Nous nous excusons préalablement auprès du lecteur si des écueils de traduction ou d'articulation subsistent et perturbent la fluidité de la lecture.

L'ensemble des figures est produit par l'auteur, sauf la figure 12 (source xkcd). La grande majorité des figures est *directement* reproducible, c'est à dire pouvant être obtenue par execution des scripts. L'ensemble des codes sources, des modèles à l'interprétation des résultats et à cette propre rédaction, est disponible de manière ouverte avec l'ensemble de son historique atomique (*commits*) sur le dépôt du projet¹. L'ensemble des jeux de données produits dans ce cadre est ouvert, et l'ensemble des données utilisées sont ouvertes ou rendues ouvertes (de manière agrégée correspondant au niveau d'utilisation par les modèles dans le cas d'une base tierce fermée).

Ce mémoire en lui-même a été relu par les lecteurs suivants (ordre alphabétique) : Arnaud Banos (AB), Clémentine Cottineau (CC), Florent Le Néchet (FL), Cinzia Losavio (CL), Sébastien Rey (SR), dans l'esprit d'une revue ouverte : en suivant les commits successifs à <https://github.com/JusteRaimbault/ThesisMemoire>, l'utilisation de commandes spécifiques permet de retracer l'ensemble du processus de revue.

Les noms en Mandarin (villes, lieux, personnes, etc.) sont transcrits en système *pinyin*.

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

PUBLICATIONS

Les publications et communications suivantes contiennent la majorité du contenu de cette thèse. Les sources sont précisément mentionnées en introduction de chaque chapitre. Les traductions sont assurées par l'auteur le cas échéant.

PUBLICATIONS

Raimbault, J. (2017). An Applied Knowledge Framework to Study Complex Systems. A. Chapoutout, D. Krob, A. Roussel, F. Stephan, eds. Complex Systems Design & Management, Dec 2017, Paris, France. pp.31-45, 2017, *Proceedings of the 8th International Conference on Complex Systems Design & Management*.

Raimbault, J. & Bergeaud, A. (2017). The Cost of Transportation : Spatial Analysis of Fuel Prices in the US, forthcoming in *Transportation Research Procedia, EWGT 2017*. arxiv :1706.07467

Raimbault, J. (2017). Identification de causalités dans des données spatio-temporelles, *Sageo 2017 Proceedings*. arXiv :1709.08684

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

Raimbault, J. (2017). A Discrepancy-Based Framework to Compare Robustness Between Multi-attribute Evaluations. In *Complex Systems Design & Management* (pp. 141-154). Springer International Publishing.

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

Raimbault, J. (2016). Generation of Correlated Synthetic Data, *Actes des Journées de Rochebrune 2016*.

Raimbault, J. (2015). Models Coupling Urban Growth and Transportation Network Growth : An Algorithmic Systematic Review Approach, forthcoming in *ECTQG 2015 proceedings*. arxiv :1605.08888

DOCUMENTS DE TRAVAIL

Raimbault, J. (2017). Calibration of a Density-based Model of Urban Morphogenesis. *En revue pour PloS ONE*. arXiv preprint arXiv :1708.06743.

Raimbault, J. (2017). Exploration of an Interdisciplinary Scientific Landscape. *En revue pour Scientometrics*. arXiv preprint arXiv :1712.00805.

Raimbault, J. (2017). Indirect Evidence of Network Effects in a System of Cities. *En revue pour Environment and Planning B*.

Raimbault J., Cottineau C., Le Texier M., Le Néchet F. & Reuillon R. (2017). Space Matters : extending sensitivity analysis to initial spatial conditions in geosimulation models. *En revue pour Environment and Planning B*.

Banos A., Chasset P.-O., Commenges A. Cottineau C., Pumain D. & Raimbault J. (2018). Where do you mean ? A spatialised bibliometrics approach of a scientific journal production. *En revue pour Big Data & Society*.

COMMUNICATIONS

Complexity, Complexities and Complex Knowledge, *Geodiversity International Workshop, Paris, October 2017*.

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

Raimbault 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*

Cottineau C., Raimbault 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. & Raimbault 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. In *Treizièmes Rencontres de ThéoQuant, Besançon, Mai 2017*

Un Cadre de Connaissances pour une Géographie Intégrée. In *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, 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, Sep 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>). Nous remercions 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). Nous souhaitons remercier le Centre National de la Recherche Scientifique (CNRS) et l'UMR 8504 Géographie-cités pour leurs soutiens 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.

TABLE DES MATIÈRES

Introduction	3
I FONDATIONS	23
1 INTERACTIONS ENTRE RÉSEAUX ET TERRITOIRES	27
1.1 Territoires et Réseaux	29
1.2 De Paris à Zhuhai	52
1.3 Elements de terrain	70
2 MODÉLISER LES INTERACTIONS ENTRE RÉSEAUX ET TERRITOIRES	85
2.1 Modéliser les Interactions	87
2.2 Une Approche Epistémologique	107
2.3 Revue Systématique et Modélographie	124
3 POSITIONNEMENTS	141
3.1 Modélisation, données massives et calcul intensif	143
3.2 Reproductibilité	163
3.3 Positionnement Epistémologique	173
II BRIQUES ÉLÉMENTAIRES	197
4 THÉORIE EVOLUTIVE URBAINE	207
4.1 Corrélations entre forme des territoires et forme des réseaux	212
4.2 Causalités Spatio-temporelles	228
4.3 Modèle de croissance macroscopique	249
5 MORPHOGENÈSE URBAINE	273
5.1 Une Approche Interdisciplinaire de la Morphogenèse	275
5.2 Morphogenèse Urbaine par Agrégation-diffusion	291
5.3 Génération de configurations territoriales corrélées	308
III SYNTHÈSE	323
6 CO-ÉVOLUTION À L'ÉCHELLE MACROSCOPIQUE	327
6.1 Modèles existants	329
6.2 Extension dynamique du modèle d'interaction	339
7 CO-ÉVOLUTION À L'ÉCHELLE MESOSCOPIQUE	357
7.1 Modèles de Croissance de Réseau	359
7.2 Co-évolution à l'échelle mesoscopique	371
7.3 Modélisation de la Gouvernance du Système de Transport	380
IV OUVERTURE	401
8 ÉCHELLES ET ONTOLOGIES	405
8.1 Représentation territoriales	407
8.2 Equilibre Utilisateur Statique	408

8.3	Transport Routier et Déterminants des Coûts	422
9	CADRE THÉORIQUE	437
9.1	Contributions et Perspectives	439
9.2	Une Théorie Géographique	444
9.3	Un Cadre de Connaissances Appliqué	451
	Conclusion	471
V	APPENDICES	481
A	INFORMATIONS SUPPLÉMENTAIRES	483
A.1	Etudes de cas	484
A.2	Elements de Terrain	488
A.3	Epistémologie Quantitative	491
A.4	Modélographie	496
A.5	Correlations Statiques	504
A.6	Régimes de causalité	515
A.7	Effets de réseau	519
A.8	Morphogenèse par agrégation-diffusion	520
A.9	Données Synthétiques Corrélées	529
A.10	Exploration du modèle SimpopNet	531
A.11	Modèle de co-évolution macroscopique	532
A.12	Heuristiques de génération de réseau	538
A.13	Co-évolution à l'échelle mesoscopique	541
A.14	Modélisation de la gouvernance du système de transport	542
B	DÉVELOPPEMENTS MÉTHODOLOGIQUES	551
B.1	Modèles stochastiques de croissance urbaine	553
B.2	Sensibilité des Lois d'Echelle Urbaines	557
B.3	Génération de Données Synthétiques Corrélées	561
B.4	Robustesse d'une évaluation multi-attributs	564
B.5	Un Cadre pour les Systèmes Socio-techniques	580
B.6	Exploration d'un paysage scientifique interdisciplinaire	592
C	DÉVELOPPEMENTS THÉMATIQUES	615
C.1	Ponts entre Géographie et Economie	617
C.2	CybergeoNetworks : une analyse bibliométrique multi-dimensionnelle spatialisée	618
C.3	Classification sémantique des brevets	642
C.4	Communication scientifique par la gamification	667
C.5	Données synthétiques corrélées : Séries temporelles financières	672
C.6	Modélisation multi-scalaire des dynamiques résidentielles	679
D	DONNÉES	689
D.1	Données de Traffic du Grand Paris	689
D.2	Graphes topologiques des Réseaux Routiers	689
D.3	Interviews	690
D.4	Données Synthétiques et résultats de simulations	691

E OUTILS	693
E.1 Softwares and Packages	694
E.2 Architecture and Sources for Algorithms and Models of Simulation	695
E.3 Tools and Workflow for an open Reproducible Research	700
F QUANTITATIVE ANALYSIS OF THESIS REFLEXIVITY	703

TABLE DES FIGURES

FIGURE 1	Projets de transport successifs du Grand Paris	56
FIGURE 2	Impact du Grand Paris Express sur l'accessibilité	57
FIGURE 3	Corrélations retardées empiriques	62
FIGURE 4	Gain d'accessibilité permis par le HZMB . . .	67
FIGURE 5	Réseau à grande vitesse en Chine	74
FIGURE 6	TOD à Hong-Kong et Zhuhai	76
FIGURE 7	Algorithme de revue systématique	111
FIGURE 8	Réseau de citations	117
FIGURE 9	Motifs d'interdisciplinarité	121
FIGURE 10	Méthodologie de la revue systématique . . .	127
FIGURE 11	Types de couplages	129
FIGURE 12	Usage naïf de la fouille de données	155
FIGURE 13	Distance des diagramme de phase à la référence	161
FIGURE 14	Exemples de diagrammes de phase	161
FIGURE 15	Reproductibilité et visualisation	166
FIGURE 16	Distribution spatiale des morphologies	216
FIGURE 17	Distribution spatiale des indicateur de réseau .	221
FIGURE 18	Corrélations Spatiales	223
FIGURE 19	Variation des corrélations avec l'échelle . . .	224
FIGURE 20	Séries temporelles auto-régressives	237
FIGURE 21	Correlations dans le modèle RDB	240
FIGURE 22	Identification de régimes d'interactions	242
FIGURE 23	Evolution des mesures de réseau	245
FIGURE 24	Corrélations retardées en Afrique du Sud . . .	247
FIGURE 25	Corrélation de la croissance en fonction de la distance	260
FIGURE 26	Interface du modèle d'interaction	261
FIGURE 27	Révélation d'effets de réseau	262
FIGURE 28	Calibration du modèle de gravité	263
FIGURE 29	Valeurs des paramètres calibrés	264
FIGURE 30	Calibration du modèle complet	267
FIGURE 31	Exemple de formes urbaines générées	298
FIGURE 32	Transitions de formes générées	301
FIGURE 33	Dépendance au chemin	303
FIGURE 34	Calibration du modèle	305
FIGURE 35	Exploration par PSE	307
FIGURE 36	Espace faisable des corrélations entre morphologie et réseau	313
FIGURE 37	Génération de configurations couplées	314
FIGURE 38	Comportement du modèle	336
FIGURE 39	Corrélations	338

FIGURE 40	Schématisation du modèle de co-évolution macroscopique	340
FIGURE 41	Comportement temporel du modèle de co-evolution	344
FIGURE 42	Comportement agrégé du modèle de co-evolution	345
FIGURE 43	Correlations dans le modèle abstrait	347
FIGURE 44	Corrélations empiriques pour le système de villes français	350
FIGURE 45	Fronts de Pareto	353
FIGURE 46	Evolution des paramètres optimaux	354
FIGURE 47	Example de réseau auto-renforçant	355
FIGURE 48	Exemple de génération de réseau biologique .	363
FIGURE 49	Exemples de réseaux	366
FIGURE 50	Espace topologique faisable	368
FIGURE 51	Comparaison aux réseaux réels	369
FIGURE 52	Morphogenèse mesoscopique	372
FIGURE 53	Calibration du modèle de morphogenèse . .	375
FIGURE 54	Régimes de causalité	378
FIGURE 55	Formes de réseau obtenues pour différents niveaux de gouvernance	392
FIGURE 56	Application de Lutecia au Delta de la Rivière des Perles	393
FIGURE 57	Calibration du modèle Lutecia	395
FIGURE 58	Application web pour les données de trafic .	412
FIGURE 59	Variabilité spatiale des plus courts chemins .	413
FIGURE 60	Variabilité des temps de trajet	414
FIGURE 61	Stabilité temporelle de la centralité	416
FIGURE 62	Auto-corrélation spatiale	418
FIGURE 63	Prix moyen par Contés	427
FIGURE 64	Autocorrelation spatiale	428
FIGURE 65	Résultats des analyses GWR	431
FIGURE 66	Réseau de citations de la Théorie Evolutive Urbaine	457
FIGURE 67	Réseau complet des domaines de connaissance	461

LISTE DES TABLEAUX

TABLE 1	Transports en commun dans le Delta de la Rivière des Perles	66
TABLE 2	Processus d'interaction entre réseaux et territoires	82
TABLE 3	Synthèse des approches de modélisation	105
TABLE 4	Proximités lexicales stationnaires	112
TABLE 5	Communautés sémantiques	119

TABLE 6	Type de modèles	132
TABLE 7	134
TABLE 8	Synthèse des processus modélisés. Ceux-ci sont classés par échelle, type de modèle et discipline.	137
TABLE 9	Relation croisées entre indicateurs de réseau et morphologiques	226
TABLE 10	Espace des paramètres du modèle d'interaction	256
TABLE 11	Valeurs de l'AIC empirique	270
TABLE 12	Résumé des paramètres du modèle de morphogenèse	296
TABLE 13	326
TABLE 14	Sensibilité à l'espace	335
TABLE 15	Résumé des paramètres de croissance de réseau	364
TABLE 16	Résumé des paramètres du modèle LUTECIA	389
TABLE 17	Prix des carburants	425
TABLE 18	Régressions au niveau du comté	433



INTRODUCTION

INTRODUCTION

La machine à brouillard du Plateau de Saclay serait-elle le seul artefact in-temporel dans cet environnement métropolitain qui se cherche toujours ? Projeltons nous en 2100, dans cette banlieue sud parfois sordide de ce qui sera toujours Paris. Les bouleversements locaux ont bien eu lieu, mais pas de la façon attendue, le climat local étant toujours férus de ce fameux brouillard. Par contre, l'environnement urbain et la relation à la ville sont entièrement conditionnés par une grande proximité aux lignes de transport lourd : la disparition des moyens de transport thermiques, puis de l'ensemble des véhicules légers par échec technologiques des alternatives électriques, ont exacerbé le rôle des lignes de train ou de metro existantes. Les densités ont progressivement augmenté autour des gares pour produire d'impressionnantes complexes de tours, tandis que les espaces péri-urbains se vidaient progressivement. Les infrastructures de transport sont quant à elles restées quasiment à l'identique après 2030, le peu de ressources disponibles étant dédié à leur entretien, et leur extension étant conjointement sortie rapidement des agendas politiques. Ce plateau est alors rempli de bâtiments à l'abandon, puisqu'il attend toujours ce tronçon du Grand Paris Express qui n'aura finalement jamais été réalisé. La nature reprend peu à peu ses droits.

Ce pitch pour film d'anticipation à petit budget a pour avantage de nous révéler l'existence de processus complexes intriqués à différentes échelles de temps et d'espace dans la fabrique des villes : le développement historique du réseau ferroviaire en région parisienne a conditionné les évolutions futures et le RER B a suivi l'ancienne Ligne de Sceaux, le plan de DELOUVRIER pour le développement régional et son execution partielle, sont des éléments d'explication de la structure du réseau parisien de transports en commun qui conditionne fortement le développement urbain dans notre scenario ; les processus de relocalisation au sein de l'espace de la métropole, liés à une plus ou moins grande nécessité de proximité ou d'accessibilité selon les modes de transports utilisés, participent à l'évolution urbaine ; dans le cas du plateau de Saclay des processus de planification spécifiques à différents niveaux jouent un rôle crucial dans la différentiation du territoire.

La liste pourrait être ainsi continuée indéfiniment, chaque approche apportant sa vision mature correspondant à un corpus de connaissances scientifiques dans des disciplines diverses comme la géographie, l'économie urbaine, les transports. Cette anecdote est suffisante pour faire ressentir la complexité des systèmes territoriaux que nous étudierons. Notre but ici est de se plonger dans cette complexité, et en particulier donner un point de vue original sur l'étude des relations entre réseaux de transport et territoires. Le choix de cette posi-

tion sera largement discuté dans une partie thématique, nous nous concentrerons à présent sur l'originalité du point de vue que nous allons prendre.

DE LA POSITION GÉNÉRALE

L'ambition de cette thèse est de ne pas avoir d'ambition a priori. Cette entrée en matière, rude en apparence, contient à différents niveaux les logiques sous-jacentes à notre processus de recherche. Au sens propre, nous nous plaçons tant que possible dans une démarche constructive et exploratoire, autant sur les plans théoriques et méthodologiques que thématique, mais encore proto-méthodologique (outils appliquant la méthode) : si des ambitions unidimensionnelles ou intégrées devaient émerger, elles seraient conditionnées par l'arbitraire choix d'un échantillon temporel parmi la continuité de la dynamique qui structure tout projet de recherche. Au sens structurel, l'auto-référence qui soulève une contradiction apparente met en exergue l'aspect central de la réflexivité dans notre démarche constructive, autant au sens de la récursivité des appareils théoriques, de celui de l'application des outils et méthodes développés au travail lui-même ou que de celui de la co-construction des différentes approches et des différents axes thématiques. Le processus de production de connaissance pourra ainsi être lu comme une métaphore des processus étudiés. Enfin, sur un plan plus enclin à l'interprétation, cela suggérera la volonté d'une position délicate liant une conscience politique dont la nécessité est intrinsèque aux sciences humaines (par exemple ici contre l'application technocratique des modèles, ou pour le développement d'outils luttant pour une science ouverte) à une rigueur d'objectivité plus propre aux autres champs abordés, position forçant à une prudence accrue.

CONTEXTE SCIENTIFIQUE : PARADIGMES DE LA COMPLEXITÉ

Pour une meilleure introduction du sujet, il est nécessaire d'insister sur le cadre scientifique dans lequel nous nous positionnons. Ce contexte est crucial à la fois pour comprendre les concepts épistémologiques implicites dans nos questions de recherche, et aussi pour être conscient de la variété de méthodes et outils utilisés. La science contemporaine prend progressivement le tournant de la complexité dans de nombreux champs que nous illustrerons par la suite, ce qui implique une mutation épistémologique pour abandonner le réductionnisme² strict qui a échoué dans la majorité de ses tentatives de

² De manière schématique, le réductionnisme consiste en la position épistémologique que les systèmes sont entièrement compréhensibles à partir des éléments fondamentaux les constituant et des lois régissant leur évolution. Les niveaux supérieurs n'ont ni autonomie ni pouvoirs causaux irréductibles.

synthèse [anderson1972more]. Arthur a rappelé récemment [arthur2015complexity] qu'une mutation des méthodes et paradigmes en était également un enjeu, de par la place grandissante prise par les approches computationnelles qui remplacent les résolutions purement analytiques généralement limitées en possibilités de modélisation et de résolution. La capture des *propriétés émergentes* par des modèles de systèmes complexes est une des façons d'interpréter la philosophie de ces approches.

Ces considérations sont bien connues des Sciences Humaines et Sociales (qualitatives et quantitatives) pour lesquelles la complexité des agents et systèmes étudiés est une des justifications de leur existence : si les humains étaient effectivement des particules, on pourrait s'attendre à ce que la majorité des disciplines les prennent comme objet d'étude n'aient jamais émergé puisque la thermodynamique aurait alors résolu la majorité des problèmes sociaux³. Elles sont au contraire moins connues et acceptées en sciences "dures" comme la physique : [laughlin2006different] développe une vision de la physique à la même position de "frontière des connaissances" que d'autre champs plus récents qui pourrait sembler en être encore à leur genèse. La plupart des connaissances actuelles concernent des structures classiques simples, alors qu'un grand nombre de systèmes présentent des propriétés *d'auto-organisation*, au sens où les lois microscopiques ne sont pas suffisantes pour inférer les propriétés macroscopiques du système à moins que son évolution ne soit entièrement simulée (plus précisément cette vision peut être prise comme une définition de l'émergence sur laquelle nous reviendrons par la suite, or des propriétés auto-organisées sont par nature émergentes). Cela correspond au premier cauchemar du Démon de Laplace développé dans [deffuant2015visions].

A la croisée de positionnements épistémologiques, de méthodes et de champs d'application, les *Sciences de la complexité* se concentrent sur l'importance de l'émergence et de l'auto-organisation dans la plupart des phénomènes réel, ce qui les place plus proche de la frontière des connaissances que ce que l'on peut penser pour des disciplines classiques (LAUGHLIN, op. cit.). Ces concepts ne sont pas récents et avaient déjà été mis en valeur par [anderson1972more]. On peut aussi interpréter la Cybernétique comme un précurseur des Sciences de la Complexité en la lisant comme un pont entre technologie et sciences cognitives [wiener1948cybernetics], et surtout en développant les notions de retroaction et de contrôle.

Plus tard, la Synergétique [haken1980synergetics] a posé les bases d'approches théoriques des phénomènes collectifs en physique. Les causes possibles de la croissance récente du nombre de travaux se

³ Bien que cette affirmation soit elle-même discutable, les sciences physiques classiques ayant également échoué à prendre en compte l'irréversibilité et l'évolution de Systèmes Complexes Adaptatifs comme le souligne [prigogine1997end].

réclamant d'approches complexes sont nombreuses. L'explosion de la puissance de calcul en est certainement une vu le rôle central que jouent les simulations numériques [[varenne2010simulations](#)]. Elles peuvent aussi être à chercher auprès de progrès en épistémologie : introduction de la notion de perspectivisme [[giere2010scientific](#)], réflexions plus fine autour de la nature des modèles [[varenne2013modeliser](#)]⁴. Les potentialités théoriques et empiriques de telles approches jouent nécessairement un rôle dans leur succès⁵, comme le confirme les domaines très variés d'application (voir [[newman2011complex](#)] pour une revue très générale), comme par exemple la Science de Réseaux [[barabasi2002linked](#)] les Neurosciences [[koch1999complexity](#)] ; les Sciences Humaines et Sociales, dont la Géographie [[manson2001simplifying](#)][[pumain1997pour](#)] ; la Finance avec les approches éconophysiques [[stanley1999econophysics](#)] ; l'Ecologie [[grimm2005pattern](#)]. La Feuille de Route des Systèmes Complexes [[2009arXiv0907.2221B](#)] propose une double lecture des travaux en Complexité : une approche horizontale faisant la connexion entre champs d'étude par des questions transversales sur les fondations théoriques de la complexité et des faits stylisés empiriques communs, et une approche verticale, dans le but de construire des disciplines intégrées et les modèles multi-scalaires hétérogènes correspondants. L'interdisciplinarité est ainsi cruciale pour notre contexte scientifique.

INTERDISCIPLINARITÉ

Il est important d'insister sur le rôle de l'interdisciplinarité dans la position de recherche prise ici. Il s'agit autant d'un travail en Géographie Théorique et Quantitative qu'en Modélisation de Systèmes Complexes, étant finalement les deux à la fois selon le point de vue que prendra le lecteur. En ce sens, nous le réclamons de la *Science des Systèmes Complexes* que nous tenterons de positionner comme discipline propre à travers cette implémentation précise⁶. Ce n'est pas sans risques d'être lu avec méfiance voir défiance par les tenants des disciplines classiques, comme des exemples récents de malentendus ou conflits ont récemment illustré [[dupuy2015sciences](#)]. Il faut se rappeler l'importance de la spirale vertueuse de BANOS entre disciplinarité et interdisciplinarité [[banos2013pour](#)]. Celle-ci doit nécessairement impliquer différents agents scientifiques, et il est compli-

⁴ Dans ce cadre, les progrès scientifiques et épistémologiques ne peuvent pas être dissociés et peuvent être vus comme étant en co-évolution, au sens d'une forte interdépendance et d'une adaptation mutuelle.

⁵ Même si l'adoption de nouvelles pratiques scientifiques peut par ailleurs être biaisé par l'imitation et le manque d'originalité [[dirk1999measure](#)], ou de façon plus ambiguë, par des stratégies de positionnement indépendante des stratégies de connaissance, puisque le combat pour les fonds est un obstacle croissant à une recherche saine [[bollen2014funding](#)].

⁶ Un niveau de lecture abstrait du travail dans son ensemble apportera des informations sur la production de connaissance elle-même, comme nous le développerons en [9.3](#).

qué pour un agent de se positionner dans les deux branches ; notre fond scientifique devra nous permettre de ne pas de nous positionner uniquement dans la *disciplinarité géographique* (même si celle-ci sera simultanément une composante cruciale) mais bien aussi dans celle des Systèmes Complexes (qui est interdisciplinaire, voir 3.3 pour contourner la contradiction apparente), et notre sensibilité scientifique et épistémologique nous pousse à faire de même.

L'évolution scientifique des sciences de la complexité, qui est vue par certains comme une révolution [colander2003complexity], ou même comme *un nouveau type de science* [wolfram2002new], pourrait affronter des difficultés intrinsèques dues aux comportements et a-priori des chercheurs en tant qu'être humains. Plus précisément, le besoin d'interdisciplinarité qui fait la force des Sciences de la Complexité pourrait devenir une de ses grandes faiblesses, puisque la structure fortement en silo de la science peut avoir des impacts négatifs sur les initiatives impliquant des disciplines variées. Nous n'évoquons pas les problèmes de sur-publication, quantification, compétition, qui sont plus liés à des questions de Science Ouverte et de son éthique, tout aussi de grande importance mais d'une autre nature. Cette barrière qui nous hante et que nous pourrions ne pas surmonter, a pour plus évident symptôme des *divergences culturelles disciplinaires*, et les conflits d'opinion en résultant. Ce drame du malentendu scientifique est d'autant plus grave qu'il peut en effet détruire totalement certains progrès en interprétant comme une falsification des travaux qui traitent une question toute différente. L'exemple récent en économie d'un travail sur les inégalités liées aux hauts revenus présenté dans [aghion2015innovation], et dont les conclusions ont été commentées comme s'opposant aux thèses de [piketty2013capital], est typique de ce schéma. Ce second se concentre sur la construction de bases de données propres sur le temps long pour les revenus et montre empiriquement une récente accélération des inégalités de revenus, son modèle visant à lier ce fait stylisé avec l'accumulation de capital a été critiqué comme trop simpliste. D'autre part, [aghion2015innovation] montrent par des analyses économétriques que s'il existe bien un lien de causalité de l'innovation vers les inégalités de haut salaires, l'innovation accroît cependant la mobilité sociale, étant donc également moteur de réduction des inégalités. D'où des conclusions divergentes sur le rôle des capitaux personnels dans une économie, notamment sur leur relation ambiguë à l'innovation. Mais des *point de vue* ou *interprétations* différentes ne signifient pas une incompatibilité scientifique, et on pourrait même imaginer rassembler ces deux approches dans un cadre et modèle unifié, produisant des interprétations possiblement similaires et potentiellement encore nouvelles. Une telle approche intégrée aura de grandes chances de contenir plus d'information (selon la façon dont le couplage est opéré) et être une avancée scientifique. Cette expérience de pensée illustre les

potentialités et la nécessité de l'interdisciplinarité. Dans une autre veine assez similaire, [2017arXiv170105627H] ré-analyse des données biologiques d'une expérience de 1943 qui prétendait confirmer l'hypothèse des processus d'évolution Darwiniens par rapport aux processus Lamarckiens, et montrent que les conclusions ne tiennent plus dans le contexte actuel d'analyse de données (avances énormes sur la théorie et les possibilités de traitement) et scientifique (avec d'autres nombreuses preuves de nos jours des processus Darwiniens) : c'est un bon exemple de malentendu sur le contexte, et la manière selon laquelle le cadre de travail à la fois technique et thématique influence fortement les conclusions scientifiques. Nous développons à présent divers exemples révélateurs de la manière dont des conflits entre disciplines peuvent être dommageables.

Comme déjà mentionné, DUPUY et BENGUIGUI soulignent dans [dupuy2015sciences] le fait que dans le domaine de l'urbanisme, ont récemment éclaté des conflits ouverts entre les tenants classiques des disciplines et des nouveaux arrivants, en particulier les physiciens, même si leur entrée dans le domaine n'est pas nouvelle. La disponibilité de grand jeux de données d'un nouveau type (réseaux sociaux, données des nouvelles technologies de la communication) ont attiré l'attention d'un plus grand nombre sur des objets plus traditionnellement étudiés par les sciences humaines, puisque les méthodes analytiques et computationnelles de la physique statistique sont devenues applicables. Bien que ces travaux soient généralement présentés comme la construction d'une approche scientifique des villes, tout en discutant la nature scientifique des approches existantes, la nouveauté réelle des résultats obtenus et la non-légitimation des approches "classiques" sont discutables. Pour citer quelques exemples, [barthelemy2013self] conclut que Paris a subit une transition pendant la période d'Haussman et ses opérations de planification globale, qui sont des faits naturellement connus depuis longtemps en Histoire Urbaine et Géographie Urbaine. [chen2009urban] redécouvre que le modèle gravitaire est amélioré par l'introduction de décalages dans les interactions et dérive analytiquement l'expression d'une force d'interaction entre les villes, sans se placer dans un cadre théorique ou thématique. De tels exemples peuvent être multipliés, confirmant l'inconfort courant entre physiciens et géographes. Des bénéfices significatifs pourraient résulter d'une intégration raisonnée des disciplines [o2015physicists] mais la route semble être bien longue encore.

Des conflits similaires se rencontrent à l'interface des relations entre économie et géographie : comme le décrit [marchionni2004geographical], la discipline de la géographie économique, traditionnellement proche de la géographie, a fortement critiqué à son émergence l'approche relativement récente *Nouvelle Economie Géographique*. Celle-ci provient de l'économie et son but est la prise en compte de l'espace par les méthodes économiques classiques. Elles n'ont en fait pas les mêmes des-

seins et buts, et le conflit apparaît comme un malentendu complet vu d'un oeil extérieur. Par exemple, la Nouvelle Economie Géographique privilégiera des explications impliquant des processus économiques universels et indépendant des échelles, tandis que la Géographie Economique basera son argumentation sur les particularité locales et la contingence des processus. Les hypothèses épistémologiques sous-jacentes sont également très différentes, comme par exemple la relation au réalisme, la première étant fondée sur un réalisme abstrait pas forcément concrètement réaliste (utilisation de processus abstraits), tandis que la deuxième sera plus pragmatique. La mesure dans laquelle ces deux approches sont complémentaires ou incompatible reste toutefois une question ouverte d'après [marchionni2004geographical]. Des relations disciplinaires similaires seront rencontrées dans notre travail, comme entre la physique et la géographie. Nous développons par ailleurs en C.1 une exploration des liens entre économie et géographie du point de vue de la modélisation.

Des conflits disciplinaires peuvent aussi se manifester sous la forme d'un rejet de méthodes nouvelles par les courants dominants. Suivant FARMER [farmer2009economy], l'échec opérationnel de la plupart des approches économiques classiques pourrait être compensé par un usage plus systématique de la modélisation et simulation basées agent. L'absence de résolution analytique qui est inévitable pour l'étude de la plupart des systèmes complexes adaptatifs semble rebouter la plupart des économistes. Or, [barthelemy2016structure] insiste sur la déconnexion exacerbée entre une grande partie des modèles et théories économiques et les observations empiriques, du moins dans le domaine de l'économie urbaine. Celle-ci pourrait être un symptôme de la déconnexion disciplinaire évoquée ci-dessus. Toujours en économie, [storper2009rethinking] propose aussi des changements de paradigmes par un retour à l'agent et une construction associée de théories *evidence-based*.

La finance quantitative peut être instructive pour notre propos et sujet, d'une part par les similarités de la cuisine interdisciplinaire avec notre domaine (rapport avec la physique et l'économie, champs plus ou moins "rigoureux", etc.). Dans ce domaine coexistent divers champs de recherche ayant très peu d'interactions entre eux. On peut considérer deux exemples. D'une part, les statistiques et l'économétrie sont extrêmement avancées en mathématiques théoriques, utilisant par exemple des méthodes de calcul stochastique et de théorie des probabilités pour obtenir des estimateurs très raffinés de paramètres pour un modèle donné (voir par exemple [barndorff2011multivariate]). D'autre part, l'éconophysique a pour but d'étudier des faits stylisés empiriques et inférer les lois correspondantes pour tenter d'expliquer des phénomènes économiques, par exemple ceux liés à la complexité des marchés financiers [stanley1999econophysics]. Ceux-ci incluent les cascades menant aux ruptures de marché, les propriétés fractales

des signaux des actifs, la structure complexe des réseaux de corrélation. Chacun a ses avantages dans un contexte particulier et gagnerait à des interactions accrues entre les deux domaines.

Ces divers exemples pris au fil du vent sont de brèves illustrations du caractère crucial de l'interdisciplinarité et de sa difficulté à pratiquer. Sans presque exagérer, on pourrait imaginer l'ensemble des chercheurs se plaindre de mauvaises ou difficiles expériences d'interdisciplinarité, avec un retour largement positif lors des rares succès. Nous allons tenter par la suite d'emprunter ce chemin étroit, empruntant des idées, théories et méthodes de diverse disciplines, dans l'idéal de la construction d'une connaissance intégrée.

PARADIGMES DE LA COMPLEXITÉ EN GÉOGRAPHIE

Pour revenir à notre anecdote introductive, nous nous concentrerons sur l'étude d'un objet thématique qui sera les systèmes territoriaux : à l'échelle microscopique, les agents peuvent bien être vus comme éléments constitutifs fondamentaux du territoire, qui émergera comme processus complexe à différentes échelles. Plus généralement, il s'agit par commencer de brosser une revue du rôle de la complexité en géographie. Les géographes sont naturellement familiers avec la complexité, puisque l'étude des interactions spatiales est l'un de leurs objets de prédilection. La variété de champs en géographie (géomorphologie, géographie physique, géographie environnementale, géographie humaine, géographie de la santé, etc. pour en nommer certains) a sûrement joué un rôle clé dans la constitution d'une pensée géographique subtile, qui considère des processus hétérogènes et multi-scalaires.

PUMAIN rappelle dans [[pumain2003approche](#)] une histoire subjective de l'émergence des paradigmes de la complexité en géographie, que nous restituons ici. La cybernétique a produit des théories des systèmes comme celle utilisée pour les premiers modèles de dynamique des systèmes visant à simuler l'évolution de variables caractérisant un territoire, sous la forme d'équations différentielles couplées, comme [[chamussy1984dynamique](#)] l'illustre pour un modèle couplant population, emplois et stock de logements. Plus tard, le glissement vers les concepts de criticalité auto-organisée et d'auto-organisation en physique ont conduit aux développements correspondants en géographie, comme [[anders1992systeme](#)] qui témoigne de l'application des concepts de la synergétique aux dynamiques des systèmes urbains.

Enfin, les paradigmes actuels des systèmes complexes ont été introduits par plusieurs entrées relativement indépendantes. On peut nommer parmi celles-ci les concepts issus des fractales, les automates cellulaires, le *Scaling*, et la Théorie Evolutive des Villes. Nous revoyons brièvement ces approches ci-dessous.

L'étude de la nature fractale de la forme urbaine a été introduite par [batty1986fractal], plus tard synthétisée par [batty1994fractal] et a eu de nombreuses applications jusqu'à des développements plus récents comme [keersmaecker2003using] pour l'analyse de la forme urbaine ou [tannier:hal-00860260] pour l'élaboration de planifications urbaines durables.

La Théorie du *Scaling* a par ailleurs été importée de la physique et de la biologie et des relations allométriques pour expliquer les lois d'échelle urbaine comme propriétés universelles liées au type d'activité : infrastructure et économies d'agglomération (scaling infralinéaire) ou résultante d'un processus d'interactions sociales (scaling supralinéaire), et suppose les villes comme versions à l'échelle l'une de l'autre [bettencourt2007growth]. Nous n'utiliserons pas explicitement ces deux approches mais celles-ci restent sous-jacentes dans les paradigmes utilisés⁷.

Les automates cellulaires, introduits en géographie par TOBLER [couclelis1985cellular], sont une autre entrée des approches complexes pour la modélisation urbaine. BATTY en propose une synthèse jointe avec les modèles basés agents et les fractales dans [batty2007cities]. Ce type de modèle prendra une place modeste mais non négligeable dans notre travail.

Une autre introduction de la complexité en géographie fut pour le cas des systèmes urbains à travers la théorie évolutive des villes de PUMAIN. Nous nous placerons plus particulièrement dans la lignée de celle-ci et la développons ainsi avec plus de détails. En interaction intime avec la modélisation dès ses débuts (le premier modèle Simpop décrit par [sanderson1997simpop]) rentre dans le cadre théorique de [pumain1997pour]), cette théorie vise à comprendre les systèmes de villes comme des systèmes d'agents adaptatifs en coévolution, aux interactions multiples, avec différents aspects mis en valeur comme l'importance de la diffusion des innovations.

La série des modèles Simpop [pumain2012multi] a été conçue pour tester différentes hypothèses de la théorie, comme par exemple le rôle des processus de diffusion de l'innovation dans l'organisation du système urbain. Ainsi, des régimes sous-jacent différents ont été mis en évidence pour les systèmes de ville en Europe et aux Etats-unis [bretagnolle2010comparer].

A d'autres échelles de temps et dans d'autres contextes, le modèle SimpopLocal [schmitt2014modelisation] a pour but d'étudier les conditions pour l'émergence de systèmes urbains hiérarchiques à partir d'établissements disparates. Un modèle minimal (au sens de paramètres nécessaires et suffisants) a été isolé grâce à l'utilisation de calcul intensif via le logiciel d'exploration de modèles Open-Mole [schmitt2014half], ce qui était un résultat impossible à atteindre de manière analytique pour un tel type de modèle complexe. Les pro-

⁷ Par exemple, les lois d'échelles ont une place privilégiée dans l'application de la Théorie Evolutive [pumain2006evolutionary].

grès techniques d'OpenMole [reuillon2013openmole] ont été menés simultanément avec les avances théoriques et empiriques.

Les avancées épistémologiques ont également été cruciales dans ce cadre, comme REY le développe dans [rey2015plateforme], et de nouveaux concepts comme la modélisation incrémentale [cottineau2015incremental] ont été découverts, avec de puissantes applications concrètes : [cottineau2014evolution] l'applique sur le système de villes soviétique et isole les processus socio-économiques dominants, par un test systématique des hypothèses thématiques et des fonctions d'implémentation. Des directions pour le développement de telles pratiques de Modélisation et Simulation en géographie quantitative ont récemment été introduits par BANOS dans [banos2013pour]. Il conclut par neuf principes⁸, parmi lesquels on peut citer l'importance de l'exploration intensive des modèles computationnels et l'importance du couplage de modèles hétérogènes, qui sont avec d'autre principes tel la reproductibilité au centre de l'étude des systèmes complexes géographiques selon le point de vue décrit précédemment. Nous nous positionnerons en grande partie dans l'héritage de cette ligne de recherche, travaillant de manière conjointe sur les aspects théoriques, empiriques, épistémologiques et de modélisation.

VILLES, SYSTÈMES DE VILLES, TERRITOIRES

Entrons à présent dans le vif du sujet pour construire progressivement la problématique précise qui s'inscrira dans le contexte global développé jusqu'ici. Nos objets géographiques élémentaires (au sens de précurseurs dans notre genèse théorique) sont la *Ville*, le *Système de Villes*, et le *Territoire*, que nous allons à présent définir.

Un élément central des systèmes socio-géographiques est l'objet *Ville*, sur lequel nous nous positionnons pour une cohérence épistémologique propre. La question de la définition de la ville a fait couler beaucoup d'encre. [robin1982cent] montre par exemple que REYNAUD avait déjà conceptualisé la ville comme lieu central d'un espace géographique, permettant agrégation et échanges, théorie qui sera reformulée par CHRISTALLER comme *Théorie des Lieux Centraux*. Cette définition théorique est rejoints par la conception de PUMAIN qui considère la ville comme une entité spatiale clairement identifiable, constituée d'agents sociaux (élémentaires ou non) et d'artefacts techniques, et qui est l'incubateur du changement social et de l'innovation [pumain2010theorie]. Nous prendrons cette définition dans notre travail. Il faut toutefois garder à l'esprit que la définition concrète d'une ville en terme d'entités géographiques et d'étendue spatiale est problématique : des définitions morphologiques (c'est à

⁸ Cela doit-il devenir les dix commandements ? RENÉ DOURSAT soulignait l'absence du dernier commandement de BANOS, l'essence intrinsèque de notre entreprise est peut être en partie liée à sa recherche.

dire se basant sur la forme et la distribution du bâti), fonctionnelles (se basant sur l'utilisation des fonctions urbaines par les agents, par exemple par aire de déplacement domicile-travail dominant), administratives, etc., sont partiellement orthogonales et plus ou moins adaptées au problème étudié [guerois2002commune]. Récemment, un certain nombre d'études ont montré la forte sensibilité des lois d'échelles urbaines⁹ aux délimitations choisies pour l'estimation, pouvant entraîner une inversion des propriétés qualitatives attendues (voir par exemple [arcaute2015constructing]). Les variations des exposants estimés en fonction de paramètres de définition, comme effectué par [2015arXiv150707878C], peut être interprété comme une propriété plus globale et une signature du système urbain.

Cela confirme la nécessité de considérer les villes dans leur système, et l'importance de la notion de *Système Urbain*¹⁰. Un système urbain peut être considéré comme un ensemble de villes en interaction, dont les dynamiques seront plus ou moins fortement couplées. [berry1964cities] considère les villes comme "*systèmes dans des systèmes de villes*", appuyant sur le caractère multi-scalaire (au sens d'échelles emboîtées ayant un certain niveau d'autonomie) et nécessairement complexe, conception reprise et étendue par la Théorie Evolutionniste des Villes détaillée précédemment (voir aussi 9.2). Le terme de *Système de Villes* sera utilisé lorsque l'on pourra clairement identifier des villes comme sous-systèmes, et on parlera de Système Urbain de manière plus générale (une ville elle-même étant un système urbain).

Enfin, sous-jacente à la compréhension des dynamiques des systèmes urbains intervient la notion de *Territoire*. Polymorphe et correspondant à des visions multiples, celle-ci, que nous développerons en profondeur en 1.1, peut être définie de manière préliminaire simplement. Le territoire désigne alors la distribution spatiale des activités urbaines, des agents les exerçant ou les développant, et des artefacts techniques, dont l'infrastructure, les supportant, ainsi que la superstructure¹¹ qui leur est associée¹².

⁹ Les lois d'échelle consistent en une régularité statistique observable au sein d'un ensemble de ville, reliant par exemple une variable caractéristique Y_i à la population P_i sous la forme d'une loi puissance $Y_i = Y_0 \cdot (P_i/P_0)^\alpha$.

¹⁰ Concernant la définition d'un système, on pourra la prendre en toute généralité comme un ensemble d'éléments en interaction, présentant une certaine structure déterminée par celle-ci, et possédant un certain niveau d'autonomie avec son environnement. Il peut s'agir d'une autonomie majoritairement ontologique dans le cas d'un système ouvert, ou d'une autonomie réelle dans le cas d'un système fermé.

¹¹ Nous comprenons la superstructure au sens marxiste, c'est à dire la structure organisationnelle et l'ensemble des idées d'une société, incluant les structures politiques.

¹² Le lien entre le Territoire et la Ville, ou le Système de Ville, sera également creusé plus loin lors de la construction approfondie du concept.

RÉSEAUX, INTERACTIONS ET CO-ÉVOLUTION

Une caractéristique fondamentale des systèmes urbains et des territoires est leur inscription simultanée dans l'espace et le temps, qui transparaît dans leur dynamiques spatio-temporelles, à de multiples échelles. La notion de *processus* au sens de [hypergeo], c'est à dire l'enchaînement dynamique de faits aux propriétés causales¹³, permet de capturer les relations entre composantes de ces dynamiques, et est ainsi une notion clé pour une compréhension partielle de ces systèmes. Toute compréhension partielle sera associée au choix d'échelles, qui doit être comprise ici au sens opérationnel (caractéristiques physiques), et d'une *ontologie* qui correspond à la spécification des objets réels étudiés¹⁴. Nous allons à présent spécifier ces concepts abstraits, en introduisant les Réseaux, leurs *Interactions* avec les territoires et leur approche par la *Co-évolution*.

Une ontologie particulière retiendra notre attention : au sein des territoires émergent des Réseaux *Physiques*, qui peuvent être compris selon [dupuy1987vers] comme la matérialisation d'un ensemble de connexions potentielles entre agents du territoire. La question de l'implication de ces réseaux et de leur dynamique dans les dynamiques territoriales, qu'on peut synthétiser comme *interactions entre réseaux et territoires*, a fait l'objet d'abondants débats scientifiques et techniques, notamment dans le cas des réseaux de transport. Nous reviendrons sur la nature et le positionnement de ceux-ci aux Chapitres 1 et 2, mais nous pouvons d'ores et déjà prendre certaines de difficultés soulevées comme point de départ de notre questionnement. L'un des aspects récurrents est celui du *mythe des effets structurants*, consacré par [offner1993effets] en critique d'une utilisation exagérée par les planificateurs et les politiques d'un concept scientifique dont les fondements empiriques sont encore discutés. La question fondamentale sous-jacente que nous reformulons est la suivante : *dans quelle mesure est-il possible d'associer des dynamiques territoriales à une évolution de l'infrastructure de transport ?* On peut poser la question de manière ré-

¹³ Nous prendrons la causalité au sens de causalité circulaire dans les systèmes complexes, qui considère des cycles d'entrainement entre phénomènes, ou des structures plus complexes. La causalité linéaire, c'est à dire un phénomène entraînant un autre, est un cas particulier idéalisé de celle-ci. Nous reviendrons en détail sur la notion de causalité et sur ses différentes approches par les géographes en Section 4.2

¹⁴ Plus précisément, nous utilisons la définition de [livet2010] qui couple l'approche ontologique du point de vue de la philosophie, c'est à dire "l'étude de ce qui peut exister", et celui de l'informatique qui consiste à définir les classes, les objets et leurs relations qui constituent la connaissance d'un domaine. Cet usage de la notion d'ontologie biaise naturellement notre recherche vers des paradigmes de modélisation, mais nous prenons la position (développée en détails plus loin) de comprendre toute construction scientifique comme un *modèle*, rendant la frontière entre théories et modèles moins pertinentes que pour des visions plus classiques. Toute théorie doit faire des choix sur les objets décrits, leur relations et les processus impliqués, et contient donc une ontologie dans ce sens.

ciproque, et même la généraliser : quels sont les processus capturant les interactions entre ces deux objets ?

Une approche permettant de poser différemment le problème est la notion de *co-evolution*, utilisée en Théorie Evolutive pour qualifier les processus fortement couplés¹⁵ d'évolution des villes comme utilisé par [paulus2004coevolution], et appliquée aux relations entre réseaux et villes par [bretagnolle:tel-00459720]¹⁶. Cette dernière distingue une phase "d'adaptation mutuelle" entre réseaux et villes, correspondant à une dynamique dans laquelle des effets causaux sont clairement attribuables à l'un sur le développement de l'autre (par exemple, les nouvelles lignes de transport répondent à une demande croissante induite par la croissance urbaine, ou inversement la croissance urbaine est favorisée par une nouvelle connectivité au réseau), de la phase de co-évolution, qu'elle définit comme une "interdépendance forte" (p. 150) dans laquelle les rétroactions jouent un rôle privilégié et "la dynamique du système de ville n'est plus contrainte par le développement des réseaux de transport" (p. 170). Ces boucles de rétroaction et cette interdépendance mutuelle, vus dans leur perspective dynamique, correspondent à des relations causales circulaires (au sens donné plus haut) difficiles à séparer. Nous prendrons comme définition préliminaire de la co-évolution entre deux composantes d'un système *l'existence d'un couplage fort, correspondant généralement à des relations causales circulaires*.

PROBLÉMATIQUE

Ce cadre permet de capturer un certain degré de complexité, mais reste cependant flou ou trop général dans sa caractérisation, à la fois théorique et empirique. Nous ferons ici le pari de mettre à l'épreuve et d'approfondir cette approche, pour éclaircir ses apports potentiels pour la compréhension des interactions entre réseaux et territoires. La clarification d'une part de ce qu'elle signifie et d'autre part de

¹⁵ On parlera de *couplage* de systèmes ou de processus pour désigner la constitution d'un système englobant les éléments couplés, par l'émergence de nouvelles interaction ou de nouveaux éléments. La définition de la nature et de la force d'un couplage est une question ouverte, et nous utiliserons la notion de manière intuitive, pour désigner un plus ou moins grand niveau d'interdépendance entre les sous-systèmes couplés.

¹⁶ [paulus2004coevolution] transfère directement le concept biologique de coévolution (qui consiste en une interdépendance forte entre deux espèces dans leurs trajectoires évolutives, et qui en fait correspond à l'existence d'une *niche écologique* constituée par les espèces comme nous le développerons plus loin en 9.2), et parle de villes qui "se concurrencent, s'imitent, coopèrent". Ce transfert reste flou (sur les échelles temporelles impliquées, le statut des objets qui co-évoluent) et finalement non exploré. Des trajectoires similaires ne peuvent suffire à exhiber des interdépendances fortes comme il affirme en conclusion, celles-ci pouvant être fortuites. De plus, le transfert de concepts entre disciplines est une opération pour laquelle prudence doit être de mise (nous illustrerons cela par l'étude interdisciplinaire de la morphogenèse, concept initialement biologique, en Chapitre 5).

son existence empirique sera un noeud gordien de notre démarche. Notre problématique générale se décompose alors en deux axes complémentaires :

1. Comment définir et/ou caractériser les processus de co-évolution entre réseaux de transports et territoires ?
2. Comment modéliser ces processus, à quelles échelles et par quelles ontologies ?

Le deuxième aspect découle de notre positionnement scientifique, qui postule l'utilisation de la modélisation, et plus particulièrement de la simulation de modèle, comme un instrument fondamental de connaissance des processus au sein des systèmes complexes.

ORGANISATION GÉNÉRALE

Nous proposons de répondre à la problématique ci-dessus par la stratégie suivante. Une première partie posera les fondations nécessaires, en précisant les définitions, concepts et objets étudiés, en dessinant le paysage scientifique gravitant autour de la question, et en raffinant le positionnement épistémologique. Cette partie est composée de trois chapitres :

1. Un premier chapitre développe la question des interactions entre réseaux et territoires, d'un point de vue théorique mais aussi en les illustrant par des études de cas et des éléments de terrain. Il permet de situer la notion de co-évolution à la fois de manière concrète et abstraite.
2. Un deuxième chapitre se charge d'une manière similaire de classifier le positionnement au regard de la modélisation de la coévolution. L'état de l'art est complété par une cartographie des disciplines scientifiques concernées et par une modélographie, c'est à une classification et décomposition systématique d'un corpus de modèles afin de comprendre les ontologies utilisées et de possibles déterminants de celles-ci.
3. Une troisième chapitre développe notre positionnement épistémologique, qui s'avère avoir une influence considérable sur les choix de modélisation qui seront opérés par la suite. Nous y développons les questions liées au pratiques de modélisation, de *datamining* et de calcul intensif, des questions de reproducibilité, et des considérations épistémologiques plus générales intrinsèques aux systèmes étudiés.

De ces analyses complémentaires se dégagent deux positionnements thématiques correspondant à deux échelles de modélisation, peu explorés pour notre question particulière : la Théorie Evolutive des

Villes qui induit une modélisation macroscopique au niveau du système de ville, et la Morphogenèse Urbaine qui permet de considérer les liens entre forme et fonction à l'échelle mesoscopique. La deuxième partie s'attèlera donc à construire les briques élémentaires à partir de ces approches, qui serviront par la suite à la construction des modèles :

4. Le quatrième chapitre traite de différents aspects impliqués par la Théorie Evolutive. Le caractère non-stationnaire des processus dans l'espace est un élément crucial, que nous démontrons empiriquement dans une première section par l'étude des corrélations spatiales entre forme urbaine et topologie du réseau routier pour l'Europe et la Chine. Ensuite, la notion de causalité circulaire est explorée, et nous développons une méthode permettant d'isoler ce qu'on appelle des *régimes de causalité*, c'est à dire des configurations typiques d'interaction capturées par les motifs de corrélation retardée. Celle-ci est testée sur données synthétiques et données réelles dans le cas de l'Afrique du Sud, où l'on démontre un effet des politiques de segregation sur les interactions réseaux-territoires elle-mêmes. Cette première partie du chapitre complète de manière empirique la caractérisation de la coévolution ébauchée en première partie. Enfin, nous construisons un modèle de système urbain basé sur les interactions entre villes, qui permet de démontrer indirectement l'existence d'effets de réseau, qui postule cependant un réseau fixe.
5. Le cinquième chapitre creusera la notion de *Morphogenèse*, en commençant par en proposer un point de vue cohérent au travers de différentes disciplines la mobilisant, afin d'en dégager une caractérisation se reposant sur l'émergence d'une architecture par relations causales circulaires entre forme et fonction. Cette précision sera cruciale dans la nature des modèles mis en place. Une deuxième section développe un modèle simple de croissance urbaine prenant en compte la distribution de la population seule, et capturant les forces contradictoires de concentration et de dispersion. Nous démontrons sa capacité à reproduire des formes urbaines existantes à partir des données de forme urbaine calculées précédemment. Il est ensuite couplé séquentiellement à un modèle de génération de réseau, ce qui permet d'exhiber un large spectre de corrélations potentiellement générées.

A ce stade, nous batissons dans la troisième partie sur les fondations et avec les briques élémentaires notre construction fondamentale, qui consiste en différents modèles (ou famille de modèles) de coévolution, que nous différencions selon les deux approches considérées. Toujours dans une logique d'approches parallèles et complé-

mentaires, nous élaborons les développements des deux chapitres précédents, dans deux chapitre modélisant la co-évolution :

6. Le sixième chapitre développe un modèle de co-évolution à l'échelle macroscopique. Dans un premier temps, nous explorons de manière systématique l'unique modèle analogue existant. Nous développons ensuite le modèle par extension du modèle d'interaction déjà introduit. Son exploration systématique révèle sa capacité à produire différents régimes de co-évolution, certains témoignant de causalités circulaires. Il est également calibré sur le système de villes français sur le temps long, sur données de population et de réseau ferroviaire, ce qui permet d'inferer des informations indirectes sur les processus impliqués.
7. Le septième chapitre s'intéresse aux modèles de morphogenèse urbaine capturant les processus de co-évolution. La question des heuristiques de génération de réseau est d'abord traitée, en comparant les potentialités de diverses méthodes. Dans une démarche de multi-modélisation, celles-ci sont ensuite intégrées dans une famille de modèle de morphogenèse, que l'on calibre sur les indicateurs de forme urbaine et de topologie de réseau, au premier ordre (valeurs des indicateurs) et au second ordre (matrices des corrélations). Nous ébauchons ensuite un modèle plus complexe, visant à intégrer les processus de gouvernance dans la croissance du réseau de transport. Celui-ci est exploré de manière préliminaire.

Après avoir démontré les capacités de nos deux approches à capturer certains aspects de la co-évolution et d'informer les processus correspondants, nous procédons à une ouverture dans une dernière partie :

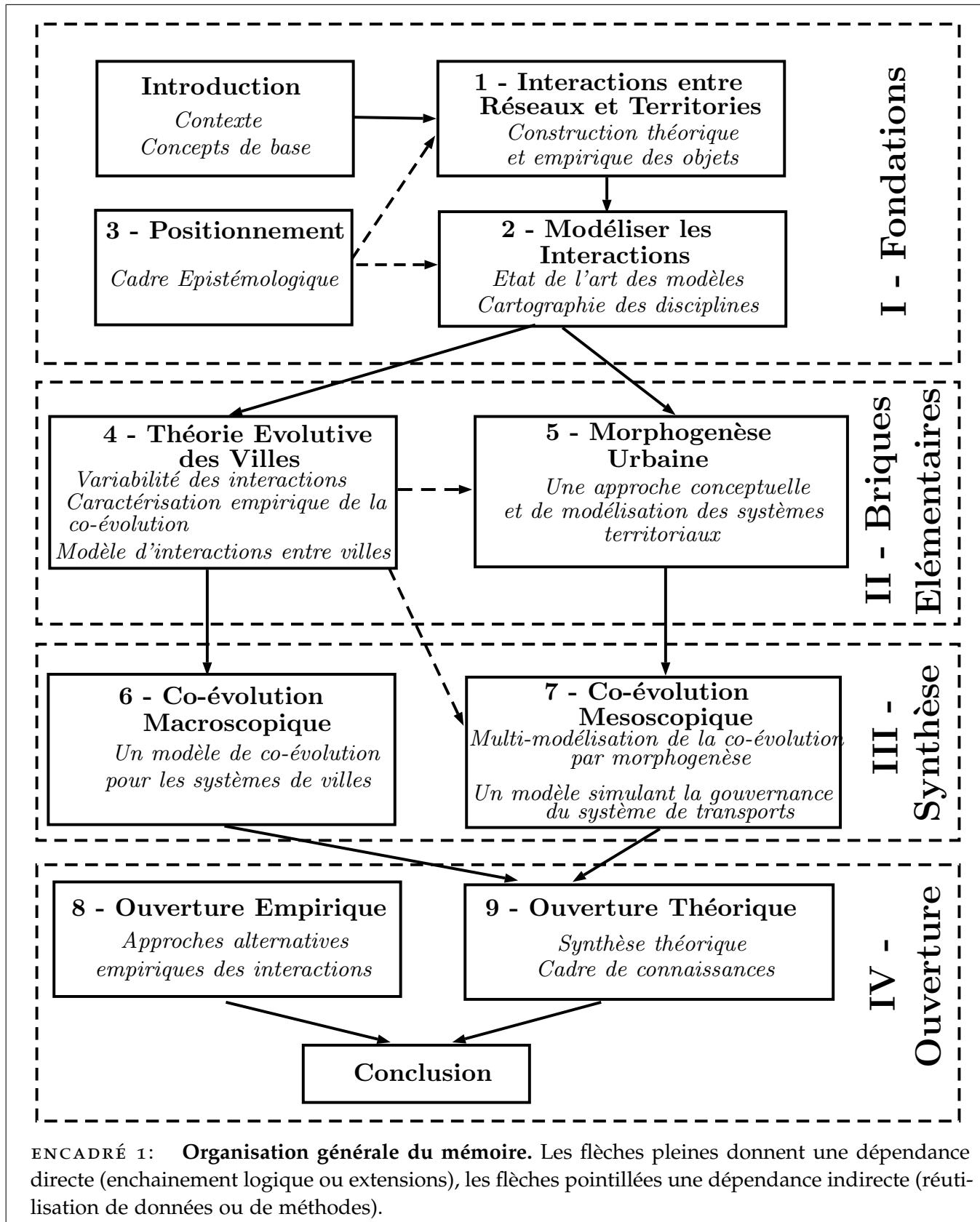
8. Le huitième chapitre est consacré à une ouverture par des analyses empiriques, visant à explorer une possible extension des échelles et des ontologies. L'analyse des flux de traffic routier en Ile-de-France, correspondant à une échelle microscopique des interactions entre réseau et territoire, révèle une nature chaotique à ces échelles et questionne la pertinence de leur modélisation. L'analyse spatio-temporelle des prix du carburant aux Etats-Unis, qui capturent indirectement l'interaction entre le système socio-économique et le réseau routier, confirme d'une part l'existence d'échelles spatiales typiques et de régimes locaux d'interaction, et d'autre part la superposition de processus territoriaux bottom-up et de processus de gouvernance top-down.
9. Le neuvième et dernier chapitre consiste en une ouverture théorique et épistémologique. Nous esquissons une réconciliation théorique de la morphogenèse et de la théorie évolutive, dans laquelle la co-évolution est centrale. Ce développement pourrait

poser les bases d'une théorie et de modèles multi-échelle pour la co-évolution. Nous développons enfin dans une démarche réflexive un cadre de connaissance pour l'étude des systèmes complexes, à la fois produit et précurseur de l'ensemble de notre démarche.

Nous résumons cette organisation, ainsi que les dépendances directes ou indirectes entre les différents chapitres, dans l'encadré 1 ci-dessous.

★ ★

★



ENCADRÉ 1: **Organisation générale du mémoire.** Les flèches pleines donnent une dépendance directe (enchaînement logique ou extensions), les flèches pointillées une dépendance indirecte (réutilisation de données ou de méthodes).

Première partie

FONDATIONS

Cette partie pose les fondations de notre démarche, en reconstruisant la question de manière théorique et par l'illustration de cas d'étude, puis en dressant un panorama scientifique de ses approches existantes en modélisation. Nous développons également notre positionnement épistémologique aux implications pratiques importantes.

INTRODUCTION DE LA PARTIE I

Un voyage, la découverte d'une ville, de nouvelles rencontres, un partage d'idées : autant de processus qui impliquent une générativité cognitive et une interaction complexe entre nos représentations, nos actions et l'environnement. La construction d'une connaissance scientifique n'échappe pas à ces règles. On pourrait alors voir dans l'objet étudié lui-même, prenons la ville et ses agents, une allégorie du processus de production de connaissance sur l'objet. Comme Romain Duris qui débarque dans l'Auberge Espagnole, et découvre ces rues inconnues que plus tard on aura parcouru cent fois, où on aura vécu mille choses : on débarque dans un monde de concepts, d'approches, de points de vues complémentaires sur des choses qui ne sont pas la même chose. Cette discrépance ontologique est finalement tout aussi présente dans nos représentations de l'espace urbain : Oven Street c'est un des centres de la connaissance pour le membre de Géocités ; c'est le centre de Paris, donc de la France, donc du Monde pour le fier autochtone du 6ème ; c'est le marché Saint-Germain et le shopping de luxe globalisé pour le touriste international ; c'est un morceau d'histoire pour l'élève des Ponts pour qui cela évoque le temps des Saint-pères. Des objets, des concepts, compris et définis par de multiples disciplines et agents producteurs de connaissance : parle-t-on finalement vraiment de la même chose ? Comment tirer parti de cette richesse de points de vue, comment intégrer la complexité permise par cette diversité ? Apporter des éléments de réponse suppose une démarche constructive, générative et autant inclusive que possible. Les choix sont toujours plus éclairés si on a un aperçu d'un maximum d'alternatives. Le trader qui habite son loft en haut des mid-levels et travaille dans son building à deux pas entre deux rails, connaît bien Hong-Kong, mais un seul parmi ses multiples visages, et il lui sera difficilement concevable qu'existe une misère à Kowloon, dont les habitants ne conçoivent pas le Hong-Kong éphémère mais parfois cyclique des travailleurs temporaires du mainland, qui eux ne conçoivent pas les difficultés administratives et financières de migrants de Thaïlande ou d'Inde, l'ensemble étant encore moins concevable pour un étudiant parisien égaré. Mais c'est justement l'égarement qui à dose appropriée sera source d'une connaissance plus large : les fourmis établissent leurs optimisations extrêmement précises à partir d'une marche qu'on peut considérer comme aléatoire. Les algorithmes génétiques, mais encore plus les processus d'évolution biologiques ancrés dans le physique, reposent sur un subtil compromis entre ordre et désordre, entre signal et bruit, entre stabilités et perturbations. Se perdre pour mieux se retrouver fait l'essence et le charme du voyage, qu'il soit physique, conceptuel, social. Finalement, pas de comparaison possible entre une orientation au Caylar ou sur la montagne de Bange à un ennui rectiligne en forêt d'Orléans.

Cet intermède littéraire soulève des problèmes fondamentaux induits par une exigence d'interdisciplinarité et la volonté de construction d'une connaissance complexe intégrative. Dans un premier temps, la réflexivité et la mise en relation d'une perspective prise avec un certain nombre d'autres perspectives existantes est nécessaire pour la pertinence de celle-ci. Il s'agit donc de construire solidement les concepts et spécifier les références empiriques, afin de préciser la problématique et ses objectifs *de manière endogène*. D'autre part, le cadre épistémologique de la démarche se doit d'être précisé. Ci-dessus est finalement imagée une approche *perspectiviste*, qui est une position épistémologique particulière que nous préciserons ici. De plus, le statut des démonstrations est conditionné par la conception des méthodes et des outils, qui est particulière dans le cas des modèles de simulation.

Cette partie répond à ces contraintes, en posant les *fondations* nécessaires à la suite de notre démarche. En terrain relativement mouvant, celles-ci devront dans certains cas être particulièrement profondes pour une stabilité de l'édifice global : ce sera par exemple le cas de l'état de l'art qui mobilisera des techniques d'épistémologie quantitative. Nous rappelons qu'elle s'organise de la manière suivante :

1. Le premier chapitre construit les concepts et objets de manière théorique, et dégage un large éventail d'approches possibles aux interactions entre réseaux de transport et territoires.
2. Le second chapitre développe les différentes approches de modélisation des interactions entre réseaux et territoires. Il établit un état de l'art, structuré par une typologie établie précédemment. Il dresse ensuite le paysage scientifique des disciplines concernées, et cherche les caractéristiques des modèles propres à chaque discipline ainsi que des possible déterminants de celles-ci dans une modélographie.
3. Le troisième chapitre est relativement indépendant et précise nos positions épistémologiques. Il permet notamment de situer la complexité dans laquelle nous cherchons à nous placer, de spécifier ce qui peut être attendu d'une démarche de modélisation et de quelle façon, et de donner une définition plus large du concept de coévolution.

* * *

*

INTERACTIONS ENTRE RÉSEAUX ET TERRITOIRES

Pour mieux visualiser les notions de causalités circulaires dans les systèmes complexes, et pourquoi celles-ci peuvent conduire à des paradoxes en apparence, l'image fournie par DIDEROT dans [diderot1965entretien] est éclairante : “*Si la question de la priorité de l’œuf sur la poule ou de la poule sur l’œuf vous embarrassé, c’est que vous supposez que les animaux ont été originaiement ce qu’ils sont à présent*”. En voulant traiter naïvement des questions similaires induites par notre problématique introduite précédemment, les causalités au sein de systèmes complexes géographiques peuvent être présentées comme un problème “de poule et œuf” : si un effet semble causer l’autre et réciprocement, est-il possible et même pertinent de vouloir isoler les processus correspondants, s’ils font en fait partie d’un système plus large qui évolue à d’autres échelles ? Une vision réductrice, qui consisterait à attribuer des rôles systématiques à l’une composante ou l’autre, s’oppose à l’idée suggérée par DIDEROT qui rejoint celle de *co-évolution*. L’un des enjeux est donc de dresser un aperçu des processus d’interactions entre réseaux et territoires, afin de préciser la définition de la co-évolution, ce qui sera fait à l’issue d’un travail similaire pour les approches par la modélisation, à la fin de la première partie.

Ce chapitre doit être lu comme la construction introduisant nos objets et positions d’étude, et sera complété par une revue de littérature exhaustive sur le sujet précis de la modélisation des interactions, qui fera l’objet du Chapitre 2. Dans une première section 1.1, nous préciserons l’approche prise de l’objet territoire, et dans quelle mesure celui-ci naturellement implique la considération des réseaux de transport pour la compréhension des dynamiques couplées. Cela permet de construire un cadre de lecture définissant les systèmes territoriaux, particulièrement adapté à notre approche par la co-évolution. Ces considérations abstraites seront illustrées par des cas d’étude empiriques dans la deuxième section 1.2, choisis très différents pour comprendre les enjeux d’universalité sous-jacents : la métropole du Grand Paris et le Delta de la rivière des Perles en Chine. Enfin, dans la troisième section 1.3, des éléments d’observation de terrain effectués en Chine préciseront et complexifient la construction de ce cadre théorique et empirique.

★ ★

★

Ce chapitre est entièrement inédit.

1.1 TERRITOIRES ET RÉSEAUX

Nous commençons par une construction plus précise des concepts mobilisés, qui permet de comprendre comment les concepts de territoire et de réseau sont rapidement en interdépendance forte, impliquant une importance ontologique des interactions entre les objets correspondants. Nous verrons que les territoires impliquent l'existence de réseaux, mais que réciproquement ceux-ci les influencent également. Un développement plus particulier sur les propriétés des réseaux de transport permet d'amener progressivement une vision précise de la *co-évolution*, que nous prendrons jusque là dans son sens préliminaire donné précédemment, c'est à dire l'existence de relations causales circulaires entre réseaux de transports et territoires.

1.1.1 Territoires et Réseaux, intimement liés dès leur définition

Territoires : une approche par les systèmes de villes

Le concept¹ de *Territoire*, que nous avons introduit précédemment par ceux de Ville et de Système de Ville, sera central à nos raisonnements et nécessite d'être approfondi et enrichi. En Ecologie Spatiale, un groupe d'agents ou plus généralement un écosystème occupe une certaine étendue spatiale [[tilman1997spatial](#)], qu'on peut identifier comme notion de territoire. Les territoires des sociétés humaines impliquent des dimensions supplémentaires, par exemple par l'importance de leur représentations sémiotiques². Celles-ci jouent un rôle significatif dans l'émergence des constructions sociétales, dont la genèse est profondément liée à celle des systèmes urbains. Selon [[raffestin1988reperes](#)], la *Territorialité Humaine* est "la conjonction d'un processus territorial avec un processus informationnel", ce qui implique que l'occupation physique et l'exploitation de l'espace par les sociétés humaines sont complémentaires des représentations (cognitives et matérielles) de ces processus territoriaux, qui influent en retour sur leur évolution.

En d'autres termes, à partir de l'instant où les constructions sociales déterminent la constitution des établissements humains, les structures sociales abstraites et concrètes joueront un rôle dans l'évolution des territoires, et ces deux objets seront intimement liés. Des exemples de tels liens se retrouvent à travers la propagation d'informations et de représentations, par des processus politiques, ou encore par la correspondance plus ou moins effective entre territoire vécu et territoire perçu. Un territoire est ainsi compris comme une structure

¹ Nous utiliserons le terme *concept* pour des connaissances construites, plutôt que celui de *notion*, qui suivant [[raffestin1978construits](#)] est plus proche d'une information empirique.

² c'est à dire des signes marquants les territoires et leur sens, mais aussi leur représentations, cartographiques par exemple

sociale organisée dans l'espace, qui comprend ses artefacts concrets et abstraits.

Cette approche du territoire rejoint la définition préliminaire que nous en avons prise, et vient alors la renforcer. L'approche de RAFFESTIN insiste sur le rôle des villes comme lieu de pouvoir (au sens d'un lieu rassemblant des processus décisionnel et de contrôle socio-économique) et de création de richesse au travers des échanges et interactions³ (sociaux, économiques). La ville n'a cependant pas d'existence sans son hinterland, ce qu'on interpréter comme le *territoire d'une ville*⁴. Cette correspondance permet de lire l'ensemble des territoires au prisme du système de villes, comme développé par la Théorie Evolutive des Villes [pumain2010theorie]. Celle-ci interprète les villes comme des systèmes complexes auto-organisés, qui agissent comme des médiateurs du changement social : par exemple, les cycles d'innovation s'initialisent au sein des villes et se propagent entre elles (voir C.3 pour une entrée empirique sur la notion d'innovation) : cela permet de comprendre le territoire comme un espace des flux, ce qui permettra d'introduire la notion de réseau comme nous le verrons plus loin. Les villes sont par ailleurs vues comme des agents compétitifs qui co-évoluent [paulus2004coevolution], ce qui permet de préfigurer également l'importance de la co-évolution pour les dynamiques territoriales.

On a ainsi deux approches complémentaires du territoire qui nous permettent de considérer des territoires humains structurés par les systèmes de villes⁵.

³ Une interaction sera comprise dans son sens le plus général, comme une action réciproque de plusieurs entités l'une sur l'autre. Celle-ci peut être physique, informationnelle, transformer les entités, etc. Voir [morin1976methode] pour une construction complète et complexe du concept, en lien intime avec celui d'organisation.

⁴ Même si une correspondance exacte entre territoires et villes n'est probablement qu'une simplification de la réalité, puisque les territoires peuvent s'entremêler à différentes échelles, selon différentes dimensions. Une lecture par lieux centraux de type CHRISTALLER [banos2011christaller] permet de se faire une image conceptuelle de cette correspondance. Des définitions fonctionnelles comme celles des aires urbaines de l'Insee, qui définit l'aire autour d'un pôle dépassant une taille critique (10000 emplois) par les communes dont un seuil minimal d'actifs travaillent dans le pôle (40%) - voir <https://www.insee.fr/fr/metadonnees/definition/c2070>, est une approche possible. La sensibilité des propriétés du système urbain à ces paramètres est testée par [2015arXiv150707878C]. La définition de la ville est alors intimement liée à celle de ses territoires, et celle du système urbain à l'ensemble des territoires.

⁵ Ces visions complémentaires du territoire peuvent également être enrichies par une perspective historique. [di1998espace] procède à une analyse historique des différentes conceptions de l'espace (qui aboutissent entre autres à l'espace vécu, l'espace social et l'espace classique de la géographie) et montre comment leur combinaison forme ce que RAFFESTIN décrit comme territoires. [giraut2008conceptualiser] rappelle les différents usages récents qui ont été faits de la notion de territoire, de la géographie culturelle où il a plus été utilisé par effet de mode, à la géopolitique où c'est un terme bien spécifique lié aux structures de gouvernance, en passant par des utilisations où il sert plus de concept, et dégage l'aspect interdisciplinaire d'un objet capturant une certaine complexité des systèmes étudiés.

Par ailleurs, un aspect central des établissements humains qui a une longue tradition d'étude en géographie, et qui est directement relié au concept de territoire, est celui des *réseaux*. Nous allons préciser leur définition et voir comment le passage de l'un à l'autre est intrinsèque aux approches que nous en prenons.

Définition des réseaux

Un *réseau* doit être compris au sens large d'une mise en relation entre entités d'un système, qui peuvent être vus comme relations abstraites, liens, interactions. [haggett1970network] postule que l'existence d'un réseau est nécessairement liée à celle de flux⁶, et rappelle la représentation topologique sous forme de graphe de tout système géographique dans lequel circulent des flux entre des entités ou des lieux qui sont abstraits sous la forme de noeuds, reliés par des liens. Les liens du réseau disposent alors d'une *capacité*, qui traduit leur capacité à transporter les flux (qui peut également être définie de manière équivalence comme *impédance*). L'analyse topologique révèle déjà un certain nombre de propriétés du système, mais [haggett1970network] précise l'importance de la spatialisation du réseau, incluse dans les propriétés de ses noeuds (localisation) et de ses liens (localisation, impédance), pour la compréhension des dynamiques dans le réseau (flux) ou du réseau lui-même (croissance du réseau). Cette spécificité a été rappelée par [barthelemy2011spatial] qui met en perspective les domaines empiriques concernés par les réseaux spatiaux, certains modèles de croissance de réseau, et certains modèles de processus dans les réseaux : par exemple, les structures topologiques, ou les processus de diffusion seront très contraints par le caractère spatial.

Pour approfondir le concept de réseau en appuyant sur sa forte interdépendance avec celui de Territoire, nous reprenons [dupuy1987vers] qui propose des éléments pour une "théorie territoriale des réseaux" s'inspirant du cas concret d'un réseau de transport urbain. Cette théorie distingue les *réseaux réels*⁷ et les *réseaux virtuels*, eux-mêmes induits entre autre par la configuration territoriale. Les réseaux réels sont la matérialisation de réseaux virtuels. Plus précisément, un territoire est caractérisé par de fortes discontinuités spatio-temporelles induites par la distribution non-uniforme des agents et des ressources. Ces discontinuités induisent naturellement un réseau d'interactions potentielles entre les éléments du système territorial, notamment des agents et des ressources. [dupuy1987vers] désigne ces interactions po-

6 On définit le flux comme un échange matériel (personnes, marchandises, matières premières) ou immatériel (information) entre deux entités.

7 Les réseaux réels contiennent une catégorie qu'on peut désigner comme réseaux concrets, matériels ou physiques - nous utiliserons ces termes de manière interchangeable par la suite, à laquelle les réseaux de transport appartiennent; d'autres catégories comme les réseaux sociaux sont également des réseaux réels sur lesquels nous ne nous attarderons pas.

tentielles comme *projets transactionnels*. Celles-ci induisent la notion de *potentiel d'interaction*, c'est à dire une propriété de l'espace dont les interactions dérivent⁸. Par exemple, de nos jours les actifs ont besoin d'accéder à la ressource qu'est l'emploi, et des échanges économiques s'effectuent entre les différents territoires qui peuvent être plus ou moins spécialisés dans les productions de différents types.

Des réseaux aux réseaux réels

Dans certains cas, un réseau potentiel peut se matérialiser en réseau réel. La question sous-jacente est alors de savoir si le champ de potentiel des territoires est en partie à l'origine de cette matérialisation, si celle-ci est totalement indépendante, ou si la dynamique des deux est fortement couplée, en d'autres termes en co-évolution. La matérialisation résultera généralement de la combinaison de contraintes économiques et géographiques avec des motifs de demande, de manière non-linéaire. Un tel processus est loin d'être immédiat, et conduit à de forts effets de non-stationnarité et de dépendance au chemin⁹ : l'extension d'un réseau existant dépendra de la configuration précédente, et selon les échelles de temps impliquées, la logique et même la nature des opérateurs, c'est à dire des agents participant à sa production, peut avoir évolué.

Les exemples de trajectoires concrètes peuvent être très variées : [kasraian2015development] montre par exemple dans le cas de la Randstad sur le temps long, une première période pendant laquelle le réseau ferré s'est développé pour suivre le développement urbain, tandis que des effets inverses ont été constaté plus récemment. A une échelle urbaine sur le temps long, la dépendance au chemin est montrée pour Boston par [block2012hysteresis] puisque l'environnement bâti et la distribution de la population apparaissent comme fortement dépendants des lignes de tramway antérieures même lorsqu'elles n'existent plus : la façon dont la ligne de transport change l'espace urbain s'opère dans les dynamiques immédiates mais aussi sur le temps long par des effets de renforcement ou à cause de l'inertie du bâti par exemple.

Ainsi, l'existence d'un territoire humain implique nécessairement la présence de réseaux d'interactions abstraites, et les réseaux concrets sont cruciaux pour transporter les individus et les ressources (incluant les réseaux de communication puisque l'information est une ressource essentielle [morin1976methode]), mais les processus d'éta-

⁸ Etant donné tout champ vectoriel de classe C^1 sur \mathbb{R}^3 , le théorème d'HELMOLTZ fournit un potentiel vecteur et un potentiel scalaire dont ce champ dérive par rotationnel et gradient. Cela justifie dans le cas particulier d'un tel point de vue formel le passage d'un champ d'interaction entre agents à un champ de potentiel.

⁹ La non-stationnarité spatiale consiste en la dépendance de la structure de covariance des processus à l'espace, tandis que la dépendance au chemin traduit le fait que les trajectoires prises par le passé influencent fortement les trajectoires actuelles du système.

blissement de ceux-ci sont difficiles à identifier de manière générale. Notre choix ontologique de positionnement dans la théorie de DUPUY, donne une place privilégiée aux relations entre réseaux et territoires, puisqu'il induit dans la construction des objets même une imbrication complexe entre ceux-ci.

Le statut du réseau par rapport au territoire est d'autre part fortement conditionné par le contexte socio-économique et technologique. Selon DURANTON [**duranton1999distance**], un facteur influençant la forme des villes pré-industrielles était la performance des réseaux de transport. Les progrès technologiques, conduisant à une baisse des coûts de transport, ont induit un changement de régime, ce qui a mené à une prépondérance du marché foncier dans la formation des villes (et par conséquent un rôle des réseaux de transport qui déterminent les prix par l'accessibilité), et plus récemment à une importance croissante des réseaux de télécommunication ce qui a induit une "tyrannie de la proximité" puisque la présence physique n'est pas remplacable par une communication virtuelle [**duranton1999distance**].

Cette approche territoriale des réseaux semble naturelle en géographie, puisque les réseaux sont étudiés conjointement avec des objets géographiques qu'ils connectent, en opposition aux travaux théoriques sur les réseaux complexes qui les étudient de manière relativement déconnectée de leur fond thématique [**ducruet2014spatial**].

Des réseaux qui façonnent les territoires ?

Cependant les réseaux ne sont pas seulement une manifestation matérielle de processus territoriaux, mais jouent également leur rôle dans ces processus comme leur évolution peut influencer l'évolution des territoires en retour. Il émerge alors une difficulté intrinsèque : il n'est pas évident d'attribuer des mutations territoriales à une évolution du réseau et réciproquement la matérialisation d'un réseau à des dynamiques territoriales précises, et différents facteurs exogènes rentrent par ailleurs en compte, comme le prix de l'énergie ou les technologies existantes dans le cas de l'effet du réseau sur les territoires par exemple. Dans le cas des *réseaux techniques*, une autre désignation des réseaux concrets donnée dans [**offner1996reseaux**], de nombreux exemples de tels retroactions peuvent être mis en évidence : une accessibilité accrue peut être un facteur favorisant la croissance urbaine, ou bien l'interconnexion de différents réseaux de transport permet une extension significative de la portée des déplacements. A une plus petite échelle, des changements de l'accessibilité peuvent induire des relocalisations de différentes composantes urbaines. Ces retroactions des réseaux sur les territoires n'agissent pas nécessairement sur des composantes concretes : CLAVAL montre dans [**claval1987reseaux**] que les réseaux de transport et de communication contribuent à la représentation collective d'un territoire en

agissant sur un sentiment d'appartenance, qui peut alors jouer un rôle crucial dans l'émergence d'une dynamique régionale fortement cohérente. Développons d'abord plus en détail les possibles influences des réseaux sur les territoires.

La confusion autour de possibles relations causales simples a nourri un débat scientifique encore actif aujourd'hui. La question sous-jacente repose sur des attributions plus ou moins déterministes d'impact d'infrastructures ou d'un nouveau mode de transport sur des transformations territoriales. On peut trouver des précurseurs de ce raisonnement dès les années 1920 : MCKENZIE, de l'école de Chicago, parle dans [burgess1925city] des " modifications des formes du transport et de la communication comme facteurs déterminants des cycles de croissance et de déclin [des territoires]" (p. 69). Des méthodologies pour identifier ce qui est alors nommé *effets structurants* des réseaux de transport ont été développées pour la planification dans les années 1970 : [bonnafous1974methodologies] situe le concept d'effet structurant dans le cadre d'une logique d'utilisation de l'offre de transport comme outil d'aménagement (les alternatives étant le développement d'une offre pour répondre à une congestion du réseau, et le développement simultané d'une offre et d'un aménagement associé). Ces auteurs identifient du point de vue empirique des effets directs d'une nouvelle offre sur le comportement des agents, sur les flux de transport et des possibles inflexions sur les trajectoires socio-économiques des territoires concernés. [bonnafous1974detection] développe une méthode pour identifier de tels effets par modifications de la classe des communes dans une typologie établie *a posteriori*. Plus récemment, [bonnafous2014observatoires] a proposé la mise en place *d'observatoires permanents* des territoires pour rendre plus robustes ce type d'analyse, en permettant un suivi continu de l'évolution des territoires les plus concernés par l'emprise d'une nouvelle infrastructure.

Selon [offner1993effets] qui reprend des idées déjà évoquées par [francois1977autorou] par exemple, il s'est par la suite développé un usage non raisonné et hors contexte de ces méthodes par les planificateurs et les politiques qui les mobilisaient généralement pour justifier des projets de transports de manière technocratique : justifiant d'un effet direct d'une nouvelle infrastructure sur le développement local (par exemple économique), les élus sont en mesure de demander des financements et de légitimer leur action auprès des contribuables. [offner1993effets] insiste sur la nécessité d'un positionnement critique sur ces enjeux, rappelant qu'il n'existe pas de démonstration scientifique d'un effet qui serait systématique. Une édition spéciale de l'Espace Géographique sur ce débat [espacegeo2014effets] a rappelé d'une part que de telles croyances étaient encore largement présentes aujourd'hui dans les milieux opérationnels de la planification, ce qui peut s'expliquer par exemple par le besoin de justifier l'action publique, et d'autre

part qu'une compréhension scientifique des relations entre réseaux et territoires est encore en pleine construction.

Une illustration concrète d'actualité permet de se faire une image de cette instrumentalisation : les débats en juillet 2017 relatifs à l'ouverture des LGV Bretagne et Sud-Ouest ont montré toute l'ambiguïté des positions, des conceptions, des imaginaires à la fois des politiques mais aussi du public : inquiétude quant à la spéculation sur l'immobilier dans les quartiers de gare, questionnements des pratiques de mobilité quotidienne mais aussi sociale¹⁰. La complexité et la portée des sujets montrent bien la difficulté d'une compréhension systématique d'effets du transport sur les territoires.

Une vision intégrative : les Systèmes Territoriaux

Cet aperçu introductif, des territoires aux réseaux, nous permet ainsi de clarifier notre approche des systèmes territoriaux qui sera sous-jacente dans l'ensemble de la suite. Une prise en compte des diverses rétroactions potentielles des réseaux pour la compréhension des territoires est suggérée par un retour à la citation de Diderot ayant introduit le sujet devrait aider à ce point, au sens où il ne faut pas considérer le réseau ni les territoires comme des systèmes indépendants qui s'influencerait soit l'une soit l'autre par des relations causales en sens unique, mais comme des composantes fortement couplées d'un système plus large, et donc étant en relations causales circulaires. Selon les composantes ainsi que l'échelle considérées, différentes manifestations de celles-ci pourront être observables, et il existera des cas où il y apparemment influence de l'une sur l'autre, d'autres où les influences sont simultanées, ou encore d'autres ou aucune relation n'est observable de manière significative.

Comme nous avons mis en exergue le rôle des réseaux dans de nombreux aspects des dynamiques territoriales, nous proposons une définition des systèmes territoriaux les incluant explicitement. Nous considérons un *Système Territorial* comme un *territoire humain qui contient à la fois des réseaux d'interactions et des réseaux réels*. Les réseaux réels, et plus particulièrement les réseaux concrets¹¹, sont une composante à part entière du système, jouant dans les processus d'évolution, au travers de multiples retroactions avec les autres composantes à plusieurs échelles spatiales et temporelles.

¹⁰ Voir par exemple http://www.liberation.fr/futurs/2017/07/02/immobilier-plus-de-parisiens-comment-les-bordelais-voient-l-arrivee-de-la-lgv_1580776, ou 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 pour une réaction "à chaud" de divers acteurs locaux, témoignant d'un impact au minimum sur les représentations. Par exemple, les Bordelais semblent craindre l'arrivée de Parisiens en recherche d'un logement moins cher et de meilleures conditions de vie, ce qui pourrait augmenter les prix au moins aux environs de la gare.

¹¹ Qui comme nous l'avons vu précédemment sont des réseaux réels matérialisés.

Le réseau n'est pas nécessairement une composante en tant que telle du territoire, mais bien du *Système Territorial* en notre sens¹². Cette vision rejoint le positionnement de [dupuy1985systemes] qui introduit le territoire comme "produit d'une dialectique" entre composantes territoriales et réseaux. Notons le raccourci sémantique pour désigner les composantes du système territorial qui ne sont pas les réseaux et qui interagissent avec celui-ci, par le terme de territoire. Celles-ci dépendent des ontologies et des échelles considérées, comme nous le verrons par la suite, et peuvent aller des agents microscopiques aux villes elle-mêmes. Comme nous le verrons aussi par la suite (voir 2.1), il existe des paradigmes où ce raccourci n'est pas fait, comme dans le cas particulier des interactions entre transport et usage du sol ou les entités sont spécifiques. Mais il est fait si on reste à un cadre plus général, comme en témoigne l'un des ouvrages de référence sur le sujet [offner1996reseaux]¹³. Nous assumerons également ce raccourci de langage, en désignant par *interactions entre réseaux et territoires* ou *co-évolution entre réseaux et territoires*, les interactions ou la co-évolution entre les réseaux physiques et les composantes qu'ils relient, au sein d'un système territorial et donc d'un territoire.

1.1.2 *Les réseaux de transport, catalyseurs privilégiés des interactions*

Nous précisons à présent le cas particulier des réseaux de transport et développons des concepts spécifiques associés qui joueront un rôle prépondérant dans la précision de notre problématique.

Caractéristiques et spécificités des réseaux de transport

Centraux aux discussions déjà évoquées sur les effets structurants des réseaux, les réseaux de transports jouent un rôle significatif dans l'évolution des territoires, mais il n'est évidemment pas question de

¹² Ce choix ontologique n'est pas anodin et appuie la dialectique entre réseaux et territoires. Partant de l'époque lointaine où les réseaux physiques n'existaient pas, l'émergence d'un territoire humain, que nous supposons équivalent à un réseau d'interactions, induit la mise en place de la dialectique diachronique complexe entre réseaux physiques et territoires humains. On peut ainsi lire la genèse du système territorial comme une boucle morinienne [morin1976methode], dans laquelle on entre par le territoire initial puis qui se boucle du réseau physique aux composantes territoriales pour former le système territorial (donc le territoire dans la majorité des cas) de la manière récursive suivante :

Territoire initial → Territoire = Configuration territoriale → Réseau physique



¹³ Lorsque [amar1985essai] propose un modèle conceptuel de morphogenèse des réseaux, il désigne les composantes territoriales par "Le Monde", ce qui n'apporte pas de solution au problème sémantique. Le parti pris de garder le territoire, au sein du territoire, suggère une récursivité, et donc une complexité dans la générativité du système [morin1976methode]. La mobilisation du concept de morphogenèse à partir du Chapitre 5 suggère que cette récursivité serait plus que fortuite, mais bien intrinsèque au problème.

leur attribuer des effets causaux déterministes. On parlera de manière générale de réseau de transport pour désigner l'entité fonctionnelle permettant un déplacement des agents et des ressources au sein et entre les territoires¹⁴. Même si d'autres types de réseaux sont également fortement impliqués dans l'évolution des systèmes territoriaux (voir par exemple les débats sur l'impact des réseaux de communication sur la localisation des activités économiques), les réseaux de transport conditionnent d'autres types de réseaux (logistique, échanges commerciaux, interactions sociales concrètes pour donner quelques exemples) et sont une entrée privilégiée en rapport aux motifs d'évolution territoriale, en particulier dans nos sociétés contemporaines pour lesquelles les réseaux de transport jouent un rôle privilégié [bavoux2005geographie]. Nous nous concentrerons ainsi par la suite uniquement sur les réseaux de transport.

Le développement du réseau français à grande vitesse est une illustration du rôle des réseaux de transport sur les politiques de développement territorial. Présenté comme une nouvelle ère de transport sur rail, il s'agit d'une planification au niveau de l'Etat de lignes totalement nouvelles et relativement indépendantes de par leur vitesse deux fois plus élevée, selon la lecture de [zem bri 1997fondements]. La grande vitesse a été défendue par les acteurs politiques entre autres comme central pour le développement. L'articulation faible de ces nouveaux réseaux avec le réseau classique et avec les territoires locaux est à présent observé comme une faiblesse structurelle [zem bri 1997fondements] (c'est à dire conséquence de la structure du réseau tel qu'il a été planifié dans le Schéma Directeur de 1990), et des impacts négatifs sur certains territoires, comme par la suppression de dessertes intermédiaires sur les lignes classiques empruntées par le TGV, qui contribue à un accroissement de l'effet tunnel¹⁵ ont été montrés [zem bri 2008contribution]. Une revue faite dans [bazin2011grande] confirme qu'aucune conclusion générale sur des effets locaux d'une connection à une ligne à grande vitesse ne peut être tirée, bien que ce sésame garde une place conséquente dans les imaginaires des élus¹⁶. Le développement des différentes Lignes à Grande Vitesse s'inscrit dans des contextes territoriaux très différents, et il est dans tous les cas délicat d'interpréter des processus en les sortant de leur contexte : par exemple, les lignes LGV Nord et LGV Est s'inscrivent dans des échelles européennes plus

¹⁴ On désigne ainsi à la fois l'infrastructure, mais aussi ses conditions d'exploitation, le matériel roulant, les agents exploitants.

¹⁵ L'effet tunnel désigne le processus de télescopage du territoire traversé par une infrastructure, celle-ci n'étant utilisable à partir de celui-ci.

¹⁶ Mais des conclusions particulières existent dans certains cas : par exemple un effet positif de la LGV Sud-Est sur la fréquentation touristique de villes moyennes intermédiaires comme Montbard ou Beaune [bonnafous1987regional]; ou le positionnement de Lille comme métropole européenne dans lequel les connexions LGV ont joué [giblin2004lille].

vastes que la LGV Bretagne ouverte en juillet 2017¹⁷. Les effets de l'ouverture d'une ligne peuvent s'étendre au delà des seuls territoires directement concernés : [l2014contribution] montre par l'utilisation d'indicateurs issus de la *Time Geography*¹⁸ (mesurant une quantité de temps de travail disponible dans le cadre d'un aller-retour journalier) que la ligne Tours-Bordeaux a des répercussions potentielles dans le Nord et l'Est de la France. Ces exemples illustrent la manière dont les réseaux de transport peuvent avoir des effets à la fois directs et indirects, positifs ou négatifs, et à différentes échelles, ou bien aucun effet sur les dynamiques territoriales.

Des processus dépendant des échelles

La question des échelles temporelles et spatiale concernées a été jusqu'ici abordée de manière auxiliaire aux concepts introduits. Nous proposons à présent de les intégrer de manière structurelle à notre raisonnement, c'est à dire guidant le développements de nouveaux concepts. Ainsi, les concepts de *Mobilité*, d'*Accessibilité*¹⁹, puis de *Dynamique structurelle sur le temps long*, correspondent chacun à des échelles de temps et d'espace décroissantes : intra-urbain et journalier, métropolitain et décennal, régional (au sens large et flexible de la portée d'un système de villes) et centennal. La correspondance que nous postulons ici entre échelles de temps et échelles d'espace, loin d'être évidente, sera montrée lors du développement de chacun de ces concepts. Par contre, la prise en compte d'échelles multiples est importante, comme le montre [RIETVELD1994329] par une revue des approches économiques des interactions, qui appuie la différence entre l'intra-urbain et l'intra-régional : à grande échelle, différentes méthodes (modèles ou approches qualitatives) donnent des résultats très différents quant à l'impact du stock d'infrastructure, tandis qu'à petite échelle, l'impact positif du stock global sur la productivité est a priori non discutable.

Transports et Mobilité

La notion de mobilité et l'ensemble des approches associées, capturent en partie nos questionnements à grande échelle. Nous définissons la mobilité de manière générale comme un déplacement d'agents

¹⁷ La ligne LGV Nord relie Paris à Lille puis Calais (ouverte entièrement en 1997), et s'inscrit dans la liaison avec Londres, Bruxelles et Amsterdam. La LGV Est relie Paris à Strasbourg (ouverte partiellement en 2007, puis entièrement en 2016) et permet de desservir le Luxembourg et l'Allemagne. La LGV Bretagne, ouverte en 2017, est le tronçon de la LGV Ouest vers Rennes et sa desserte est uniquement bretonne [zembris2010new]

¹⁸ La *Time Geography*, introduite par le géographe suédois T. HÄGERSTRAND, s'intéresse majoritairement aux trajectoires des individus dans le temps et l'espace, et de leurs implications dans les interactions avec l'environnement [chardonnel2007time].

¹⁹ L'accessibilité, comme nous le verrons, se définit à plusieurs échelles, mais nous privilierons ce terme pour les paysages d'accessibilité à l'échelle métropolitaine.

territoriaux dans l'espace et le temps. Elle relève des motifs d'utilisation des réseaux de transport. [hall2005reconsidering] introduit un cadre théorique permettant une typologie des pratiques de mobilité. En particulier, il montre une décroissance rapide de la fréquence des déplacements avec la portée spatiale et la durée, et donc que les motifs "micro-micro" (pour échelle temporelle journalière et échelle spatiale intra-urbaine), qu'on désigne par *mobilité quotidienne*, sont majoritaires. Cela ne signifie pas pour autant une absence de lien avec d'autres échelles : d'une part les motifs de mobilité sont très fortement conditionnés par la distribution des activités comme l'illustre [lee2015relating], mais également corrélés à la structure sociale [camarero2008exploring], qui évoluent tous deux à des échelles de temps d'un ordre différent (supérieur à la dizaine d'année, donc au moins un ordre de grandeur de différence). Ainsi, infrastructure et superstructure déterminent pratiques de mobilité, donnant un rôle important aux réseaux de transports dans celles-ci.

Réciproquement, les motifs d'utilisation des réseaux de transport sont le produit des dynamiques de mobilité quotidiennes, et ceux-ci s'y adaptent, tout en induisant des relocalisations des actifs et emplois : il existe une co-évolution entre transports et composantes territoriales aux échelles microscopiques et mesoscopiques, qui sont un objet d'étude à part entière. Par exemple, [fusco2004mobilite] révèle une influence²⁰ de la mobilité sur la structure urbaine, l'offre d'infrastructure et ses propriétés ayant cependant des effets simultanément sur la mobilité et sur la structure urbaine. Dans le cas des réseaux autoroutiers, [faivr2003] rappelle la nécessité de construire un cadre d'analyse dépassant la logique des effets structurants sur le temps long, et montre également des interactions à petite échelle propres à la mobilité sur lesquelles des conclusions plus systématiques peuvent être établies, comme une évolution des pratiques de mobilité impliquant une utilisation différente du réseau de transport. Nous avons donc à grande échelle une première interdépendance forte entre réseaux de transports et territoires, une première échelle de co-évolution.

Enfin, il est important de garder à l'esprit la forte contingence des concepts mobilisés ici. La co-construction du concept de mobilité et des solutions techniques modélisant celle-ci dans un but opérationnel, a été illustrée par [commenges:tel-00923682] pour le contexte français, qui révèle entre autre une application peu adaptée au contexte français de cadres et méthodes importés des Etats-Unis. Cette contingence signifie que le choix des concepts même dépend de déterminants plus larges que leur utilité directe, et suggère une inscription systémique globale dans le *Système Territorial*.

²⁰ Qui est interprétée comme causale au sens des réseaux Bayesiens.

Transports et Accessibilité

Le concept d'*Accessibilité* est fondamental pour notre question, puisqu'il se positionne à la croisée même des réseaux et des territoires. Basée sur la possibilité d'accéder un lieu par un réseau de transport (pouvant prendre en compte la vitesse, la difficulté de se déplacer), elle est généralement définie comme un potentiel d'interaction spatiale²¹ [**bavoux2005geographie**]. Elle a été introduite sous cette forme initialement par [**hansen1959accessibility**], dans un but d'application à la planification. Diverses formulations et formalisations d'indicateurs correspondants ont été proposées. Il a été montré que celles-ci rentrent dans le même cadre théorique. En effet, [**weibull1976axiomatic**] développe une approche axiomatique de l'accessibilité, c'est à dire proposant de la caractériser à partir d'un nombre minimal d'hypothèses fondamentales (les axiomes). [**miller1999measuring**] reprend ce cadre et montre qu'il englobe trois façons classiques de comprendre l'accessibilité. Celles-ci sont respectivement celle basée sur la *Time Geography* et les contraintes, celle sur les mesures d'utilité pour l'utilisateur, et celle sur un temps de trajet moyen. Les mesures correspondantes sont dérivées dans un cadre mathématique unifié, ce qui permet un lien à la fois théorique et opérationnel entre des approches du concept a priori différentes.

On peut voir dans un premier temps dans quelle mesure des motifs d'accessibilité induisent une évolution du réseau. Ce concept est souvent utilisé comme un outil de planification ou comme une variable explicative de localisation des agents par exemple, puisqu'il s'agit par exemple d'un bon indicateur pour la quantité de personnes affectées par un projet de transport.

Les débats récents sur la planification du *Grand Paris Express* [**mangin2013paris**], cette nouvelle infrastructure de transport métropolitaine planifiée pour les vingt prochaines années, a révélé l'opposition entre une vision de l'accessibilité comme nécessaire pour désenclaver des territoires désavantagés, et une vision de l'accessibilité comme moteur du développement économique pour des zones déjà dynamiques, les deux n'étant pas forcément compatibles car correspondent à des corridors de transport différents. L'un était initialement porté par l'Etat dans la perspective des pôles de compétitivité, l'autre par la région dans une perspective d'équité territoriale. Ces deux logiques répondent bien sûr à des objectifs différents à plusieurs niveaux, et la solution choisie doit former un compromis. Nous reviendrons sur cet exemple précis du Grand Paris en détails par la suite.

Cet exemple permet de suggérer un effet des motifs de potentiels sur l'évolution du réseau : même si celui-ci passe par des structures sociales complexes (nous y reviendrons aussi en détail plus loin),

²¹ et souvent généralisée comme une *accessibilité fonctionnelle*, par exemple les emplois accessibles aux actifs d'un lieu. Les potentiels d'interaction spatiale s'exprimant dans les lois gravitaires peuvent aussi être compris de cette façon.

il existe de nombreuses situations où une croissance du réseau de transport (qui peut se manifester par une évolution topologique, c'est à dire l'ajout d'un lien, mais aussi une évolution des capacités des liens) est directement ou indirectement induite par une distribution d'accessibilité [zhang2007economics]. Ce phénomène peut concerner des modifications fondamentales du réseau comme des modifications mineures : [rouleau1985villages] étudie l'évolution sur le temps long (de 1800 à 1980) des villages satellites à Paris qui ont été progressivement intégrés à son tissu urbain et montre à la fois une persistance de la trame viaire et parcellaire, mais aussi des évolutions locales répondant à des logiques de connectivité par exemple, tout en s'inscrivant dans un cadre d'évolution globale plus complexe (comme dans le cas d'Haussmann). Nous désignerons ce processus abstrait de réponse du réseau à une demande de connectivité par *rupture de potentiel*²².

Un autre processus significatif est l'impact d'une évolution de l'accessibilité par relocalisations sur les motifs d'utilisation du réseau, et particulièrement la congestion, induisant une modification de la capacité (flux pouvant être porté par les liens du réseau) : ce phénomène est montré dans le cas de Beijing par [yang2006transportation], qui révèle des modifications d'impédance (vitesse effective dans le réseau routier) allant jusqu'à 30%. Il peut être mis en correspondance avec les processus liés à la mobilité, même si on se situe ici plutôt dans des échelles meso-meso, c'est à dire une évolution du réseau et des relocalisations sur des temporalités de l'ordre de la dizaine d'année (le réseau étant plus lent, de l'ordre de la vingtaine d'années), et sur des échelles spatiales métropolitaines²³.

Réciproquement, une évolution du réseau implique une reconfiguration immédiate de la distribution spatiale des accessibilités (au sens de l'ensemble des approches existantes, puisque toutes mobilisent le réseau), et aussi potentiellement des transformations territoriales sur une plus longue durée : on rejoint finalement le débat des effets structurants que nous avons déjà commenté. On a déjà vu que l'accessibilité co-évolue²⁴ avec les pratiques de mobilité, ce qui suppose un effet à cette échelle. Concernant les relocalisations et la distribution des populations, il existe des cas où il est en effet possible d'attribuer à la

²² En analogie avec le phénomène de *dielectric breakdown*, ou décharge partielle, qui correspond au passage du courant dans un isolant quand la différence de potentiel électrique est trop grande.

²³ qui correspondent à des étendues spatiales de 100 à 200km, mais à diverse réalités urbaines. Une métropole sera une ville d'importance dans un système de villes à grande échelle, et sera vue avec son territoire fonctionnel (par exemple Paris et une grande partie de l'Ile-de-France). L'émergence de nouvelles formes métropolitaines, comme les *Mega-city-regions* qui sont composés de métropoles de taille comparable, sur une faible étendue spatiale, et en très forte interaction, complique cette question de l'échelle. Nous reviendrons sur ces objets en 1.2.

²⁴ Le concept s'applique a priori à diverses échelles, ce qui sera confirmé par la définition plus précise que nous prendrons à la fin de cette première partie.

croissance du réseau des dynamiques des territoires, que nous allons développer par la suite.

[duranton2012urban] montrent ainsi à une échelle de temps moyenne de 20 ans pour les Etats-unis, par l'utilisation de variables instrumentales²⁵, que la croissance de l'accessibilité dans une ville cause une croissance de l'emploi. Sur une échelle temporelle similaire, mais à l'échelle spatiale du pays pour la Suède, [johansson1993infrastructure] montre que l'accessibilité locale ("intra-régionale") et globale ("inter-régionale") explique la croissance de la production et la productivité des entreprises. [doi:10.1080/01441647.2016.1168887] procède à une revue systématique des études empiriques des impacts à moyen terme des infrastructures de transport, et montre qu'une densification urbaine à proximité des nouvelles infrastructures est très probable, celle-ci étant résidentielle dans le cas d'une infrastructure ferroviaire et pour les emplois et l'activité industrielle et commerciale dans le cas d'une infrastructure routière²⁶. De même, on peut montrer des effets forts de la présence d'infrastructures pour des types particuliers d'usage du sol : [nilsson2016measuring] l'illustre par exemple pour les fast food dans deux villes aux Etats-Unis, en montrant statistiquement que l'accès à une infrastructure importante induit une agrégation spatiale des commerces.

Ces derniers exemples suggèrent l'existence potentielle d'effets de l'accessibilité, et donc du réseau, sur les dynamiques territoriales. Dans certain cas, les effets structurants sont ainsi présents. Mais ceux-ci sont toujours liés au contexte précis ainsi qu'aux échelles. Cela nous permet de faire la transition vers les concepts liés aux dynamiques des systèmes urbains sur le temps long.

Transports et Systèmes Urbains

La troisième entrée conceptuelle sur les interactions entre réseaux et territoires, et qui sera particulièrement liée à l'idée de co-évolution, est celle par les systèmes urbains, à petite échelle spatiale et sur le temps long. Nous désignerons le concept par *Dynamique structurelle du système urbain*.

La Théorie Evolutive des Villes considère les systèmes de villes comme des systèmes de systèmes à de multiples échelles, du niveau microscopique intra-urbain, au niveau macroscopique du système entier, par le niveau mesoscopique de la ville [pumain2008socio]. Ces

²⁵ La méthode des variables instrumentales permet de dégager des relations causales entre une variable explicative et une variable expliquée. Le choix d'une troisième variable, appelée variable instrumentale, soit être fait tel que celle-ci n'influence que la variable explicative mais pas la variable expliquée, en quelque sorte un choc exogène.

²⁶ Les études revues couvrent majoritairement la seconde moitié du 20ème siècle et l'Europe, les Etats-Unis et l'Asie de l'Est. Il est donc important de garder à l'esprit que même relativement générales, les conclusions doivent toujours être contextualisées.

systèmes sont complexes, dynamiques, et adaptatifs : leur composants *co-évoluent* et le système répond à des perturbations intérieures ou extérieures par des modifications de sa structure et de sa dynamique. Nous développerons longuement les multiples implications de cette approche tout au long de notre travail, et retenons ici les processus d'interactions entre villes. Ces interactions consistent en des échanges informationnels ou matériels, et la diffusion de l'innovation en est une composante cruciale [**pumain2010theorie**]. Elles sont nécessairement portées par les réseaux physiques, et plus particulièrement les réseaux de transport. On s'attend ainsi du point de vue théorique à une interdépendance forte entre villes et réseaux de transport à ces échelles, c'est à dire à une coévolution.

Du point de vue empirique, celle-ci a déjà été mise en valeur : [**bretagnolle:tel-00459720**] souligne une corrélation croissante dans le temps entre la hiérarchie urbaine et la hiérarchie de l'accessibilité temporelle pour le réseau ferroviaire français (a priori plus claire pour cette mesure que pour les mesures intégrées d'accessibilité soumises à l'auto-corrélation comme nous le verrons en 4.2). Celle-ci est un marqueur de rétroactions positives entre le rang urbain et la centralité de réseau. Différents régimes dans le temps et l'espace ont été identifiés : pour l'évolution du réseau ferroviaire français, une première phase d'adaptation du réseau à la configuration urbaine existante a été suivie par une phase de coévolution, au sens où les relations causales sont devenues difficiles à identifier. L'impact de la contraction de l'espace-temps par les réseaux sur le potentiel de croissance des villes avait déjà été montré pour l'Europe par des analyses exploratoires dans [**bretagnolle1998space**].

Les résultats de modélisation par [**bretagnolle2010comparer**], et plus particulièrement les paramétrisations différentes du modèle Simpop2²⁷, montrent que l'évolution du réseau ferroviaire aux Etats-unis a suivi une dynamique bien différente, sans diffusion hiérarchique, donnant forme localement à la croissance urbaine dans certains cas. Ce contexte particulier de conquête d'un espace vierge d'infrastructures implique un régime spécifique pour le système territorial. D'autres contextes révèlent des impacts différents du réseau à court et long terme : [**berger2017locomotives**] étudient l'impact de l'établissement du réseau ferroviaire suédois sur la croissance des populations urbaines, de 1800 à 2010, et trouvent un effet causal immédiat de la croissance de l'accessibilité sur la croissance de la population, suivi sur le temps long d'une forte inertie de la hiérarchie des populations. Dans chaque cas, on a bien existence de *dynamiques structurelles* sur le temps long, qui correspondent aux dynamiques lentes de la structure

²⁷ La structure générique du modèle Simpop2 est la suivante [**pumain2008socio**] : les villes sont caractérisées par leur population et leur richesse ; produisent des biens selon leur profil économique ; les interactions entre villes produisent des échanges, déterminés par les fonctions d'offre et demande ; les populations évoluent selon la richesse après échanges.

du système urbain, et témoignent en ce sens d'*effets structurants sur le temps long* comme le souligne [pumain2014effets].

Il s'agit bien de différencier ces derniers des effets structurants sujets des débats mentionnés précédemment. Au niveau du système urbain, il est pertinent de suivre globalement des trajectoires qui étaient possibles, et localement l'effet a nécessairement un aspect probabiliste. D'autre part, il faut mettre l'accent sur le rôle de la dépendance au chemin pour les trajectoires des systèmes urbains : par exemple la présence en France d'un système préalable de villes et de réseau (routes postales) a fortement influencé le développement du réseau ferré, ou comme [berger2017locomotives] l'a montré pour la Suède. De même, [doi:10.1068/b39089] souligne l'importance des événements historiques dans les dynamiques couplées du réseau routier et des territoires, choc historiques pouvant être vus comme exogènes et induisant des bifurcations du système qui accentuent l'effet de la dépendance au chemin. Ainsi, pour ces dynamiques de structure sur le temps long, des prévisions ne sont guère envisageables.

Cette troisième approche nous a permis de dégager un point de vue complémentaire de la co-évolution, à une autre échelle.

Des liens entre échelles suggérés par les Lois d'Échelle

Notre grille de lecture par échelles progressives, qui permet de dégager une assez bonne correspondance entre échelle spatiale et temporelle, ainsi que d'y associer les concepts adaptés, ne capture bien sûr pas l'ensemble des processus possibles : ceux qui seraient fondamentalement multi-échelles, par exemple en impliquant l'émergence de leur propre niveau intermédiaire, ne sont pas évoqués. Ceux-ci sont importants et nous y reviendrons ci-dessous. Dans un premier temps, nous proposons d'effectuer un lien conceptuel entre les échelles par l'intermédiaire des *lois d'échelle* (que nous comprenons au sens général donné en introduction). Ce lien permet en particulier de dépasser une lecture réductrice par cloisonnement d'échelle.

Les réseaux de transport sont par essence hiérarchiques, cette propriété dépendant des échelles dans lesquelles ils sont intégrés, et se manifestant par l'émergence de lois d'échelle pour leurs propriétés. Par exemple, [10.1371/journal.pone.0102007] montrent empiriquement des propriétés de loi d'échelle pour un nombre conséquent d'aires métropolitaines à travers la planète. Or les lois d'échelle révèlent la présence de hiérarchies dans un système, comme pour la hiérarchie de tailles dans les systèmes de villes exprimée par la loi de Zipf [nitsch2005zipf] ou d'autres lois d'échelles urbaines [2013arXiv1301.1674A ; 2015arXiv151000902B], ce qui suggère une structure particulière pour ces systèmes. On peut s'attendre à la retrouver dans les processus d'interaction eux-mêmes. La topologie du réseau de transport suit de telles lois pour la distribution de ses mesures locales comme la centralité [samaniego2008cities], celles-ci étant directement liées au

Echelle	Echelle spatiale	Echelle temporelle	Concept	Référence
Micro	Intra-urbaine (10km)	Journalière (1j)	Mobilité	[hall2005reconsidering]
Meso	Métropolitaine (100km)	Décade (10ans)	Accessibilité	[wegeren2004land]
Macro	Régionale (500km)	Siècle (100ans)	Dynamique structurelle	[pumain1997pour]

motifs d'accessibilité à différentes échelles. De plus, la topologie du réseau fait partie des facteurs induisant la hiérarchie d'usage, se retrouvant dans les externalités négatives de congestion, en relation avec la distribution spatiale de l'usage du sol [Tsekeris2013]. Ainsi, la considération des lois d'échelles pour les réseaux de transport, et plus généralement pour les systèmes territoriaux, est dans un premier temps une signature de la complexité de ces systèmes, et permet dans un second temps un lien implicite entre les échelles.

Echelles : synthèse

Pour rappeler notre cadre de lecture par échelles, nous proposons le tableau suivant :

Les appellations ainsi que les ordres de grandeur des échelles temporelles et spatiales sont évidemment indicatifs, de même que les concepts clés qui sont en fait ceux qui nous ont permis d'entrer dans ces échelles. Nous donnons également des références illustrant des cadres conceptuels correspondant. Ce tableau nous sera toutefois utile pour garder à l'esprit les échelles typiques auxquelles nous ferons référence.

Processus : synthèse

A ce stade, nous pouvons d'ores et déjà proposer une synthèse préliminaire des processus d'interaction que nous avons introduit. Une typologie plus exhaustive sera possible à l'issue du chapitre.

Ainsi, des composantes territoriales peuvent agir sur les réseaux de transport par :

- Impact des motifs de mobilité sur les impédances et les capacités
- Rupture de potentiel, émergence de centralités
- Sélection hiérarchique de l'accessibilité
- Effets systémiques structurels et bifurcations

Réciproquement, des processus où les propriétés des réseaux agissent sur les territoires incluent :

- Relocalisations induites par des contraintes de mobilité

- Changement d'usage du sol du à une infrastructure de transport
- Motifs d'accessibilité induits par les réseaux, pouvant induire des relocalisations
- Interactions entre territoires portées par les réseaux, incluant l'effet tunnel lorsque celles-ci sont télescopées

Ces différents processus n'ont pas tous le même statut d'abstraction ni les mêmes échelles. Nous avons de plus volontairement occulté des processus déjà évoqués, au sein desquels le couplage est plus fort et pour lesquels la circularité est déjà présente dans l'ontologie, comme les processus liés à la planification. Nous allons détailler à présent ceux-ci, ce qui nous permettra par la suite de raffiner la liste ci-dessus et de la présenter sous forme de typologie après l'avoir enrichie par des études empiriques.

1.1.3 Des interactions à la co-évolution

A ce stade, nous avons identifié des processus d'interaction entre réseaux de transport et territoires jouant un rôle significatif dans la complexité des systèmes territoriaux. Dans le cadre de l'approche d'un système territorial par la définition donnée lors de la construction première des concepts, cette question peut être reformulée comme l'étude de systèmes territoriaux réticulaires, avec une emphase sur le rôle des systèmes de transports. On a vu que l'étendue des échelles spatiales et temporelles va de celle de la mobilité quotidienne (micro-micro) à des processus sur le temps long dans les systèmes de villes (macro-macro), avec la possibilité de combinaisons intermédiaires. La précision des échelles particulièrement pertinentes fera l'objet de la majorité des préliminaires (Partie 1) et des fondations (Partie 2), jusqu'au Chapitre 5 qui conclura les fondations. Etendons à présent cette liste et donnons des exemples concrets précisant la complexité des interactions.

Importance du contexte géographique

La mise en contexte de notre question dans un cadre bien particulier révèle l'importance de la prise en compte du contexte géographique. L'exemple des territoires de montagne, où les contraintes de ressources et de déplacement sont fortes, montre la richesse des situations possibles lorsqu'un schéma générique est mis en contexte dans un cas particulier.

Par exemple, sur des territoires de montagne français comparables, [berne2008ouverture] montre que les réactions à un même contexte d'évolution du réseau de transport peuvent mener à des dynamiques

territoriales très diverses, certains trouvant de forts bénéfices de l'accessibilité accrue, d'autres au contraire devenant plus fermés. Dans le même cadre, ces potentiels processus antagonistes sont examinés plus en détail par [bernier2007dynamiques], pour lesquels il propose un typologie basée sur le potentiel d'ouverture à la fois des dynamiques territoriales et des dynamiques des réseaux : par exemple, un territoire peut présenter de riches opportunités d'attractivité, par exemple des opportunités touristiques, tout en gardant une faible accessibilité. Réciproquement, il donne l'illustration des contraintes douanières pouvant freiner le potentiel d'ouverture d'une infrastructure performante.

En écho aux approches par systèmes de villes, [torricelli2002traversees] montre comment dans ce contexte il est possible de faire un lien entre nature des flux de transport et développement local du système urbain : les villes de montagne ont d'abord émergé comme point de passage sur les chemins de col, puis ont perdu de leur importance avec l'avènement des routes. L'arrivée du chemin de fer a pu les redynamiser, par le tourisme et l'industrie, et enfin l'autoroute a encore plus récemment induit une déstructuration par des effets de périurbanisation par exemple. Ainsi, les dynamiques structurelles sur le temps long sont particulières, en conséquence du contexte géographique.

Processus de planification

Comme nous l'avons déjà suggéré, les potentiels impacts des dynamiques territoriales sur les réseaux impliquent des processus à plusieurs niveaux. Ainsi, les projets d'infrastructure sont généralement planifiés²⁸, afin de répondre à certains objectifs fixés par des acteurs souvent institutionnels. Ces objets nous amènent progressivement vers le concept de gouvernance, mais prenons d'abord un instant pour illustrer des projets planifiés.

L'exemple de l'échec de planification de l'aéroport de Ciudad Real en Espagne montre que la réponse d'une infrastructure planifiée n'est pas systématique. Les explications à celui-ci découlent très probablement d'une combinaison complexe de multiples facteurs, difficiles à séparer. [otamendi2008selection] prédisait avant l'ouverture de l'aéroport une gestion complexe due à la dimension des flux attendus et propose un modèle approprié, or les ordres de grandeurs de flux effectifs étaient plus proches des milliers que des millions planifiés et l'aéroport a rapidement fermé. Il est compliqué de savoir la raison de l'échec, s'il s'agit de l'optimisme quand au polycentrisme régional (l'aéroport est à mi-chemin de Madrid et Séville), la non-réalisation

²⁸ Nous parlerons de *planification* en général, urbaine, territoriale, d'un projet d'infrastructure, pour désigner la conception volontaire d'un projet et d'un plan par un acteur d'aménagement, dans le but de transformer l'espace selon certaines motivations propres à l'acteur et à ses interactions avec les autres acteurs.

de la gare sur la ligne à grande vitesse, ou des facteurs purement économiques.

[[heddebaut:hal-01355621](#)]²⁹ montrent pour l'impact des infrastructures sur le long terme, dans le cas du tunnel sous la Manche³⁰, par une analyse des investissements et des politiques dans le temps, que les effets effectivement constatés pour la région Nord-Pas-de-Calais comme un gain de centralité et de visibilité au niveau Européen, sont en fort décalage avec les discours justifiant le projet, et que le renouvellement des acteurs implique un non-accompagnement du projet sur le long terme, rendant son impact plus hasardeux : on rejoint l'idée défendue par BRETAGNOLLE dans [[espacegeo2014effets](#)] selon laquelle des "effets de structure" effectivement existent mais que ceux-ci se manifestent sur le temps long en termes de dynamiques systémiques pour lesquelles une vision locale courte n'a que peu de sens. A l'échelle intra-urbaine, [[fritsch2007infrastructures](#)] prend l'exemple du Tramway de Nantes pour montrer, par une étude localisée des transformations urbaines à proximité d'une nouvelle ligne, que les dynamiques de densification urbaine sont en décalage avec ce qu'en attendaient les élus et planificateurs, c'est à dire une association forte entre proximité à la ligne et densification.

Ces exemples confirment que la compréhension des effets des territoires sur les infrastructures impliquent la prise en compte de la notion de *gouvernance*.

Gouvernance

Le développement d'un réseau de transport nécessite des acteurs disposant à la fois des moyens concrets et économiques de mener à bien la construction, et d'autre part ayant la légitimité de mener ce développement. Il s'agit donc nécessairement d'acteurs de la superstructure sociale, pouvant être différents niveaux de pouvoirs publics, parfois associés à des acteurs privés. Le concept de *gouvernance*, que nous comprenons comme la gestion d'une organisation disposant de ressources communes dans des buts liés à l'intérêt de la communauté concernée (pouvant être définis de différentes façons, par exemple de manière *top-down* par les acteurs de gouvernance ou de manière *bottom-up* par consultation des agents concernés par la décision), est alors essentiel pour comprendre l'évolution des projets de transports et donc des réseaux de transport. Nous parlerons de *gouvernance territoriale* lorsque les décisions concernent directement ou indirectement des composantes de systèmes territoriaux.

²⁹ Le possible jeu de mot par le titre ambigu sur l'existence du "Tunnel effect" rappelle l'effet tunnel, qui réside en la non-interaction d'une infrastructure sur un territoire le traversant sans s'y arrêter.

³⁰ Mis en service en 1994 entre Calais en France et Folkestone au Royaume-uni, ce tunnel ferroviaire sous-marin de 50km permet une liaison physique entre l'Europe continentale et le Royaume-uni.

Par exemple, [[offner200territorial](#)] illustre les difficultés posées par la dérégulation de certains services publics en réseau quant aux compétences territoriales des autorités, et propose l'émergence d'une régulation locale pour un nouveau compromis entre réseaux et territoires.

Certains aspects de la gouvernance territoriale peuvent avoir un impact déterminant sur le développement des infrastructures de transport. Illustrons ceux-ci pour des cas particuliers d'application de *modèles urbains*³¹. [[deng2007potential](#)] montre dans le cas des villes Chinoises que les nouvelles directives en terme de logement peuvent fortement détériorer la performance des infrastructures, et que des dispositions spécifiques doivent être prises pour anticiper ces externalités négatives. Celles-ci concernent notamment les dispositions en termes de *Transit Oriented Development* (TOD). Le TOD est une approche particulière de l'aménagement urbain visant à articuler développement de l'offre de transport en commun et développement urbain. Il s'agit en quelque sorte d'une co-évolution volontaire de la part des développeurs (autorités administratives et/ou de planification), dans laquelle l'articulation est pensée et planifiée. Nous reviendrons sur le TOD lors d'études empiriques par la suite.

Ces concepts ne sont pas nouveaux, puisqu'ils étaient implicites par exemple dans l'aménagement des villes nouvelles en Ile-de-France, sous une forme différente puisque celles-ci étaient également fortement zonées (c'est à dire planifiées en zones relativement cloisonnées et mono-fonctionnelles) et dépendantes de l'automobile pour certains quartiers [[es119](#)]. [[l2012ville](#)] est un exemple de projet européen ayant exploré des mises en pratiques de paradigmes du TOD : des détails d'aménagement comme un réseau de qualité pour les modes actifs à courte portée sont cruciaux pour une concrétisation des principes. Par exemple, [[lhostis:hal-01179934](#)] utilise une analyse multi-critères³² pour comprendre les facteurs déterminants dans la sélection des stations de la ville planifiée, incluant densité urbaine et temps d'accès aux stations. [[LIU2014120](#)] montre que si certaines politiques de planification, en particulier en France, ne se réclament pas directement de cette approche, leurs caractéristiques sont très similaires comme le révèle le cas de Lille.

L'articulation entre transport et aménagement doit souvent être opérée de façon fortement couplée pour parvenir aux objectifs recherchés, d'autant plus que le projet est spécialisé : [[laroque2002paris](#)] rappelle l'anecdote du metro SK de Noisy-le-Grand montre un cas

³¹ Au sens de la planification, c'est à dire de schémas conceptuels génériques permettant de guider une démarche de planification.

³² Dans le cadre de l'aide à la décision pour la planification des infrastructures de transport, l'analyse multi-critère est une alternative aux analyses coût-bénéfices (qui comparent des projets en agrégeant un coût généralisé) qui permet de prendre en compte de multiple dimensions, souvent contradictoires (par exemple coût de construction et robustesse pour un réseau), et obtenir des solutions optimales au sens de Pareto

de dépendance complète de la fonctionnalité du transport à l'aménagement local. Pour desservir un projet de complexe de bureau, une ligne spécifique avec une matériel roulant léger est construite pour faire le lien avec la gare RER de Mont-d'Est. Le projet immobilier avortera alors que la ligne est inaugurée en 1993, celle-ci sera d'abord entretenue régulièrement puis laissée à l'abandon sans jamais avoir été ouverte au public.

Ainsi, les processus de gouvernance, qui peuvent se décliner de plusieurs manières, comme ceux de planification, ou plus spécifiques de TOD, jouent un rôle important dans les interactions entre réseaux de transports et territoires. Ceux-ci s'ajoutent à notre panorama, étant d'un type particulier car impliquant leur propre niveau d'émergence et une forte autonomie.

Co-évolution des réseaux de transport et des territoires

Cette construction progressive nous a permis de souligner la complexité des interactions entre réseaux et territoires, ce qui suggère la pertinence de l'ontologie particulière de la *co-évolution* comme nous l'avons définie en introduction. [levinson2011coevolution] souligne la difficulté de la compréhension de la co-évolution entre transport et usage du sol en termes de causalités circulaires, en partie à cause des différentes échelles de temps impliquées, mais aussi par l'hétérogénéité des composantes. [offner1993effets] parle de congruence, qu'on peut comprendre comme une dynamique systémique impliquant des corrélations fortuites ou non, ce qui serait une vision préliminaire de la co-évolution.

La nécessité de dépasser les approches réductrices des effets structurants, tout en capturant la complexité des interactions entre réseaux et territoires par leur co-évolution, est confirmée par le cas des effets économiques des trains à grande vitesse : [Blanquart2017] procède à une revue à la fois théorique et empirique, incluant la littérature grise, des études de cas spécifiques, et conclut, au delà des retombées directes liées à la construction sur lesquelles il y a consensus, que les effets propres sur un temps plus longs paraissent aléatoires. Cela témoigne en fait de situation locales complexes, un grand nombre d'aspect conjoncturels entrant en jeu dans la production d'effets, qu'on ne peut alors pas attribuer seulement au transport. Cette revue confirme par ailleurs le décalage entre les discours politiques et techniques prévalant aux projets de transports et les analyses effectives a posteriori révélée par [bazin:hal-00615196]. [bazin2007evolution] procède d'autre part à une étude ciblée du marché immobilier à Reims en anticipation de l'arrivée du TGV Est. En procédant à une analyse diachronique pour chaque année entre 1999 et 2005, par quartier, des prix immobiliers et de la provenance des acheteurs (Franciliens ou locaux), ils concluent que seul des opérations très localisées pouvaient être di-

rectement reliées au TGV, l'ensemble du marché répondant à une dynamique globale indépendante.

Ainsi, notre aperçu constructif, large et voulu circulaire, des interactions entre réseaux de transports et territoires, confirme la pertinence de cette notion de *co-évolution* d'une part, mais suggère un approfondissement et une clarification de celle-ci. Nous nous appliquerons dans la section suivante à approfondir de manière empirique différents aspects abordés ici.

★ ★

★

1.2 DE PARIS À ZHUHAI

Nous développons dans cette section des cas d'étude géographique à l'échelle métropolitaine comme nous l'avons définie précédemment. Nous les choisissons très différents pour maximiser la diversité des processus potentiellement identifiables (puisque comme nous l'avons montré le contexte géographique est crucial). Il s'agit de la métropole du Grand Paris, et de la mega-région urbaine du Delta de la Rivière des Perles dans le sud de la Chine.

L'objectif de cette section est de spécifier, préciser, illustrer, enrichir, l'aperçu des processus de co-évolution que nous avons établi de manière générale. La géographie ne peut tirer de conclusion générales, dans les cas où celles-ci sont pertinentes, sans études de cas particuliers bien précis. Dans l'application d'un modèle générique à un ensemble de territoires, on cherchera les déviations au modèle, qu'il s'agira alors d'expliquer par des raisonnements géographiques, signifiant une forte implication avec le lieu en particulier. Notre démarche est similaire : si nous pouvons raccrocher nombre de concepts développés à un cas d'étude, ceux-ci seront nécessairement enrichis³³.

1.2.1 *Le Grand Paris : histoire et enjeux*

La région parisienne est une bonne illustration de la complexité des interactions entre réseaux de transports et territoires. La période temporelle pertinente pour notre question court de la fin du 19ème siècle à nos jours. Nous proposons, après une présentation brève du contexte, de rappeler l'histoire du développement des transports en Ile-de-France, qui permet de révéler ses articulations avec l'urbanisme, en particulier les enjeux liés à la planification du réseau de transport. Nous traiterons ensuite le présent et le futur du Grand Paris, d'abord concernant l'émergence d'une nouvelle structure de gouvernance au niveau de la métropole, puis les projets de transport récents impliqués, mettant l'exemple au cœur de notre problématique. Nous ferons finalement une incursion plus détaillée dans une analyse empirique des relations entre variables territoriales et différentiels d'accessibilité pour les projets de transport, préfigurant certains des développements méthodologiques que nous mènerons par la suite.

Contexte

Le contexte spatial est l'échelle intermédiaire d'une région métropolitaine globalement monocentrique. Précisons cette structure spatiale. Si la métropole prise jusqu'à la moyenne couronne (c'est à dire l'étendue correspondant environ au noyau urbain central bâti de manière

33 Et possiblement connectés par le transfert de la structure du système particulier à la structure de la connaissance.

continue) possède un certain niveau de polycentrisme³⁴, notamment grâce à l'effet des villes nouvelles, devenues d'importants pôles d'emplois locaux [berroir2005contribution].

Le rôle des différentes infrastructures de transport dans les différentes dynamiques économiques en Ile-de-France n'est pas trivial, comme le montre [PADEIRO201344] qui cherche à expliquer statistiquement la croissance de l'emploi entre 1993 et 2008 dans les moyennes et petites communes franciliennes en fonction de la proximité à une infrastructure : les effets dépendent à la fois du mode (autoroute ou aéroport) mais aussi du secteur économique considéré. Réciproquement, les développements successifs des projets de transport, s'opèrent de manière généralement discontinue dans le temps. Comme nous le détaillerons par la suite, ils sont liés à des dynamiques de planification et des processus de gouvernance qu'il convient de comprendre de manière conjointe aux dynamiques territoriales. La métropole parisienne témoigne ainsi de relations complexes entre territoires et réseaux.

Réseau de Transport du Grand Paris

L'histoire du développement du réseau de transport de la métropole francilienne est rappelée dans [larroque2002paris]. La particularité centralisatrice française a conduit à une structure particulière du réseau ferré à l'échelle nationale, mais aussi à l'échelle régionale. La domination de Paris a en effet fortement marqué la structuration du réseau de transport au cours des différentes périodes historiques où il a subi des évolutions conséquentes. [larroque2002paris] décomposent la seconde moitié du vingtième siècle en trois périodes.

Avant 1975, la distribution de l'accessibilité des actifs aux emplois est clairement centralisée et le centre de Paris fortement congestionné. La mise en place du réseau RER entre 1975 et 1988 permet grâce à la construction conjointe des Villes Nouvelles une articulation entre transport et urbanisme et un certain niveau de polycentrisme. [larroque2002paris] rappellent toutefois que les réalisations dans cette période sont en décalage croissant avec la demande réelle de transport. La période qui suivra 1988 jusqu'à 2000, année marquée par l'alternance politique, consistera surtout en le renouvellement des acteurs et l'élaboration de nouvelles stratégies, comme en témoigne le Schéma Directeur de 1994. Les développements du réseau sur cette période n'induisent aucun changement majeur de la distribution spatiale de l'accessibilité,

34 Le polycentrisme, en opposition au monocentrisme, signifie qu'il est possible d'identifier différents centres dans un système urbain. La façon de définir un centre dépendra de l'échelle et des phénomènes considérés : il peut s'agir par exemple de l'existence de différents pôles d'emplois de taille comparable à l'échelle intra-métropolitaine. De la même façon que le concept est polymorphe, les façons de le mesurer quantitativement sont multiples et complémentaires [servais2004polycentrisme].

malgré la réalisation de l'interconnexion centrale du RER D, de la ligne 14 et du RER E.

Les schémas directeurs successifs conduisent au SDRIF de 2013 [sdrif2013]. Ceux-ci préfigurent le futur réseau du *Grand Paris Express*, dont un fort impact est attendu en termes de cohésion territoriale en favorisant les liaisons de banlieue à banlieue qui sont les plus problématique dans le réseau actuel. De plus, le schéma est volontairement intégré, par densification autour des gares et articulation des opérations d'aménagement et des nouvelles infrastructures. Cet aspect d'intégration des réseaux dans les territoires et des territoires par les réseaux se retrouve bien dans la communication publique de l' Autorité Organisatrice des Transports (ancien STIF, devenu Ile-de-France Mobilités)³⁵. On retrouve donc l'importance des processus de gouvernance dans l'articulation des réseaux de transport et des territoires dans l'exemple de l'Ile-de-France au cours du temps.

D'autres processus déjà mentionnés se manifestent également, sous différentes formes. Par exemple, le rôle de la dépendance au chemin dans les trajectoires du système territorial est illustré par [larroque2002paris] qui montre l'inertie due aux choix techniques successifs lorsque ceux-ci rencontrent un succès : le choix initial d'un réseau métropolitain intra-muros, la mise en place du réseau RER, la politique de tarification par zones de la carte orange à la fin des années 90, sont autant de décisions sur des domaines divers mais ayant chacune leur part significative dans les développements postérieurs possibles. Ces auteurs montrent également comment les décisions concernant le réseau de transport en commun peuvent induire, par mauvaise couverture ou performance du réseau de transport en commun, l'émergence de processus d'interactions où le couple usage de la voiture et périurbanisation³⁶ est favorisé, à l'image de l'*automobile city* décrite par [newman1996land].

[padeiro:tel-00438092] rappelle que le prolongement des lignes de métro en proche banlieue a toujours été restreint, renforçant le rôle de la ville Paris dans les relations entre le territoire métropolitain et les réseaux. Par ailleurs, il montre que les polarisations urbaines (adaptation du bâti et de la composition socio-économique) autour des stations au delà du périphérique sont pour leur partie socio-économique des dynamiques antérieures qu'accompagne alors l'arrivée du métro : dans ce cas, il n'y a pas d'effet structurant à proprement parler.

³⁵ Voir par exemple l'actualité du 4 octobre 2017 sur <https://www.iledefrance-mobilites.fr/actualites/un-reseau-de-transports-qui-grandit/> qui souligne que "Avec 29 km de réseau supplémentaires et l'ouverture de 28 points de desserte, les territoires se rapprochent", témoignant de l'importance de l'accessibilité pour les territoires, notion par ailleurs floue. Les mêmes orientations de discours se retrouvent pour les différents projets d'extension ou de construction de nouvelles lignes.

³⁶ Le périurbain fait partie des nouvelles formes d'urbanisation, et consiste en des territoires intermédiaires entre urbain et rural, bénéficiant d'une bonne accessibilité mais présentant un faible densité et des habitats individuels majoritairement.

Vers une gouvernance métropolitaine

Au contexte métropolitain décrit précédemment correspond une complexité de la structure de gouvernance. En particulier, les développements actuels, à la fois du réseau de transport et des projets d'aménagement, coïncident avec l'émergence d'un nouveau niveau de gouvernance, intermédiaire entre communes et départements d'une part et Région et Etat d'autre part. On peut se demander dans quelle mesure cette émergence est reliée aux dynamiques d'interactions entre territoires, et comment celle-ci influera sur les interactions entre territoires et réseaux. [gilli2009paris] proposent en 2009 un diagnostic de la situation institutionnelle de la région parisienne, et des pistes pour une approche couplée entre gouvernance et aménagement. Ils mettent en valeur la préfiguration de "l'instauration d'un acteur collectif métropolitain", qui correspond à la métropole du Grand Paris qui sera inaugurée 7 ans plus tard, puisque le conseil métropolitain est mis en place fin 2016.

La mise en place de ce nouveau niveau de gouvernance a été disséquée plus récemment toujours par [gilli2014gouverner], où il la situe dans un contexte socio-économique et des autres niveaux de gouvernance (Etats, Région, intercommunalités) plus large. Cela lui permet de dresser un diagnostic territorial qui fournit des éléments explicatifs à son émergence : en perte de vitesse sur le plan de l'aménagement par rapport à ses dynamiques passées, mais aussi sur le plan social au vu d'inégalités socio-économiques locales très fortes, la métropole a besoin de se réinventer, et ce nouveau souffle se cristallise naturellement dans le Grand Paris, c'est à dire que, comme il conclut, "l'avenir de Paris est sa banlieue". Cette initiative se concrétise par la convergence d'une part des initiatives et du volontarisme des élus locaux, et d'autre part d'une redéfinition du rôle de l'Etat, voulue centralisatrice jusqu'en 2012 puis laissant la place libre à la gouvernance métropolitaine avec l'alternance politique en 2012. même si les projets lancés et les financements restent les mêmes dans les grandes lignes : le projet du Grand Paris Express est un compromis entre la solution voulue par l'Etat et celle poussée par la région. Suivant [desjardins2016grand], si la structure de gouvernance métropolitaine est aujourd'hui toujours relativement impuissante, et si l'oubli de l'aspect social du développement métropolitain est toujours très présent, ces mutations témoignent toutefois d'un changement structurel profond dans l'organisation de la région. Nous détaillons à présent le projet de transport du Grand Paris Express.

Projet du Grand Paris Express : vers un rééquilibrage des accessibilités ?

La région métropolitaine de Paris est en train de connaître de grandes mutations, avec la mise en place d'une gouvernance métropolitaine et de nouvelles infrastructures de transport. La construction d'un réseau

Légende

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

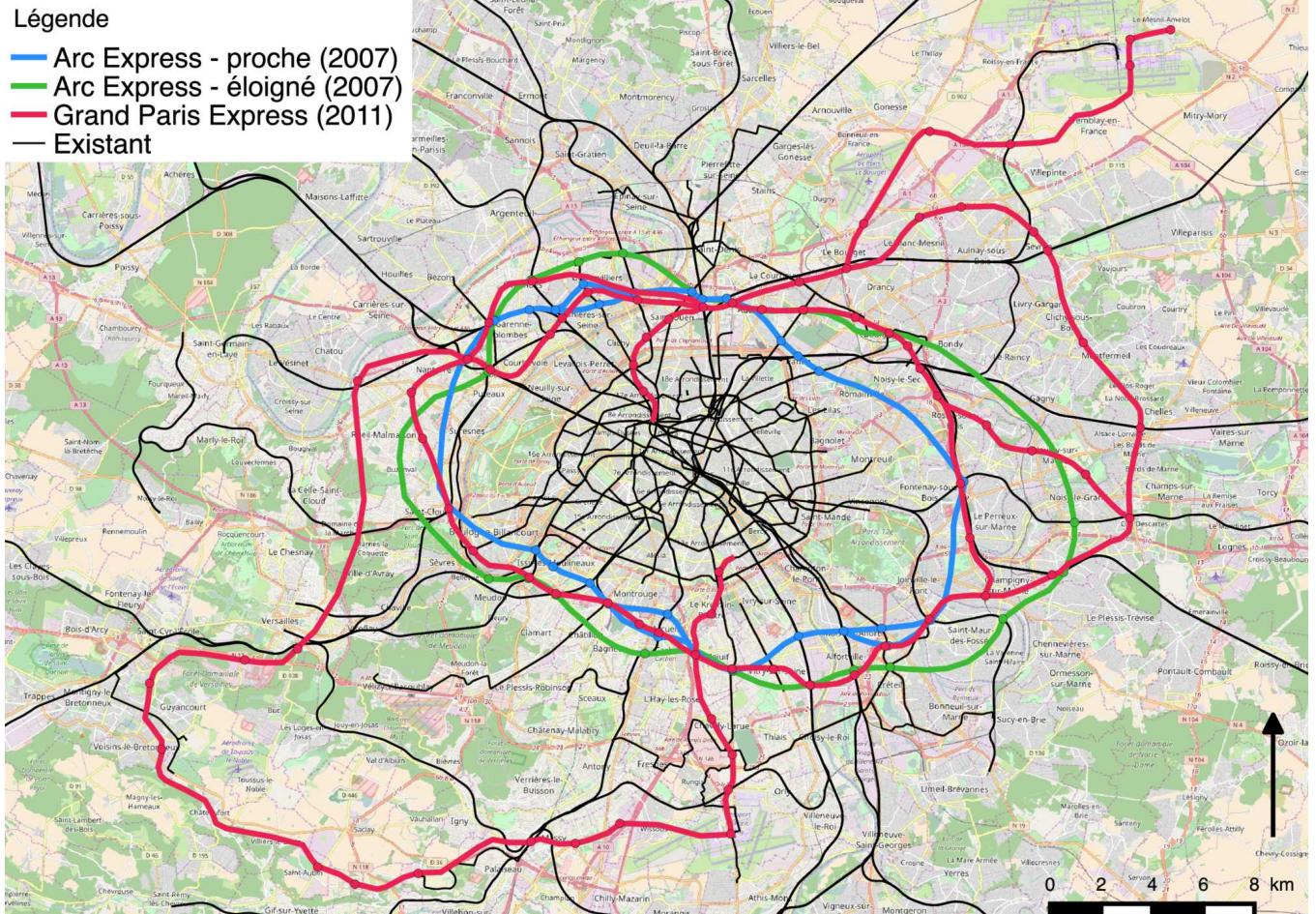


FIGURE 1 : Projets de transport successifs de la métropole du Grand Paris. Nous montrons les deux alternatives du projet Arc Express porté par la région, et le Grand Paris Express (GPE) porté par l'état, et dont le tracé final résulte d'un compromis entre l'état et la région. Le Réseau du Grand Paris, précurseur du GPE, n'est pas montré ici pour des raisons de visibilité à cause de sa proximité avec celui-ci. Le fond de carte, donné pour indication, a pour source OpenStreetMap.

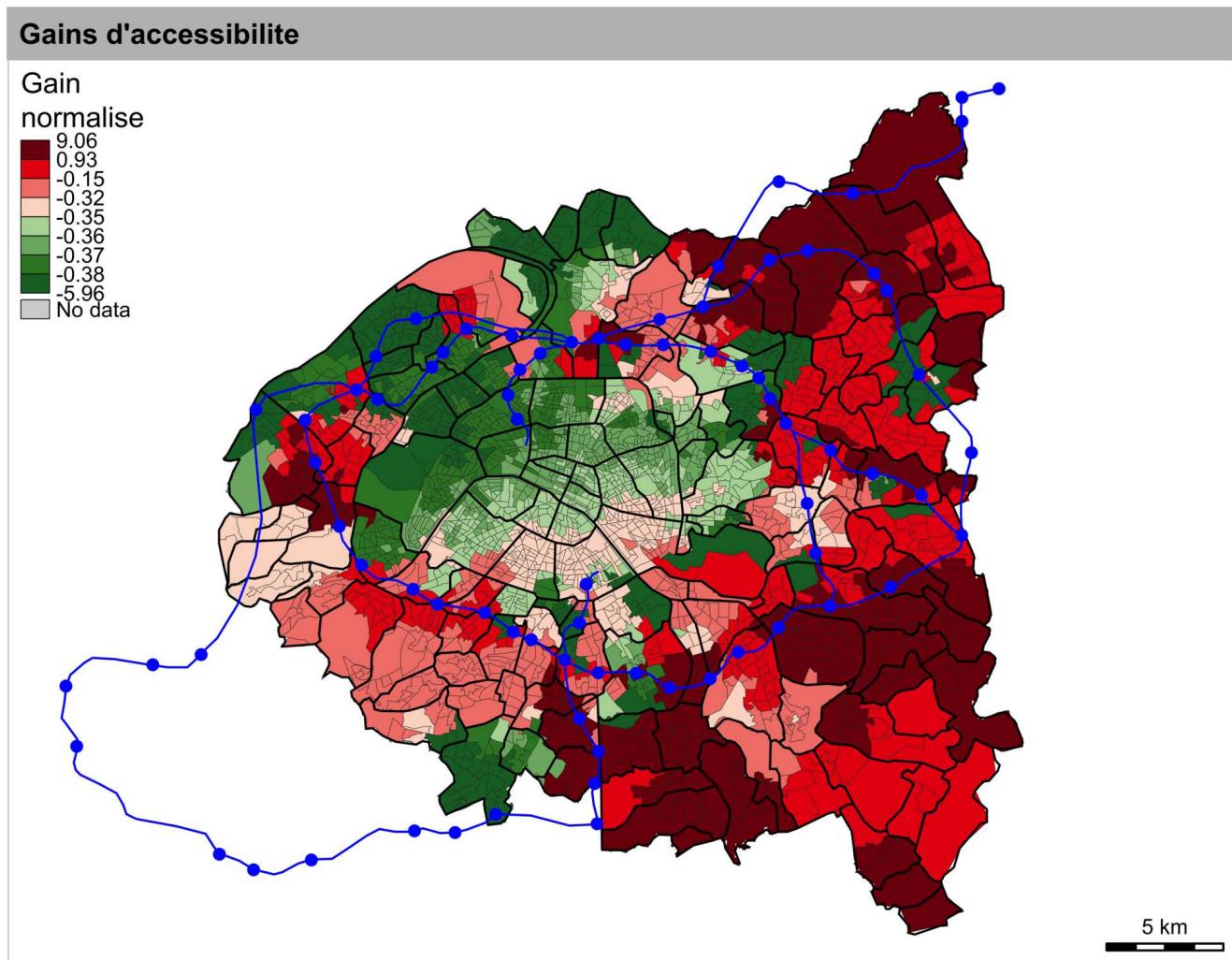


FIGURE 2 : Impact des lignes du GPE sur l'accessibilité temporelle. La carte donne, pour les départements de la petite couronne et Paris (75, 92, 93, 94) les gains d'accessibilité temporelle, définie pour chaque Iris (unité statistique infra-communale élémentaire) comme le temps moyen de trajet en transport en commun vers l'ensemble des centroïdes des autres communes pondéré par la population de destination. Le gain est calculé comme la différence d'accessibilité avec et sans Grand Paris Express. Nous montrons le gain normalisé, c'est à dire centré (de moyenne nulle) et réduit (écart-type unitaire). En bleu, les lignes et nouvelles gare du GPE. On observe les gains les plus forts majoritairement à l'Est, en cohérence avec la littérature existante comme [beaucire2013grand]. Les sillons territoriaux des lignes de RER (A à l'ouest, D et B au nord, B au sud) présentent des gains relativement faible car déjà très accessibles.

de métro en rocade permettant des liaisons de banlieue à banlieue répond à un besoin ancien, et a mené à plusieurs propositions sur lesquelles se sont opposés l'Etat et la Région au tournant des années 2010 [**desjardins2010bataille**]. Le projet Arc Express [**stif2007arc**], porté par la Région et plus axé sur une égalité des territoires, contrastait avec les propositions initiales de Réseau du Grand Paris visant à relier des “clusters d’excellence” en dépit d’un possible effet tunnel. La solution finalement adoptée (voir le dernier schéma directeur [**sdrif2013**]) est un compromis et permet un rééquilibrage est-ouest de l’accessibilité [**beaucire2013grand**]. La Fig. 1 cartographie les différents projets.

Les impacts immédiats d’une nouvelle infrastructure de transport en termes d’accessibilité, c’est-à-dire la transformation de la distribution spatiale des différentes accessibilités, concernent généralement des territoires bien plus larges que les zones où la ligne et ses stations sont implantées : les motifs d’accessibilité sont dus aux propriétés topologiques du réseau et celles-ci sont fortement discontinues en fonction de la structure du graphe. Illustrons le cas des lignes du Grand Paris Express et de leur impact direct sur l’accessibilité régionale. Nous cartographions en Fig. 2 les gains d’accessibilité temporelle permis par ce projet sur les départements métropolitains (75, 92, 93 et 94), avec le tracé le plus récent pour l’ensemble des lignes. L’accessibilité temporelle est calculée pour chaque Iris i de la manière suivante : avec les populations des communes P_j , t_0 un paramètre de durée typique d’un déplacement (que nous fixons à une heure [**zahavi1980regularities**]), t_{ij} le temps de trajet en transports en commun entre le centroïde de i et celui de la commune j , nous prenons une moyenne pondérée définie comme

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

Cette expression permet de bien avoir un potentiel d’accessibilité, et la pondération par la population permet de ne pas biaiser l’indicateur par des trajets potentiellement négligeables en proportion des trajets totaux.

On observe, conformément à l’analyse de [**beaucire2013grand**], un rééquilibrage des différentiels d’accessibilité entre Est et Ouest. A distance égale du centre, l’accessibilité est plus basse pour la Seine-Saint-Denis et le Val-de-Marne que pour les Hauts-de-Seine, c’est à dire que ces départements ont potentiellement plus de difficultés pour accéder au reste de la métropole. La carte des gains temporels moyens montre les gains plus grands également pour ces deux départements. Des communes socio-économiquement défavorisées comme Aulnay sont bénéficiaires des plus grands gains de temps. La ligne 16 permet en effet un désenclavement significatif du nord-est de la

Seine-Saint-Denis [**desjardins2016grand**]. La création de liaisons de banlieue à banlieue est un aspect majeur de ce désenclavement et est voulue comme un moteur de l'émergence de nouvelles centralités, vers une métropole toujours plus polycentrique, dans la lignée de la politique d'aménagement des villes nouvelles, pour ne plus parler de proche banlieue mais de quartiers faisant partie intégrante du Grand Paris. Les effets peuvent cependant être mitigés selon les zones : [**l2013grand**] montrent que le Grand Paris Express induira un accès direct à un plus grand nombre d'emplois pour un nombre significatif de chômeurs en petite couronne, mais que les écarts avec la grande couronne seront accentués et qu'il existe des risques de décrochage de certaines communes lointaines mal desservies.

L'un des enjeux cruciaux pour la construction du Grand Paris est de veiller à ne pas obtenir une métropole à plusieurs vitesses, et de tirer parti de la connectivité accrue à plusieurs échelles (internationale, nationale, régionale, métropolitaine) pour réduire les inégalités territoriales plutôt que les accroître³⁷. Le nouveau réseau semble contribuer à cette dynamique, sous condition d'un développement territorial coordonné, permettant la concrétisation des gains immédiats d'accessibilité en terme de transformation territoriale. Il n'existe pas de méthode pouvant prévoir celle-ci de manière déterministe comme nous l'avons déjà développé. Il est cependant possible d'analyser rétrospectivement de manière empirique les couplages entre variables territoriales et variables liées, pour essayer de mettre en valeur quantitativement les phénomènes de co-évolution. Nous proposons à présent d'illustrer cette démarche.

Lier dynamiques territoriales et construction du Grand Paris Express

L'un des enjeux de notre travail par la suite sera de clarifier empiriquement des situations dans lesquelles des dynamiques fortement couplées relevant de cette problématique pourront être mises en évidence puis à travers des modèles d'isoler des processus et des conditions permettant telle ou telle situation. Nous proposons d'approfondir l'illustration du GPE, tout en introduisant une approche possible pour lier dynamique territoriale et celle du nouveau réseau anticipé.

Des aspects très variés des territoires sont concernés par l'interaction avec les réseaux. Dans nos études précédentes, les aspects économiques et financiers du foncier et l'immobilier n'ont pas été considérés. Il s'agit cependant d'éléments cruciaux des dynamiques territoriales et sont étudiés de manière intensive dans des champs comme l'analyse territoriale ou l'économie urbaine : par exemple, [**homocianu:tel-00359302**] étudie les choix résidentiels des ménages pour comprendre les interactions entre usage du sol et transport.

³⁷ Rappelons qu'une inégale répartition des agents et des ressources générera des différences de potentiel plus grandes qu'une distribution uniforme, celles-ci pouvant alors être liées à l'évolution du réseau

Mode	Vitesse moyenne
RER	60km.h ⁻¹
Transilien	100km.h ⁻¹
Metro	30km.h ⁻¹
Tramway	20km.h ⁻¹

Nous proposons ici d'utiliser entre autres une base de données de transactions immobilières pour la région parisienne sur les 20 dernières années, avec une granularité temporelle de 2 ans et coordonnées spatiales exactes. [guerois2009dynamique] l'utilise par exemple pour établir une typologie des dynamiques spatiales du marché immobilier parisien.

Cette étude plus précise peut être comprise comme une recherche de signes précurseurs de rupture de potentiels du réseau : en effet, si des dynamiques territoriales intrinsèques anticipent l'arrivée d'une nouvelle station de transports en commun, les implications seront bien différentes du cas où celle-ci conduit ces variables après sa construction. L'interprétation en termes "d'effets structurants" sera notamment très différente. Nous appliquons ici la méthode de causalités spatio-temporelles développée en 4.2. Nous proposons d'étudier les relations entre différentiel d'accessibilité pour chaque projet, et variables liées au foncier (transactions immobilières) et socio-économiques, afin de voir s'il est possible de capturer un lien entre les différentiels d'accessibilité et les différentiels de dynamisme territoriaux. En effet, les liens entre nouvelles lignes et évolution du foncier sont parfois remarquables [damm1980response].

Les données des transactions immobilières sont fournies par la base BIENS (Chambre des Notaires d'Ile de France, base propriétaire). Le nombre de transactions utilisables après nettoyage est de 862360, se répartissant sur l'ensemble des IRIS, pour une plage temporelle couvrant de 2003 à 2012 incluses. Les données par IRIS pour population et revenu (revenu médian et indice de Gini) proviennent de l'INSEE. Les données de réseau ont été vectorialisées à partir des cartes des projets (voir Fig. 1 pour les projets). Les temps de trajets sont calculés par transport en commun uniquement, avec des valeurs standard pour les vitesses moyennes des différents modes [larroque2002paris], que nous résumons dans le tableau suivant :

La matrice des temps est calculée depuis l'ensemble des centroïdes des IRIS vers l'ensemble des centroïdes des communes. Ceux-ci sont reliés au réseau par des connecteurs à la gare la plus proche, de vitesse 50km.h⁻¹ (trajet en voiture). Les analyses sont implémentées intégralement en langage R [rcoreteam] et l'ensemble des données, du

code source et des résultats sont disponibles sur un dépôt git ouvert³⁸.

Nous calculons pour chaque projet, le différentiel ΔT_i d'accessibilité temporelle de trajet à partir de chaque IRIS en comparaison à celui dans le réseau sans le projet, où accessibilité temporelle est définie par $T_i = \sum_j \exp -t_{ij}/t_0$ avec j communes, t_{ij} temps de trajet, et t_0 paramètre d'atténuation. Nous ne pondérons pas ici par la population des communes de destination contrairement à l'accessibilité Z_i utilisée précédemment, pour être certain de ne pas capturer d'autocorrélation pour la population ou de corrélations entre population et variables territoriales que nous étudions. A chaque projet est associée une date³⁹, correspondant environ à l'année d'annonce mature du projet, restant toutefois arbitraire car difficile d'une part à déterminer précisément, un projet n'émergeant pas d'un coup du jour au lendemain, et d'autre part pouvant correspondre à des réalités différentes d'apprentissage du projet par les différents agents économiques (nous faisons donc l'hypothèse réductrice mais nécessaire d'une diffusion sur la majorité des agents dans un temps inférieur à l'année).

Le lien entre différentiels d'accessibilité et variations des variables territoriales est effectué par l'étude des corrélations retardées. Cette méthode sera développée en détail en 4.2, mais nous n'avons pas besoin d'entrer dans les détails techniques ici. L'idée est la suivante : si deux variables présentent une forte corrélation avec un certain retard temporel, on a une notion faible de causalité, les variation de la variable précurseur pouvant être à l'origine de celles de la variable non-décalée dans le temps (on dit faible, car il est toujours possible que les corrélations soient fortuites bien sûr).

Nous étudions les corrélations retardées de ΔT_i avec les variations ΔY_i des variables socio-économiques suivantes : population, revenu médian, indice de Gini des revenus, prix moyen des transactions immobilières et montant moyen des crédits immobiliers. La corrélation est estimée en retardant l'accessibilité, c'est à dire en estimant $\rho[\Delta T_i(t - \tau), \Delta Y_i(t)]$. Un test de Fisher est effectué pour chaque estimation, et la valeur est fixée nulle si celui-ci n'est pas significatif ($p < 0.05$ de manière classique). L'étude avec accessibilité généralisées au sens de Hansen [[hansen1959accessibility](#)] (pondérée par les populations à la destination, ou les populations à l'origine et les emplois à la destination) a également été menée mais moins intéressante car très peu sensible à la composante mobilité (réseau et atténuation)

³⁸ A l'adresse

<https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/>

SpatioTempCausality/GrandParis. Les données de la base BIENS ne peuvent être fournies pour raison de fermeture contractuelle de la base.

³⁹ 2006 pour Arc Express, 2008 pour le Réseau du Grand Paris, 2010 pour le Grand Paris Express

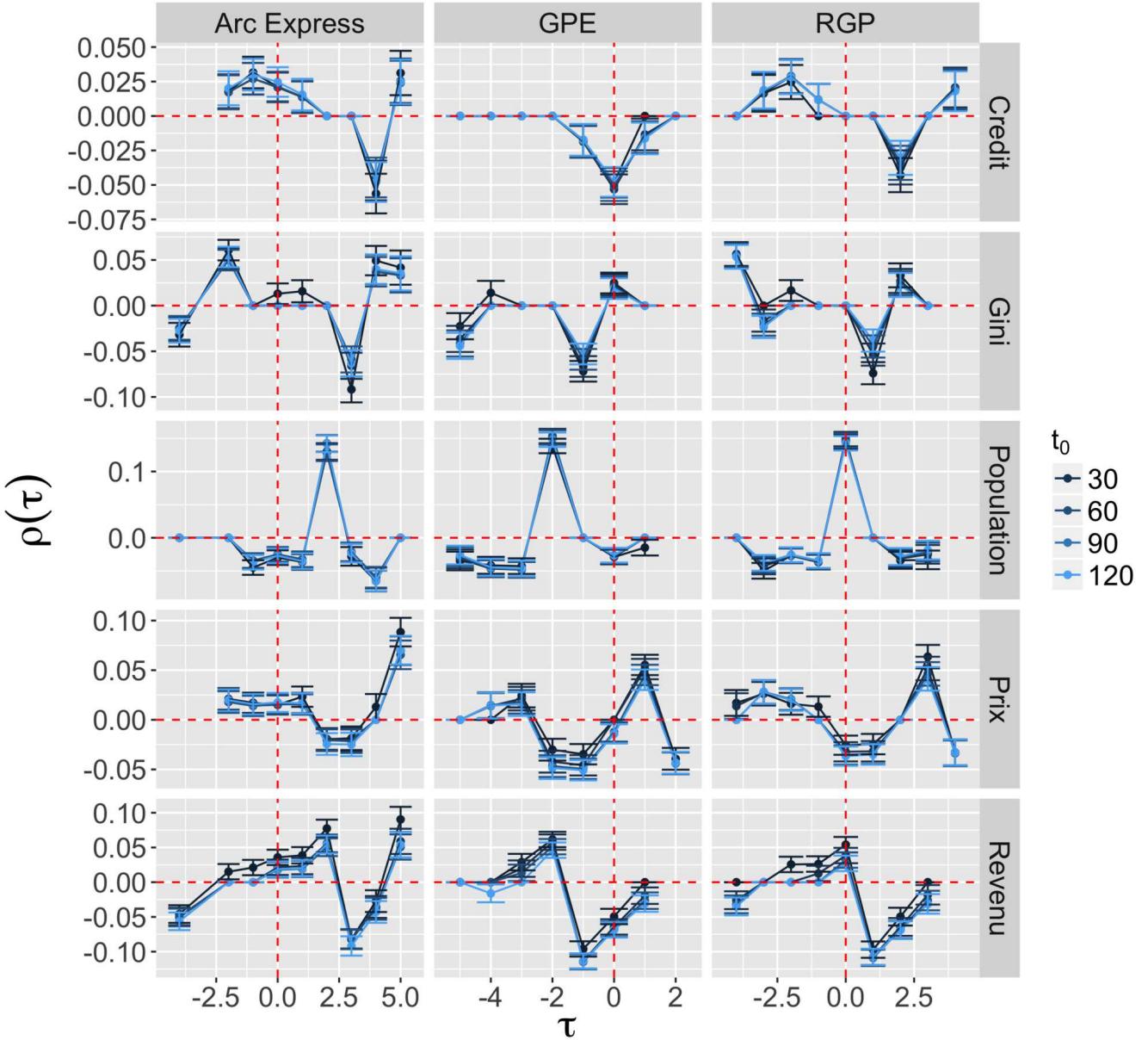


FIGURE 3 : Corrélations retardées empiriques entre différentiel d'accessibilité et variables territoriales. Les graphiques donnent la valeur de la corrélation retardée $\rho(\tau)$ en fonction du retard τ , entre le différentiel d'accessibilité en temps de trajet moyen ΔT_i , pour chaque projet (en colonnes : Arc Express, Grand Paris Express (GPE), Réseau du Grand Paris (RGP)) et le différentiel des différentes variables socio-économiques et de transactions immobilières ΔY_i (en lignes : valeur des crédits immobiliers (Crédit), Prix moyen des transactions immobilières (Prix), Revenu médian (Revenu), Indice de Gini des revenus (Gini), Population), pour différentes valeurs du paramètre d'atténuation t_0 . Les barres d'erreur donnent l'intervalle de confiance à 95%. Les lignes rouges pointillées aident à la lecture : elles permettent horizontalement de voir si les corrélations sont significatives, verticalement de voir la valeur du retard optimal. Par exemple, la lecture de la première ligne suggère que les projets anciens ont causé une baisse des crédits immobilier accordés dans les iris dont l'accessibilité a suivi une croissance positive, et que ces variables sont synchronisées pour le GPE.

par rapport aux variables elle-mêmes, informe uniquement sur des relations entre celles-ci et n'est donc pas présentée ici.

Nous présentons en Fig. 3 les résultats pour l'ensemble des réseaux et variables. La lecture s'effectue de la façon suivante : pour une variable et un projet donnés, la courbe $\rho(\tau)$ peut présenter des maxima pour une valeur $\tau_m > 0$ ou $\tau_m < 0$. Cette corrélation maximale correspond à un retard donnant une "synchronisation maximale" entre les deux variables, et le signe du retard donne le sens de la causalité entre les deux variables.

Il est remarquable tout d'abord de noter l'existence d'effets significatifs (au sens de corrélations significative et d'un intervalle de confiance à 95% ne contenant pas 0) pour l'ensemble des variables. Des valeurs plus basses du paramètre t_0 donnent des corrélations plus fortes en valeur absolue, révélant une possible plus grande importance de l'accessibilité locale sur les dynamiques territoriales. Le comportement de la population montre un pic très détaché correspondant à 2008, laissant supposer un impact du plus vieux projet d'Arc Express sur la croissance de la population. Sous cette hypothèse, l'effet des autres projets serait alors fallacieux de par leur proximité dans les grands tronçons. Cela impliquerait d'ailleurs que les zones où ils diffèrent fondamentalement comme le Plateau de Saclay ne soient que très peu sensibles au projet de transport, confirmant l'aspect artificiel planifié du développement de ce territoire.

Concernant les revenus, on observe un comportement similaire mais négatif, ce qui impliquerait un appauvrissement lié à l'augmentation de l'accessibilité, mais qui semble toutefois s'accompagner d'une baisse des inégalités puisque le coefficient de Gini présente également une corrélation négative dans les retards positifs. Enfin, comme attendu les prix immobiliers sont tirés par l'arrivée potentielle des nouveaux réseaux, effet qui disparait à deux ans pour le Grand Paris Express, suggérant une bulle immobilière passagère dans les quartiers autour. Nous démontrons ainsi l'existence de liens de correlations retardées complexes qu'on nomme causalités en ce sens, entre dynamiques territoriales et dynamiques anticipées des réseaux. Une compréhension plus fine des processus à l'oeuvre est au delà de la portée de cet étude préliminaire, car supposerait par exemple des études de terrain qualitatives ou des études de cas ciblées.

Cette étude suggère des effets potentiels de la modification d'accès-sibilité due au projets du Grand Paris, puisque certains effets révélés peuvent être liés à des politiques d'aménagement anticipant également le nouveau réseau. On suggère ainsi une existence réelle des processus d'effet du réseau sur les territoires, puisque la majorité des retards optimaux sont positifs.

1.2.2 Le Delta de la Rivière des Perles

Nous changeons à présent de région géographique, de structure urbaine, de période temporelle, pour évoquer un autre cas d'étude pertinent en Chine. La région parisienne étendue peut être vue comme un ensemble cohérent⁴⁰ : il s'agirait d'une *Mega-région urbaine*, concept que nous allons à présent définir et développer pour l'instance particulière du Delta de la Rivière des Perles.

Nouveaux régimes urbains et Mega-région urbaine

La notion de megalopolis a été introduite par [gottmann1964megalopolis] pour désigner l'émergence d'agglomérats urbains à une échelle non-existante auparavant. Elle est à l'origine du concept de *Mega-city Region* (MCR) consacré par [hall2006polycentric]. Sur le cas Européen, ils dégagent des ensembles de métropoles fortement connectées par rapport aux flux de mobilité, aux connections entre entreprises, qui forment ce qu'il appellent des *Mega-city Region* polycentriques (par exemple la Randstad aux Pays-bas, la région Rhin-Ruhr en Allemagne). Les caractéristiques sont une certaine proximité géographique des centres, une forte intégration par les flux, et un certain niveau de polycentrisme. Il s'agit d'une forme urbaine inédite par le passé, dont l'émergence semble liée aux processus de globalisation.

Ce concept est toujours plus d'actualité avec l'apparition récente de nouveaux types d'urbanisation, notamment par l'urbanisation accélérée dans des pays à forte croissance économique et en mutation très rapide comme la Chine [swerts2015megacities].

Le second cas que nous développons ici rentre dans cette catégorie : le Delta de la Rivière des Perles (PRD) est une des illustrations classiques de la structure d'une MCR fortement polycentrique. Historiquement initialement composé de Guangzhou uniquement, le développement de Hong-Kong puis la mise en place des Zones Economiques Spéciales dans le cadre des politiques d'ouverture de Deng Xiaoping, a conduit à un développement extrêmement rapide de Shenzhen, et dans une moindre mesure de Zhuhai. La province du Guangdong dans lequel le PRD se situe intégralement a actuellement le plus fort PIB régional de Chine, et la MCR regroupe une population d'environ 60 millions (les estimations fluctuant fortement selon la définition prise de la MCR et la prise en compte de la population flottante). Le phénomène de migration des campagnes est très présent dans la région et une ville comme Dongguan a par exemple basé son économie sur des manufactures employant ces travailleurs migrants.

C (CL) : Le développement de Zhuhai a aussi été rapide, mais le modèle de développement adopté est différent de celui de Shenzhen (dev. de l' industrie lourde interdi à Zhuhai) —> . Il faut que tu précise (dans une parenthèse qu'il s'agit de deux ZES)

⁴⁰ [gill2005bassin] rappelle l'importance de l'hinterland du Bassin Parisien et l'importance de ne pas considérer l'hypercentre de manière isolée, et considérer ainsi la MCR qui inclut un certain nombre de centres urbains importants à une heure de Paris : Chartres, Orléans, Rouen, Reims et Lille grâce à la grande vitesse.

Gouvernance de la MCR

[Ye2014200] analyse les actions de gouvernance métropolitaine à l'échelle de centres de la MCR, et plus particulièrement comment les communes de Guangzhou et Foshan ont progressivement accru leur coopération pour former une zone métropolitaine intégrée, pouvant ainsi fortement influencer le développement des transports par exemple et permettant la mise en place d'un réseau connecté. Une forte tension entre des processus émergents par le bas, et un dirigisme d'état relativement fort en Chine, se répercutant de l'État central, au gouvernement provincial jusqu'aux gouvernements locaux, a permis la mise en place d'une telle structure. La compétition avec les autres villes de la MCR reste très forte, et la logique d'intégration (au sens d'articulation entre les différents centres, d'interactions et de flux entre ceux-ci) de la MCR est partiellement guidée par la région seulement. La nature particulière des ZES de Shenzhen et Zhuhai, liée aux relations privilégiées avec les Zones Administratives Spéciales de Hong-Kong et Macao, qui n'ont été réintégrées à la République Populaire qu'à la fin du millénaire et conservent un certain niveau d'indépendance en termes de gouvernance, complique encore les jeux d'acteurs au sein de la région. La question de la correspondance entre certains niveaux de gouvernance et des processus urbains est épingleuse : [liao2017ouverture] interprète les transferts progressifs des initiatives économiques du pouvoir central vers les autorités locales comme une forme de gouvernance multi-niveaux.

C (CL) : Ajouter plus haut "Zone Economique Spéciale" (ZES)

C : district panyu

Gouvernance des transports

Dans le cadre des transports pour la MCR, il n'existe pas d'autorité spécifique à cette échelle pour l'organisation des transports (mais bien des entités au niveau de l'Etat, de la Province et des Communes), et chaque commune gère indépendamment le réseau local, tandis que les connections entre villes sont assurées par le réseau de train national. Cela conduit à des situations particulières dans lesquelles des zones se retrouvent très enclavées, avec une hétérogénéité très forte localement. Ainsi, la pointe sud de la ville de Guangzhou qui sert d'accès direct à la mer, est plus proche géographiquement du centre de Zhongshan, mais un lien direct par transports en commun est difficile à envisager, alors que la zone est bien reliée au centre de Guangzhou par la ligne de métro. Une situation similaire s'observe au terminus de la ligne 11 de Shenzhen, pour le quartier limitrophe de Dongguan, ce dernier étant très peu accessible en transports en commun⁴¹. Cette situation serait cependant transitoire, étant donné les infrastructures déjà en construction et celles planifiées sur un plus

⁴¹ Voir la carte 4 pour les localisations, celle-ci donnant par ailleurs les accessibilités par réseau routier.

TABLE 1 : **Transports en commun dans le Delta de la Rivière des Perles.** Nous donnons les populations en 2010 issues de [yearbook2013guangdong]. Les kilométrages sont issus des différents documents de planification pour les metros de Guangzhou [guangzhou2016metro], Shenzhen [shenzhen2016plan] et Dongguan [dongguan2017ditie] et pour le tramway de Zhuhai [zhuhai2016tram]. Zhongshan n'est pas incluse car exploite un système de bus en site propre mais pas d'infrastructure lourde.

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)	1.2Mio	10km	173km

long terme : le métro de Shenzhen, qui couvre aujourd’hui 285km, est planifié pour atteindre jusqu’à 30 lignes et une longueur d’environ 1100km⁴² en 2030 comme déclaré par le plan officiel de la ville [shenzhen2016plan]. Il est clair que ces développements suivent pour la majorité un développement urbain existant, une question cruciale est la volonté et la capacité à contenir l’étalement urbain et structurer les futurs développements autour de ce nouveau réseau, dans l’esprit d’une intégration volontaire entre urbanisme et transport de type *Transit Oriented Development* que nous avons introduit précédemment. Différents terminus seront connectés au metro de Dongguan, et de nouvelles lignes intercités structureront les déplacements de plus longue portée, ce qui fera du Delta dans un horizon temporel proche une MCR relativement bien intégrée en termes de transports en communs. Pour se donner une idée du développement du réseau dans les années à venir, la Table 1 donne la taille des réseaux planifiés dans les différentes villes d’ici 2030.

C (CL) : D'où viennent ces chiffres ? Je ne connaît pas les chiffres exactes des autres villes mais je suis sûre des chiffres de Zhuhai car j'ai récemment travaillé dessous. La population permanente de Zhuhai en 2010 selon les chiffre officielle était de : 1,562,530 (hukou pop. : 1,048,398; Pop. Flottante : 514,132) * www.stats-zh.gov.cn

Impact du Pont Zhuhai-Hong-Kong-Macao

Un projet majeur d’infrastructure de transport dans la région est le pont-tunnel fermant l’embouchure du Delta, reliant Zhuhai et Macao à Hong-Kong (HZMB). La longueur de la traversée est de 36,5km, ce qui en fait un ouvrage d’art exceptionnel [hussain2011hong]. L’ouverture au traffic a été retardée de plusieurs années et est prévue finalement pour fin 2017⁴³. [zhou2016medium] montre que les changements de motifs d’accessibilité attendus pour l’Ouest du Delta sont re-

42 A titre de comparaison, le réseau Transilien a une longueur avoisinant les 1300km en incluant les lignes RER, ce qui pourrait les rendre comparable, mais il faut garder à l'esprit que l'Ile-de-France a une surface de 12000km² contre 2000km² pour Shenzhen. Cela implique pour Shenzhen une densité de desserte bien plus haute, correspondant aux zones de fortes densité urbaine, si bien que le plan prévoit 70% de transit par métro à l'horizon 2030.

43 voir <http://www.hzmb.org/cn/default.asp>

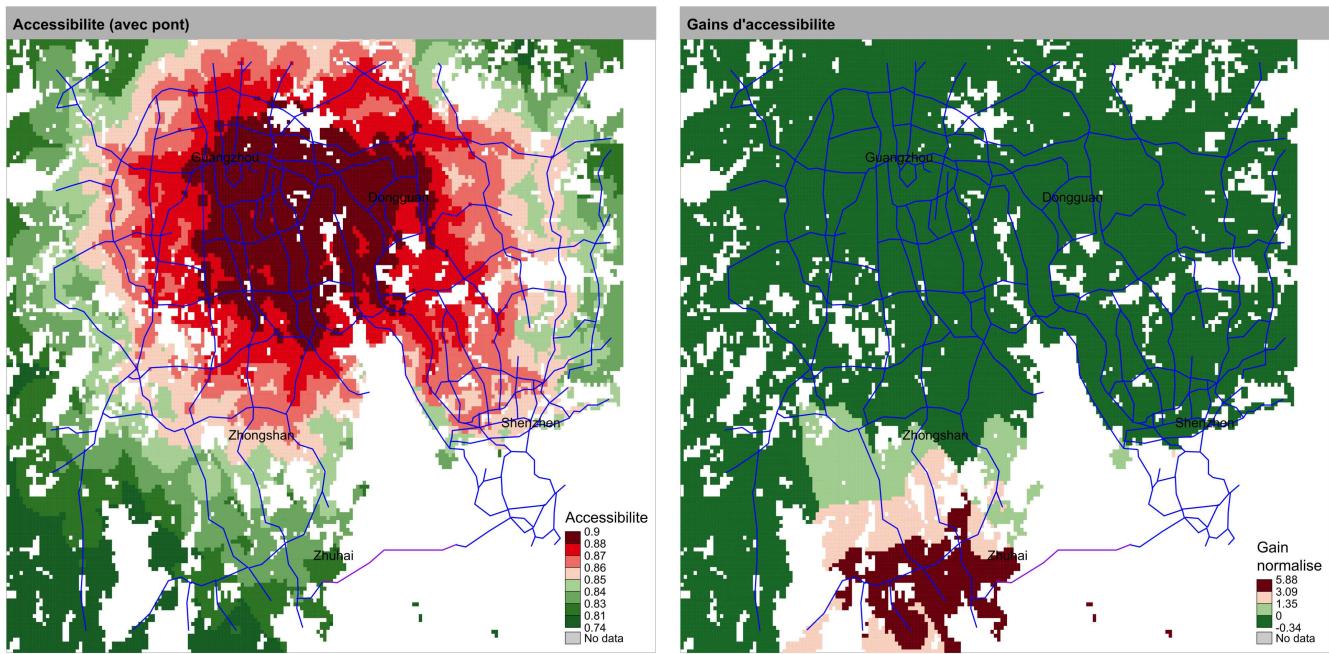


FIGURE 4 : Gain d'accessibilité permis par le HZMB dans le Delta de la Rivière des Perles, pour le territoire de Chine continentale. (Gauche) Accessibilité à la population Z_i ; (Droite) Gains normalisés d'accessibilité. La population de Hong-Kong est prise en compte dans les points de destination. Le réseau autoroutier (2017) est cartographié en bleu et le nouveau lien du pont en violet.

lativement forts, et ceux-ci peuvent potentiellement induire de fortes bifurcations dans les trajectoires des villes. La nécessité du projet est défendue par les différents porteurs du projet (Province du Guangdong, Région Administrative Spéciale de Hong-Kong, Région Administrative Spéciale de Macao) par des arguments de fort bénéfice économique dans le cadre des politiques d'ouverture, ainsi que par un bénéfice social pour l'Ouest notamment. Par exemple, Zhuhai se positionne comme un nouveau pivot entre Hong-Kong et l'ouest. L'équilibrage d'accessibilité, au sens de la diminution des inégalités spatiales d'accessibilité, s'opère cependant pour le mode routier uniquement, ce qui conduit à questionner ses impacts potentiels : d'une part l'accès à l'automobile reste réservé à une partie de la population seulement, d'autre part les effets négatifs de la congestion peuvent rapidement modérer les gains d'accessibilité. Ces gains d'accessibilité sont cartographiés suivant la même méthode que précédemment, et montrés avec l'accessibilité Z_i elle-même en Fig. 4.

Les impacts à moyen et long terme du pont sont ainsi difficiles à estimer. [wu2012impact] trouve des motifs similaires à ceux que nous estimons, c'est à dire un bénéfice significatif pour Zhuhai (et Hong-Kong que nous n'avons pas pris en compte), ainsi que des effets immédiats de modification de traffic et des impacts économiques liés au péage ou à l'accroissement du tourisme. Ils postulent surtout la po-

sition de Zhuhai-Macao comme un nouveau pivot dans la région. Si cela est vérifiable immédiatement en termes de centralité et d'accessibilité, il n'est pas dit que cette nouvelle position influence particulièrement la trajectoire socio-économique de Zhuhai. Un accompagnement politique particulier passant par une collaboration accrue entre Hong-Kong, Zhuhai et Macao sera importante [zhou2016medium]. Des effets économiques immédiats sont attendus, comme une augmentation des résidents de Zhuhai travaillant à Hong-Kong (les habitants de Zhuhai sont les seuls de la région à bénéficier d'une carte spéciale leur permettant de se rendre régulièrement dans les Zones Administratives Spéciales⁴⁴), mais le contraire, comme des investissements de Hong-Kong vers l'ouest du Delta n'ont pas de raison d'être systématiques : le premier cas prolonge la dynamique déjà existante avec Macao, le second est à construire en grande partie. Ainsi, cet exemple est un cas typique de notre problématique générale.

Perspectives

Une piste d'exploration passant par la modélisation consiste à poser le problème différemment et de chercher comprendre la dynamique du système métropolitain de manière intégrée, c'est à dire comme un système territorial en notre sens, dans lequel le couplage fort entre territoire et réseau est opéré par une ontologie propre des entités de gouvernance. Celle-ci sera l'objet de la section ??.

Cette deuxième étude plus brève nous a permis de mettre en valeur une structure de gouvernance fondamentalement différente, mais la même idée d'un projet de transport considérable modifiant profondément les motifs d'accessibilité. Les attentes des acteurs quant aux mutations territoriales potentiellement induites sont comparables au sens qu'une forte attente est mise dans le projet.

1.2.3 Comparabilité des études de cas

Nous avons étudié ici deux cas de développement métropolitain et de projets d'infrastructures dans leurs cadres. La possibilité de transfert des modèles urbains (comme le TOD), au sens de l'applicabilité de cadres génériques à des contextes géographiques différents, est généralement délicate. La synthèse de conclusions empiriques issus de cas d'étude très éloignés l'est également.

La particularité Est-asiatique a déjà été montrée pour la structure économique, et comment celle-ci ne peut être interprétée de manière simple par une séparation des processus microscopiques et macroscopiques comme certaines lectures rapides et idéologiquement orientée ont pu le faire, comme la vision de la Banque Mondiale [amsden1994isn]. La comparabilité de systèmes urbains est une question ouverte au

⁴⁴ Source : sortie de terrain du 06/11/2016 avec C. Losavio (voir A.2).

centre des enjeux de la Théorie Evolutive Urbaine. Celle-ci est liée au caractère ergodique de ces systèmes : l'hypothèse d'ergodicité postule que la trajectoire d'une ville dans le temps capture l'ensemble des états urbains possibles, et ainsi que les différentes villes sont différentes manifestations du même processus stochastique à différentes périodes. Dans ce cas, un ensemble de villes permettrait de se faire une idée des trajectoires temporelles. Intuitivement ce n'est pas le cas, et les systèmes urbains seraient plutôt non-ergodiques [**pumain2012urban**]. Empiriquement, cette non-correspondance entre statistiques globales et dynamiques individuelles des villes est montrée pour des données de traffic par [**2017arXiv171009559D**]. Ainsi il s'agira de rester prudent pour la généralisation des conclusions, autant empiriques que théoriques, ou issues de la modélisation.

★ ★

★

1.3 ELEMENTS DE TERRAIN

Cette section propose d'illustrer la problématique des interactions entre réseaux de transports et territoires, et plus particulièrement leur complexité et la diversité des situations possibles déjà perceptibles de manière qualitative (et également subjective dans un second temps) à l'échelle microscopique, par des exemples concrets de terrain. Le terrain géographique est le Delta de la Rivière des Perles en Chine, dans la province du Guangdong, que nous avons décrit ci-dessus, et plus particulièrement en grande partie la ville de Zhuhai. L'objectif est d'enrichir notre répertoire par des situations concrètes, de voir si celles-ci peuvent être associées aux processus génériques que nous avons déjà dégagé, ou si d'autres se manifestent aux échelles d'observation.

Nous assumons le terme de *Terrain géographique*, en toute conscience des débats épistémologiques que peuvent poser l'utilisation de celui-ci. En effet, on extrait des observations de lieux expérimentés, dans le cadre d'une problématique particulière [[retaille2010terrain](#)]. Notre démarche appuiera aussi sur le rôle des représentations, souligné comme forme à part entière de terrain par [[lefort2012terrain](#)], lorsque nous prendrons une position subjective.

Dans le cadre du projet européen Medium⁴⁵, visant à une approche interdisciplinaire de la soutenabilité pour les villes Chinoises en se concentrant sur les villes moyennes, cette ville a été choisie comme cas d'étude. Lorsque la source n'est pas explicitement précisée, les observations proviennent du travail de terrain, pour lequel des compte-rendus narratifs sont disponibles en Annexe A.2. Le protocole de sélection des objets et des lieux, ainsi que le protocole d'observation, sont donnés aussi. Le format des compte-rendus narratifs est "à-la-volée" suivant les recommandations de [[goffman1989fieldwork](#)] pour la prise de notes en terrain d'immersion notamment, tandis que la position volontairement subjective rejoint [[ball1990self](#)] qui souligne l'importance de la réflexivité pour tirer des conclusions rigoureuses à partir d'observations qualitatives de terrain duquel le chercheur est partie intégrante⁴⁶.

C (CL) : Tu devrais ajouter une note sur la définition de villes moyennes utilisée dans le projet (voir interview de Natacha sur le site de MEdium)

⁴⁵ Le projet Medium, qui met en partenariat des universités européennes et chinoises, s'intitule "New pathways for sustainable urban development in China's medium-sized cities". Il vise à étudier la soutenabilité selon un prisme interdisciplinaire et multidimensionnel, dans le cas de zones urbaines en forte croissance. Trois villes moyennes chinoises ont été choisies comme cas d'étude. Voir <http://mediumcities-china.org/> pour plus d'information.

⁴⁶ La considération du chercheur comme *sujet* en relation avec son objet d'étude n'implique pas dans notre cas de rétroaction du chercheur sur le système vu l'ampleur de celui-ci dans le cas d'un réseau de transport à l'échelle d'une ville, mais bien un conditionnement des observations par une subjectivité dont il s'agit de se détacher dans l'exploitation postérieure du matériel d'observation, mais qu'ignorer ne peut qu'augmenter les biais.

1.3.1 Développement d'un réseau de transport

L'objectif du travail de terrain est donc d'observer les multiples facettes et couches d'un système de transport public complexe et en mutation permanente, ses liens avec les opérations urbaines visibles, et dans quelle mesure ceux-ci témoignent de processus d'interaction entre réseaux et territoires. La portée des observations s'étend sur Zhuhai comme illustration des transports locaux mais aussi ponctuellement sur d'autres régions en Chine. Ces observations ont une logique propre en comparaison de la modélisation des réseaux de transport ou l'analyse de données, comme des études d'accessibilité ou des modèles d'interaction entre usage du sol et transport, qui seront menés par la suite. En effet, celles-ci échouent généralement à capturer des aspects à une grande échelle, souvent directement liés à l'utilisateur, qui peuvent devenir cruciaux au regard de l'utilisation effective du réseau. Par exemple, la multi-modalité⁴⁷ peut être rendue efficace en pratique par l'émergence de modes de transports auto-organisés informels, ou la mise en place de nouveaux modes comme le vélo en partage, ce qui résout le "problème du dernier kilomètre" [liu2012solving], qui semble être souvent négligé dans la planification de zones nouvellement développées en Chine. Au contraire, des détails pratiques comme la réservation des tickets ou les délais d'enregistrement à l'embarquement peuvent influencer considérablement les motifs d'usage.

Différents voyages sur le territoire Chinois ont été effectués pour observer les manifestations concrètes du développement du réseau à grande vitesse. Depuis 2008, la Chine a établi le plus grand réseau de HSR du monde à partir de zéro, qui a connu un grand succès et dont les lignes sont actuellement saturées. Celui-ci répond à des motifs de demande primaires en termes de taille de ville, montrant qu'il a été planifié de telle façon que le réseau réponde à des dynamiques territoriales. Son fort usage montre l'impact du réseau sur la mobilité, possible précurseur de mutations territoriales.

Pour montrer dans quelle mesure les territoires peuvent affecter le développement des réseaux de manière diverse, prenons un exemple particulier, lié au développement du tourisme, qui correspond à une dimension particulière qui a été prise en compte dans la planification. Ainsi, la ligne entre Guangzhou et Guiyang (axe nord-ouest précurseur de la future liaison directe Guangzhou-Chengdu) a vu la construction de stations spécifiques au développement du tourisme, comme Yangshuo dans le Guangxi, dont la fréquentation a alors for-

⁴⁷ La multi-modalité consiste en la combinaison de différents modes de transports : routier, train, métropolitain, tramway, bus, modes doux, etc., dans un motif de mobilité. Un système de transport multimodal consiste en la superposition des couches modales, et celles-ci peuvent plus ou moins bien s'articuler pour la production de trajets optimaux selon de multiples objectifs (coût, temps, coût généralisé, confort, etc.) qui eux-même dépendent de l'individu, du motif de déplacement.

tement augmenté. Un an après l'ouverture de la gare, le lien routier majeur avec la ville est toujours en construction, montrant que les différents réseaux réagissent différemment aux contraintes à différents niveaux. Un grand nombre de trains s'y arrêtent toutefois le week-end - plus d'un par heure, et sont remplis plus de deux semaines en avance. De nouveaux motifs de mobilité peuvent être induits par cette nouvelle offre, comme l'illustre l'interview d'une habitante de Guangzhou faite à Yangshuo, qui était venue pour un court week-end avec ses collègues, dans le cadre d'un voyage de "team-building" financé par sa startup en technologie de l'information. Ces nouvelles pratiques de mobilité sont montrées par une deuxième interview d'une habitante de Pékin rencontrée à Emeishan, envoyée par son entreprise de Design Industriel pour un court passage à Chengdu pour une formation dans une filiale locale. L'entreprise privilégie le train à grande vitesse, et celle-ci a récemment accru ses pratiques de mobilité pour ses salariés.

Une stratégie similaire peut s'observer concernant la desserte de destinations touristiques pour la ligne Chengdu-Emeishan. L'objectif principal de cette ligne est pour l'instant de desservir les destinations touristiques très fréquentée d'Emeishan et de Leshan. Cependant, le lien manquant entre Leshan et Guiyang est déjà bien avancé dans sa construction et complétera le lien direct entre Guangzhou et Chengdu. Cela révèle des dynamiques diachroniques et complémentaires de développement du réseau en fonction des propriétés de territoires. Cette ligne fait partie du squelette structurant des "8+8" reformulées récemment par le gouvernement central⁴⁸, et les territoires traverses en attendant beaucoup comme le montre [lu2012chengdu] pour la ville de Yibin à mi-chemin entre Chengdu et Guiyang.

Nous observons également des mutations conjointes du réseau ferroviaire et de la ville. Nous illustrons ainsi en Fig. 5 l'insertion du HSR dans ses territoires. Des effets directs du réseau sont liés au développement de quartiers totalement neufs aux alentours des nouvelles gares, parfois dans une démarche de type TOD - nous y reviendrons plus en détail. De plus, des effets indirects plus subtils sont suggérés par des indices comme la promotion des opérations par la publicité. Celle-ci montre les attentes socio-économiques envers le réseau et les agents locaux qui se doivent contribuer à son succès : les publicités vantant les mérites de la grande vitesse, et la vente d'appartement dans des opérations immobilières associées. Cette dynamique semble contribuer à la construction d'une "classe moyenne" et au rôle qu'elle doit jouer dans le dynamisme des territoires [rocca2008power]⁴⁹. L'in-

⁴⁸ Il s'agit du plan général pour les futures lignes à grande vitesse, réactualisé récemment pour comprendre 8 parallèles nord-sud et 8 autres est-ouest, complétant les 4+4 déjà réalisées.

⁴⁹ Construction, comme le souligne JEAN-LOUIS ROCCA, autant concrète car relevant de certaines réalités objectives, qu'imaginaire dans les discours universitaires et politiques, qui construisent l'objet simultanément à son étude ou son utilisation.

sertion des lignes dans les territoires semble dans certains cas forcée, comme le montre la gare de Yangshuo qui exploite l'opportunité du tourisme offerte par le passage de la ligne dans une zone très peu peuplée mais très attractive par ses paysages, ou les nouvelles opérations immobilières peu accessibles par leur prix à Zhuhai.

Enfin, il est important de noter que le développement du réseau répond simultanément à différents types de contexte territoriaux. Des branches à courte portée du nouveau réseau à grande vitesse, comme la ligne Guangzhou-Zhuhai, peuvent être vues comme à l'intermédiaire entre un service à longue distance et un transport régional de proximité, en fonction de la modularité des motifs de desserte. Cette ligne s'inscrit ainsi dans des interactions urbaines à longue portée (le service Zhuhai-Guiyang étant par exemple assuré) et dans des interactions au sein de la Mega-city Region, l'essentiel de la desserte étant des trains pour Guangzhou. A cela s'ajoute le réseau de train classique qui conserve un certain rôle dans les interactions territoriales : certaines connexions requièrent l'utilisation des deux réseaux et des transports urbains, comme la liaison entre Zhuhai et Hong-Kong expérimentée par voie terrestre seulement⁵⁰.

1.3.2 *Implémentation du TOD : des illustrations contrastées*

Le développement simultané du réseau de transport et de l'environnement urbain est directement observable sur le terrain. Le réseau urbain local et les opérations de développement immobilières sont planifiés en étroite conjonction avec le nouveau réseau de train : le tramway de Zhuhai, pour lequel une unique ligne est aujourd'hui ouverte et en phase de test, vise à participer à une approche par "Transit-oriented Development" (TOD)⁵¹ du développement urbain qui vise à favoriser l'utilisation des transports publics et une ville avec moins d'automobiles, comme voulu par exemple par le Comité de Planification de la *High-Tech Zone* en charge du développement autour de la gare nord de Zhuhai. L'observation des alentours de la gare de Tangjia, également construite dans le même esprit, une certaine atmosphère de désertion et une organisation peu pratique peut mener au questionnement de l'efficacité de l'approche. Cela suggère également une certaine nature auto-prophétique du projet, comme suggéré par les publicités pour un nouvel immobilier à vendre, apuyant sur l'importance de la présence de la ligne ferroviaire. Toute une narration incitant les acteurs locaux et les individus à s'impli-

C (CL) : Tu as cité au paravant -dans la partie que j'ai lu - le "TOD" mais par un lecteur - comme moi - qui ne sait pas de quoi il s'agit il faudra (au moins que tu ne l'aises déjà fait dans une partie précédente de la thèse) préciser que tu parles du "Transit-oriented Development"

⁵⁰ À la suite du Typhon Hato le 23/08/2017, les liaisons maritimes avec le centre de Hong-Kong et l'aéroport international ont été interrompues pour une grande partie du delta, et n'ont été rétablies pour Zhuhai que début novembre 2017.

⁵¹ Voir les travaux préliminaires de consultation pour la planification, comme par exemple <https://wenku.baidu.com/view/b1526461ff00bed5b8f31d01.html> pour le contexte du nouveau quartier de Xiaozhen, à l'ouest de Xiangzhou.



FIGURE 5 : Manifestations locales des mutations induites par le nouveau réseau à grande vitesse. (Haut Gauche)
 Gare à grande vitesse de Tangjia, sur la commune de Zhuhai. La publicité monumentale pour une opération immobilière vante les mérites d'une proximité au réseau, qui est également utilisée comme un argument pour des prix plus élevés ; (**Haut Droite**) Ligne à grande vitesse à Zhuhai, arrêt de bus déserté et projet immobilier en cours de réalisation dans une zone difficilement accessible : cette frange urbaine est en contact direct avec le milieu rural de l'autre côté de la ligne, et excentrée de la ville ; (**Bas Gauche**) La gare de Yangshuo sur la ligne Guangzhou-Guiyang, dont la principale fonction est le développement de cette destination touristique qui base la majorité de son économie sur ce domaine ; (**Bas Droite**) Publicité pour la grande vitesse dans le Sichuan, à la gare de l'aéroport international de Chengdu sur la ligne vers Leshan et Emeishan. Le train quitte la ville futuriste pour survoler la campagne, rappelant l'effet tunnel des territoires intermédiaires télescopés par la grande vitesse.

quer autour du TOD semble être utilisée par différents acteurs du développement.

D'autres observations de terrain, comme dans les Nouveaux Territoires (*New Territories*) à Hong-Kong, témoignent d'un TOD efficace et réalisant son objectif, avec une complémentarité entre transport lourd et tramway local léger, ainsi qu'une grande densité urbaine autour des gares. Ces observations rappellent la complexité des trajectoires urbaines couplées au développement du réseau, et qu'il s'agit d'être prudent avant de tirer toute conclusion générale à partir de cas particuliers. Nous résumons en Fig. 6 la comparaison des deux cas de TOD évoqués ci-dessus, sous forme de schéma synthétique des grandes lignes urbanistiques de chacune des zones. A Hong-Kong, les zones urbaines ont été planifiées conjointement avec la ligne du MTR (transport lourd) et les multiples lignes de tramway léger [hui2005study]. L'infrastructure du transport léger et l'organisation des missions permettent de rejoindre rapidement la gare la plus proche, distribuant une accessibilité très uniforme pour l'ensemble des quartiers du territoire. Au contraire à Zhuhai, le village de Tangjia est ancien, antérieur même à l'ensemble du reste de Zhuhai, qui s'est développé sans articulation particulière avec les infrastructures de transport. Le tracé du tramway, qui vient d'ouvrir, complète le tracé de la nouvelle ligne ferroviaire, dans un but de reorganisation du nord de Zhuhai, et en particulier la High-tech Zone qui s'étend de la gare du Nord (Zhuhai Bei) à Tangjia. Actuellement, l'organisation urbaine est fortement marquée par cette mise en place déphasée, puisque l'accessibilité en transport en commun est toujours relativement faible, les lignes de bus étant sujettes à une congestion croissante due à la forte augmentation du nombre d'automobiles. Par ailleurs, la mise en place du Tramway a été laborieuse, de par l'utilisation d'une technologie par troisième rail au sol importée d'Europe, et qui n'avait jamais été testée dans de telles conditions d'humidité⁵², ce qui a poussé à une remise en question du plan du réseau dans son ensemble.

Cet exemple de terrain nous démontre ainsi que (i) sous la même qualification existent des processus très différents, extrêmement dépendants aux particularités géographiques, politiques, économiques ; et que (ii) la mise en place d'un territoire fonctionnel en termes d'accessibilité nécessite une articulation fine qui semble résulter d'une approche de planification intégrée réalisée sur le temps long.

1.3.3 Observation Flottante

Nous proposons finalement d'ébaucher une entrée qualitative et subjective d'un certain type, pour suggérer une façon de compléter nos connaissances et mieux cerner les processus de manière concrète.

⁵² Source : communication personnelle avec Yinghao Li, juillet 2017.



FIGURE 6 : Analyse comparative de deux implémentations du TOD en PRD. A une échelle comparable, nous synthétisons la configuration urbaine de Yuenlong (元朗) et Tuenmun (屯門), Hong-Kong New Territories (香港, 新界), à gauche, et de Xinwan, Xiangzhou, Zhuhai (珠海, 香洲, 新湾), à droite, qui contient la High-Tech zone de Zhuhai dans sa partie nord en particulier. Les configurations témoignent de dynamiques d'articulation différentes, et des temporalités de construction décalées, révélant ainsi diverses réalités sous la notion de TOD. Une première interprétation serait que celle-ci est efficace si la trajectoire du système territorial complet (aménagement urbain et réseau de transport) est infléchie tôt dans sa genèse, tandis qu'un système avec un certain niveau de maturité sera plus inerte. Trad. : 到香港 - vers Hong-Kong; 到广州 - vers Guangzhou; 到珠海 - vers Zhuhai.

L'entrée prise suit la méthode *d'observation flottante*, introduite à l'interface de l'anthropologie et la sociologie par [petonnet1982observation], avec l'ambition de fonder une anthropologie urbaine, au sens de l'étude des comportements humains au sein d'un environnement urbain. Il ne s'agit pas exactement de la même idée que l'anthropologie de l'espace de Choay [choay2009pour] qui explore la direction inverse, c'est à dire le propre des sociétés humaines de façonner l'espace, et la capacité de construire un environnement bâti à différentes échelles par l'architecture et l'urbanisme. Notre contexte méthodologique est le suivant Répondant à un besoin de mouvement que le sédentaire éprouve facilement, le chercheur se place au centre du processus de production de connaissances, nous citons, en "rest[ant] en toute circonstance vacant et disponible, à ne pas mobiliser l'attention sur un objet précis, mais à la laisser flotter afin que les informations la pénètrent sans filtre, sans a priori, jusqu'à ce que des points de repère, des convergences, apparaissent et que l'on parvienne alors à découvrir des règles sous-jacentes". Cette méthode peut servir d'étude préliminaire pour fixer des protocoles et grilles précises d'entretien : elle est par exemple utilisée justement au sujet du transport par [de2012deplacements]. Nous nous en servons dans notre cas comme méthode d'extraction de faits stylisés, afin d'informer des exemples de processus d'interactions directement visibles.

MÉTHODE Les mouvements pendulaires à échelle intra-métropolitaine sont nécessairement vécus d'une façon particulière en comparaison à d'autres lieux géographiques et à d'autres échelles sur le même lieu. Et si une façon d'appréhender des faits stylisés particuliers était alors d'effectuer l'analogie d'une étude de perturbation sur le système, mais en prenant comme référentiel l'observateur lui-même ? Il s'agirait de faire porter un choc sur une situation "d'équilibre", puis de se laisser flotter au gré du courant pour appréhender la réaction et certains mécanismes qu'il aurait été difficile de considérer en suivant sa routine. Une expérience naturelle causée par une perturbation des transports (qui en région francilienne est bien courante, dans tous les cas plus qu'en Chine) est un événement provoquant une expérience naturelle, au sens où le chercheur peut capturer des situations et réactions individuelles particulières. Notre méthodologie est relativement simple : déambuler dans les transports en commun, avec ou sans but et de manière ou non aléatoire, mais en essayant sur chaque trajet de maximiser les opportunités de mise en situation ou de capture d'évènement, typiquement en évitant un trajet de routine⁵³. La répétition de l'expérience visera également à maximiser la couverture spatiale, temporelle, de situation. Une production traçable est en théorie nécessaire à chaque itération, qu'il s'agisse de description factuelle, de

53 Cette contrainte sera respectée dans notre cas pour le Guangdong, mais pas pour l'Ile-de-France.

description perçue, de semi-synthèse. Celle-ci permet a posteriori de voir les stratifications successives du vécu et des expériences d'observation progressivement raffinées dans leur contexte, et de tracer ainsi la genèse des idées induites. Nous faisons le choix de retranscrire l'aspect subjectif, voir maximiser celui-ci dans les synthèses générales des observations, afin d'appuyer cet aspect en contraste avec la suite de notre travail qui sera relativement déconnecté du sujet menant la recherche, et en echo avec les recommandations de [ball1990self] pour la place de la subjectivité dans la recherche ethnographique de terrain.

De par le choix de la méthode, les résultats de cette sous-section portent majoritairement sur les transports. Les interactions avec les territoires seront perçues majoritairement dans les pratiques de mobilité observées.

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 connaisseurs 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 quelque 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.

ENCADRÉ 2: Une expérience en observation flottante en région parisienne.

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. A 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.

ENCADRÉ 3: Une expérience en observation flottante, Guangdong, Zhuhai.

RÉSULTATS 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 Appendice A.2. 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 autre 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 grande 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 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 négatives⁵⁴.
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 épique 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 conclusion 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 onto-logiques et épistémologiques que nous prendrons par la suite.

⁵⁴ Cet effet étant par ailleurs nécessairement en interdépendance forte avec les propriétés culturelles, qui est en fait une composante fondamentale des territoires.

★ ★

★

TABLE 2 : Processus d'interaction entre réseaux et territoires. Nous synthétisons les processus selon les échelles et selon la typologie de sens.

	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 structuelle ; Bifurcations

SYNTHÈSE DES PROCESSUS ÉTUDIÉS

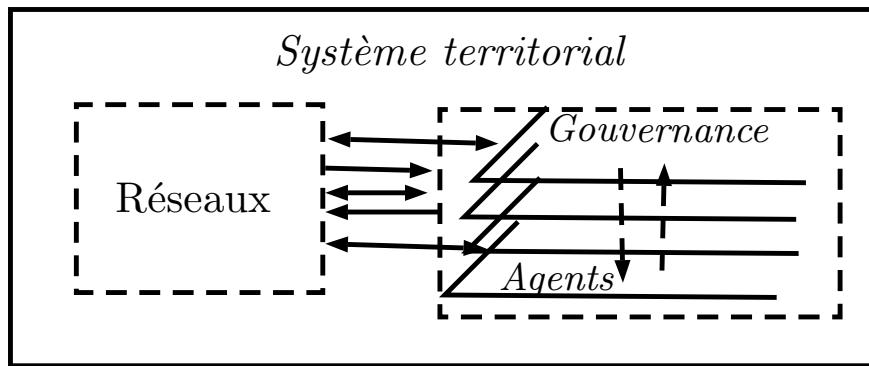
Nous concluons ce chapitre introductif par une synthèse et une mise en perspective des processus d'interaction identifiés par l'analyse théorique, empirique et la littérature. Celle-ci permettra de situer les revues des entreprises de modélisation auxquelles nous procéderons dans le chapitre 2, puis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

Une entrée par les échelles

Une première entrée pour synthétiser les processus abordés consiste à les considérer par échelle. On a 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-même multi-échelle (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, on a mis en évidence des processus pour lequel 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 2.

Une entrée par les acteurs

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 suivant :



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'autre aspects du territoire avec lesquels ils sont aussi couplés fortement.

Cette entrée peut être mise en perspective avec le cadre conceptuel de [le2010approche], 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 : 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 Chapitre2, et sera d'autre part complétée et précisée à l'issue de celui-ci.

CONCLUSION DU CHAPITRE

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. A différentes échelles temporelles typiques, l'année, la décennie et le siècle, correspondent des échelles spatiales : métropolitaine, régionale 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. A 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 [offner1993effets]. Par ailleurs, les multiples situations géographiques suggère 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 Evolutive 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 [west2017scaling] 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 (voir 9.3 pour un développement des domaines de connaissance et du rôle du modèle). Comme le rappelle J.P. MARCHANT¹, “*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 (voir notre positionnement scientifique en 3). Sans développer 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 dé-

¹ Communication personnelle, Mai 2017.

² Même si dans une relecture à la lumière de 9.3 ce positionnement n'a pas de sens puisque notre démarche contenait déjà des modèles à partir du moment où elle était scientifique.

marches de 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 citation 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-analayse (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 [raimbault2015models], puis pour sa deuxième partie la méthodologie introduite par [raimbault2016indirect] et développée dans [raimbault2017exploration] ainsi que les outils de [bergeaud2017classifying] ; et enfin est inédit pour sa dernière partie.

2.1 MODÉLISER LES INTERACTIONS

2.1.1 Modélisation en Géographie Quantitative

Histoire

La modélisation joue en Géographie Théorique et Quantitative un rôle fondamental. [cuya2014analyse] procède à une analyse spatio-temporelle du mouvement de la Géographie Théorique et Quantitative en langue française et souligne l'émergence de la discipline comme une combinaison d'analyses quantitatives (e.g. analyse spatiale et pratiques de modélisation et de simulation) et de construction théoriques. Cette dynamique est datée à la fin des années 70, et est intimement liée à l'utilisation et l'appropriation des outils mathématiques [pumain2002role]. L'intégration de ces deux composantes permet la construction de théories à partir de faits stylisés empiriques, qui produisent à leur tour des hypothèses théoriques pouvant être testées sur les données empiriques. Cette approche est née sous l'influence de la *New Geography* dans les pays Anglo-saxons et en Suède. Concernant la modélisation urbaine en elle-même, d'autre champs que la géographie ont proposé des modèles de simulation à peu près à la même période. Par exemple, le modèle de LOWRY, développé par [lowry1964model] dans un but appliqué immédiat à la région métropolitaine de Pittsburgh, suppose un système d'équations pour la localisation des actifs et des emplois dans différentes zones. Ce modèle a été une pierre angulaire de la modélisation urbaine, puisque comme le montre [goldner1971lowry] il avait déjà moins d'une dizaine d'années plus tard un conséquent héritage de développements conceptuels et opérationnels³. Des modèles relativement similaires sont toujours largement utilisés aujourd'hui.

Simulation de modèle et calcul intensif

Une histoire étendue de la genèse des modèles de simulation en géographie est faite par REY dans [rey2015plateforme] avec une attention particulière pour la notion de validation de modèles (nous reviendrons sur la place de ces aspects dans notre travail en 3). L'utilisation de ressources de calcul pour la simulation de modèles est antérieure à l'introduction des paradigmes de la complexité actuels, remontant par exemple à FORRESTER, informaticien qui a été pionnier des modèles d'économie spatiale inspirés par la cybernétique⁴. Avec l'augmentation des potentialités de calcul, des transformations

³ [goldner1971lowry] fait l'hypothèse que ce succès est du à la combinaison de trois facteurs : une possibilité d'application opérationnelle directe, une structure causale du modèle simple à appréhender (les actifs se localisent en fonction des emplois), et un cadre flexible pouvant être étendu ou adapté.

⁴ Celle-ci, ainsi que le courant systémique, sont comme nous l'avons déjà développé précurseurs des paradigmes actuels de la complexité.

épistémologiques ont également suivi, avec l'apparition de models explicatifs comme outils expérimentaux. REY compare le dynamisme des années soixante-dix quand les centres de calcul furent ouverts aux géographes à la démocratisation actuelle du Calcul Haute Performance⁵. Aujourd'hui, cette facilité d'accès consiste entre autres à du calcul sur grille dont l'utilisation est rendue transparente, c'est à dire sans besoin de compétences techniques pointues liées au mécanismes de la distribution des calculs.. Ainsi [schmitt2014half] donnent un exemple des possibilités offertes en termes de calibration et de validation de modèle, réduisant le temps de calcul nécessaire de 30 ans à une semaine - ces techniques jouent un rôle clé pour les résultats que nous obtiendrons par la suite. Cette évolution est également accompagnée par une évolution des pratiques [banos2013pour] et techniques [10.1371/journal.pone.0138212] de modélisation.

La modélisation, et en particulier les modèles de simulation, est vue par beaucoup comme une brique fondamentale de la connaissance : [livet2010] rappelle la combinaison des domaines empirique, conceptuel (théorique) et de la modélisation, avec des retroactions constructives entre chaque. Un modèle peut être un outil d'exploration pour tester des hypothèses, un outil empirique pour valider une théorie sur des jeux de données, un outil explicatif pour révéler des causalités et ainsi des processus internes au système, un outil constructif pour construire itérativement une théorie conjointement avec celle des modèles associés. Ce sont des exemples de fonctions parmi d'autres : [varenne2010simulations] propose une classification des diverses fonctions d'un modèle. Nous considérons la modélisation comme un instrument fondamental de connaissance des processus au sein d'un système, plus particulièrement dans notre cas au sein d'un système complexe adaptatif. Nous rappelons ainsi que notre question de recherche s'intéressera aux *modèles dont l'ontologie contient une part non négligeable d'interactions entre réseaux et territoires*.

2.1.2 Modéliser les territoires et réseaux

Développons à présent un aperçu des différentes approches modélisant des interactions entre réseaux de transport et territoires. Remarquons de manière préliminaire une forte contingence des constructions scientifiques sous-jacentes à celles-ci. En effet, selon [bretagnolle2002time], “les idées des spécialistes de la planification cherchant à donner des définitions des systèmes de ville, depuis 1830, sont étroitement liées aux transformations des réseaux de communication”. Le contexte historique (et donc socio-économique et technologique) conditionne fortement les

⁵ Le développement des premiers modèles de simulation urbaine coïncide avec l'ouverture des premiers centres de calcul aux sciences humaines, comme le rappelle par ailleurs PUMAIN (entretien du 31/03/2017) par exemple pour l'implémentation du modèle d'entropie d'ALLEN.

théories formulées. Cela implique que les ontologies et les modèles correspondants proposés par les géographes et les planificateurs sont fortement liés aux préoccupations historiques courantes, ce qui limite nécessairement leur portée théorique et/ou opérationnelle. Au delà de la question de la définition du système qui joue également un rôle central, on comprend bien l'impact que peut avoir cette influence sur la portée des modèles développés. Dans une vision perspectiviste de la science [**giere2010scientific**] de telles limites sont l'essence de l'entreprise scientifique, et comme nous démontrerons dans le chapitre 9 leur combinaison et couplage dans le cas de modèles est généralement une source de connaissance.

L'entrée que nous proposons ici pour dresser un aperçu des modèles est complémentaire de celle prise au Chapitre 1, en regardant par objet principal (c'est à dire les relations Réseau → Territoire, Territoire → Réseau et Territoire ↔ Réseau)⁶.

Le cadre de lecture des échelles est également celui proposé au Chapitre 1, sachant que nous ne nous intéressons pas aux échelles microscopiques par choix de ne pas considérer la mobilité quotidienne. On a ainsi schématiquement des échelles temporelles et spatiales mesoscopiques et des échelles macroscopiques.

Nous avons vu que la correspondance à des échelles temporelles et spatiales n'est pas systématique (voir la typologie provisoire à double entrée des processus). Par contre, celle à des domaines particuliers et à des acteurs l'est plus. Cette revue de littérature est donc orientée dans cette seconde direction.

Territoires

Le courant principal s'intéressant à la modélisation de l'influence du réseau de transport sur les territoires se trouve dans le champ de la planification, à des échelles spatiales et temporelles moyennes (les échelles de l'accessibilité métropolitaine que nous avons développées ci-dessus). Des modèles en géographie à d'autres échelles, comme les modèles Simpop déjà évoqués [**pumain2012multi**], ne supposent pas une ontologie particulière pour le réseau de transport, et s'ils incluent des réseaux entre les villes comme porteur des échanges, ne permettent cependant pas d'étudier en particulier les relations entre réseaux et territoires. Nous reviendrons plus loin sur des extensions pertinentes pour notre question. Revoyons pour commencer un contexte de modèles plus proches des études de planification.

⁶ Nous rappelons la signification de cette notation introduite au Chapitre 1 : une flèche directe signifie des processus qu'on peut attribuer relativement de manière univoque à l'origine, tandis qu'une flèche réciproque suppose intrinsèquement l'existence d'interactions réciproques, généralement en coinidence avec l'émergence d'entités jouant un rôle dans celles-ci.

MODÈLES LUTI Ces approches sont désignées de manière générale comme *modèles d'interaction entre usage du sol et transport (LUTI, pour Land-Use Transport Interaction)*. Il est entendu par usage du sol la répartition des activités territoriales, généralement réparties en typologies plus ou moins précises (par exemple logements, industrie, tertiaire, espace naturel). Ces travaux peuvent être difficiles à cerner car liés à différentes disciplines scientifiques⁷. Leur principe général est de modéliser et simuler l'évolution de la distribution spatiale des activités, en prenant le réseau de transport comme contexte et déterminant significatif des localisations. Pour comprendre le cadre conceptuel sous-jacent à la majorité des travaux, l'Encadré 4 résume celui issu de [wegener2004land]⁸.

Par exemple, du point de vue de l'Economie Urbaine, les propositions de tels modèles existent depuis un certain temps : [putman1975urban] rappelle le cadre d'économie urbaine où les principales composantes sont les emplois, la démographie et le transport, et passe en revue des modèles économiques de localisation qui s'apparentent au modèle de LOWRY déjà mentionné.

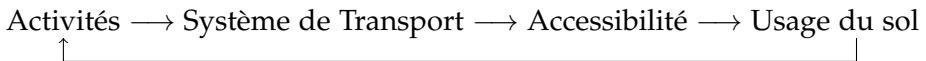
[wegener2004land] donne plus récemment un état de l'art des études empiriques et de modélisation sur ce type d'approche des interactions entre usage du sol et transport. Le positionnement théorique est plutôt proche des disciplines de la socio-économie des transports et de la planification (voir les paysages disciplinaires dressés en 2.2). [wegener2004land] compare et classe dix-sept modèles, parmi lesquels aucun n'inclut une évolution endogène du réseau de transport sur les échelles de temps relativement courtes (de l'ordre de la décennie) des simulations. On retrouve bien la correspondance avec les échelles typiquement mesoscopiques établies précédemment. Une revue complémentaire est faite par [chang2006models], élargissant le contexte avec l'inclusion de classes plus générales de modèles, comme des modèles d'interactions spatiales (parmi lesquels l'attribution du traffic et les modèles à quatre temps), les modèles de planification basés sur la recherche opérationnelle (optimisation des localisations des différentes activités, généralement résidences et emplois), les modèles microscopiques d'utilité aléatoire, et les modèles de marché foncier.

Afin de donner une meilleure intuition de la logique sous-jacente à certains modèles Luti, nous détaillons en Encadré 5 et en Encadré 6 la

⁷ Nous prenons le parti ici de rassembler de nombreuses approches ayant la caractéristique commune de modéliser principalement l'évolution de l'usage du sol, sur des échelles temporelles et spatiales moyennes. L'unité et le positionnement relatif de ces approches couvrant de l'économie à la planification, sont une question ouverte qui n'a à notre connaissance jamais été traitée de front. Le travail mené en 2.2 donne des pistes de réponse par une approche d'épistémologie quantitative.

⁸ Un cadre plus général que nous avons déjà développé, qui permet de faire le pont avec notre cadre, est celui de [le2010approche], qui replace le triptyque Système de Transport/Système de localisation/Système d'activités en relation avec les agents : agents demandeurs, agents aménageurs, facteurs externes.

[wegener2004land] pose un cadre général théorique et empirique pour les modèles d’interaction entre transport et usage du sol. Les quatre concepts mobilisés sont l’usage du sol, la localisation des activités, le système de transport et la distribution de l’accessibilité. Un cycle d’effets circulaires sont résumés dans la boucle suivante :



Le système de transport est supposé à *infrastructure fixe*, c'est à dire que les effets de la distribution des activités sont ceux sur *l'utilisation* du système de transport (donc liés à la *mobilité* dans notre cadre plus général) : choix modal, fréquence des voyages, longueur des voyages.

Les effets théoriquement attendus sont classés selon les directions de la relation (*Usage du sol* → *Transport* ou *Transport* → *Usage du sol*, ainsi qu’une boucle *Transport* → *Transport*, qui fait partie des éléments ignorés dans notre cas), et par ailleurs par facteur agissant (densité résidentielle, d’emplois, localisation, accessibilité, coûts de transport) ainsi que par aspect affecté (longueur et fréquence des voyages, choix de mode, densités, localisations). On peut par exemple citer :

- *Usage du sol* → *Transport* : une densité résidentielle minimale est nécessaire pour l’efficience du transport public, une concentration des emplois implique des voyages plus long, les villes plus grandes ont une part modale plus importante pour les transports en commun.
- *Transport* → *Usage du sol* : une forte accessibilité implique des prix plus élevés et un développement accru pour le résidentiel, les entreprises se localisent pour une meilleure accessibilité aux moyens de transport à grande échelle.
- *Transport* → *Transport* : les lieux avec une bonne accessibilité produiront plus et de plus longs voyages, le choix modal et le coût de transport sont fortement corrélés.

Ces effets théoriques sont par ailleurs comparés aux observations empiriques, qui pour la plupart donnent la manière d’implémentation du processus. Certains ne sont pas observés en pratique, tandis que la plupart sont en accord avec les attentes théoriques.

Commentaire 1 : Un cadre uniscalaire ? Ce cadre prend en compte schématiquement deux échelles principales, celle de la mobilité quotidienne et celle de la localisation des activités. Sachant qu’en pratique les comportements de mobilité sont généralement pris en compte sous forme de flux moyens, il se réduit souvent à une unique échelle mesoscopique. Dans tous les cas, il ne permet pas de tenir compte de dynamiques sur le temps plus long comprenant l’évolution de l’infrastructure du système de transport ou des dynamiques structurelles des systèmes de villes sur le temps long.

Commentaire 2 : Une vision systématique des effets structurants ? Par ailleurs, les pourfendeurs de la rhétorique des effets structurants trouveront en ce cadre leur bête noire, puisque les effets directs de l’accessibilité sur l’usage du sol puis sur la localisation des activités sont postulés ici. Ces critiques pourront être refoulées par l’observation qu’il s’agit des effets attendus théoriques, et que le cadre est mis en perspective des effets empiriques effectivement observés. Il sera à cependant à prendre avec précaution, en le situant toujours en terme de contexte et d’échelles.

ENCADRÉ 4: Cadre conceptuel des interactions entre transport et usage du sol selon [wegener2004land].

Le modèle Pirandello®^a est présenté dans [delons:hal-00319087] comme l'une des premières tentatives de développement de modèle Luti opérationnel en France. Le modèle se base sur quatre processus économiques fondamentaux : le marché du foncier et l'offre de logement, la mobilité résidentielle des ménages, l'attribution des destinations de déplacement, le choix modal. Le modèle est statique, c'est à dire calcule un équilibre pour les distributions spatiales des actifs et des emplois, ainsi que pour les flux de transport.

Les processus fondamentaux pris en compte et leur implémentation sont les suivants :

- Les choix résidentiels des ménages se basent sur une fonction d'utilité prenant en compte (i) un terme de confort en Cobb-Douglas de la surface et du revenu, corrigée par une préférence linéaire pour les logements individuels ; (ii) un terme d'accessibilité basé sur le coût généralisé (agrégation du coût de transport et du temps, avec un prix du temps) ; (iii) le prix du logement et de la taxe locale en fonction de la surface de logement ; (iv) un effet fixe par revenu et par zone ; et (v) un terme aléatoire supposé suivre une loi de Gumbel. Les probabilités de localisation pour une tranche de revenus sont alors données par un modèle de choix discret étant donné cette utilité.
- Les prix du logement sont formés selon une loi d'échelle de la population.
- Un système d'enchère local répond à la demande obtenue précédemment, en fonction d'une offre de logement exogène.
- Les entreprises se localisent en maximisant leur profit, fonction de la productivité (Cobb-Douglas du salaire et de l'accessibilité) et du prix du foncier, sous contrainte d'une distribution spatiale fixée du nombre d'emplois, de la surface de bureaux, et de la production totale de la région.
- Le transport est pris en compte par un modèle à quatre étapes, qui distribue les choix modaux et les choix de destination par un modèle de choix discrets, et les flux assignés selon un équilibre de Wardrop (voir 8.2), ce qui permet d'ajuster les valeurs de l'accessibilité étant donné une distribution spatiale des activités.

Le mécanisme de combinaison de ces différents processus pour obtenir un équilibre global est détaillé par [kryvobokov2013comparison], et consiste à l'établissement de trois sous-équilibres à différentes échelles : flux de transport (donnant les coûts) sur le court terme, localisation et prix de l'immobilier sur le moyen terme, prix du foncier et terrain disponibles (fixés de manière exogène pour l'ensemble de la période modélisée).

Commentaire : Equilibre, modèle opérationnel et calibration. Un certain nombre de remarques peuvent être faites à ce modèle, les plus importantes pour notre approche étant : (i) l'hypothèse d'équilibre peut être un outil puissant pour comprendre la structure des attracteurs du système, mais n'a pas de fondement empirique, et encore moins pour le couplage d'équilibres à différentes échelles ; (ii) ainsi, la nature opérationnelle du modèle est discutable, puisque l'étude de l'impact de scénarios sur les déplacements des attracteurs permet difficilement d'inférer sur les dynamiques locales du système ; et (iii) les sous-modèles sont calibrés plus ou moins rigoureusement et relativement séparément, or les conditions d'un calibrage par décomposition sont une question ouverte encore peu explorée et liée à la nature du couplage de modèles. En notre sens, un tel modèle micro-fondé serait dans tous les cas en meilleure cohérence avec une philosophie de modélisation générative dynamique et de parcimonie (voir 3.1).

^a L'origine du nom n'est pas donnée, mais suggère fortement l'influence de ses créateurs originaux V. PIRON et J. DELONS.
ENCADRÉ 5: **Le modèle Pirandello.**

Le modèle Nendum2D, décrit en détail dans [viguie2014downscaling], se concentre sur la localisation des actifs et leur interaction avec la rente foncière et les promoteurs immobiliers : il s'agit d'un modèle inspiré du modèle de Fujita-Ogawa [fujita1982multiple], héritant de la littérature en Economie Urbaine.

Les processus inclus dans le modèle sont, chacun ayant une échelle de temps particulière fixée par un paramètre :

- Les ménages font un compromis entre surface de logement et budget disponible hors coûts de transports et loyer, suivant une fonction de Cobb-Douglas pour l'utilité correspondante. Ce processus induit une dynamique pour la surface des logements en fonction de la distance au centre.
- Ils se relocalisent pour avoir une utilité moyenne plus grande que la moyenne.
- Les loyers évoluent pour maximiser l'occupation ou en réponse à une demande extérieure.
- De nouveaux bâtiments sont construits par des promoteurs cherchant à maximiser leur profits.

Ce modèle est dynamique et simule l'évolution de ces différentes variables dans l'espace (la formulation ci-dessous est monocentrique, une variante polycentrique et prenant en compte une distribution exogène d'emplois existent) et dans le temps. Son échelle spatiale est métropolitaine, et l'échelle de temps peut s'étendre d'une échelle moyenne (décade) à du temps plus long (siècle), sachant que cette dernière est peu crédible puisque qu'elle garde statique nombreuses autres composantes du système urbain.

Commentaire : extension des ontologies. Le couplage de Nendum avec un modèle d'attribution de transport, le modèle Modus^a, vise à inclure la retroaction de la congestion dans le système de transport sur les coûts, et donc sur la localisation et la structure urbaine. Des questions fondamentales se dégagent des premières expériences de couplage :

- Le schéma directeur est-il vraiment utile, puisqu'il ne semble qu'accompagner des dynamiques déjà présentes. En d'autres termes, *le processus de gouvernance est-il endogène ?* Le Sdrif capture-t-il en fait une dynamique intrinsèque sur le temps plus long ?
- Le couplage des modèles pose en lui-même des difficultés techniques, de communication entre des modules déjà implémentés dans différents langages et de convergence du modèle couplé en un nombre raisonnable d'itérations.
- Il pose d'autre part des difficultés ontologiques : chaque modèle inclut des mécanismes opposés pour la même ontologie (effet d'agrégation contre congestion pour la distribution de la population). La question se pose s'il faut rajouter spécifiquement une ontologie de couplage (par exemple des équations spécifiques intégrant ces effets contradictoires), pour permettre d'une part une meilleure convergence, d'autre part une meilleure cohérence ontologique.

^a Dans le cadre du projet en cours de réalisation ANR VITE ! (voir <http://www.agence-nationale-recherche.fr/Projet-ANR-14-CE22-0013>).

ENCADRÉ 6: **Le modèle Nendum.**

structure, les ontologies et les hypothèses de deux modèles développés dans le cas spécifique de l’Île-de-France (permettant d’une part la comparaison entre les deux et d’autre part donnant un écho au développement thématique de 1.2). Même pour des ontologies très proches (prix immobiliers, localisation des ménages), on voit la variété d’hypothèses possibles et de problématiques soulevée par les modèles.

DES MODÈLES OPÉRATIONNELS TRÈS VARIÉS La variété des modèles existants a conduit à des comparaisons opérationnelles : [paulley1991overview] rendent compte d’un projet comparant différents modèles appliqués à différentes villes. Leurs résultats permettent d’un part de classifier des interventions en fonction de leur impact sur le niveau d’interaction entre transport et usage du sol, et d’autre part de montrer que l’effet des interventions dépend fortement de la taille de la ville et de ses caractéristiques socio-économiques.

Les ontologies des processus, et notamment sur la question de l’équilibre, sont aussi variées. Les avantages respectifs d’une approche statique (calcul d’un équilibre statique de la localisation des ménages pour une certaine spécification de leur fonctions d’utilité) et d’une approche dynamique (simulation hors équilibre des dynamiques résidentielles) a été étudié par [kryvobokov2013comparison], dans un cadre métropolitain sur des échelles de temps de l’ordre de la décennie. Les auteurs montrent que les résultats sont globalement comparables et que chaque modèle a son utilité selon la question posée.

Différents aspects du même système peuvent être traduits par divers modèles, comme le montre par exemple [wegener1991one], et le trafic, les dynamiques résidentielles et d’emploi, l’évolution de l’usage du sol en découlant, influencée aussi par un réseau de transport statique, sont généralement pris en compte. [iacono2008models] couvre un horizon similaire avec un développement supplémentaire sur les modèles à automates cellulaires d’évolution d’usage du sol et les modèles à base d’agents. La portée temporelle d’application de ces modèles, de l’ordre de la décennie, et leur nature opérationnelle les rend utiles pour la planification, ce qui est assez loin de notre souci d’obtenir des modèles explicatifs de processus géographiques. En effet, il est souvent plus pertinent pour un modèle utilisé en planification d’être lisible comme outil d’anticipation, voire de communication, que d’être fidèle aux processus territoriaux au prix d’une abstraction.

PERSPECTIVES POUR LES LUTI [timmermans2003saga] émet des doutes quant à la possibilité de modèles d’interaction réellement intégrés, c’est à dire produisant des motifs de transports endogènes et se détachant d’artefacts comme l’accessibilité dont l’influence du caractère artificiel reste à établir, notamment à cause du manque de

données et une difficulté à modéliser les processus de gouvernance et de planification. Il est intéressant de noter que les priorités actuelles de développement des modèles LUTI semblent centrées sur une meilleure intégration des nouvelles technologies et une meilleur intégration avec la planification et les processus de prise de décision, par exemple via des interfaces de visualisation comme le propose [JTLU611]. Ils ne cherchent pas à s'étendre à des problématiques de dynamiques territoriales incluant le réseau sur de plus longues échelles par exemple, ce qui confirme la portée et la logique d'utilisation et de développement de ce type de modèles.

Une généralisation de ce type d'approche à une plus grande échelle, comme celle proposée par [russo2012unifying], consiste au couplage du LUTI à l'échelle mesoscopique à des modèles macroéconomiques à l'échelle macroscopique⁹. Ceux-ci ne considèrent pas l'évolution du réseau de transport de manière explicite mais s'intéressent seulement aux motifs abstraits d'offre et demande. L'économie urbaine a développé des approches spécifiques similaires dans leur démarche : [masso2000] décrit par exemple un modèle intégré couplant développement urbain, relocation et équilibre des flux de transports.

Ainsi, nous pouvons synthétiser ce type d'approche, qu'on pourra désigner par abus de langage *approche LUTI*, par les caractéristiques fondamentales suivantes : (i) Modèles visant à comprendre une évolution du territoire, dans le contexte d'un réseau de transport donné ; (ii) Modèles dans une logique de planification et d'applicabilité, étant souvent impliqués eux-mêmes dans les prises de décision ; et (iii) Modèles à des échelles moyennes, dans l'espace (métropole) et dans le temps (décade).

Croissance du Réseau

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

⁹ [russo2012unifying] généralise en fait le cadre des LUTI pour proposer un cadre d'interaction entre Economie Spatiale et Transports (*Spatial Economics and Transport Interactions*). Celui-ci inclut les LUTI à l'échelle urbaine, et au niveau national les modèles macroéconomiques simulant production et consommation, compétition des activités, production du stock d'offre de transport. Les modèles de transport supposent toujours réseau fixe et établissent des équilibres au sein de celui-ci, ce qui implique une petite échelle spatiale et une courte échelle temporelle.

vue économique ; (ii) soit les composantes territoriales jouant en effet sur le réseau sont stables au échelles longues considérés.

La croissance de réseaux est l'objet de démarches de modélisation qui cherchent à expliquer la croissance des réseaux de transport. Ils prennent généralement un point de vue *bottom-up* et endogène, c'est-à-dire cherchant à mettre en évidence des règles locales qui permettraient de reproduire la croissance du réseau sur de longues échelles de temps (souvent le réseau routier). Comme nous allons le voir, il peut s'agir de la croissance topologique (création de nouveaux liens) ou la croissance des capacités des liens en relation avec leur utilisation, selon les échelles et les ontologies considérées. Nous distinguons pour simplifier des grands courants disciplinaires s'étant intéressé à la modélisation de la croissance des réseaux de transport : ceux-ci sont liés respectivement à l'économie des transports, la physique, la géographie des transports et la biologie.

On rejoint ainsi partiellement la classification de [xie2009modeling], qui propose une revue étendue de la modélisation de croissance des réseaux, dans une perspective d'économie des transports mais en élargissant à d'autres champs. [xie2009modeling] distingue des grands courants disciplinaires ayant étudiés la croissance des réseaux de transport : la géographie des transports a développé très tôt des modèles basés sur des faits empiriques mais qui ont visé à reproduire la topologie plutôt que sur les mécanismes ; les modèles statistiques sur des cas d'étude fournissent des conclusions très mitigées sur les relations causales entre croissance du réseau et demande (la croissance étant dans ce cas conditionnée aux données de demande) ; les économistes ont étudié la production d'infrastructure à la fois d'un point de vue microscopique et macroscopique, généralement non spatialisés ; la science des réseaux a produit des modèles stylisés de croissance de réseau qui se basent sur des règles topologiques et structurelles plutôt que des règles se reposant sur des processus correspondant à des réalités empiriques.

ECONOMIE Les économistes ont proposé des modèles de ce type : [zhang2007economics] passe en revue la littérature en Economie des Transports sur la croissance des réseaux, rappelant les trois aspects principalement traités par les économistes sur le sujet, qui sont la tarification routière, l'investissement en infrastructures et le régime de propriété, et propose finalement un modèle analytique combinant les trois. Ces trois classes de processus relèvent d'une interaction entre les agents économiques microscopiques (utilisateurs du réseau) et les agents de gouvernance. Les modèles peuvent inclure une description détaillée des processus de planification, comme [levinson2012forecasting] qui combine des enquêtes qualitatives et des statistiques pour paramétrier un modèle de croissance de réseau. [xie2009jurisdictional] compare l'influence relative des processus de croissance centralisés

(planification par une structure de gouvernance) et décentralisés (croissance locale ne rentrant pas dans le cadre d'une planification globale). [levinson2003induced] procède à une étude empirique des déterminants de la croissance du réseau routier pour les *Twin Cities* aux Etats-Unis (Minneapolis-Saint-Paul), établissant que les variables basiques (longueur, changement dans l'accessibilité) ont le comportement attendu, et qu'il existe une différence entre les niveaux d'investissement, impliquant que la croissance locale n'est pas affectée par les coûts, ce qui peut correspondre à une équité des territoires en termes d'accessibilité. Ces données sont utilisées par [zhang2016model] pour calibrer un modèle de croissance de réseau qui superpose les décisions d'investissement aux motifs d'utilisation du réseau. [yerra2005emergence] montre avec un modèle économique basé sur des processus d'autorenforcement (c'est à dire incluant une rétroaction positive des flux sur la capacité) et incluant une règle d'investissement basée sur l'attribution du trafic, que des règles locales sont suffisantes pour faire émerger une hiérarchie du réseau routier à usage du sol fixé. Une synthèse de ces travaux gravitant autour de LEVINSON est faite dans [xie2011evolving].

PHYSIQUE La physique a introduit récemment des modèles de croissance des réseaux d'infrastructure, em s'inspirant largement de cette littérature économique : un modèle très similaire au dernier cité est donné par [louf2013emergence] avec des fonctions coûts-bénéfices plus simples mais obtenant une conclusion similaire. Etant donné une distribution de noeuds (villes)¹⁰ dont la population suit une loi puissance, deux villes seront connectées par un lien routier si une fonction d'utilité coût-bénéfice, combinant linéairement flux gravitaire potentiel et coût de construction¹¹, a une valeur positive. Ces hypothèses locales simples suffisent à faire émerger un réseau complexe et des transitions de phase en fonction du paramètre de poids relatif dans le coût, conduisant à l'apparition de la hiérarchie. [zhao2016population] applique ce modèle de manière itérative pour connecter des zones intra-urbaines, et montre que la prise en compte des populations dans la fonction de coût change significativement les topologies obtenues.

Une autre classe de modèles, proche dans leur idée des modèles procéduraux, se basent sur des processus d'optimisation géométrique locale, et visent à ressembler à des réseaux réels dans leur topologie. [2016arXiv160906470B] étudie ainsi un modèle de croissance d'arbre

¹⁰ On se trouve ici dans un cas où l'hypothèse de non-évolution des population des villes tandis que le réseau s'établit itérativement trouve peu de support empirique ou thématique, puisqu'on a montré que réseau et villes avaient des échelles de temps d'évolution comparables. Ce modèle produit donc plus à proprement parler un *réseau potentiel* étant donné une distribution de villes, et il est à interpréter avec précaution.

¹¹ Ce qui donne une fonction de coût de la forme $C = \beta/d_{ij}^\alpha - d_{ij}$, où α et β sont des paramètres

appliqué aux pistes de fourmis, dans lequel coût de maintenance et coût de construction influencent tous les deux les choix de nouveau lien. Le modèle de morphogenèse de [[courtat2011mathematics](#)] qui utilise un compromis entre réalisation des potentiels d'interaction et coût de construction, ainsi que des règles de connectivité, reproduit de manière stylisée des motifs réels des réseaux de rues. Un modèle très proche est décrit dans [[rui2013exploring](#)], tout en incluant des règles supplémentaires pour l'optimisation locale (prise en compte du degré pour la connection de nouveaux liens). La conception optimale de réseau, plutôt pratiquée par l'ingénierie, utilise des paradigmes similaires : [[vitins2010patterns](#)] explore l'influence de différentes règles d'une grammaire de formes (notamment les motifs de connection entre les liens de différents niveaux hiérarchiques) sur les performances de réseaux générés par algorithme génétique.

Détaillons les mécanisme de l'un de ces modèles de croissance géométrique. [[barthelemy2008modeling](#)] décrit un modèle basé sur une optimisation locale de l'énergie qui génère des réseaux routiers à l'aspect globalement crédible. Le modèle suppose des "centres", qui correspondent à des noeuds d'un réseau routier, et des segments de route dans l'espace reliant ces centres. Le modèle part de centre initiaux connectés, et procède par itérations pour simuler la croissance du réseau de la façon suivante :

1. De nouveaux centres sont ajoutés aléatoirement suivant une distribution de probabilité exogène, aux pas de temps multiple d'une durée fixée.
2. Le réseau croît suivant une règle de minimisation de coût : les centres sont groupés par projection sur le réseau; chaque groupe fait croître un segment de longueur fixée dans la direction moyenne vers l'ensemble du groupe à partir de la projection (sauf si celle-ci est nulle, un segment croît alors en direction de chaque point).

Ce modèle est ajusté pour que les aires des parcelles délimitées par le réseau suivent une loi d'échelle avec un exposant similaire à celui observé pour la ville de Dresde. Il a l'avantage d'être simple, d'avoir peu de paramètres (distribution de probabilité pour les centres, taille des tronçons construits), de reposer sur des règles locales crédibles. Cette dernière propriété est à double tranchant, puisqu'on peut alors s'attendre à ce que le modèle ne puisse capturer que peu de complexité, en négligeant de nombreux processus mis en valeur au chapitre 1 comme la gouvernance.

RÉSEAUX BIOLOGIQUES Enfin, une approche originale et intéressante pour la croissance des réseaux est le réseau biologique. Cette approche appartient au champ de l'ingénierie morphogénétique, qui vise à concevoir des systèmes complexes artificiels inspirés de sys-

tèmes complexes naturels et sur lesquels un contrôle des propriétés émergentes est possible [doursat2012morphogenetic]. Les *Machines Physarum*, qui sont des modèles d'une moisissure auto-organisée (*slime mould*) ont été prouvés comme résolvant de manière efficiente des problèmes difficiles (au sens de leur complexité computationnelle, voir 3.3) comme des problème de routage [tero2006physarum] ou des problèmes de navigation NP-complets comme le Problème du Voyageur de Commerce [zhu2013amoeba]. Ces propriétés permettent à ces systèmes de produire des réseaux ayant des propriétés de coût-robustesse Pareto-efficiences [tero2010rules] qui sont typiques des propriétés empiriques des réseaux réels, et de plus relativement proches en forme de ceux-ci (sous certaines conditions, voir [adamatzky2010road]).

Ce type de modèles peut être d'intérêt dans notre cas puisque les processus d'auto-renforcement basés sur les flots sont analogues aux mécanismes de renforcement de lien en économie des transports. Ce type d'heuristique a été testé pour générer le réseau ferré Français par [mimeur:tel-01451164], faisant un pont intéressant avec les modèles d'investissement de LEVINSON décrits précédemment¹².

MODÉLISATION PROCÉDURALE Finalement, nous pouvons mentionner d'autres tentatives comme [de2007netlogo ; yamins2003growing], qui sont plus proches de la modélisation procédurale [lechner2004procedural ; watson2008procedural] et pour cette raison n'ont que peu d'intérêt pour notre cas puisqu'ils peuvent difficilement être utilisés comme modèles explicatifs¹³. La modélisation procédurale consiste à générer des structures à la manière des grammaires de forme¹⁴, mais celle-ci se concentre généralement sur la reproduction fidèle de forme locale, sans tenir compte des propriétés macroscopiques émergentes. Les

¹² Sachant que pour cette étude, les critères de validation appliqués restent toutefois limités, soit à un niveau inadapté aux faits stylisés étudiés (nombre d'intersection ou de branches) soit trop générales pouvant être produit par n'importe quel modèle (longueur totale et pourcentage de population desservie), et relèvent de critère de forme typique de la modélisation procédurale qui ne peuvent que difficilement rendre compte des dynamiques internes d'un système comme développé précédemment. De plus, prendre pour validation externe la production d'un réseau hiérarchique découle d'une exploration incomplète de la structure et du comportement du modèle, puisque celui-ci par ses mécanismes d'attachement préférentiel doit mécaniquement produire une hiérarchie. Ainsi, une attention particulière devra être donnée au choix des critères de validation.

¹³ Suivant [varenne2017theories], un modèle explicatif permet de produire une explication à des régularités ou des lois observées, par exemple en suggérant des processus pouvant en être à l'origine. Si les processus du modèle sont explicitement dissociés d'une ontologie raisonnable, ceux-ci ne peuvent être explications potentielles. Nous donnerons en 3.1 un développement de cette notion dans le cadre d'une réflexion plus générale sur l'épistémologie de la modélisation.

¹⁴ Une grammaire de forme est un système formel (c'est à dire un ensemble de symboles initiaux, les axiomes, et un ensemble de règles de transformation) qui agit sur des objets géométriques. Partant de motifs initiaux, elles permettent de générer des classes d'objets

classifier comme modèles de morphogenèse n'est pas correct et correspond à une incompréhension des mécanismes du *Pattern Oriented Modeling* [**grimm2005pattern**]¹⁵ d'une part et de l'épistémologie de la Morphogenèse d'autre part (voir 5.1). Nous utiliserons ce type de modèle (mélange d'exponentielles pour produire une densité de population par exemple) pour générer des données synthétiques initiales uniquement pour faire tourner d'autres modèles complexes (voir 3.1 et 5.3).

2.1.3 Modéliser la co-évolution

Nous pouvons à présent nous intéresser aux modèles intégrant dynamiquement le paradigme Territoire ↔ Réseau, qui on le rappelle suppose qu'un conditionnement de l'un par l'autre n'est pas identifiable. Les ontologies utilisées, comme nous le verrons, couplent¹⁶ souvent des éléments de réseau avec des composantes territoriales, mais cette position n'est pas une nécessité et certains éléments peuvent être hybrides (par exemple une structure de gouvernance du système de transport peut relever simultanément des deux aspects). Dans notre lecture des modèles, ces différentes spécifications se dégageront naturellement.

Nous désignerons largement par modèle de co-évolution les modèles de simulation qui incluent un couplage des dynamiques de la croissance urbaine et du réseau de transport. Ceux-ci sont relativement rares, et pour la plupart au stade de modèles stylisés. Les efforts étant assez disparates et dans des domaines très variés, il y a peu d'unité dans ces approches, si ce n'est l'abstraction de l'hypothèse d'interdépendance entre réseaux et caractéristiques du territoire dans le temps. Nous proposons de les passer en revue toujours avec la grille de lecture des échelles.

Echelle microscopique et mesoscopique

MODÈLES GÉOMÉTRIQUES [**achibet2014model**] décrit un modèle de co-évolution à une très grande échelle (échelle du bâtiment), dans lequel l'évolution du réseau et des bâtiments sont tous les deux régis par un agent commun, influencé différemment par la topologie du réseau et la densité de population, qui peut être compris comme un agent développeur. Le modèle permet de simuler une extension urbaine auto-organisée et de produire des configurations de quartier. Bien qu'il couple fortement composantes territoriales (bâtiments) et

¹⁵ Le *Pattern Oriented Modeling* consiste à chercher à expliquer des motifs observés, généralement à plusieurs échelles, dans une démarche *bottom-up*. La modélisation procédurale n'en relève pas, puisqu'elle vise à reproduire et non à expliquer.

¹⁶ Nous rappelons la définition du couplage de modèle, qui correspond à celle de couplage de système ou de processus donnée en introduction : il s'agit de la constitution d'un modèle qui est une extension de chacun des modèles initiaux simultanément.

réseau routier, les résultats présentés ne permettent pas de tirer de conclusion sur les processus de co-évolution en eux-même.

Une généralisation du modèle d'optimisation locale géométrique décrit précédemment a été développé dans [barthelemy2009co], et cherche à capturer la co-évolution entre topologie du réseau et densité de ses noeuds. La localisation de nouveaux noeuds est influencée à la fois par la densité et la centralité, permettant de boucler le couplage fort. Plus précisément, Le fonctionnement global du modèle est le même, ainsi que la règle de croissance du réseau. Les centres se localisent quant à eux selon une fonction d'utilité qui est une combinaison linéaire de la centralité de chemin moyenne dans un voisinage et de l'opposée de la densité (dispersion due aux prix plus élevés en fonction de la densité). Cette utilité permet de définir la probabilité de localisation des nouveaux centres suivant un modèle de choix discrets. Le modèle permet de montrer que l'influence de la centralité accentue les phénomènes d'agrégation (notamment par une résolution analytique sur une version en une dimension du modèle), et reproduit par ailleurs des profils exponentiels décroissants pour la densité (loi de Clarke), observés empiriquement.

[ding2017heuristic] introduit un modèle de co-évolution entre différentes couches du réseau de transport, et montre l'existence d'un paramètre de couplage optimal en terme d'inégalités de centralité pour la conception d'un réseau : si on assimile le réseau routier à granularité très fine à une distribution de population, ce modèle se rapproche du précédent modèle de co-évolution entre réseau de transport et territoire.

MODÈLES ÉCONOMIQUES [levinson2007co] prend une approche économique plus riche du point de vue des processus de développement de réseau impliqués, similaire à un modèle à quatre étapes (c'est-à-dire incluant une génération de flux origine-destination et une attribution du traffic dans le réseau) qui inclut coût de transport et congestion, couplé avec un module d'investissement routier qui simule les revenus des péages pour les agents qui construisent, et un module d'évolution d'usage du sol qui simule les relocalisations des actifs et des emplois. Les expériences d'exploration de ce modèle montrent que l'usage du sol et le réseau en co-évolution conduisent à des retroactions positives renforçant les hiérarchies. Elles sont cependant loin d'être satisfaisantes pour deux raisons : la topologie du réseau n'évolue pas à proprement parler puisque seules les capacités et les flux changent dans le réseau, ce qui signifie que des mécanismes plus complexes (comme la planification de nouvelles infrastructures) sur de plus longues échelles de temps ne sont pas pris en compte. [li2016integrated] a récemment étendu ce modèle par l'ajout de prix immobiliers endogènes et d'une heuristique d'optimisation par algorithme génétique pour les agents décideurs.

D'un autre point de vue, [levinson2005paving] est aussi présenté comme un modèle de co-évolution mais repose sur un modèle prédictif à chaîne de Markov, et donc plus proche d'une analyse statistique que d'un modèle de simulation basé sur des processus. [rui2011urban] décrit un modèle dans lequel le couplage entre usage du sol et la topologie du réseau est fait par un paradigme faible, l'usage du sol et l'accessibilité n'ayant pas de retroaction sur la topologie du réseau, le modèle d'usage du sol étant conditionné à la croissance du réseau autonome.

AUTOMATES CELLULAIRES Un modèle hybride simple exploré et appliqué à un exemple stylisé de planification de la répartition fonctionnelle d'un nouveau quartier dans [raimbault2014hybrid], repose sur les mécanismes d'accès aux activités urbaines pour la croissance des établissements avec un réseau s'adaptant à la forme urbaine. Les règles pour la croissance du réseau sont trop simples pour capturer des processus plus élaborés qu'une simple connection systématique (comme une rupture de potentiel par exemple), mais le modèle produit à une petite échelle une large gamme de formes urbaines qui reproduisent les motifs typiques des établissements humains. Ce modèle s'inspire de [moreno2012automate] pour ses mécanismes de base mais permet une génération de formes bien plus larges par la prise en compte des fonctions urbaines.

A ces échelles relativement grandes, s'étendant de l'échelle urbaine à celle métropolitaine, les mécanismes de localisation de population influencée par l'accessibilité couplés à des mécanismes de croissance de réseau optimisant certaines fonctions semblent être la règle pour ces modèles : de la même façon, [wu2017city] couplent un Automate Cellulaire de diffusion de population à un réseau optimisant un coût local dépendant de la géométrie et de la distribution de population.

Des modèles répondant à des problématiques assez lointaines peuvent par ailleurs être reliés à notre question : par exemple, de manière conceptuelle, une certaine forme de couplage fort est également utilisé dans [bigotte2010integrated] qui par une approche de recherche opérationnelle propose un algorithme de design de réseau pour optimiser l'accessibilité aux services, prenant en compte à la fois la hiérarchie du réseau et celle des centres connectés.

Ainsi, les modèles de co-évolution aux échelles microscopique et mesoscopiques suivent globalement la structure suivante : (i) processus de localisation ou relocalisation des activités (actifs, bâtiments) influencés par leur propre distribution et les caractéristique du réseau ; (ii) evolution du réseau, topologique ou non, répondant à des règles très diverses : optimisation locale, règles fixes, planification par des agents décideurs. Cette diversité suggère la nécessité de prendre en compte la superposition de multiples processus régissant l'évolution du réseau.

Modélisation de Systèmes Urbains

A une échelle macroscopique, la co-évolution est parfois prise en compte dans des modèles de systèmes urbains. [baptiste1999interactions] propose de coupler un modèle de croissance urbaine basé sur les migrations (introduit par l'application de la synergétique au système de ville par [sanderson1992systeme]) avec un mécanisme d'auto-renforcement des capacités pour le réseau routier sans modification topologique. Plus précisément, les principes généraux du modèle sont les suivants :

- Des indicateurs d'attractivité et de répulsion permettent pour chaque ville de déterminer des taux d'émigration et d'immigration et de faire évoluer les populations
- La topologie du réseau est fixée dans le temps, mais les capacités des liens évoluent. La règle est une augmentation de la capacité lorsque le flux dépasse celle-ci par un seuil donné comme paramètre pendant un certain nombre d'itérations. Les flux sont affectés par modèle gravitaire d'interaction entre les villes.

Sa dernière version est présentée par [baptistemodeling]. Les conclusions générales qui peuvent être tirées de ce travail sont que ce couplage permet de faire émerger une configuration hiérarchique¹⁷ et que l'ajout du réseau produit un espace moins hiérarchique, permettant à des villes moyennes de bénéficier de la rétroaction du réseau de transport.

Le modèle proposé par [blumenfeld2010network] peut être vu comme un pont entre l'échelle mesoscopique et les approches de type système urbain, puisqu'il simule les migrations entre villes et la croissance du réseau induite par une rupture de potentiel lorsque les détours sont trop grands. Dans la continuité des modèles Simpop pour modéliser les systèmes de villes, [schmitt2014modelisation] décrit le modèle SimpopNet qui vise à précisément intégrer les processus de co-évolution dans les systèmes de villes à longue échelle temporelle, typiquement par des règles pour un développement hiérarchique du réseau comme fonction des dynamiques des villes, couplées à celles-ci qui dépendent de la topologie du réseau. Malheureusement le modèle n'a pas été exploré ni étudié de manière plus approfondie, et de plus est resté au niveau de modèle jouet. [cottineau2014evolution] propose une croissance endogène des réseaux de transport comme la dernière brique de construction du cadre de modélisation MARIUS, mais cela reste à un niveau conceptuel puisque cette brique n'a pas

¹⁷ Mais on sait par ailleurs que des modèles plus simples, un attachement préférentiel uniquement par exemple, permettent de reproduire ce fait stylisé. Le modèle doit avoir pour objectif de répondre à des problématiques plus larges, comme la compréhension fine des processus de co-évolution, ce qui n'est pas fait ici. Cependant, l'un de ses objectifs opérationnels est par ailleurs rempli, par l'application à la France et l'étude de l'impact d'un projet de Ligne à Grande Vitesse, rappelant les multiples fonctions possibles d'un modèle (voir 3.1).

encore été spécifiée ni implémentée. Il n'existe à notre connaissance pas de modèle empirique ou appliqué à un cas concret se basant sur une approche de la co-évolution par les systèmes urbains vus par la Théorie Evolutive des Villes.

On voit bien l'opposition aux principes épistémologiques de l'économie géographique : [fujita1999evolution] introduisent par exemple un modèle évolutionnaire capable de reproduire une hiérarchie urbaine et une organisation typique de la Théorie des Places Centrales [banos2011christaller] mais repose toujours sur la notion d'équilibres successifs, et surtout considère un modèle "à-la-Krugman" c'est à dire un espace à une dimension, isotrope, et dans lequel les agents sont répartis de manière homogène¹⁸. Cette approche peut être instructive sur les processus économiques en eux-mêmes mais plus difficilement sur les processus géographiques, puisque ceux-ci impliquent un déroulement des processus économiques dans l'espace géographique dont les particularités spatiales qui ne sont pas prise en compte dans cette approche sont essentielles. Notre travail s'attachera à montrer dans quelle mesure cette structure de l'espace peut être importante et également explicative, puisque les réseaux, et encore plus les réseaux physiques induisent des processus dépendants au chemin spatio-temporel et donc sensibles aux singularités locales et propices aux bifurcations induites par la combinaison de celles-ci et de processus à d'autres échelles (par exemple la centralité induisant un flux).

A l'échelle macroscopique, les modèles existants se basent sur des évolutions des agents (souvent les villes) conséquence de leurs interactions, portées par le réseau, tandis que l'évolution du réseau peut répondre à différentes règles : auto-renforcement, rupture de potentiel. La structure générale est globalement la même qu'à des échelles plus grandes, mais les ontologies restent fondamentalement différentes.

Synthèse

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 3.

¹⁸ L'absence d'espace réel n'est pas un problème dans cette approche économique qui vise à comprendre des processus hors-sol. Dans notre cas, la structure de l'espace géographique n'est pas séparable, et même au cœur des problématiques qui nous intéressent.

TABLE 3 : **Synthèse des approches de modélisation.** Le type donne le sens de la relation; la classe est le champ scientifique dans lequel le modèle se place; les échelles correspondent à nos échelles simplifiées; les fonctions sont données au sens de 3.1; nous donnons enfin le type de résultat qu'ils fournissent et les paradigmes utilisés.

Type	Classe	Echelle Temporelle	Echelle Spatiale	Fonction	Résultats	Paradigmes
Réseaux → Territoires	LUTI	Moyenne	Mesoscopique	Planification, Prédition	Simulation de l'usage du sol	Economie urbaine
Territoires → Réseaux	Economie des Réseaux	Moyenne	Mesoscopique	Explication	Rôle de processus économiques	Economie, 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	Economie des Réseaux	Moyenne	Mesoscopique	Explication	Effets de renforcement	Economie
	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

Une co-évolution négligée ?

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.

* * *

*

2.2 UNE APPROCHE EPISTÉMOLOGIQUE

Nous avons eu un aperçu large de différents types de modèles prenant en compte les interactions entre réseaux et territoires, ainsi que les disciplines et problématiques associées. Ces aspects très différents suggèrent un cloisonnement fort des disciplines. Il reste de plus difficile de situer les modèles potentiels de co-évolution dans cette nébuleuse. Il est légitime de se demander quelles sont les relations existantes et potentielles entre les différentes approches ? Quelles domaines peuvent être passés inaperçus bien que complémentaires ?

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

- Suivant [[commenges:tel-00923682](#)], 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.

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.

Une manière d'explorer cette hypothèse et de répondre aux questions précédentes consiste en une étude épistémologique que nous proposons de mener de manière quantitative et systématique. Cette approche est complémentaire de l'analyse de littérature précédente, et permet à la fois de la contextualiser et de la systématiser. Il faut par ailleurs garder en tête l'idée que l'étude des raisons de la rareté des modèles nous informera nécessairement sur les modèles eux-mêmes et les questions reliées à leur construction : la *connaissance de la connaissance* [[morin1986methode](#)] accroît la connaissance.

Une étude préliminaire a pour but de confirmer la pertinence d'une approche d'épistémologie quantitative, en suggérant une forte isolation des disciplines. Celle-ci est menée par un algorithme de revue systématique algorithmique, qui reconstruit des corpus de références par exploration de voisinage sémantiques, c'est à dire la récupération itérative de références voisines dans leur contenu sémantique principal. Nous procédons ensuite à une analyse de réseaux, couplant

réseau de citation et réseau sémantique, pour préciser les contours des disciplines impliquées. Nous suggérons finalement des possibles extensions vers de l'apprentissage non-supervisé et la fouille de texte complets pour une extraction automatique de la structure de modèles par exemple.

Commençons par situer le contexte des analyses en *épistémologie quantitative*¹⁹ que nous proposons de mener.

2.2.1 *Epistémologie quantitative*

Les méthodes possibles pour des entrées quantitatives en épistémologie sont nombreuses. Une bonne illustration de la variété des approches est donnée par l'analyse de réseau. En utilisant des caractéristiques topologiques du réseau de citation, un bon pouvoir prédictif pour les motifs de citation est par exemple obtenu par [2013arXiv1310.8220N]. Les réseaux de co-auteurs peuvent également être utilisés pour des modèles prédictifs [2014arXiv1402.7268S]. Une approche par réseau multi-couches est proposée dans [omodei2017evaluating], qui utilise des réseaux bipartites d'articles et de chercheurs, afin de produire des mesures d'interdisciplinarité en utilisant des mesures de centralité généralisées. Les disciplines peuvent être stratifiées en couches pour révéler des communautés entre celles-ci et ainsi des motifs de collaboration [2015arXiv150601280B]. Les réseaux de mots-clés sont utilisés dans d'autres champs comme en Economie de l'innovation : par exemple, [choi2014patent] propose une méthode pour identifier les opportunités technologiques par la détection de mots-clés importants du point de vue des mesures topologiques. De façon similaire, [shibata2008detecting] utilise une analyse topologique du réseau de citation pour détecter des fronts de recherche émergents.

Revues systématique

Avec l'avènement des nouveaux moyens techniques et des nouvelles sources de données, la revue de littérature classique tend à se coupler à des revues automatiques. Des techniques de revue systématique ont été développées, des revues qualitatives aux meta-analyses quantitatives qui permettent de produire des nouveaux résultats par combinaison d'études existantes [rucker2012network]. Passer sous silence certaines références peut même être considéré comme une erreur scientifique dans le contexte de l'émergence des systèmes d'in-

¹⁹ 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.

formation qui par l'accès plus aisé à l'information rend difficilement justifiable l'omission de références clés [lissacksubliminal]²⁰.

Interdisciplinarité

Le développement d'approches interdisciplinaires est de plus en plus nécessaire pour la plupart des disciplines, à la fois pour la découverte de nouvelles connaissances mais aussi pour l'impact sociétal des découvertes, comme le rappelle récemment le volume spécial de la revue Nature (**natureInterdisc**). [banos2013pour] suggère que leur développement doit s'insérer dans une spirale subtile entre et au sein des disciplines. Une autre façon de voir ce phénomène est de le comprendre comme l'émergence de champs verticalement intégrés²¹ de manière conjointe aux questions horizontales comme détaillé dans la feuille de route des Systèmes complexes ([2009arXiv0907.2221B]).

Il existe naturellement de multiples points de vue sur ce qu'est exactement l'interdisciplinarité (de nombreux d'autres termes comme la trans-disciplinarité ou la cross-disciplinarité existent aussi) et cela dépend en fait des domaines impliqués : des disciplines hybrides apparuées récemment (voir par exemples celles soulignées par [baïs2010praise] comme l'astro-biologie, ou d'autres plus proche de notre champ comme la géomatique) sont une bonne illustration du cas où les intrications sont très fortes, tandis que des champs comme "l'urbanisme" dont les définitions sont multiples montrent dans quelle mesure l'intégration horizontale est nécessaire et comment de la connaissance transversale peut être produite. Les interactions entre les disciplines ne sont pas toujours faciles, comme le montre les malentendus lorsque les sujets sur la ville ont été récemment introduits aux physiciens comme [dupuy2015sciences] le rappelle, malentendus dont les effets peuvent être négatifs s'ils conduisent à des conflits ou à une ignorance de connaissances déjà établies par un autre domaine.

Ces questions font partie de la compréhension des processus de production de connaissance, i.e. la *Connaissance de la connaissance* comme [morin1986methode] la présente, dans laquelle les perspectives *evidence-based*, qui impliquent des approches quantitatives, jouent un rôle important. Ces paradigmes peuvent être compris comme une *épistémologie quantitative*. Des mesures quantitatives de l'interdisciplinarité ferraient pour cette raison partie d'une approche multidimensionnelle de l'étude de la science, qui va en quelque sorte "au delà de la bibliométrie" [cronin2014beyond]. La préoccupation de cette section

²⁰ Tout en restant conscient que même avec une méthode systématique, il est impossible d'être absolument exhaustif. L'objectif est d'augmenter autant que possible la couverture, dans l'idée d'une approche inclusive de multiples points de vue, comme le propose notre positionnement épistémologique de perspectivisme donné en 3.3.

²¹ C'est à dire intégrant, généralement entre les échelles, différentes branches d'un champ : par exemple la biologie intégrative [liu2005systems] vise à des ponts entre approches génomiques, approches physiologiques, approches écologiques, en tirant parti de l'intégration des méthodes : expérimentation, modélisation, simulation.

se positionne dans ce champ de recherche. Nous passons d'abord en revue les approches existantes à la mesure de l'interdisciplinarité.

Les définitions de l'interdisciplinarité elle-même et les indicateurs pour la mesurer ont déjà été traités par un vaste corpus de littérature. [huutoniemi2010analyzing] rappelle la différence entre les approches *multi-disciplinaires* (une agrégation de travaux de différentes disciplines) et *interdisciplinaires* (qui implique un certain niveau d'intégration). Ils construisent un cadre qualitatif pour classifier différents types d'interdisciplinarité, et distinguent par exemple les interdisciplinarités empiriques, théoriques et méthodologique. L'aspect multi-dimensionnel de l'interdisciplinarité est confirmé même au sein d'un champ spécifique comme la littérature [austin1996defining]. Une première façon de quantifier l'interdisciplinarité d'un ensemble de publications est de regarder la proportion de disciplines hors d'une discipline principale dans lesquelles elles sont publiées, comme [rinia2002impact] fait pour l'évaluation de projets en physique, de manière complémentaire au jugement d'experts. [porter2007measuring] désigne cette mesure comme *spécialisation*, et la compare avec une mesure d'*intégration* donnée par l'étendue des citations faites par un article au sein des différentes *Subject Categories* (classification du *Web of Knowledge*), qui est également appelé indice de *Rao-Stirling*. [lariviere2010relationship] l'utilise sur un corpus du *Web of Science* pour montrer l'existence d'un niveau d'interdisciplinarité intermédiaire optimal pour l'impact en termes de citations sur une fenêtre de 5 ans post-publication. Un travail équivalent est fait dans [lariviere201410], qui se concentre sur l'évolution des mesures sur une longue portée temporelle. L'influence des données manquantes sur cet index est étudié par [moreno2016uncertainty], qui fournit un cadre étendu qui prend en compte l'incertitude. L'utilisation de réseaux a également été proposée : [porter2009science] combine l'indice d'intégration avec une technique de cartographie qui consiste en la visualisation de réseaux synthétiques construits par les co-citations entre disciplines. [leydesdorff2007betweenness] montre que la centralité de chemin est un indicateur pertinent d'interdisciplinarité, lorsqu'un environnement de citation pertinent est considéré.

2.2.2 Revue Systématique Algorithmique

Nous proposons de procéder de manière préliminaire à une revue de la littérature systématique et algorithmique. Un algorithme itératif formel pour construire des corpus de références à partir de mots-clés initiaux, basé sur l'analyse textuelle, est développé et mis en oeuvre. Nous étudions ses propriétés de convergence et procédons à une analyse de sensibilité. Nous l'appliquons ensuite à des requêtes représentatives de notre question spécifique, pour lesquelles les résultats tendent à confirmer l'hypothèse d'isolation relative des disciplines.



FIGURE 7 : **Architecture globale de l'algorithme.** A partir d'un ensemble de mots-clés initiaux, on construit un corpus par requête au catalogue, duquel on extrait des nouveaux mots-clés par analyse textuelle. On itère alors en boucle jusqu'à obtenir un corpus fixe ou dépasser un nombre maximal fixé d'itérations.

Tandis que la majorité des études en bibliométrie se reposent sur les réseaux de citation [2013arXiv1310.8220N] ou les réseaux de co-auteurs [2014arXiv1402.7268S], nous proposons d'utiliser un paradigme moins exploré, basé sur l'analyse textuelle, introduit par [chavalarias2013phylomemetic], 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.3, avec les détails de son implémentation et des analyses de sensibilité. Sa logique est donnée 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.

Nous partons de cinq différentes requêtes initiales qui ont été manuellement extraites des divers domaines identifiés dans la bibliographie²², afin de comparer les corpus obtenus pour chaque requête.

²² Qui sont “cityANDsystemANDnetwork”, “land-useANDtransportANDinteraction”, “networkANDurbanANDmodeling”, “populationANDdensityANDtransport”, “transportationANDnetworkANDurbanANDgrowth”. Ce choix inclut les approches

TABLE 4 : Matrice symétrique des proximités lexicales entre les corpus finaux, définies comme la somme des co-occurrences pondérées de mots-clés finaux entre corpus, normalisé par le poids total des mots-clés finaux. La taille des corpus finaux est donnée par W . Les valeurs obtenues pour les proximités sont considérablement faibles par rapport à la valeur maximale 1, ce qui confirme que les corpus sont éloignés de manière significative.

Corpus	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

Après avoir construit les corpus, nous étudions leur cohérence lexico-comme un indicateur de réponse à notre question initiale. De grande distances devraient confirmer l'hypothèse formulée ci-dessus, i.e. que des disciplines auto-centrées pourraient être à l'origine d'un manque d'intérêt pour des modèles co-évolutifs. La table 4 montre les valeurs de la proximité lexique relative, que nous définissons par un indice de similarité d'ensemble pondéré donné par

$$d(I, J) = \frac{\sum_{k_i \in I, k_j \in J} \mathbb{I}_{k_i = k_j} \cdot (s(k_i) + s(k_j))}{\sum_{k_i \in I} s(k_i) + \sum_{k_j \in J} s(k_j)}$$

pour les corpus I, J , et avec s fonction strictement positive donnant une mesure d'importance des mots au sein des corpus fournie par la méthode d'extraction des mots-clés (voir A.3). Ses valeurs sont significativement faibles en comparaison à la valeur de référence 1 pour des corpus égaux (la mesure s'interprète comme une proportion de mots en co-occurrence), ce qui tend à confirmer notre hypothèse²³.

La constatation d'un faible nombre de modèles qui simulent la co-évolution des réseaux de transport et de l'usage du sol urbain pourrait être due à l'absence de communication entre les disciplines

par systèmes de villes, les approches LUTI, les approches de croissance de réseau. Il ne peut bien sûr être exhaustif. Cette étude étant préliminaire on admet de travailler potentiellement sur des échantillons. Par exemple, l'utilisation de "co-evolution" n'est pas concluante car trop peu d'articles utilisent cette formulation. De même, la question de la langue conditionne les résultats : une requête en Français conduit à des niches linguistiques finalement assez pauvres en diversité, et nous faisons ainsi uniquement des requêtes en anglais. L'approche par hyper-réseau développée plus loin sera elle multilingue.

²³ Pour situer ces résultats de manière relative, il faudrait un modèle nul (c'est à dire générateur de corpus avec distributions sémantiques similaires mais sans structure de corrélation entre mots) avec des corpus aléatoires par exemple, ce qui pourrait faire l'objet de développements futurs.

scientifiques étudiant différents aspects du problème. D'autres explications possibles qui en sont proches peuvent par exemple être le manque de cas d'application concrets de tels modèles vu les échelles temporelles mises en jeu et donc l'absence de financement propre - ce qui n'est pas si loin de l'absence d'une discipline y consacrant certains de ses objets. Cette question des portées et des échelles des modèles fera l'objet de la meta-analyse à la section suivante 2.3. Ainsi, nous avons proposé ici une méthode algorithmique pour donner des éléments de réponse par l'extraction de corpus basée sur l'analyse textuelle, dont les résultats numériques semblent aller dans le sens d'une compartmentalisation des disciplines (au sens particulier utilisé ici d'une distance sémantique entre corpus niches). Cette analyse était relativement limitée dans la portée de ses résultats, notamment par le faible nombre de requêtes et un certain nombre d'incertitudes intrinsèques, mais est suffisante pour produire un diagnostic, à savoir (i) une structure disciplinaire fortement marquée peut être extraite de l'analyse de corpus, et (ii) l'utilisation d'outils sémantique permet une extraction d'information endogène. Fort de ce diagnostique préliminaire, nous proposons d'approfondir l'analyse par une variation et extension de la méthode employée.

2.2.3 *Bibliométrie indirecte*

Comme décrit précédemment, l'analyse sémantique des corpus finaux ne contient pas la totalité de l'information sur les liens entre disciplines ni sur les motifs de propagation de la connaissance scientifique comme ceux contenus dans les réseaux de citations par exemple. De plus, la collection des données dans l'algorithme précédent est sujette à convergence vers des thèmes relativement auto-cohérents de par la structure propre de la méthode. On pourrait obtenir plus d'information sur les motifs sociaux de choix ontologiques pour la modélisation en étudiant les communautés dans des réseaux plus larges, ce qui correspondrait plus à des disciplines (ou des sous-disciplines selon le niveau de granularité). Nous proposons de reconstruire les disciplines autour de notre thématique, pour obtenir une vue plus précise du paysage scientifique sur notre sujet et des liens entre disciplines. Une contribution fondamentale de cette section consiste en la construction de jeux de données hybrides à partir de sources hétérogènes, et les développement des outils associés qui peuvent être réutilisés et améliorés pour des applications similaires. Cette démarche peut être vue comme une bibliométrie indirecte²⁴, puisqu'on cherche

²⁴ La bibliométrie, ou scientométrie lorsqu'elle est appliquée en particulier à la science comme dans notre cas, consiste en la mesure et la qualification des motifs de production de connaissance par l'intermédiaire de leur supports directement observables (productions scientifiques, fonctionnement des institutions, relations sociales entre chercheurs, etc.) [cronin2014beyond]. Cet ouvrage rappelle que ce domaine est en pleine mutation et dresse une carte des nouvelles approches.

à reconstruire une information endogène et à extraire des relations entre différentes dimensions.

Contexte

L'approche développée ici couple exploration et analyse de réseau de citation avec analyse textuelle, dans le but de cartographier le paysage scientifique dans le voisinage d'un corpus donné. Le contexte est particulièrement intéressant pour la méthodologie développée. Premièrement, le sujet étudié est très large et par essence interdisciplinaire. Deuxièmement, les données bibliographiques sont difficiles à obtenir, soulevant la question de comment la perception d'un horizon scientifique peut être déterminée par les acteurs de la dissémination et donc loin d'être objective, rendant les solutions techniques comme celle développée ici en conséquence des outils cruciaux pour une science ouverte et neutre.

Notre approche combine une analyse des communautés sémantiques (comme fait dans [[2016arXiv160208451P](#)] pour les articles en physique mais sans extraction des mots-clés, ou par [[2015arXiv151003797G](#)] pour un analyse des réseaux sémantiques de débats politiques) avec celle du réseau de citations pour extraire par exemple des mesures d'interdisciplinarité. Cette contribution se démarque des travaux précédents quantifiant l'interdisciplinarité puisqu'elle ne suppose pas de domaines *a priori* ou une classification des références considérées, mais reconstruit par le bas les champs via l'information sémantique endogène. [[nichols2014topic](#)] introduit une approche similaire, utilisant le modèle d'extraction de thématiques *Latent Dirichlet Allocation*²⁵ pour caractériser l'interdisciplinarité de récompenses dans des sciences précises.

Données

Notre approche implique des spécifications pour le jeu de données utilisé, à savoir : (i) couvrir un voisinage conséquent du corpus étudié dans le réseau de citation afin d'avoir une vue la moins biaisée possible du paysage scientifique ; (ii) avoir au moins une description textuelle pour chaque noeud. Pour cela, nous rassemblons et complions les données de sources hétérogènes en utilisant une architecture et implémentation spécifiques, décrites en Appendice B.6. Pour simplifier, nous dénommons *référence* toute production scientifique standard²⁶ qui peut être citée par une autre (articles de journaux, livre, chapitre de livre, article d'actes, communication, etc.) et contient des

²⁵ Le modèle LDA, introduit par [[blei2003latent](#)], suppose les documents comme produits par des thèmes sous-jacents, avec une distribution de Dirichlet pour leur composition ainsi que pour la distribution des mots par thèmes. Son estimation donne la composition des thèmes en terme de mots-clés.

²⁶ Ce qui est bien sûr sujet à débat, voir nos discussions en ouverture sur l'évolution des modes de communication scientifique.

informations de base (titre, résumé, auteurs, année de publication). Nous travaillons par la suite sur le réseau des références.

CORPUS INITIAL Notre corpus initial est construit à partir de l'état de l'art établi en 2.1. Sa composition complète est donnée en Appendice A.3. 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 [bergeaud2017classifying].

DONNÉES DE CITATION Le réseau de citations est reconstruit à partir de Google Scholar qui est souvent l'unique source des citations entrantes [noruzi2005google] puisqu'en science humaines les ouvrages ne sont pas systématiquement référencés par les bases fournissant des services (payants) comme le réseau de citation.²⁷ Nous sommes conscient des biais possibles de l'utilisation de cette source unique (voir par exemple [bohannon2014scientific])²⁸, mais ces critiques sont dirigées vers les résultats de recherche plutôt que les comptes de citations. Nous récoltons ainsi les références *citantes* à profondeur deux, c'est à dire les références citant le corpus initial et celles citant celles-ci. Le réseau obtenu contient $V = 9462$ références correspondant à $E = 12004$ liens de citation. Concernant les langues, l'anglais représente 87% du corpus, le français 6%, l'espagnol 3%, l'allemand 1%, complété par des langues comme le mandarin pouvant être indéfinies (la détection de celui-ci étant peu fiable).

DONNÉES TEXTUELLES Pour mener l'analyse sémantique, une description suffisamment conséquente est nécessaire. Nous collectons pour cela les résumés pour le réseau précédent. Ceux-ci sont disponibles pour environ un tiers des références, donnant $V = 3510$ noeuds avec description textuelle.

Résultats

RÉSEAU DE CITATIONS 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 $\gamma = 0.0013^{29}$. 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 :

²⁷ Par exemple, le journal Cybergeo n'est indexé dans le *Web of Science* que depuis mai 2016, suite à des négociations ardues et non sans contrepartie.

²⁸ ou <http://iscpif.fr/blog/2016/02/the-strange-arithmetic-of-google-scholars>

²⁹ Pour référence, [batagelj2003efficient] 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.

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 0.66^{30} , 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 le tableau suivant :

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 2.1³¹.

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

³⁰ 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.

³¹ 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.

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.

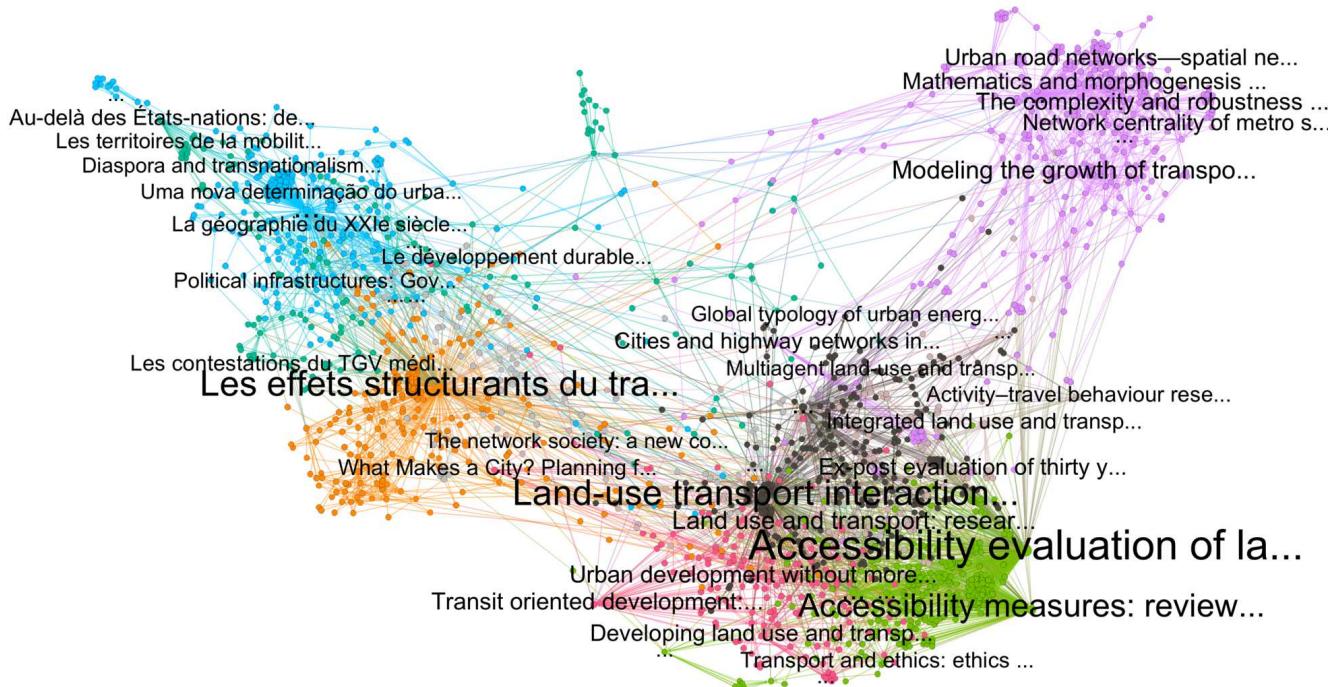


FIGURE 8 : Réseau de citations. Nous visualisons les références ayant au moins deux liens, par un algorithme de force-atlas. Les couleurs donnent les communautés décrites dans le texte. En orange, bleu, turquoise : géographie urbaine, géographie des transports, sciences politiques ; en rose, noir, vert : planning, accessibilité, LUTI ; en violet : réseaux spatiaux (physique et économie).

COMMUNAUTÉS SÉMANTIQUES L'extraction des mots-clés est faite suivant une heuristique inspirée de [chavalarias2013phylogenetic]. La description complète de la méthode et de son implémentation est donnée en Appendice 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.

La topologie du réseau brut ne permet pas l'extraction claire de communautés, en particulier à cause de hubs qui correspondent à des termes fréquents commun à de nombreux champs (e.g. model, space). Ces mots sont utilisés de manière comparable dans l'ensemble des champs étudiés, et ne portent pas d'information pour les séparer³². Nous faisons l'hypothèse que ces termes à fort degré ne portent pas d'information particulière sur des classes données et peuvent ainsi être filtrés étant donné un seuil de degré maximal k_{\max} (on s'intéresse alors à ce qui fait la spécificité de chaque domaine). De la même manière, les liens avec un poids faibles sont considérés comme du bruit et filtrés selon un seuil de poids minimal θ_w . La méthode générique permet de plus une filtration préliminaire des mot-clés, complémentaire à la filtration topologique, par fréquence d'apparition dans les documents $[f_{\min}, f_{\max}]$, à laquelle les résultats ne sont pas sensibles dans notre cas. L'analyse de sensibilité des caractéristiques du réseau filtré, notamment de sa taille, modularité et structure des communautés, est donnée en A.3. Nous choisissons des valeurs de paramètres permettant une optimisation multi-objectifs entre modularité et taille du réseau, $\theta_w = 10$, $k_{\max} = 500$, par le choix d'un point compromis sur un front de Pareto, qui donne un réseau sémantique de taille ($V = 7063$, $E = 48952$). Celui-ci est visualisé en Appendice A.3.

Nous récupérons ensuite les communautés dans le réseau par un clustering de Louvain standard sur le réseau filtré optimal. On obtient 20 communautés pour une modularité de 0.58. Celles-ci sont examinées à la main pour être nommées, les techniques de désignation automatique [yang2000improving] ne sont pas assez élaborées pour faire la distinction implicite entre champs thématiques et méthodologiques par exemple (en fait entre les domaines de connaissance, voir 9.3) qui est une dimension supplémentaire que nous ne traitons pas ici, mais nécessaire pour avoir des désignations parlantes. Les communautés sont décrites en Table 5. On voit tout de suite la complémentarité avec l'approche par citations, puisque se dégagent ici à la fois des sujet d'étude (High Speed Rail, Maritime Networks), des domaines et méthodes (Networks, Remote Sensing, Mobility Data Mining), des domaines thématiques (Policy), des méthodes pures (Agent-based Modeling, Measuring). Ainsi, une référence peut mobiliser plusieurs de ces communautés. On a de plus une granularité plus fine de l'information. L'effet du langage est puissant puisque la géographie française se distingue en une catégorie séparée (des analyses poussées pourraient être envisagées pour mieux comprendre le phénomène et en tirer parti : sous-communautés, reconstruction d'un réseau spécifique, études par traduction ; mais celles-ci sont hors de propos dans cette étude exploratoire). On constate l'importance des réseaux, des problématiques de sciences politiques et socio-économiques. Nous

³² Mais en porteraient si l'on comparait un corpus de géographie quantitative et un corpus de musicologie par exemple.

TABLE 5 : **Description des communautés sémantiques.** On donne leur taille, leur proportion en quantité de mots-clés (sous la forme de *multi-stems*) cumulés sur l'ensemble du corpus, et des mots-clés représentatifs sélectionnés par degré maximal.

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

mobiliserons la première catégorie dans la plupart des modèles développés, mais en gardant en tête l'importance des problématiques liées à la gouvernance, nous réaliserons un travail spécifique en ??.

MESURES D'INTERDISCIPLINARITÉ La distribution des mots clés dans les communautés permettent de définir une mesure d'interdisciplinarité au niveau de l'article. La combinaison des couches de citation et sémantique dans l'hyperréseau fournit des mesures d'interdisciplinarité au second ordre (motifs sémantiques des cités ou des citants), que nous n'utiliserons pas ici à cause de la taille modeste du réseau de citation (voir B.6 et ??). Plus précisément, une référence i peut être vue comme un vecteur de probabilités sur les classes sémantiques j , qu'on notera sous forme matricielle $\mathbf{P} = (p_{ij})$. Celles-ci sont estimées simplement par les proportions de mots-clés classifiés dans chaque classe pour la référence. Une mesure classique d'interdiscipli-

narité [bergeaud2017classifying] est alors $I_i = 1 - \sum_j p_{ij}^2$. Soit A la matrice d'adjacence du réseau de citation, et soit I_k les matrices de sélection des lignes correspondants à la classe k de la classification de citation : $Id \cdot \mathbb{1}_{c(i)=k}$, telle que $I_k \cdot A \cdot I_{k'}$ donne exactement les citations de k vers k' . La proximité de citation entre les communautés de citation est alors définie par $c_{kk'} = \sum I_k \cdot A \cdot I_{k'}/\sum I_k \cdot A$. On définit la proximité sémantique en définissant une matrice de distance entre références par $D = d_{ii'} = \sqrt{\frac{1}{2} \sum (p_{ij} - p_{i'j})^2}$ puis la proximité sémantique par $s_{kk'} = I_k \cdot D \cdot I_{k'}/\sum I_k \sum I_{k'}$.

Nous montrons en Fig. 9 les valeurs de ces différentes mesures, ainsi que la composition sémantique des communautés de citation, pour les classes sémantiques majoritaires. La distribution de I_i montre que les articles gravitant dans le domaine du LUTI sont les plus interdisciplinaires dans les termes utilisés, ce qui pourrait être lié à leur caractère appliqué. Les autres disciplines sont dans des motifs similaires, à part la géographie et la planification des infrastructures qui présentent des distributions quasi-uniformes, témoignant de l'existence de références très spécialisées dans ces classes. Ce n'est pas nécessairement étonnant vu les sous-champs pointus exhibés (sciences politiques par exemple, et de même les études prospectives type coût-bénéfices sont très étriquées). Ce premier croisement des couches nous confirme les spécificités de chaque champ. Concernant les compositions sémantiques, la plupart agissent comme validation externe vu les classes majoritaires. Le champ le moins concerné par les problèmes socio-économiques est la planification des infrastructures, ce qui donnera du grain à moudre aux détracteurs de la technocratie. Les questions de changement climatique et durabilité sont relativement bien réparties. Enfin, les ouvrages géographiques concernent en majorité des problèmes de gouvernance.

Les matrices de proximité confirment la conclusion obtenue précédemment en termes de citation, les partages étant très faibles, les plus hautes valeurs étant jusqu'à un quart de la planification vers la géographie et des LUTI vers le TOD (mais pas l'inverse, les relations peuvent être à sens unique). Hors, les proximités sémantiques montrent par exemple que LUTI, TOD, Accessibility et Networks sont proches dans leur termes, ce qui est logique pour les trois premiers, et confirme pour le dernier que les physiciens se basent majoritairement sur les méthodes des ces champs liés au planning pour légitimer leur travaux. La géographie est totalement isolée, sa plus proche voisine étant la planification des infrastructures. Cette étude est très utile pour notre propos, puisqu'elle montre des domaines cloisonnés partageant des termes et donc a priori des problématiques et sujet commun. Les domaines ne se parlent pas toute en parlant des langues pas si lointains, d'où la pertinence accrue de vouloir accorder leur partitions dans nos travaux : nos modèles devront mobiliser des éléments, ontologies et échelles de ces différents champs.

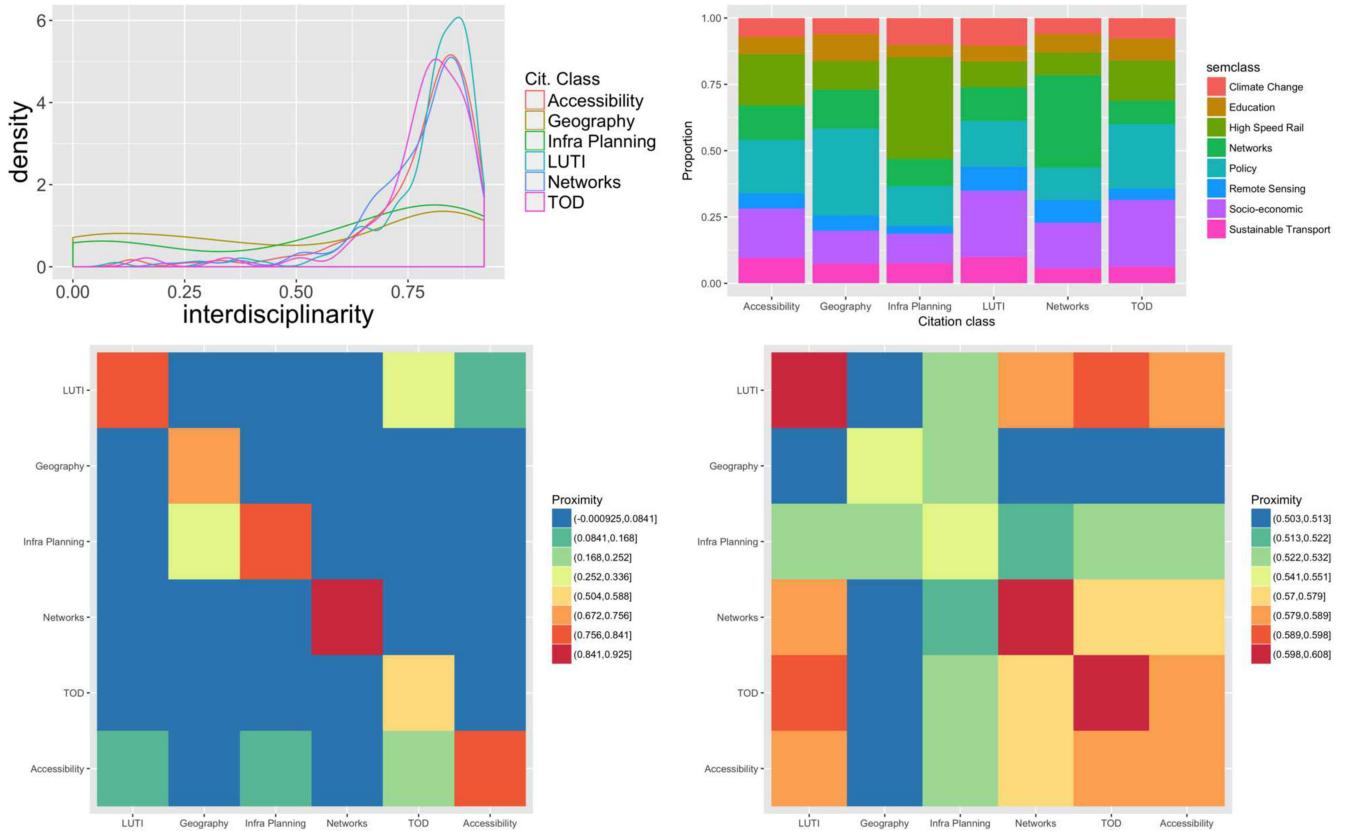


FIGURE 9 : Motifs d'interdisciplinarité. (*Haut Gauche*) Distribution statistique des I_i par classes de citations, en d'autre termes répartition des niveaux d'interdisciplinarité au sein des classes de citation; (*Haut Droite*) Composition sémantiques des classes de citation : pour chaque classe de citation (en abscisse), la proportion de chaque classe sémantique (en couleur) est donnée; (*Bas Gauche*) Matrice de proximité de citation $c_{kk'}$ entre classes de citations; (*Bas Droite*) Matrice de proximité sémantique $s_{kk'}$ entre classes de citations.

Nous concluons cette analyse par une approche plus robuste pour quantifier les proximités entre couches de l'hyperréseau. Il est ais  de construire une matrice de corr ation entre deux classifications, par les corr ations de leur colonnes. Nous d finissons les probabilit s P_C toutes gales  1 pour la classification de citation. La matrice de correlation de celle-ci avec P s'tend de -0.17  0.54 et a une moyenne de valeur absolue de 0.08, ce qui est significatif par rapport  des classifications al atoire puisque un bootstrap  b = 100 r p titions avec les matrices m lang es donne un minimum  -0.08 ± 0.012, un maximum  0.11 ± 0.02 et une moyenne absolue  0.03 ± 0.002. Cela montre que les classifications sont compl mentaires et que cette compl mentarit  est significative statistiquement par rapport  des classifications al atoires. L'ad quation de la classification s mantique par rapport au r seau de citation peut galement tre quantifi e par la modularit  multi-classes [[nicosia2009extending](#)] (voir [C.3](#) pour une d finition math matique), qui traduit la probabilit  qu'un lien soit d   la classification tudi e, en prenant en compte l'appartenance simultan e  de multiples classes. Ainsi, la modularit  multi-classes des probabilit s s mantiques pour le r seau de citation est de 0.10, ce qui d'une part est significativement signe d'ad quation, un bootstrap toujours  b = 100 donnant une valeur de 0.073 ± 0.003, qui reste limit e vu la valeur maximale fix e par les probabilit s de citations dans leur propre r seau qui donnent une valeur de 0.81, ce qui confirme d'autre part la compl mentarit  des classifications.

Nous avons ainsi dress  dans cette section un aper u des disciplines en relation avec notre sujet, ainsi que leur relations. Il s'agira dans la section suivante de comprendre avec plus de d tail leur "contenu", c'est--dire les moyens mobilis s pour r soudre les probl mes rencontr s.

Discussion

Donnons bri vement des directions d'extension de l'analyse que nous venons de mener ainsi que des implications pour le positionnement pist mologique de notre travail.

Vers une mod lisation des th mes et une extraction automatique du contexte

Une direction possible pour renforcer cette analyse en pist mologie quantitative serait de travailler sur les textes complets des r f rences contenant des efforts de mod lisation des interactions entre r seaux et territoires, avec le but d'extraire automatiquement les th matiques des articles. Des m thodes plus adapt es pour les long texte que celle utilis e ici incluent par exemple l'Allocation Latente de Dirichlet [[blei2003latent](#)]. L'id e serait de proc der  une sorte de mod lographie automatique, tendant la m thodologie de mod lographie d velopp e par [[schmitt2013modelographie](#)], pour extraire des carac-

téristiques telle les ontologies, l'architecture ou la structure des modèles, les échelles ou même des valeurs typiques des paramètres. Il n'est pas clair dans quelle mesure la structure des modèles peut être extraite de leur description dans un article, et cela dépend sûrement de la discipline considérée. Par exemple dans un champ relativement cadré comme la planification des transports, l'utilisation d'une ontologie pré-définie (dans le sens d'un dictionnaire) et d'une grammaire floue pourrait être efficace vu les conventions assez strictes dans la discipline. En géographie théorique et quantitative, au-delà de la barrière de la diversité des formalisations possibles pour une même ontologie, l'organisation de l'information est sûrement plus délicate à appréhender par de l'apprentissage non-supervisé à cause de la nature plus littéraire de la discipline : les synonymes et les figures de style sont généralement la norme pour l'écriture d'un bon niveau en sciences humaines, rendant plus floue une possible structure générique de la description des connaissances.

Réflexivité

La méthodologie que nous avons développé ici est efficace pour offrir des instruments de réflexivité, c'est à dire qu'elle peut être utilisée pour étudier notre approche elle-même. Une de ses applications, hors de celle à la revue scientifique Cybergeo dans la perspective de Science Ouverte (voir Appendice B.6), sera à notre propre corpus de références, dans le but de révéler des possibles directions de recherche ou problématiques exotiques. Il est éventuellement possible de le faire de manière dynamique, grâce à l'historique de git qui permet de récupérer n'importe quelle version de la bibliographie à une date donnée sur les trois ans écoulés. Il s'agira aussi de comprendre nos motifs de production de connaissance afin de contribuer à 9.3. Le développement détaillé est fait en Appendice F.

* * *

*

2.3 REVUE SYSTÉMATIQUE ET MODÉLOGRAPHIE

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 meta-analyse, c'est à dire une tentative d'explication de certaines caractéristiques des modèles par des modèles statistiques.

2.3.1 *Revue systématique et Meta-analyse*

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 meta-analyse par réseau, voir [rucker2012network]). 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. Nous adoptons pour cela le protocole 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 examen 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 mots-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. On obtient 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-hub³³ 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.
- Les revues, et en fait les éditeurs, semblent influencer différemment les référencements, augmentant potentiellement le biais

³³ <http://sci-hub.cc/>

de requête. La littérature grise ainsi que les pre-prints sont pris en compte différemment selon les champs.

- Le passage en revue manuel des grand corpus permet de pas louper des “poids lourds” qui auraient pu être omis en amont [**lissacksubliminal**]. 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 [**bastian201oseventy**]. 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 en elle-même 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 [9.3](#) 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. On pourrait 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 [8.2](#)), 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.

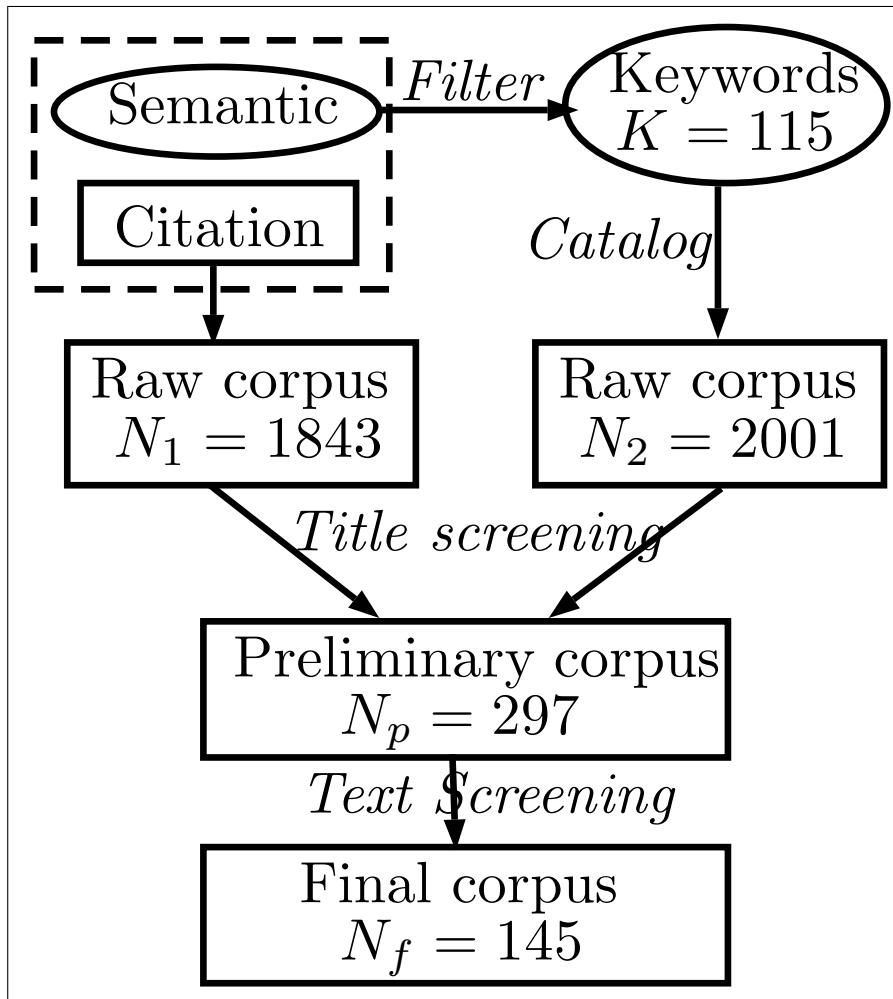


FIGURE 10 : Méthodologie de la revue systématique. Les rectangles désignent des corpus de références, les ellipses des corpus de mot-clés, et les pointillés les corpus initiaux. A chaque étape est donnée la taille du corpus.

2.3.2 Modélographie

Nous passons à présent à une analyse mixte basée sur ce corpus, inspirée par les résultats des sections précédentes notamment pour la classification. Elle a pour but d'extraire et de décomposer précisément les ontologies, échelles et processus, puis d'étu-

dier des liens possibles entre ces caractéristiques des modèles et le contexte dans lequel ils ont été introduits. Il s'agit ainsi de la meta-analyse en quelque sorte, que nous désignerons ici par modélographie. Pour ne pas froisser les puristes, il ne s'agit en effet pas d'une meta-analyse à proprement parler car nous ne combinons pas des analyses proches pour extrapoler des résultats potentiels d'échantillons plus grand. Notre démarche est proche de celle de COTTINEAU dans [2016arXiv160606162C] 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 convaincus de la pertinence d'affiner de telles techniques (voir 9.3.3) 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 couplage 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.
- Echelle de temps maximale.
- Echelle 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.

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.

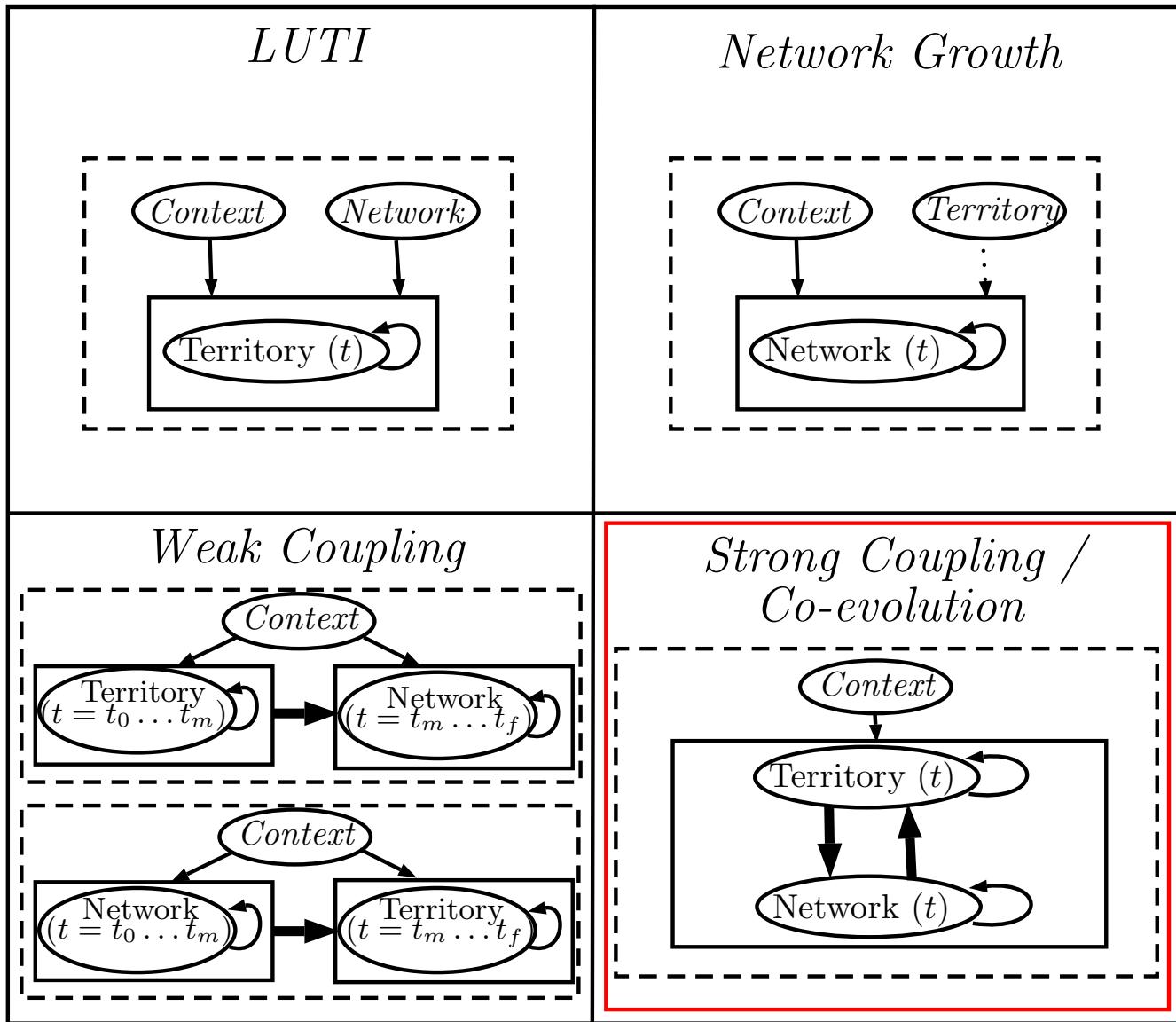


FIGURE 11 : Représentation schématique de la distinction entre différents types de modèles couplant territoires et réseaux. Les ontologies sont représentés par des ovales, les sous-modèles par les boîtes pleines, les modèles par les boîtes pointillées, les couplages par les flèches. Nous surlignons en rouge l'approche qui sera l'objectif final de notre travail.

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).

Processus et cas d'étude

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 on notera 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 lutti, ou aux investissements d'infrastructure de réseau. On observe 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 minoritaire, les approches de type système 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.

Caractéristiques du corpus

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.

Modèles étudiés

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 6 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 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.

TABLE 6 : **Types de modèles étudiés selon les différentes classifications.** Tables de contingence de la variable discrète donnant le type de modèle (réseau, territoire ou couplage fort), pour la classification a priori, la classification sémantique et la classification de citation.

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 planning	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	0
strong	0	0	0	2	5	0	0
territory	13	1	6	18	2	3	0

Echelles étudiées

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 ($\chi^2 = 50, p = 0.08$).

Régressions classiques

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'en-

semble des modèles possible 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 [igel2005multi], 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 AICc³⁴). 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 Appendice A.3 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 Tab. 7. 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 Appendice A.3). 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 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.

³⁴ 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.

TABLE 7 : **Explication des caractéristiques des modèles.** Résultats de l'estimation par moindres carrés (OLS) des modèles linéaires sélectionnés, pour chacune des variables à expliquer : échelle temporelle (TEMPSCALE), échelle spatiale (SPATSCALE), indice d'interdisciplinarité (INTERDISC), année de publication (YEAR).

	<i>Variable expliquée :</i>					
	TEMPSCALE		SPATSCALE		INTERDISC	
	(1)	(2)	(3)	(4)	(5)	(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
FMETHODdro						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

Régressions par Forêts Aléatoires

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 [liaw2002classification]. 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*

Développements

Un développement possible pourrait consister en la mise en place d'une approche automatique à cette meta-analyse, du point de vue de la modélisation modulaire, combiné avec une classification du but et de l'échelle. La modélisation modulaire consiste en l'intégration de processus hétérogènes et d'implémentation de ces processus dans le but d'extraire les mécanismes donnant la meilleure proximité à des faits stylisés empiriques ou à des données [cottineau2015incremental]. L'idée serait de pouvoir extraire automatiquement la structure modulaire des modèles existants, à partir des textes complets comme proposé en 2.2, afin de classifier ces briques de manière endogène et identifier des couplages potentiels pour des nouveaux modèles.

Leçons pour la modélisation

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 spatiale 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 8 : Synthèse des processus modélisés. Ceux-ci sont classés par échelle, type de modèle et discipline.

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

SYNTHÈSE DES PROCESSUS MODÉLISÉS

Nous proposons pour fixer les idées, pour une grande partie des modèles parcourus lors de la modélographie, de procéder à un effort similaire à celui concluant l'approche thématique du Chapitre 1, c'est à dire de synthétiser les processus pris en compte par ces modèles. On ne peut 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*³⁵.

La table 8 propose cette synthèse à partir des 145 articles issus de la modélographie et pour lesquels une classification de type était possible, c'est à dire qu'il existait un modèle rentrant dans la typologie développée en 2.1. Etre complètement exhaustif relèverait d'une opé-

³⁵ En gardant en tête les choix de sélection, qui emmènent par exemple à ne pas avoir les processus de mobilité dans cette synthèse.

ration de méta-modélisation interdisciplinaire qui est bien hors de la portée de notre travail³⁶, et la liste donnée ici est indicative.

On retrouve les correspondances entre disciplines, échelles et types de modèles obtenues dans la modélographie en 2.3. Nous retirons les enseignement principaux suivants, en écho au tableau de synthèse obtenu en fin du Chapitre 1 (Table 2) :

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-même³⁷. 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. Une discrépance importante 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 positionnement théoriques de la Théorie Evolutive. Cette adéquation, plus difficile à retrouver pour l'ensemble des autres approches revues, nous suggère également de suivre cette piste.

³⁶ 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.

³⁷ Puisqu'on a plus détaillé cette étude, elle paraît même plus forte aussi, une plus grande précision permettant alors de séparer des catégories abstraites.

CONCLUSION DU CHAPITRE

La réflexivité, au sens de la reflexion de la recherche sur les facteurs influençant son contenu et sa propre structure, semble dans notre cas être nécessaire pour une appréhension claire des enjeux thématiques, méthodologiques et plus généralement scientifiques liés au processus que nous cherchons à modéliser : ceux-ci étant multi-scalaires, hybrides et hétérogènes, les angles d'approches et questionnements possibles sont nécessairement extrêmement variés, complémentaires et riche. Il pourrait s'agir d'une caractéristique fondamentale des systèmes socio-techniques, que PUMAIN formule dans [pumain2005cumulativite] 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 Economie et Géographie développée en C.1. 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 citation 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 et une meta-analyse 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. On a alors à ce stade une idée assez précise de ce qui ce fait, pourquoi et comment. L'enjeu reste de déterminer les pertinences relatives de certaines approches ou ontologies, ce qui sera le but des trois 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

POSITIONNEMENTS

Toute activité de recherche serait, selon certains acteurs de celle-ci, nécessairement politisée, de par pour commencer le choix de ses objets. Ainsi, RIOLL alerte contre l'illusion d'une recherche objective et les dangers de la technocratie [ripoll2017jig]. Nous ne rentrons 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 [gutmann2011ethics]. 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 [hacking1999social]), les implications politiques et sociales de la recherche sont bien évidemment indiscutables. Là où il y aurait matière à discussion, et nous y reviendrons en ouverture 9.3.3 car il s'agira d'une des questions ouvertes posées par notre recherche et sa démarche dans leur ensemble, serait sur la compatibilité des méthodes systématiques et *evidence-based* avec les sciences sociales, autrement dit dans quelle mesure peut-on s'extraire de certains dogmatismes encore plus marqués lors de l'usage de théorie politiques². Nous resterons 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 cer-

C (SR) : refs autre que Ripol?

C (SR) : Il faut que tu développe un peu cela dit, même si le débat est vaste, il faut peut être trouver un entre deux entre ne presque rien dire, et trop en dire. Peut être exposer les principales théories sur le sujet ?

C (SR) : Une solution serait d'inverser l'argumentation, en partant des cas présentés ci-dessous pour ensuite généraliser par des citations d'auteurs discutant ce rapport science/politique

C (SR) : num de page et/ou citation en note de bas de page pour définir le concept pour le lecteur, ref plus précise en géographie/syville comme ex Batty 1976??

C (SR) : j'éviterai cette formulation

C (SR) : cad?

C (SR) : explicite plus pour le lecteur cette articulation possible que tu perçois entre compatibilité x/y et prise de recul dogmatisme, peut-être avec un ex?

¹ <https://www.cs-dc.org/>

² MONOD montre par exemple les désastres liés aux "niaiseries épistémologiques" décluant de l'application littérale de la dialectique matérialiste marxiste à l'épistémologie du vivant.

tains critères de scientifcité [**morin1991methode**], 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 ébauchée dans les deux chapitres précédents par les choix des objets d'étude, des problématiques, et seront renforcées à mesure de la progression pour finalement être synthétisées en Chapitre 9.

Nous proposons 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 (3.2), 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. Dans une deuxième section (3.1), 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. 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 [**banos2013pour**] pourra trouver dans les deux premières sections des illustrations pratiques originales de ceux-ci, notre positionnement étant principalement dans leur lignée.

C (SR) : Théoreme de l'entonnoir de banos, ne vaudrait il mieux pas partir du plus large et finir sur le plus resserré ? Ce qui permettrait de dire en quoi les deux premières sections actuelles (reproductibilité, calcul) servent un tout plus général (complexité en géo) ?

★ ★

*

Ce chapitre est composé de divers travaux. La première section est inédite. La deuxième section rend compte pour sa première partie du contenu théorique de [raimbault2016cautious], et pour sa deuxième partie des idées présentées dans [cottineau2017initial]. La troisième section reprend dans sa première partie les bases épistémologiques de [raimbault:halshs-01505084] approfondies par [raimbault2017knowledge], est inédite pour sa deuxième partie et rend compte de [raimbault2017complex] pour sa dernière partie.

3.1 MODÉLISATION, DONNÉES MASSIVES ET CALCUL INTENSIF

Nous nous positionnons à présent sur les questions liées à l'utilisation de la modélisation, des données massives et du calcul intensif, ce qui induit aussi par extension une réflexion sur les méthodes d'exploration de modèles. Il n'est pas évident que ces nouvelles possibilités soient nécessairement accompagnées de mutations épistémologiques profondes, et nous montrons au contraire que leur utilisation nécessite plus que jamais un dialogue avec la théorie. Implicitement, cette position préfigure le cadre épistémologique pour l'étude des Systèmes Complexes dont nous donnons le contexte en 3.3 et que nous formalisons en ouverture 9.3.

Les points développés ici couvrent certains enjeux cruciaux liés aux entreprises de modélisation, et peuvent être de nature épistémologique, théorique, ou pratique. Nous tenterons tout d'abord de répondre à la question du pourquoi de la modélisation. Nous nous positionnerons ensuite sur des questions plus techniques liées à l'utilisation des ressources de calcul émergentes et des nouvelles données. Enfin, le dernier point est méthodologique, et illustre à la fois les deux premiers tout en introduisant une nouvelle méthode d'exploration de modèles.

3.1.1 *Pourquoi modéliser ?*

Nous développons dans un premier temps le rôle de la modélisation dans notre démarche de production scientifique. Les modèles ont en apparence des rôles divers selon les disciplines : un modèle en physique découle d'une théorie, permet de la confronter à l'expérimentation et devra être validé par ses pouvoirs prédictifs avec de fortes exigences, tandis qu'en science sociales computationnelles on se contentera souvent de la reproduction de faits stylisés généraux. Un modèle statistique sera composé d'hypothèses sur des relations entre variables et sur la distribution statistique d'un terme d'erreur, et les valeurs des coefficients obtenus seront interprétées même si la mesure d'ajustement est très faible. Il s'agit donc ici de préciser dans quelle logique nos travaux de modélisation se placeront³, quels seront leurs ressorts et objectifs.

Fonctions des modèles

Comme nous venons de l'évoquer, le terme de *modèle* a de multiples sens, et implique différentes réalités, pratiques, utilisations (on peut supposer une ontologie propre aux modèles qui deviennent des ob-

³ Si ce travail pourra paraître redondant, laborieux et superflu aux habitués des modèles de géosimulation, il est crucial dans notre logique d'ouverture disciplinaire, afin d'une part d'éviter tout malentendu sur le statut des résultats, d'autre part d'encourager un dialogue dans le cas d'utilisations très différentes des modèles.

jets réels, au moins lorsque ceux-ci sont implémentés). Une façon d'en proposer une sorte de typologie est de procéder à celle de leurs fonctions, comme le fait [varenne2017theories], en se basant sur l'étude de diverses disciplines (biologie, géographie, sciences sociales). Cette classification est à notre connaissance la plus exhaustive qui existe. VARENNE distingue donc cinq grandes classes de fonctions des modèles⁴, qui vont de manière croissante dans leur intégration à une pratique sociale :

1. Fonction de perception et d'observation : rendre accessible un objet inobservable à la perception (modèle physique d'une molécule), permettre des expérimentations, une mémorisation, la lecture et la visualisation de données.
2. Fonction d'intelligibilité : description de motifs, précision des ontologies, conception par la prédiction, explication et compréhension de processus⁵.
3. Fonction d'aide à la théorisation : formulation, interprétation, illustration d'une théorie, test de cohérence interne (les schémas déductifs induisent-ils des résultats de simulation de modèles contradictoires ou cohérents ?), applicabilité, calculabilité (dans le cas de schémas numériques permettant d'approcher les solutions d'équations), co-calculabilité (couplage de théories et modèles).
4. Fonction de communication sociale : communication scientifique, concertation, action avec les acteurs (*stakeholders*⁶).
5. Fonction de prise de décision : aide à la décision, action, action auto-réalisatrice dans un système abstrait (modèles de pricing en finance).

Il est clair que chaque discipline va avoir sa propre relation à ces différentes fonctions, que certaines seront privilégiées, d'autre non accessibles ou sans pertinence pour les objets étudiés ou questions posées. En physique par exemple, les aspects de validation des théories et l'existence de modèles prédictifs d'une très grande précision

⁴ Les grandes classes de fonctions sont déclinées en classes précises qui sont au nombre de 21. Nous ne les détaillons pas ici, mais donnons une synthèse décrivant les grandes classes.

⁵ La compréhension est plus générale que l'explication, car suppose une reconstruction de la structure du système et un usage déductif, c'est à dire une projection et génération du système considéré dans la structure psychique le considérant [morin1980methode].

⁶ Nous ne développerons pas du tout cet aspect, mais tenons à préciser que les *stakeholder workshops* sont l'un des axes structurants du projet Medium que nous avons décrit en 1.3. Même si la percolation avec l'axe d'analyse et de modélisation des dynamiques des systèmes urbains dans lequel notre travail s'inscrit n'est pas explicite, celle-ci s'opère implicitement dans les échanges entre perspectives, et la cohabitation au sein d'un projet laisse supposer des perspectives futures plus intégrées.

sont au cœur de la discipline, tandis que des branches entières des sciences sociales comme par exemple la planification urbaine sont axées sur des modèles pour la communication et la prise de décision. A cet égard, il ne faut pas négliger la nature de science sociale de l'économie et douter des visées prédictives de certaines expériences de modélisation⁷.

Cette classification des fonctions se retrouve en filigrane dans les raisons de modéliser développées en dehors de toute typologie par [epstein2008model] : celui-ci insiste pour tordre le cou à l'idée préconçue que les modèles servent uniquement à la prédiction, et introduit diverses raisons, parmi lesquelles on retrouve des fonctions d'intelligibilité (explication, mise en évidence de dynamiques, révéler la complexité ou la simplicité), de soutien à la théorie (découverte de nouvelles questions, mettre en valeur des incertitudes, suggérer des analogies), d'aide à la décision (solutions de crise en temps réel, trouver des compromis d'optimisation), et de communication (éduquer le public, entraîner les praticiens).

Dans ce cadre de classification fonctionnelle des modèles, notre travail utilisera principalement les fonctions suivantes :

- Modèles descriptif et extraction de motifs : il s'agira des diverses analyses empiriques visant à établir des faits stylisés sur les processus de co-évolution dans des cas d'étude donnés.
- Modèles à visée explicative et de compréhension : les modèles simulant des dynamiques territoriales que nous construirons, avec comme objectif l'intégration des processus de co-évolution, auront pour objectif principal l'explication de faits stylisés en lien avec des processus (par exemple : les variations de tel paramètre correspondant à tel processus expliquent tel fait stylisé), et dans l'idéal la *compréhension* des systèmes⁸.
- Modèles pour éprouver la théorie : validation interne, c'est-à-dire cohérence du comportement du modèle par rapport aux faits stylisés impliqués par la théorie, et externe, au sens de reproduction plus ou moins performante de dynamiques de cas d'études précis considérés dans le cadre d'une théorie ; ou plus

C (SR) : Tu peux citer des refs plus anciennes en géo fr, comme Besse 2000 et Lena 2000 dans Geopoint 2000. Ces refs permettent aussi de dissocier l'outil de son utilisation. L'analyse générale de Varenne est intéressante, mais c'est important de la croiser avec l'expertise que les géographes ont eux même développé en interne (son dernier bouquin sur modèles en géographie de FVarenne devrait pouvoir apporter des réponses/refs aussi), ou les SHS (voir communauté sociologue G. Manzo, ou plus largement communauté JASSS)

⁷ Même en finance à des fréquences élevées, où les signaux seraient plus raisonnablement assimilables à des systèmes physiques que des séries macroéconométriques par exemple comme en témoignent l'appropriation de ces problématiques par les physiciens, la prédictibilité reste questionnable et en tout cas limitée [campbell2007predicting].

⁸ En fait, la frontière entre explication et compréhension est floue et subjective. Il est possible de considérer qu'il existe déjà un certain niveau de compréhension lorsqu'un modèle avec un certain niveau de cohérence interne et ontologique, en lien avec des hypothèses théoriques raisonnables et relativement autonomes, permet de tirer des conclusions sur les dynamiques globales du système considéré.

généralement pour répondre à une question ou hypothèse précise.

Modélisation générative

Le *type*⁹ de modèles que nous utiliserons majoritairement dans notre travail s'apparente à de la *modélisation générative*, au sens donné par [epstein2006generative] dans son manifeste pour des *Sciences Sociales Génératives*. Le principe fondamental est de proposer d'expliquer des régularités macroskopiques comme émergentes des interactions entre entités microscopiques, en simulant l'évolution du système de manière générative¹⁰. Ce paradigme peut être rapproché de celui du *Pattern Oriented Modeling* en Ecologie [grimm2005pattern], qui vise à expliquer par production de motifs par le bas¹¹. Les modèles basés-agents, c'est à dire des modèles impliquant un certain nombre d'agents hétérogènes relativement autonomes et simulant leur interactions, sont une façon d'y parvenir.

L'utilisation de modélisation générative peut être mise en correspondance forte avec la notion d'émergence faible introduite par [bedau2002downward]¹². Un système qui présente des propriétés émergentes au sens faible suppose que les propriétés du niveau supérieur (macro) doivent être entièrement dérivées par simulation, tout en restant réductible sur les plans causaux et ontologiques. En d'autre termes, le niveau macro ne possède pas de pouvoirs causaux irréductibles, cela n'étant pas

⁹ Dans la perspective fonctionnelle, les structures, contenus et processus, c'est-à-dire la nature des modèles en eux-même (ce qui correspond à la nature et aux principes des modèles évoqués mais non classifiés par VARENNE), sont donnés comme exemples en illustration, mais une fonction donnée n'est pas restreinte à un modèle donné (bien que réciproquement certains modèles ne puissent remplir certaines fonctions). Il n'existe à notre connaissance pas de typologie générale des modèles par *type*, qu'on pourrait alors définir en termes d'une typologie des relations avec les autres domaines de connaissance (voir 9.3) : par exemple un modèle utilisant telle méthodologie, privilégiant tel outil, un usage particulier ou privilégié de données, etc. Dans tous les cas, les typologies ou classification existantes de modèles sont associées aux revues de littérature et synthèse propres à chaque discipline : par exemple, [harvey1969explanation] (p. 157) propose une typologie générale qui reste toutefois inspirée de et limitée à la Géographie. Les conditions de typologies interdisciplinaires sont une question ouverte, dont l'exploration dépasse largement le propos de notre travail.

¹⁰ En gardant à l'esprit que la capacité à générer est bien sûr une composante nécessaire mais pas suffisante à l'explication, comme l'illustre le débat à ce sujet autour des travaux d'Epstein synthétisé par [rey2015plateforme] (p. 154).

¹¹ En effet, le POM vise à ce que le modèle reproduise par la simulation, c'est à dire *génère*, des motifs (*patterns*) attendus à différentes échelles, constituant un laboratoire virtuel dans lequel des hypothèses peuvent être testées. Par ailleurs, la générativité d'Epstein se base sur des paradigmes similaires pour l'explication, impliquant des modèles à la complexité progressive et qui permettent le test d'hypothèse, en isolant des mécanismes suffisant pour reproduire des motifs macroskopiques.

¹² On rappelle que l'émergence faible correspond à l'émergence de propriété à un niveau supérieur qui doivent être effectivement computées par le système pour être connues.

incompatible avec l'existence de *downward causation* et son autonomie. Certains systèmes¹³ ne tombent pas dans cette catégorie à notre connaissance actuelle puisqu'on n'est pas capable de désigner des éléments microscopiques causaux. En revanche, des systèmes qu'on comprend mal mais qui se simulent eux-même et dont on est certain que l'état macro émerge des interactions microscopiques (prenons le traffic et la congestion par exemple), sont des parfaites illustrations de cette notion. Les exemples donnés par BEDAU pour démontrer son propos sont des automates cellulaires en deux dimensions, pour lesquels le rôle de la computation est évident et la *downward causation* peut être illustrée par le comportement des structures macroscopiques du jeu de la vie qui agissent rétroactivement sur les cellules. Connaitre les dynamiques de systèmes faiblement émergents nécessite par définition de les simuler, et donc de les modéliser¹⁴, cette approche est ainsi naturelle pour connaître la structure ou les processus dans un système complexe.

Le modèle comme outil de connaissance indirecte

Ainsi, nos modèles seront principalement à visée de comprehension (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 [pumain2017urban], 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

¹³ Comme le montrent la conscience en neuroscience et psychologie, ou les débats sur l'existence et l'autonomie "d'êtres sociaux" en sociologie [angeletti2015etres], pour lesquels nous pourrions ne connaître que partie seulement des éléments causaux microscopiques à savoir les individus.

¹⁴ Sur la différence entre simulation de modèle et modèle de simulation, [phan2010agent] explique dans quelle mesure ces deux notions peuvent être distinguées, mais que cela n'implique pas de différence fondamentale pour l'application concrète : la simulation d'un modèle consiste en l'opération de computation des états successifs d'un modèle dans une configuration donnée, tandis qu'un modèle de simulation est un modèle conçu pour la simulation d'un système (par exemple un modèle génératif) ou la simulation d'un autre modèle (par exemple les schémas numériques pour approcher des équations). Dans tous les cas, l'utilisation du modèle impliquera simulation d'un modèle. Ces remarques se vérifient particulièrement dans le cas de la modélisation générative. Nous utiliserons les deux de manière interchangeable par la suite.

sans tomber dans le piège de l'équifinalité¹⁵. Ainsi, l'établissement des *Calibration Profiles* du modèle SimpopLocal [reuillon2015] 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.

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 [perez2016agent], 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*¹⁶.

Comment explorer un modèle de 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 [bonhomme2017dictionnaire], 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 [grimm2014towards] 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 [reuillon2013openmole] pour les principes fondamentaux, la documentation en ligne¹⁷ pour un aperçu global des méthodes disponibles et de leur articulation dans un cadre standard, et [pumain2017innovative] pour une contextualisation des différentes méthodes. Ces travaux¹⁸ ont apporté une nombre considérable d'innovations à la fois méthodologiques, techniques, thématiques et théoriques. La philosophie d'Open-

¹⁵ 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 [von1972history]. 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.

¹⁶ Le rôle de la visée d'application des modèles est lié à la fois à une sensibilité disciplinaire, comme le domaine des Luti [wegeren2004land] qui l'est bien plus que celui de la géographie théorique et quantitative, mais aussi à une sensibilité "culturelle", comme l'illustre [batty2013new] qui montre une branche de la géographie anglo-saxonne plus proche des applications concrètes.

¹⁷ à <https://next.openmole.org/Models.html>

¹⁸ La majorité ayant été réalisés dans le cadre interdisciplinaire de l'ERC Geodiversity.

Mole s'articule autour de trois axes (voir entretien avec R. REUILLO) : 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 multi-modélisation, c'est à dire structure variable du modèle [cottineau2015modular], changement de la nature des questions posées au modèle (par exemple détermination complète de l'espace faisable [[10.1371/journal.pone.0138212](#)]), tout cela permis par l'utilisation du calcul intensif [schmitt2014half].

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 paramètres cruciaux associés, possiblement des métaparamètres, ainsi que de leur domaine thématique ; identification des indicateurs pour évaluer la performance ou le comportement du modèle.
2. Evaluation 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. Evaluation de la sensibilité aux métaparamètres, suivant la méthodologie innovante développée par la suite.
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*)¹⁹
6. Retours sur le modèle, extension et nouvelles briques de multi-modélisation, 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

C (SR) : A mon avis il manque à côté de l'identification des paramètres, l'identification des "mécanismes génératrices", et des critères d'évaluation associés, tout aussi important, qui permettront la planification dans la complexification du modèle. De façon plus générale il faut que tu introduise ce que tu entend par multi-modélisation, métaparamètres, paramètres, mécanismes, etc. en amont, de façon très claire, en abusant de citation dans le texte, ou de refs :)

¹⁹ Nous ne pratiquerons quasiment pas ce dernier point, trouvant suffisamment de réponses à nos questions avec les points précédents.

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.

Lien entre modélisation et Science Ouverte

Enfin, il est important de souligner brièvement les liens entre pratiques de modélisation et science ouverte, en parallèle du lien entre reproductibilité et Science Ouverte que nous ferons à la fin de 3.2. En fait, la Science Ouverte est composée d'un ensemble de pratiques se déclinant sur différents domaines, d'où sa ventilation logique dans nos positionnements. Pour illustrer les enjeux, nous proposons de décrire l'exemple des workflows d'exploration de modèle comme une méthode de meta-analyse de sensibilité, c'est à dire un aspect de la méthodologie appliquée ci-dessus.

Les idées de multi-modélisation et d'exploration intensive de modèle sont tout sauf nouvelles puisque OPENSHAW défendait déjà le "model-crunching" dans [openshaw1983data], mais leur utilisation effective commence seulement à émerger grâce à l'apparition de nouvelles méthodes et outils en même temps qu'une explosion des capacités de calcul : [cottineau2016back] propose une approche renouvelée de la multi-modélisation. Le couplage de modèles tel que nous l'opérons répond à des questions similaires. Dans cette lignée de recherche, la plateforme d'exploration de modèle OpenMole [reuillon2013openmole] permet d'embarquer n'importe quel modèle comme une boîte noire, d'écrire des workflow d'exploration modulables qui utilisent des méthodologies d'exploration avancées comme des algorithmes génétiques, et de distribuer de manière transparente les calculs sur des infrastructures de calcul à grande échelle comme des clusters ou grilles de calcul. Dans le cas précédent, l'outil du workflow est un outil puissant pour intégrer à la fois l'analyse de sensibilité et la meta-analyse de sensibilité, et permet de coupler n'importe quel générateur avec n'importe quel modèle de façon très directe, à la condition minimale que le modèle puisse être paramétré sur sa configuration spatiale initiale, par la donnée de meta-paramètres ou d'une configuration entière.

D'autre part, une idée des workflow est de favoriser des constructions ouvertes et collaboratives, puisque le "marketplace" d'OpenMole, directement intégré au logiciel²⁰, permet de bénéficier directement des exemples qui auront été partagés sur le dépôt collaboratif. Cela ressemble aux plateformes de partage de modèles, qui sont nombreuses pour les modèles agents par exemple, mais dans un esprit encore plus modulaire et participatif. Ainsi, certains choix épistémologiques et méthodologiques au regard de la modélisation impliquent directement un positionnement au regard de la science ouverte : la multi-modélisation et les familles de modèles, qui vont de pair avec le

²⁰ autrement accessible à <https://github.com/openmole/openmole-market>

couplage de modèle hétérogènes et multi-échelles, ne peuvent guère être viables sans des pratiques d'ouverture, de partage et de construction collaborative des modèle, comme le rappelle [banos2013pour].

Enfin, l'un des visages de la construction de connaissances ouvertes est la pédagogie. [chen2006effectiveness] propose la simulation comme outil pour apprendre aux élèves ingénieurs les processus sous-jacents aux systèmes qu'ils seront amenés à concevoir et gérer. Cet aspect est également à garder en tête de par son caractère performatif : les modèles ont alors une retroaction sur les situations réelles, ce qui complexifie encore le système considéré.

Synthèse

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 [varenne2017theories], 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 [pumain2017urban].
3. Cette façon de modéliser à la fois suppose et participe à une démarche de Science Ouverte [fecher2014open].

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

3.1.2 Pour un usage raisonné des données massives et de la computation

La *révolution des données massives* réside autant dans la disponibilité de grands jeux de données de nouveaux types variés, que dans la puissance de calcul potentielle toujours en augmentation. Même si le *tournant computationnel* ([arthur2015complexity]) est central pour une science consciente de la complexité et est sans doute la base des pratiques de modélisation futures en géographie comme [banos2013pour] souligne, nous soutenons que à la fois le *déluge de données* et les *capacités de calcul* sont dangereuses si non cadrées dans un cadre théorique et formel propre. Le premier peut biaiser les directions de recherche vers les jeux de données disponibles avec le risque de se déconnecter d'un fond théorique, tandis que le second peut occulter des résolutions analytiques préliminaires essentielles pour un usage cohérent des simulations. Nous avançons que les conditions pour la majorité

des résultats dans cette thèse sont en effet ceux mis en danger par un enthousiasme inconsidéré pour les données massives, tirant la conclusion qu'un challenge majeur pour la géocomputation future est une intégration sage des nouvelles pratiques au sein du corpus existant de connaissances.

La puissance de calcul disponible semble suivre un tendance exponentielle, comme une sorte de loi de Moore. Grace à d'une part la loi de Moore effective pour le matériel, d'autre part l'amélioration des logiciels et algorithmes, conjointement avec une démocratisation de l'accès au infrastructures de simulation à grande échelle, permet à toujours plus de temps processeur d'être disponible pour le chercheur en sciences sociales (et pour le scientifique en général, mais cette mutation a déjà été opérée depuis plus longtemps dans d'autres domaines). Il y a environ une dizaine d'année, [[gleyze2005vulnerabilite](#)] était forcé de conclure que les analyses de réseau, pour les transports publics parisiens, étaient "limitées par le calcul". Aujourd'hui la plupart des mêmes analyses seraient rapidement réglée sur un ordinateur personnel avec les logiciels et programmes appropriés : [[2015arXiv151201268L](#)] est un témoin d'un tel progrès, introduisant des nouveaux indicateurs avec une plus grande complexité de calcul, qui sont calculés sur des réseaux à grande échelle. Le même parallèle peut être fait pour les modèles Simpop : les premiers modèles Simpop au début du millénaire [[anders1997simpop](#)] étaient "calibrés" à la main, tandis que [[cottineau2015modular](#)] calibre le modèle Marius en multi-modélisation et [[schmitt2014half](#)] calibre très précisément le modèle SimpopLocal, chacun sur la grille avec des milliards de simulations. Un dernier exemple, le champ de la *Space Syntax*, a témoigné d'une longue route et de progrès considérables depuis ses origines théoriques [[hillier1989social](#)] jusqu'à ses récentes applications à grande échelle [[hillier2016fourth](#)].

Concernant les nouvelles données "massives" qui sont disponibles, il est clair que des quantités toujours plus grandes et des types toujours nouveaux sont disponibles. De nombreux exemples de champs d'application peuvent être donnés. La mobilité en est typique, puisque étudiée selon divers points de vue, comme les nouvelles données issues des systèmes de transport intelligents [[o2014mining](#)], des réseaux sociaux [[frank2014constructing](#)], ou des données plus exotiques comme des données de téléphonie mobile [[de2016death](#)]. Dans un autre esprit, l'ouverture de jeux de données "classiques" (comme les applications synthétiques urbaines, les initiatives gouvernementales pour les données ouvertes) devrait pouvoir toujours plus de métanalyses. De nouvelles façon de pratiquer la recherche et produire des données sont également en train d'émerger, vers des initiatives plus interactives et venant de l'utilisateur. Ainsi, [[2016arXiv160606162C](#)] décrit une application web ayant pour but de présenter une métanalyse de la loi de Zipf sur de nombreux jeux de données, mais en

particulier inclut une option de dépôt, à travers laquelle l'utilisateur peut télécharger son propre jeu de données et l'inclure dans la métanalyse. D'autres applications permettent l'exploration interactive de la littérature scientifique pour une meilleure connaissance d'un horizon scientifique complexe, comme [[cybergeo20](#)] fait.

Comme toujours la situation n'est naturellement pas aussi idyllique qu'elle semble être au premier abord, et l'herbe verte du pré du voisin que nous pouvons être tentés d'aller brouter se transforme rapidement en un triste fumier. En effet, les objectifs et motivations d'un grand nombre d'approches restent flous et on peut facilement s'y perdre. Des illustrations parleront d'elles-mêmes. [[barthelemy2013self](#)] introduit un nouveau jeu de données et des méthodes relativement nouvelles pour quantifier l'évolution du réseau de rues, mais les résultats, sur lesquels les auteurs semblent s'étonner, sont qu'une transition a eu lieu à Paris à l'époque d'Haussmann. Tout historien de l'urbanisme s'interrogerait sur le but exact de l'étude, puisque à la fin un sentiment étrange de réinvention de la roue flotte dans l'air. L'utilisation des ressources de calcul peut également être exagéré, et dans le cas de la modélisation multi-agent, on peut citer [[axtell2016120](#)], pour lequel l'objectif de simuler le système à l'échelle 1 : 1 semble être loin des motivations et justifications originelles de la modélisation agent, et pourrait même donner des arguments aux économistes *mainstream* qui dénigrent facilement les ABMS. D'autres anecdotes peuvent inquiéter : il existe en ligne des exemples étonnantes, comme une application web²¹ qui utilise des ressources de calcul pour simuler des distributions Gaussiennes afin de calculer pour un modèle de Gibrat, afin de calculer leur moyenne et variance, qui sont des paramètres d'entrée du modèle. En résumé, cela revient à vérifier le Théorème de la Limite Centrale. D'autre part, la distribution complète donnée par un modèle de Gibrat est entièrement connue théoriquement comme résolu e.g. par [[gabaix1999zipf](#)]. Sur ce point, nous devons partiellement être en désaccord avec le neuvième commandement de BANOS, qui rappelle que "les mathématiques ne sont pas le langage universel des modèles", ou plutôt souligner les dangers d'une mauvaise interprétation de ce principe²² : il postule que des moyens alternatifs aux mathématiques existent pour faire comprendre des processus ou des méthodes, mais précise que ceux-ci sont une porte d'entrée et ne prétend jamais qu'il est possible de se passer des mathématiques, dérive que l'exemple précédent illustre parfaitement. D'ailleurs, il est possible d'exhiber des structures mathématiques très simples, comme un simplexe en dimension quelconque, dont la visualisation "simple" est un problème ouvert. Les données fournissent aussi leur collec-

²¹ voir <http://shiny.parisgeo.cnrs.fr/gibratsim/>

²² De manière générale, les commandements de BANOS paraissent simples dans leur formulation, mais sont d'une profondeur et d'une complexité déconcertante lorsqu'on essaye d'en tirer les implications et la philosophie globale sous-jacente, et ne doivent jamais être pris à la légère.

tion de dérives. Récemment, sur la liste de diffusion de géographie francophone *Geotamtam*, un soudain engouement autour des données issues de *Pokemon Go* a semblé répondre plus à un besoin urgent et inexplicable d'exploiter cette source de données avant tous les autres, plutôt qu'à des considérations théoriques élaborées. Des jeux de données existant et précis, comme la population historique des villes (pour la France la base Pumain-INED par exemple), sont loin d'être entièrement exploités et il pourrait être plus pertinent de se concentrer sur ces jeux de données classiques qui existent déjà. De même, il faut être conscient des possibles applications de résultats basée sur des malentendus : [louail2016crowdsourcing] analyse la redistribution potentielle des transactions de carte bancaire au sein d'une ville, mais présente les résultats comme la base possible de recommandations de politiques pour une équité sociale en agissant sur la mobilité, oubliant que la forme et les fonctions urbaines sont couplés de manière complexe et que déplacer des transactions d'un endroit à un autre implique des processus bien plus complexes que des régulations directes, qui d'autant plus ne s'appliquent jamais de la façon prévue et conduisent à des résultats un peu différents. Une telle attitude, souvent observée de la part de physiciens, est très bien mise en allégorie par la figure 12 qui n'est qu'à moitié une exagération de certaines situations.

Notre principal argument est que le tournant computationnel et les pratiques de simulation seront centrales en géographie, mais peuvent également être dangereux, pour les raisons illustrées ci-dessus, i.e. que le déluge de données peut imposer les sujets de recherche et occulter la théorie, et que la computation peut éluder la construction et la résolution de modèles. Un lien plus fort est nécessaire entre les pratiques de calcul, l'informatique, les mathématiques, les statistiques et la géographie théorique. La Géographie Théorique et Quantitative est au centre de cette dynamique, puisqu'il s'agit de sa motivation initiale principale qui semble oubliée dans certains cas. Cela implique un besoin de recherche de théorie élaborées intégrées avec des pratiques de simulation conscientes. En d'autres mots, on peut répondre à des questions naïves complémentaires qui ont toutefois besoin d'être traitées une bonne fois pour toutes. Si une géographie quantitative libérée de la théorie serait possible, la réponse est naturellement non puisque cela se rapproche du piège de la fouille de données par boîte noire. Quoi qu'il soit fait par cette approche, les résultats auront un pouvoir explicatif très faible, puisqu'ils pourront mettre en valeur des relations mais pas reconstruire des processus. D'autre part, la possibilité d'une géographie quantitative purement basée sur le calcul est une vision dangereuse : même le gain de trois ordres de grandeur dans la puissance de calcul disponible ne résout pas le sort de la dimension. Prenons l'exemple des résultats de non-stationnarité obtenus en 4.1. L'utilisation de données relativement



FIGURE 12 : De l'usage naïf de la fouille de données et du calcul intensif. Source : xkcd

massives, de par les algorithmes spécialement conçus pour être capable de faire les traitements, est une condition nécessaire au résultat obtenus, mais à la fois l'échelle est les objets (c'est à dire les indicateurs calculés) sont co-déterminés par les constructions théoriques et les autres études empiriques. En effet l'absence de théorie impliquerait de ne pas connaître les objets, mesures et propriétés à étudier (e.g. le caractère multi-scalaire ou dynamique des processus), et sans résolutions analytiques, il serait souvent difficile de tirer des conclusions à partir des analyses empiriques seules concernant l'ergodicité par exemple. Rien n'est vraiment nouveau ici mais cette position doit être affirmée et tenue, précisément car notre travail se base sur ce type d'outils, essayant d'avancer sur une arête fine et fragile, avec d'un côté le vide du charlatanisme théorique infondé et de l'autre l'abîme de l'overdose technocratique dans des quantités de données folles. Plus que jamais on a besoin de théories simples mais fondées et puissantes à-la-Occam [batty2016theoretical], pour permettre une intégration saine des nouvelles techniques au sein des connaissances existantes.

3.1.3 *Etendre les analyses de sensibilité*

Contexte

Lors de l'évaluation de modèle basés sur les données, ou même de modèle plus simples partiellement basés sur les données impliquant une paramétrisation simplifiée, une issue inévitable est le manque de contrôle sur les "paramètres implicites du systèmes" (ce qui n'est pas une notion stricte mais doit être vu dans notre sens comme les paramètres régissant la dynamique). En effet, une statistique issue d'executions du modèle sur un nombre suffisant d'executions peut

toutefois rester biaisée, au sens où il est impossible de savoir si les résultats sont dus aux processus que le modèle cherche à traduire ou à une structure présente dans les données initiale. La question méthodologique fondamentale qui nous intéressera pour la suite est d'être capable d'isoler les effets propres aux processus du modèles de ceux liés à la géographie.

RATIONNELLE Bien que les modèles de simulation des systèmes géographiques en général et les modèles basés-agent en particulier représentent une opportunité considérable d'explorer les comportements socio-spatiaux et de tester une variété de scénarios pour les politiques publiques, la validité des modèles génératifs est incertaine tant que la robustesse des résultats n'a pas été établie. Les analyses de sensibilité incluent généralement l'analyse des effets de la stochasticité sur la variabilité des résultats, ainsi que les effets de variations locales des paramètres. Cependant, les conditions spatiales initiales sont généralement prise pour données dans les modèles géographiques, laissant ainsi totalement inexploré l'effet des motifs spatiaux sur les interactions des agents et sur leur interaction avec l'environnement. Dans cette partie, nous présentons une méthode pour établir l'effet des conditions spatiales initiales sur les modèles de simulation, utilisant un générateur systématique contrôlé par des meta-paramètres pour créer des grilles de densité utilisées dans les modèles de simulation spatiaux. Nous montrons, avec l'exemple d'un modèle agent très classique (le modèle Sugarscape d'extraction de ressources) que l'effet de l'espace dans les simulations est significatifs, et parfois plus grand que l'effet des paramètres eux-mêmes. Nous y arrivons en utilisant le calcul haute performance en un workflow très simple et open source. Les bénéfices de notre approche sont variés mais incluent par exemple la connaissance du comportement du modèle dans un contexte plus large, la possibilité de contrôle statistique pour régresser les sorties du modèles, ou une exploration plus fine des dérivées du modèle que par rapport à une approche directe.

ROLE DE LA DÉPENDANCE AU CHEMIN SPATIO-TEMPORELLE La dépendance au chemin spatio-temporelle est une des raisons principales rendant notre approche pertinente. En effet, un aspect crucial de la plupart des systèmes complexes spatio-temporels est leur non-ergodicité [**pumain2012urban**] (la propriété que les échantillons cross-sectionnels dans l'espace ne sont pas équivalents aux échantillons dans le temps pour calculer des statistiques comme la moyenne), qui témoigne généralement de forte dépendances au chemin spatio-temporelles dans les trajectoires. De manière similaire à ce que GELL-MANN appelle *frozen accidents* dans tout système complexe [**gell1995quark**], une configuration donnée contient des indices sur les bifurcations passées, qui peuvent avoir eu des effets considérables sur l'état du

système. Les effets temporels et cumulatifs ont été considérés dans de nombreux sous-champs géographiques et à différentes échelles géographiques, par exemple les systèmes régionaux [Wilson1981] ou l'échelle intra-urbaine [AllenSanglier1979]. L'impact de la configuration spatiale sur les dynamiques du modèle et les bifurcations spatiales a été moins étudié.

L'exemple des réseaux de transport est une bonne illustration, car leur forme spatiale et leur hiérarchie est fortement influencée par les décisions d'investissement du passé, les choix techniques, ou des décisions politiques qui ne sont parfois pas rationnelles [zembris2010new]. Certains indicateurs agrégés ne prendront pas en compte les positions et trajectoires de chaque agent (comme les inégalités totales dans le modèle Sugarscape) mais d'autres, comme dans le cas des motifs d'accèsibilité spatiale dans un système de villes, capture entièrement la dépendance au chemin et peuvent ainsi être fortement dépendants à la configuration spatiale initiale. Il n'est pas clair par exemple ce qui a causé la transition de la capitale française de Lyon à Paris dans le bas Moyen-Age, certaines hypothèses étant la reconfiguration des motifs commerciaux du Sud au Nord de l'Europe et donc une centralité accrue pour Paris due à sa position spatiale, tout en gardant à l'esprit que les centralité géographique et politique ne sont pas équivalentes et entretiennent une relation complexe [guenee1968espace]. La bifurcation induite par des facteurs socio-économiques et politiques a pris une signification profonde avec des répercussions mondiales encore aujourd'hui quand elle a été concrétisée par la configuration spatiale.

TRAVAUX EXISTANTS L'effet de la configuration spatiale sur les attributs agrégés à la zone des comportements humains a été largement discuté en géostatistiques, approximativement depuis l'introduction du *Modifiable Areal Unit Problem* (MAUP) [Openshaw1984]. Plus récemment, [Kwan2012] plaide pour un examen plus attentif de ce qui serait un *Uncertain Geographic Context Problem* (UGCoP), qui est la configuration spatiale des unités géographiques même si la taille et la délimitation des zones est la même. Au contraire, le faible nombre de considérations similaires dans la littérature traitant des modèles de simulation géographiques remet en question la généralisation de leur résultats, comme cela a été montré par exemple dans le cas des modèles LUTI [Thomasetal2017], ou des processus de diffusion étudiés par modèles basé-agents [LeTexierCaruso2017].

Méthodes

Nous détaillons à présent la méthode développée pour analyser la sensibilité des modèles de simulation aux conditions spatiales initiales. S'ajoutant au protocole usuel, qui consiste à simuler un modèle μ pour différentes valeurs de ses paramètres et faire le lien entre ces variations aux variations des résultats de simulation, nous intro-

duisons ici un générateur spatial, qui est lui-même déterminé par des paramètres et produit des ensembles de configurations spatiales initiales. Les configurations spatiales initiales sont catégorisées pour représenter des types d'espace typiques (par exemple des grilles de densité monocentriques ou polycentriques), et la sensibilité du modèle est à présent testée sur les paramètres de μ mais aussi sur les paramètres spatiaux ou les types spatiaux. Cela permet à l'analyse de sensibilité de fournir des conclusions qualitatives au regard de l'influence de la distribution spatiale sur les sorties des modèles de simulation, en parallèle des variation classiques des paramètres.

GÉNÉRATEUR SPATIAL Le générateur spatial applique un modèle de morphogenèse urbaine développé et exploré en 5.2. Pour le présenter rapidement, les grilles sont générées par un processus itératif qui ajoute une quantité de population N à chaque pas de temps, l'allouant selon un attachement préférentiel caractérisé par sa force d'attraction α . Le premier processus est ensuite lissé n fois par un processus de diffusion de force β . Les grilles sont donc générées aléatoirement par la combinaison des valeurs de ces quatre meta-paramètres α , β , n and N . Pour faciliter l'exploration, seule la distribution de densité est autorisée à varier plutôt que la taille de la grille, qui est fixée à un environnement carré 50x50 de population 100,000 unités.

COMPARER LES DIAGRAMMES DE PHASE Afin de tester l'influence des conditions spatiales initiales, nous avons besoin d'une méthode systématique pour comparer des diagrammes de phase. En effet, nous avons autant de diagramme de phase que de grilles spatiales, ce qui rend une comparaison visuelle qualitative non réaliste. Une solution est d'utiliser des procédures quantitatives systématiques. De nombreuses méthodes pourraient potentiellement être utilisées : par exemple, des indicateurs anisotropes comme la donnée de clusters et leur position dans le diagramme de phase, peuvent permettre de révéler des *meta-transitions de phase* (transition de phase dans l'espace des meta-paramètres. L'utilisation de métriques comparant des distributions spatiales, comme la *Earth Movers Distance* qui est utilisée en viion par ordinateur pour comparer des distributions de probabilité [rubner2000earth], ou la comparaison de matrices de transition agrégées de la dynamique associée au potentiel décrit par chaque distribution, est également possible. Les méthodes de comparaison de cartes, répandues en sciences environnementales, fournissent de nombreux outils pour comparer des champs en deux dimensions [visser2006map]. Pour comparer un champ spatial évoluant dans le temps, des méthodes élaborées comme les Fonctions Orthogonales Empiriques qui isolent les variations temporelles des variations spatiales, seraient applicables dans notre cas en prenant le temps comme une dimension de paramètre, mais celles-ci ont été montrées ayant une perfor-

mance similaire à la comparaison visuelle directe lorsqu'on prend la moyenne sur un ensemble de contributions crowdsourcées [[10.1371/journal.pone.0178165](#)]. Pour rester simple et car de telles considérations méthodologiques sont auxiliaire pour le propos principal de cette partie, nous proposons une mesure intuitive correspondant à la part de la variabilité inter-diagrammes relativement à leur variabilité interne. Plus formellement, cette distance est donnée par

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

où $\alpha \mapsto [\vec{x} \mapsto f_\alpha(\vec{x})]$ est l'opérateur donnant les diagrammes de phase avec \vec{x} paramètres et $\vec{\alpha}$ meta-paramètres, et d une distance entre distributions de probabilité qui peut être prise par exemple comme la distance L2 basique ou la *Earth Movers Distance*. Pour chaque valeur $\vec{\alpha}_i$, le diagramme de phase est vue comme un champ spatial aléatoire, ce qui facilite la définition des variances et de la distance.

Résultats

Sugarscape est un modèle d'extraction de ressources qui simule la distribution inégale des richesses dans une population hétérogène [[EpsteinAxtell1996](#)]. Des agents ayant différentes portées de vision et différents métabolismes collectent une ressource qui se régénère automatiquement et disponible de manière hétérogène dans le paysage initial. Ceux-ci s'établissent et collectent la ressource, ce qui mène certains d'entre eux à survivre et d'autres à périr. Les paramètres principaux du modèle sont le nombre d'agents, leur ressources minimale et maximale. Nous nous intéressons en prime à tester l'impact de la distribution spatiale, en utilisant le générateur spatial. La sortie du modèle est mesurée comme le diagramme de phase d'un index d'inégalité pour la distribution de la ressource (index de Gini). Nous étendons l'implémentation ayant initialement une distribution de richesse des agents, donnée par [[li2009netlogo](#)].

Pour l'exploration, 2,500,000 simulations (1000 points de paramètres \times 50 grilles de densité \times 50 réplications) nous permettent de montrer que le modèle est bien plus sensible à l'espace qu'à ses autres paramètres, à la fois quantitativement et qualitativement : l'amplitude des variations entre les grilles de densité est plus grande que l'amplitude dans chaque diagramme de phase, et le comportement de ces diagrammes de phase est qualitativement différents dans diverses régions de l'espace morphologique. Plus précisément, nous explorons une grille d'un espace de paramètre basique du modèle, dont les trois dimensions sont la population des agents $P \in [10; 510]$, la ressource minimale initiale par agent $s_- \in [10; 100]$ et la ressource initiale maximale par agent $s_+ \in [110; 200]$. Chaque paramètre est discréteisé en 10

valeurs, donnant 1000 points de paramètres. Nous procédons à 50 répétitions pour chaque configuration, ce qui donne des propriétés de convergence raisonnables. La distribution spatiale initiale varie parmi 50 grilles initiales, générée en échantillonnant les meta-paramètres du générateur dans un Hypercube Latin. Nous démontrons ainsi la flexibilité de notre cadre, par le couplage séquentiel direct du générateur avec le modèle. Nous mesurons la distance de l'ensemble des diagrammes de phase à 3 dimensions à un diagramme de phase de référence calculé sur l'initialisation du modèle par défaut (voir Fig. 13 pour sa position morphologique au regard des grilles générées), en utilisant l'équation 1 avec la distance L₂ pour assurer une interprétabilité directe. En effet, cela donne dans ce cas la distance au carré moyenne entre chaque points en correspondance des diagrammes, relative à la moyenne des variances de chaque. Pour cela, des valeurs plus grandes que 1 signifient que la variabilité inter-diagramme est plus importante que la variabilité intra-diagramme.

Nous obtenons une sensibilité très forte aux conditions initiales, puisque la distribution de la distance relative à la référence s'étend sur l'ensemble des grilles de 0.09 à 2.98, avec un médiane de 1.52 et une moyenne de 1.30. Cela signifie qu'en moyenne, le modèle est plus sensible aux meta-paramètres qu'aux paramètres, et que la variation relative peut atteindre jusqu'à un facteur 3. Nous montrons en Fig. 13 leur distribution dans un espace morphologique. L'espace morphologique réduit est obtenu en calculant 4 indicateurs bruts de forme urbaine, qui sont l'index de Moran, la distance moyenne, le niveau de hiérarchie et l'entropie (voir [LeNechet2015] ainsi que la section 5.2 pour une définition précise et une mise en contexte), et en réduisant la dimension avec une analyse par composantes principales pour laquelle nous gardons les deux premières composantes (92% de variance cumulée). La première mesure un "niveau d'étalement" et d'éclatement, tandis que la seconde mesure l'agrégation.²³ Nous trouvons que les grilles produisant les déviations les plus grandes sont celles avec un faible niveau d'étalement et une forte agrégation. Cela est confirmé par le comportement comme fonction des meta-paramètres, puisque des fortes valeurs de α donnent aussi une forte distance. En terme de processus du modèle, cela montre que les mécanismes de congestion induisent rapidement de plus haut niveau d'inégalités.

Nous contrôlons à présent la sensibilité en terme de comportement qualitatif des diagrammes de phase. Nous montrons en Fig. 14 les diagrammes pour deux morphologies très opposées en terme d'étalement, mais en contrôlant l'agrégation par la même valeur de PC2. Ceux-ci correspondent au cadres vert et bleu en Fig. 13. Les comportements sont relativement stables pour s_+ variant, ce qui signi-

²³ nous avons $PC1 = 0.76 \cdot \text{distance} + 0.60 \cdot \text{entropy} + 0.03 \cdot \text{moran} + 0.24 \cdot \text{slope}$ et $PC2 = -0.26 \cdot \text{distance} + 0.18 \cdot \text{entropy} + 0.91 \cdot \text{moran} + 0.26 \cdot \text{slope}$.

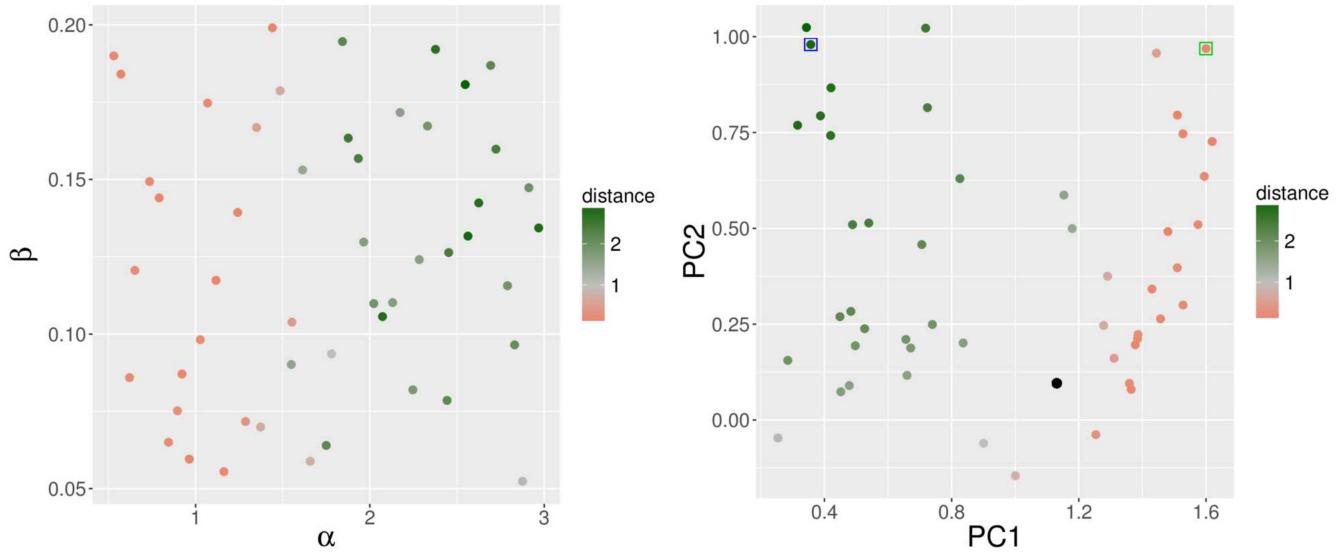


FIGURE 13 : Distance relative des diagrammes de phase à la référence pour l’ensemble des grilles. (Gauche) Distance relative comme fonction des meta-paramètres α (force de l’attachement préférentiel) et la diffusion (β , force du processus de diffusion). (Droite) Distance relative comme fonction des deux composantes principales de l’espace morphologique (voir texte). Le point rouge correspond à la configuration spatiale de référence. Les cadres verts et bleu donnent respectivement le premier et le second diagrammes particuliers montrés à la Fig. 14.

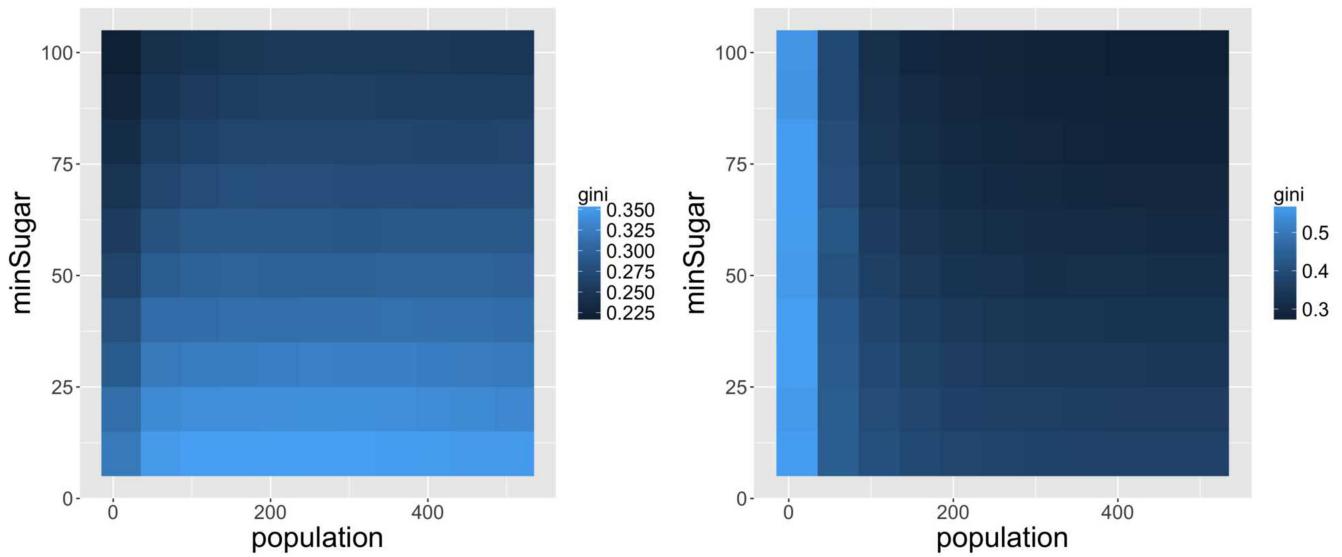


FIGURE 14 : Exemples de diagrammes de phase. Nous montrons deux diagrammes bi-dimensionnels sur (P, s_-) , obtenus à $s_+ = 110$ fixé. (Gauche) Cadre vert, obtenu avec $\alpha = 0.79$, $n = 2$, $\beta = 0.14$, $N = 157$; (Droite) Cadre bleu, obtenu avec $\alpha = 2.56$, $n = 3$, $\beta = 0.13$, $N = 128$.

fie que les agents les plus pauvres ont un rôle déterminant dans les trajectoires. Les deux exemples ont non seulement une inégalité de base très distante (le plafond du premier 0.35 est environ le plancher du second 0.3), mais leur comportement qualitatif est également radicalement opposé : la configuration étalée donne des inégalités qui décroissent quand la population décroît et qui décroissent quand la richesse minimale augmente, tandis que la concentrée donne des inégalités augmentant fortement quand la population décroît et aussi décroissantes avec la richesse minimale mais significativement seulement pour des grandes valeurs de population. Le processus est ainsi complètement inversé, ce qui aurait un impact déterminant si l'on essayait de schématiser des politiques à partir du modèle. Cet exemple confirme ainsi l'importance de la sensibilité des modèles de simulation aux conditions spatiales initiales.

* * *

*

3.2 REPRODUCIBILITÉ

La production de connaissance scientifique trouve ses fondements dans la nature cumulative et collective de la recherche, puisque les progrès sont faits lorsque, comme NEWTON l'a bien posé, on "se tient sur les épaules de géants", au sens que l'entreprise scientifique à un temps donné repose sur l'ensemble du travail précédent et qu'aucune avancée ne serait possible sans construire dessus. Cela inclut le développement de nouvelles théories, mais aussi l'extension, le test et la falsification de précédentes : l'avancée dans la construction de la tour signifie aussi la déconstruction de certaines briques obsolètes. Cet aspect de validation par les pairs et de remise en question constante est aussi ce qui légitime la Science pour une connaissance plus robuste et un progrès sociétal basés sur une connaissance d'un univers objectif, par rapport aux systèmes dogmatiques qu'ils soient politiques ou religieux [**bais2010praise**].

La reproductibilité semble être de plus en plus pratiquée de manière effective [**stodden2010scientific**] et les moyens techniques pour l'achever sont toujours plus développés (comme par exemple les outils pour déposer les données ouvertes, ou pour être transparent dans le processus de recherche comme git [**ram2013git**], ou pour intégrer la création de document et l'analyse de données comme knitr [**xie2013knitr**]), au moins dans le champ de la modélisation et de la simulation. Cependant le diable est bien dans les détails et des obstacles jugés dans un premier temps comme mineurs peuvent rapidement devenir un fardeau pour reproduire et utiliser des résultats obtenus dans des recherches précédentes. Nous décrivons deux études de cas où les modèles de simulation sont en apparence hautement reproductibles mais se révèlent vite des puzzles pour lesquels l'équilibre de temps de recherche passe rapidement sous zéro, au sens où essayer d'exploiter leur résultats coûtera plus en temps que de développer entièrement des modèles similaires.

3.2.1 *Explicitation, documentation et implémentation des modèles*

Sur le Besoin d'expliciter le modèle

Un mythe à la vie dure (auquel nous essayons en fait nous-même d'échapper) est que fournir le code source complet et les données seront une condition suffisante pour la reproductibilité, puisque la reproductibilité computationnelle complète implique un environnement similaire ce qui devient vite ardu à produire comme le montre [**2016arXiv160806897H**]. Pour résoudre ce problème, [**10.1371/journal.pone.0152686**] propose l'utilisation de conteneurs Dockers qui permet de reproduire même le comportement de logiciels avec interface graphique indépendamment de l'environnement. C'est d'ailleurs une des direction courantes de développement d'OpenMole, pour simplifier le packaging des bi-

bliothèques et des modèles en binaire (cf. R. REUILLOU dans [[raimbault2017entretiens](#)]). Dans tous les cas, le reproductibilité a des dimensions supplémentaires, il ne s'agit pas de l'objectif unique qui serait est de produire exactement les mêmes graphes et analyses statistiques, en supposant que le code fournit est celui qui a été effectivement utilisé pour produire les résultats donnés. Tout d'abord, ceux-ci doivent être autant que possible indépendants de l'implémentation [[crick2017reproducibility](#)] (c'est à dire du langage, des bibliothèques, des choix de structures de données et de type de programmation) pour des motifs clairs de robustesse. Ensuite, en relation avec le point précédent, un des buts de la reproductibilité est la réutilisation des méthodes ou résultats comme base ou modules pour une recherche future (ce qui comprend une implémentation dans un autre langage ou une adaptation de la méthode), au sens que la reproductibilité n'est pas la possibilité stricte de répliquer car elle doit être adaptable [[drummond2009replicability](#)].

Notre premier cas d'étude suit exactement ce schéma, puisqu'il a sans aucun doute été conçu pour être partagé avec la communauté et utilisé, s'agissant d'un modèle de simulation fournit avec la plate-forme de modélisation agent NetLogo [[wilensky1999netlogo](#)]. Le modèle est également disponible en ligne [[de2007netlogo](#)] et est présenté comme un outil pour simuler les dynamiques socio-économiques des résidents à bas revenus d'une ville au sein d'un environnement urbain synthétique, généré pour ressembler en terme de faits stylisés à la ville réelle de Tijuana, Mexico. Globalement, le modèle fonctionne de la façon suivante : (i) à partir de centre urbains, une distribution d'usage du sol est générée par modélisation procédurale similaire à [[lechner2006procedural](#)], c'est à dire des routes sont générées de proche en proche selon des règles géométriques et de hiérarchie locales, et un usage du sol ainsi qu'une valeur est attribué en fonction des caractéristique du patch (distance au centre, à la route); (ii) dans cet environnement urbain sont simulées des dynamiques résidentielles de migrants, qui cherchent à optimiser une fonction d'utilité dépendant du coût de la vie et de la configuration des autres migrants. A part fournir le code source, le modèle n'est que peu documenté dans la littérature ou dans les commentaires et la description de l'implémentation. Les commentaires qui suivent sont basés sur l'étude de la partie du modèle simulant la morphogenèse urbaine (setup pour la composante "dynamiques résidentielles") comme il s'agit de notre contexte global d'étude. Dans le cadre de cette étude, le code source a été modifié et commenté, dont la dernière version est disponible sur le dépôt du projet²⁴.

FORMALISATION RIGOUREUSE Une partie évidente de la construction d'un modèle est sa formalisation rigoureuse dans un cadre for-

²⁴ at <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Reproduction/UrbanSuite>

mel distinct du code source. Il n'y a bien sûr aucun langage universel pour le formuler [banos2013pour], et de nombreuses possibilités sont offertes par de nombreux champs (e.g. UML, DEVS, formulation mathématique pure), mais l'étape de formalisation précise, qui suit généralement une description plus intuitive donnant les idées et processus dominants ("rationnelle"), ne peut pas être sautée. On pourrait se dire que le code source y est équivalent, mais ce n'est pas exactement vrai car on pourrait alors ne plus distinguer certains choix d'implémentation de la structure du modèle. Aucun article ni documentation n'accompagne le modèle ici, au delà de la documentation embarquée NetLogo, qui ne décrit que de manière thématique en langage naturel les idées derrière chaque étape sans plus développer et fournit de l'information sur le rôle des différents éléments de l'interface. Comme ces éléments manquent ici, le modèle n'est guère utilisable tel quel. On pourrait nous objecter ici que la partie que nous étudions est une procédure d'initialisation et non le cœur du modèle : nous maintenons que l'ensemble des procédures doit être également documenté et implémenté avec un soin équivalent, ou pointer vers une référence extérieure dans le cas d'utilisation d'un modèle tiers, comme nous le faisons d'ailleurs pour le couplage effectué en 3.1.

Une telle formulation est essentielle pour que le modèle soit compris, reproduit et adapté ; mais elle évite également des biais d'implémentation comme

- Des éléments architecturaux dangereux : le contexte du monde est une sphère, ce qui n'est pas raisonnable pour ce modèle à l'échelle d'une ville, les mesures de proximité jouant un rôle important dans les processus de production de la forme urbaine. Les agents peuvent passer d'un côté du monde à l'autre dans la représentation euclidienne, ce qui n'est pas acceptable pour une projection en deux dimensions du monde réel. Pour éviter cela, de nombreux tests et fonctions subtiles sont utilisés, incluant des pratiques déconseillées (e.g. mort d'agents basée sur leur position pour les empêcher de sauter).
- Manque de cohérence interne : par exemple la variable de patch `land-value` (non documentée mais dont l'utilisation se reconstruit par analyse du code) utilisée pour représenter différentes quantités géographiques à différentes étapes du modèle (morphogenèse et dynamiques résidentielles), ce qui devient une incohérence interne quand les deux étapes sont couplées lorsque l'option permettant de faire croître la ville est activée.
- Erreur de code : dans un langage non typé comme NetLogo, le mélange des types peut conduire à des erreurs inattendues à l'exécution, ou même des *bugs* non détectables directement et alors plus dangereux. C'est le cas de la variable de patch `transport` dans le modèle (même si aucune erreur ne survient

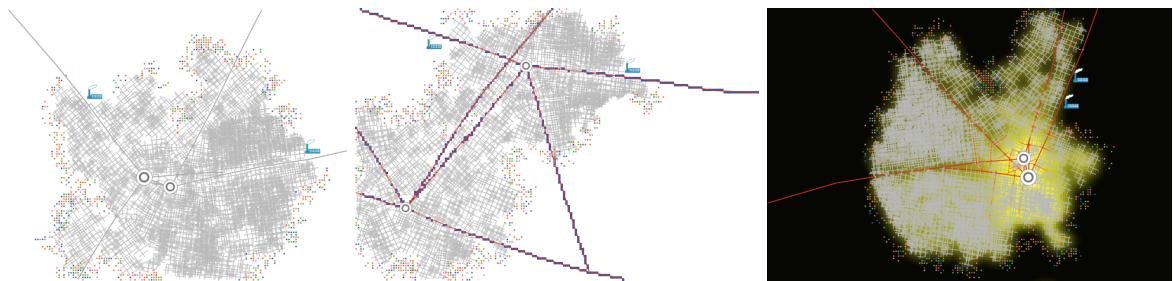


FIGURE 15 : Exemple d'amélioration simple dans la visualisation qui peut aider à appréhender les mécanismes impliqués par le modèle. (Gauche) Exemple de sortie originale ; (Centre) Visualisation des routes principales (en rouge) et de l'attribution des patches sous-jacente, qui suggère de possibles biais d'implémentation dans l'utilisation de la trace discrète des routes pour garder trace de leur position ; (Droite) Visualisation des valeurs foncières en utilisant un gradient de couleur plus lisible. Cette étape confirme l'hypothèse, par la forme de la distribution des valeurs, que l'étape de morphogenèse est un détour non-nécessaire pour générer un champ aléatoire pour lequel des simples mécanismes de diffusion devrait fournir des résultats similaires, comme détaillé dans le paragraphe sur l'implémentation. Initialement, l'interface du modèle ne permet pas ces options de visualisation, ces à dire se limite à la première image. On ne peut se rendre compte des processus en jeu pour la morphogenèse, liés aux patches de route et au valeurs foncières se diffusant.

dans la majorité des configurations depuis l'interface, ce qui est plus dangereux comme le développeur pense que l'implémentation est sûre). De tels problèmes devraient être évités si l'implémentation est faite à partir d'une description exacte du modèle.

IMPLÉMENTATION TRANSPARENTE Une implémentation totalement transparente doit être attendue, incluant une certaine ergonomie dans l'architecture et le code, mais aussi dans l'interface et la description du comportement attendu du modèle.

COMPORTEMENT ATTENDU DU MODÈLE Quelle que soit la définition, un modèle ne peut pas être réduit à sa formulation et/ou implémentation, comme le comportement attendu ou l'utilisation du modèle peuvent être vu comme des parties du modèle lui-même. Dans le cadre du perspectivisme de GIERE [giere2010scientific], la définition du modèle inclut le motif de l'utilisation mais aussi l'agent qui vise à l'utiliser. Pour cela une explication minimale du comportement du modèle et une exploration du rôle des paramètres sont fortement recommandés pour diminuer les chances de mauvais usage ou mauvaises interprétations de celui-ci. Cela inclut des graphes simples obtenus immédiatement à l'exécution sur la plateforme NetLogo, mais aussi un calcul d'indicateurs pour évaluer les sorties du modèle. Il peut aussi s'agir de visualisations améliorée pendant l'execution et l'exploration du modèle, comme le montre la figure 15.

Sur le besoin d'exactitude dans l'implémentation du modèle

Des divergences potentielles entre la description du modèle dans un article et les processus effectivement implémentés peut avoir des conséquences graves sur la reproductibilité finale. Le modèle de croissance du réseau routier donné dans [barthelemy2008modeling] est un exemple d'un tel décalage. Une implémentation stricte des mécanismes du modèle produit des résultats légèrement différents de ceux présentés dans le papier, et comme le code source n'est pas fourni nous devrions tester différentes hypothèses sur des mécanismes possibles ajoutés par le programmeur (qui semble être une règle de connexion aux intersections sous un certain seuil de distance). Des leçons qui peuvent éventuellement être tirées de cet exemple, qui rejoignent partiellement mais complètent celle tirées dans l'étude de cas précédente, sont

- la nécessité de fournir le code source
- la nécessité de fournir une description de l'architecture en même temps que le code (si la description du modèle est faite dans un langage trop loin de spécification architecturales) afin d'identifier des biais possibles d'implémentation
- la nécessité de procéder à des explorations explicites du modèle et de les détailler, ce qui dans ce cas aurait permis d'identifier de possibles biais d'implémentation.

Rendre le dernier point obligatoire pourrait assurer un risque limité de falsification puisqu'il est généralement plus compliqué de falsifier des résultats d'exploration plutôt que d'explorer effectivement le modèle. On pourrait imaginer une expérience pour tester le comportement général d'un sous-ensemble de la communauté scientifique au regard de la reproductibilité, qui consisterait en l'écriture d'un faux papier de modélisation dans l'esprit de [zilsel2015canular], dans lesquels des résultats opposés aux résultats effectifs d'un modèle donné seraient fournis, sans fournir l'implémentation du modèle. Un premier test serait de tester l'acceptation d'un papier clairement non reproductible dans divers journaux, si possible avec un contrôle sur les éléments textuels (par exemple en utilisant ou non des "buzzwords" chers au journal). Selon les résultats, une expérience plus poussée serait de fournir l'implémentation open source mais toujours avec des résultats modifiés plus ou moins fortement, afin de tester si les reviewers essayent effectivement de reproduire les résultats quand ils demandent le code (dans des capacités de calcul limitées bien sûr, le HPC n'étant pas encore largement disponibles en sciences sociales). Notre intuition est que les résultats obtenus seraient fortement négatifs, vu les difficultés rencontrées par une exigence de discipline de reproduction indépendante lors de nombreuses relectures, même pour des revues faisant de la reproductibilité une condition *sine qua non* de

la publication, les auteurs trouvant des astuces pour se dérober aux contraintes (postuler que des données de simulation ne sont pas des données, ne fournir qu'une version agrégée inutile du jeu de données utilisées, etc. ; nous reviendrons sur le rôle des données plus loin).

3.2.2 *Exploration interactive et production des résultats*

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 basé-agent intégrée à des plateformes comme NetLogo [[wilensky1999netlogo](#)] ou Gamma [[drogoul2013gamma](#)]. Un objectif similaire est sous-jacent à [[rey2015plateforme](#)], 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), la plateforme Open-Mole 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 [[holzinger2014knowledge](#)].

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 préliminaire²⁵, mais peu souvent comme permettant la production de résultats finaux : lorsque les paramètres ou dimension 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'extrac-

²⁵ que nous ne jugeons pas superficielle puisque nous les mobilisons au moins deux fois par la suite, voir 8.2 et 8.3

tion 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 [banos2001propos], 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. On revient 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 [2010AEdRv...9a0103R]²⁶. Les résultats d'un crowdsourcing de la cognition humaine peuvent rivaliser avec les techniques automatiques les plus avancées comme le montre [10.1371/journal.pone.0178165] 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, [pfaender2009spatialisation] introduit une sémiologie spécifique visant à favoriser l'exploration de grands jeux de données hétérogènes, et l'expérimente sur une application 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.

3.2.3 Perspectives

Mise en application

Encore une fois, la reproductibilité et la transparence sont des éléments essentiels incontournables de la science contemporaine, liés aux pratiques de science ouverte et d'accès ouvert. Beaucoup d'exemples (voir un récent en économie expérimentale dans [camerer2016evaluating]) dans diverses disciplines montrent le manque de reproductibilité des résultats des expériences, alors que celle-ci doit pouvoir conduire à une falsification ou à une confirmation de ces résultats. La falsification est une pratique coûteuse car demandant un certain investisse-

²⁶ 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 scientifiques : la classification d'images, dans le but d'entraîner des algorithmes supervisés, est l'exemple initial du Galaxy Zoo pour la forme des galaxies.

ment au détriment de sa propre recherche [[chavalarias2005nobel](#)]. Elle pourrait ainsi être rendue plus efficiente grâce à une transparence augmentée. Des outils spécialement dédiés à une reproductibilité directe, souvent permise par l'ouverture, devraient accroître la performance globale de la science. Mais l'accès ouvert a des impacts bien plus large que la science elle-même : [[2015arXiv150607608T](#)] montre un transfert des connaissances scientifiques accru vers la société dans le cas d'articles ouverts, notamment par des intermédiaires comme Wikipedia.

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 exemples d'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 [[10.1371/journal.pcbi.1004947](#)].

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 : [[wilson2017good](#)] donne des premières pistes. Concernant la qualité des données, de nombreux efforts sont faits pour introduire des cadres de standardisation des données : par exemple [[10.1371/journal.pone.0178731](#)] décrit un cadre conceptuel visant à guider la résolution de problème récurrent 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*²⁷, comme proposé par [[2017arXiv170706552](#)].

Données

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

²⁷ 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.

sont des sujets de recherche en eux-même, nous donnons des perspectives sur les potentiels d'une ouverture systématique des données en recherche. En géographie, les *data paper* sont une pratique inexisteante, 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 accès²⁸.

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 richesse²⁹.

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 9.3), 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 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 nécessaire.

²⁸ 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.

²⁹ 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.

Mais en suivant l'adage de Framasoft³⁰, "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*³¹ 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. On peut citer des exemples parmi les études empiriques que nous développons : 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) ; les données des stations essence utilisées en 8.3 proviennent d'une source dont la légalité ne devrait pas être creusée plus, et nous ne pouvons malheureusement pas les rendre disponibles sans prendre de risques - cet aspect n'a cependant jamais fait broncher les reviewers de l'article associé qui n'ont même pas mentionné le manque d'accès aux données.

L'ouverture implique un engagement qui fait résolument partie de nos positionnements. C'est la même idée qui soutient la construction de l'application CybergeoNetworks³², 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 meta-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.

* * *

*

³⁰ Réseau pour la promotion du logiciel libre, <https://framasoft.org/>

³¹ Dont l'index est disponible à <https://cybergeo.revues.org/28545>. Le premier article est [swerts2017database], que nous utilisons d'ailleurs en ??.

³² <http://shiny.parisgeo.cnrs.fr/CybergeoNetworks>

3.3 POSITIONNEMENT ÉPISTÉMOLOGIQUE

La dernière section de ce chapitre vise à clarifier notre positionnement épistémologique, celui-ci ayant déjà été ébauché à plusieurs occasions précédemment. Un tel positionnement n'est jamais anodin, puisqu'il conditionne fortement les démarches, les expériences et l'interprétation des résultats : comme le souligne [morin1980methode], un positionnement qui se dit objectif en rejetant toute subjectivité est bien plus biaisé qu'une approche subjective consciente.

Les points que nous souhaitons développer se placent dans une logique à la fois verticale de niveau d'abstraction et dans une logique de domaines scientifique : dans l'ordre, nous posons d'abord le contexte épistémologique général (relevant de l'histoire des sciences, à un niveau d'abstraction moyen), pour descendre en généralité pour préciser conceptuellement nos objets particuliers (épistémologie du vivant et du social), pour finalement tout remettre en perspective au niveau de la production de connaissance elle-même (épistémologie de la complexité).

3.3.1 Approche cognitive et Perspectivisme

Notre positionnement épistémologique se fonde sur une approche cognitive de la science, introduite par GIERE dans [giere2010explaining]. L'approche se concentre sur le rôle des agents cognitifs comme porteurs et producteurs de la connaissance. Son caractère opérationnel a été montré par [giere2010agent] qui étudie un modèle basé-agent de la science. Ces idées convergent avec le jeu Nobel de CHAVALARIA [chavalarias2016s] qui teste de manière stylisée l'équilibre entre production de nouvelles théories et tentative de falsification de théories existantes dans l'entreprise scientifique collective.

Ce positionnement épistémologique a été présenté par GIERE comme *perspectivisme scientifique* [giere2010scientific], dont la caractéristique principale est de considérer toute entreprise scientifique comme une *perspective* dans laquelle des *agents* utilisent des *media* (modèles) pour représenter quelque chose dans un certain but. Pour comprendre ses principes de manière plus concrète, nous pouvons le positionner sur la *check-list* du constructivisme de HACKING [hacking1999social], un outil pratique pour situer une position épistémologique. Celle-ci suppose un espace simplifié tri-dimensionnel dans lequel les dimensions sont différents aspects sur lesquels les approches réalistes et constructivistes généralement divergent. Le premier point est le niveau de contingence (dépendance au chemin du processus de construction de connaissances) : celle-ci est nécessaire dans l'approche perspectiviste qui est pluraliste et suppose des chemins parallèles de construction de connaissance. Le deuxième point mesure un "degré de constructivisme", qui est assez haut en perspectivisme car les agents produisent

la connaissance. Enfin, le dernier point qui concerne l'explication endogène ou exogène de la stabilité des théories, est fortement du côté du constructivisme, puisque cette stabilité dépend des interactions complexes entre les agents et leur perspectives et donc totalement endogène. Le perspectivisme a pour ces raisons été présenté comme un chemin intermédiaire et alternatif entre le réalisme absolu et le constructivisme sceptique [**brown2009models**]. la notion de *perspective* jouera pour nous un rôle fondamental dans le cadre développé en [9.3](#).

Cette approche mettant l'emphase sur l'auto-organisation, nous la voyons totalement compatible avec une vision anarchiste de la science comme défendue par [**feyerabend1993against**]. Celui-ci émet des doutes sur l'intérêt de l'anarchisme politique mais introduit l'*anarchisme scientifique*, qu'il ne faut pas comprendre comme un refus total de toute méthode "objective", mais d'une autorité et légitimité artificielle que certaines méthodes ou courants scientifiques pourraient vouloir prendre. Il démontre par une analyse précise des travaux de Galilée que la plupart de ses résultats étaient basés sur des croyances et que la plupart n'étaient pas accessibles avec les outils et méthodes de l'époque, et postule qu'il devrait en être de même pour certains travaux contemporains. Il n'y a donc pas de *perspective* objectivement plus légitimes que d'autres dans la mesure de leurs validation par des faits et des pairs - et même dans ces cas la légitimité doit pouvoir être discutée, car la remise en question est un fondement de la connaissance. Cela correspond exactement à la pluralité des perspectives que nous défendons.

Supposer auto-organisation et l'émergence des connaissances peut être interprété comme une priorité donnée à la construction des paradigmes *par le bas (bottom-up)*, en tentant de se distancer des préconceptions ou dogmes cadrant par le haut. En d'autres termes, il s'agit de pratiquer l'anarchisme scientifique prôné par FEYERABEND. En effet, les positions anarchistes ont trouvé un écho très cohérent dans les différents courants de la complexité, de la cybernétique à l'auto-organisation au cours du 20ème siècle [**duda2013cybernetics**]. Notre cadre de connaissances développé en [9.3](#) illustre cette émergence de la connaissance. De plus, notre volonté de réflexivité et de donner à notre travail des pistes de lecture diverses au delà de la linéarité (voir Appendice F), illustre l'application de ces principes. Les recommandations méthodologiques et les positionnements donnés précédemment dans ce chapitre pourraient sonner comme totalitaires s'ils étaient assénés de manière sèche sans contexte, mais ceux-ci sont en fait tout le contraire puisqu'ils découlent d'un dynamique récente de science ouverte qui a bien émergé par le bas, conséquence en partie de l'ouverture et de la pluralité.

3.3.2 De la Vie à la Culture

Systèmes biologiques et systèmes sociaux

Le parallèle entre les systèmes sociaux et les systèmes biologiques est souvent fait, parfois de manière plus qu'imagée comme par exemple pour la théorie du *Scaling* de WEST qui applique des équations de croissance similaires à partir des lois d'échelle, avec des conclusions inverses tout de même concernant la relation entre taille et rythme de vie [bettencourt2007growth]. Les relations d'échelle ne tiennent plus lorsqu'on essaye de les appliquer à une fourmi seule, et il faut alors l'appliquer à la fourmilière entière qui est alors l'organisme en question. En ajoutant la propriété de cognition, on confirme qu'il s'agit du niveau pertinent, puisque celle-ci possède des propriétés cognitives avancées, comme la résolution de problèmes d'optimisation spatiaux, ou la réponse rapide à une perturbation extérieure. Les organisations sociales humaines, les villes, peuvent-elles être vues comme des organismes ? [banos2013pour] file la métaphore de la *fourmilière urbaine* mais rappelle que le parallèle s'arrête assez vite. Nous allons voir cependant dans quelle mesure certains concepts de l'épistémologie de la biologie peuvent être utiles pour comprendre les systèmes sociaux que nous nous proposons d'étudier.

Nous nous basons sur la contribution fondamentale de MONOD dans [monod1970hasard], qui tente de développer les principes épistémologiques cruciaux pour l'étude du vivant. Ainsi, les organismes vivants répondent à trois propriétés essentielles qui permettent de différencier d'autres systèmes : (i) la téléconomie, c'est à dire qu'il s'agit "d'objets doués d'un projet", projet qui se reflète dans leur structure et dans celles des artefacts qu'ils produisent³³; (ii) l'importance des processus morphogénétiques dans leur constitution (voir 5.1); (iii) la propriété de reproduction invariante de l'information définissant leur structure. MONOD esquisse de plus en conclusion des pistes pour une théorie de l'évolution culturelle. La téléconomie est essentielle dans les structures sociales, puisque toute organisation essaye de satisfaire un ensemble d'objectifs, même si en général elle n'y parviendra pas et que ceux-ci co-évolueront avec l'organisation. Cette notion de multi-objectif qui est typique des systèmes complexes socio-techniques, et y sera plus cruciale que pour les systèmes biologiques.

Ensuite, nous postulons que la notion de morphogenèse est un outil essentiel pour comprendre ces systèmes, avec une définition très proche de celle utilisée en biologie. Un travail approfondi pour donner cette définition est fait en 5.1, que nous résumerons en l'existence de processus relativement autonomes guidant la croissance du système et impliquant des relations causales circulaires entre forme et fonction qui témoignent d'une architecture émergente. Pour des

³³ Qu'il ne faut pas confondre avec la téléologie, propres aux animismes, qui consiste à prêter un projet ou un sens à l'univers.

systèmes sociaux, isoler le système est plus difficile et la notion de frontière sera moins stricte que pour un système biologique, mais on retrouvera bien ce lien entre forme et fonction, comme par exemple la structure d'une organisation ayant un impact sur ses fonctionnalités.

Enfin, la reproduction de l'information est au cœur de l'évolution culturelle, par la transmission de la culture et la *mémétique*, la différence étant que le rapport d'échelle de temps entre la fréquence de transmission et les processus de croisement et de mutation ou d'autres processus non mémétiques de production culturelle est très faible, alors qu'elle est de plusieurs ordres de magnitude en biologie.

Un exemple illustre que le parallèle n'est pas toujours absurde :[2017arXiv170305917G] propose un modèle de réseau auto-catalytique pour la cognition, qui expliquerait l'apparition de l'évolution culturelle par des processus analogues à ceux s'étant produit à l'apparition de la vie, c'est à dire une transition permettant aux molécules de s'auto-entretenir et s'auto-reproduire, les représentations mentales faisant office de molécules.

Mais si les processus à l'origine sont analogues, la nature de l'évolution est bien différente par la suite, comme le montre [vanderLeeuw2009], les critères darwiniens d'évolution n'étant pas suffisant pour expliquer l'évolution de nos sociétés organisées. Il s'agit d'une complexité de nature différente dans laquelle le rôle des flux d'information est crucial (voir le rôle de la complexité informationnelle dans la sous-section suivante).

L'un des points sur lequel il s'agit également d'être attentif est la plus grande difficulté de définir les niveaux d'émergence pour les systèmes sociaux : [roth2009reconstruction] souligne le risque de tomber dans des cul-de-sac ontologiques car les niveaux ont été mal définis. Il soutient qu'il faut d'une manière générale penser au-delà de la seule dichotomie micro-macro qui est utilisée pour caricaturer les notions d'émergence faible, mais que les ontologies doivent souvent être multi-niveaux et impliquant de multiples niveaux intermédiaires.

Cette dernière question est aussi à mettre en perspective avec le problème de l'existence d'émergence forte dans les structures sociales, qui en terme sociologiques correspond à l'idée de l'existence "d'êtres collectifs" [angeletti2015etres]. MORIN distingue d'ailleurs les systèmes vivants du second type (multi-cellulaire) et du troisième types (structures sociales), mais précise que les *sujets* de ces derniers sont nécessairement inachevés [morin1980methode] (p. 852). Ainsi, les émergences du biologique au social sont analogues mais restent fondamentalement différentes.

Co-évolution

Ce positionnement sur les systèmes biologiques et sociaux trouve un écho immédiat pour le concept de co-évolution. Il provient en effet de la biologie, où il a été développé à la suite de celui d'évolution, pour être utilisé plus récemment en sciences humaines et sociales. Dans

quelle mesure le concept a-t-il été transféré? Retrouve-t-on un parallèle similaire à celui entre évolution biologique et évolution culturelle? Nous proposons pour répondre à ces questions d'apporter un bref point de vue multidisciplinaire sur la co-évolution³⁴. Nous passons par la suite en revue un large spectre de disciplines, partant de la biologie où le concept a initialement trouvé son origine pour arriver progressivement à des disciplines en relation avec les sciences du territoire.

Biologie

Le concept de co-évolution en biologie est une extension de celui bien connu d'*évolution*, qui remonte à DARWIN. [durham1991coevolution] (p. 22) rappelle les composantes et structure systémiques nécessaires pour qu'il y ait évolution³⁵ :

1. Processus de *transmission*, impliquant des unités de transmission et des mécanismes de transmission.
2. Processus de *transformation*, nécessitant des sources de variation.
3. Isolation de sous-systèmes pour que les effets des processus précédents soient observable dans des différentiations.

Ainsi, une population soumise à des contraintes (souvent synthétisée conceptuellement comme une *fitness*) qui conditionnent la transmission du patrimoine génétique des individus (transmission), et à des mutation génétiques aléatoires (transformation), sera bien en évolution dans les territoires spatiaux qu'elle occupe (isolation), et par extension l'espèce à laquelle on peut l'associer.

La co-évolution est alors définie comme un changement évolutif dans une caractéristique des individus d'une population, en réponse à un changement dans une deuxième population qui à son tour répond évolutionnairement au changement de la première, comme synthétisé par [janzen1980coevolution]. Cet auteur appuie par ailleurs

³⁴ La démarche ici est légèrement différente de celle que nous menerons en 5.1 dans le cas de la Morphogenèse, qui sera *interdisciplinaire* au sens où elle cherchera à intégrer les approches, tandis que nous restons ici dans un aperçu des concepts et donc plutôt dans du *multidisciplinaire* (voir B.6 pour des précisions sur le -*disciplinaire). Le concept de *co-évolution* étant clé pour notre travail empirique par la suite, nous en donnerons alors une caractérisation originale et prenons le parti de ne pas tomber dans le syncrétisme intégrateur pour ce concept, mais bien de l'approcher d'un *point de vue géographique*, et même plus précisément dans le cadre des systèmes territoriaux. On pourrait postuler une congruence entre la spécialisation empirique/de modélisation et celle théorique, plaçant notre processus de production de connaissance dans un profil particulier de dynamiques de Domaines de Connaissance (9.3).

³⁵ Et dans ce contexte général l'évolution n'est pas réservée à la biologie du vivant et la présence de gènes, mais aussi à des systèmes physiques vérifiant ces conditions. Nous y reviendrons plus loin.

la subtilité du concept et alerte contre ses utilisations injustifiées : la présence d'une congruence de deux caractéristiques qui semblent adaptées l'une à l'autre n'implique pas l'existence d'une co-évolution, l'une des deux espèces ayant pu s'adapter seule à une caractéristique déjà présente de l'autre.

Cette présentation brute de décoffrage mutile dans une certaine mesure la complexité réelle des écosystèmes : les populations s'insèrent dans des réseaux trophiques et des environnements, et les interactions co-évolutionnaires impliqueraient des communautés de populations d'espèces diverses, comme présenté par [strauss2005toward] sous l'appellation de co-évolution diffuse. De même, les dynamiques spatio-temporelles sont cruciales dans la réalisation de ces processus : [dybdahl1996geography] étudie par exemple l'influence de la distribution spatiale sur les motifs de co-évolution pour un escargot et son parasite, et montre qu'une vitesse de diffusion génétique dans l'espace plus grande pour le parasite conduit les dynamiques de co-évolution.

Les concepts essentiels à retenir du point de vue biologique sont ainsi : (i) existence de processus d'évolution, en particulier transmission et transformation ; (ii) dans des schémas circulaires entre populations dans le cas de la co-évolution ; et (iii) dans un cadre territorial (spatio-temporel et environnemental au sens du reste de l'écosystème) complexe.

Evolution culturelle

Ce développement sur la co-évolution nous a été amené par le parallèle entre systèmes biologiques et systèmes sociaux. L'évolution de la culture est théorisée et explorée par un champ propre, et n'est pas en reste de dynamiques co-évolutives. [Mesoudi25072017] rappelle l'état des connaissances sur le sujet et les défis à venir, comme la relation avec la nature cumulative de la culture, l'influence de la démographie dans les processus d'évolution, ou la construction de méthodes phylogénétiques permettant de reconstruire des arbres des branchements passés.

Pour donner un exemple, [carrignon2015modelling] introduit un cadre conceptuel pour la co-évolution de la culture et du commerce dans le cas de sociétés anciennes sur lesquelles on dispose de données archéologique, et propose son implémentation par un modèle basé-agent dont les dynamiques sont partiellement validées par l'étude des faits stylisés produits par le modèle. La co-évolution est bien prise ici au sens d'adaptation mutuelle de structures socio-spatiales, à des échelles de temps comparables, dans ce cadre plus général d'évolution culturelle.

L'évolution culturelle serait même indissociable de l'évolution génétique, puisque [durham1991coevolution] postule et illustre un lien fort entre les deux, qui seraient eux-même en co-évolution. [bull2000meme]

explore un modèle stylisé impliquant deux populations de répliquants (les gènes et les memes) et montre l'existence de transitions de phase pour les résultats du processus d'évolution génétique lorsque l'interaction avec le répliquant culturel est forte.

Sociologie

Le concept a été utilisé en sociologie et disciplines apparentées comme les études de l'organisation, suivant le parallèle effectué ci-dessus de la même manière que pour l'évolution culturelle. Dans le domaine de l'étude des organisations, [volberda2003co] développe un cadre conceptuel de la co-évolution inter-organisationnelle en relations avec les processus de management internes, mais déplore l'absence d'études empiriques cherchant à quantifier cette co-évolution. Dans le cadre de la gestion des systèmes de production, [tolio2010species] conceptualise un chaîne de production intelligente où produit, processus et système de production doivent être en co-évolution.

Economie géographique

En Economie Géographique, le concept de co-évolution a également largement été mobilisée. L'idée d'entités évolutionnaires en économie vient à contre-courant du courant néoclassique qui reste majoritaire, mais trouve un écho de plus en plus pertinent [nelson2009evolutionary]. [schamp2010] procède à une analyse épistémologique de l'utilisation de la co-évolution, et oppose une approche néo-Schumpeterienne de l'Economie qui considère l'émergence de populations qui évoluent à partir de règles micro-économiques (qui correspondrait à une lecture directe et relativement isolationniste de l'évolution biologique) à une approche systémique qui considérerait l'économie comme un système évolutif de manière globale (qui correspondrait à l'évolution diffuse que nous avons développé précédemment), pour proposer une caractérisation précise tombant dans le premier cas, qui suppose des *institutions* qui co-évoluent. Le plus important pour notre propos est qu'il souligne l'aspect crucial du choix des population et des entités considérées, de la zone géographique, et appuie l'importance de l'existence de relations causale circulaires.

Il est possible de donner divers exemples d'application. [doi:10.1080/00343400802662658] introduit un cadre conceptuel pour permettre de concilier nature évolutionnaire des firmes, théorie des clusters et réseaux de connaissance, dans lequel la co-évolution entre réseaux et firmes est centrale, et qui est définie comme une causalité circulaire entre différentes caractéristiques de ces sous-systèmes. [colletis2010co] introduit un cadre de co-évolution des territoires et de la technologie (questionnant par exemple le rôle de la proximité pour les innovations), qui révèle l'importance à nouveau de l'aspect institutionnel. [ter2011co] propose un cadre couplant la vision évolutionnaire des entreprises, la littérature

sur les industries et l'innovation dans les clusters, et l'approche par réseau complexe des connexions entre ces premiers dans le système territorial.

En Economie Environnementale, [kallis2007coevolution] montre que des approches “larges” (pouvant considérer la majorité des co-dynamiques comme co-évolutives) s’opposent à des approches plus strictes (dans l’esprit de la définition donnée par [schamp201020]), et que dans tous les cas une définition précise, pas forcément venant de la biologie, doit être donnée, en particulier pour la recherche d’une caractérisation empirique.

Géographie

Pour la géographie, comme nous l’avons déjà présenté en introduction, les travaux les plus proches empiriquement et théoriquement des notions de co-évolution sont étroitement liés à la Théorie Evolutive des Villes. Il n’est pas évident de tracer dans la littérature à quel moment la notion a été clairement formalisée, mais il est évident qu’elle était présente dès les fondements de la théorie comme le rappelle DENISE PUMAIN (voir D.3) : le système complexe adaptatif est composé de sous-systèmes en interdépendances complexes, souvent circulairement causales. Les premiers modèles incluent bien cette vision de manière implicite, mais la co-évolution n’est pas appuyée explicitement ou définie précisément, en termes qui seraient quantifiables ou identifiables structurellement. [paulus2004coevolution] a amené des preuves empiriques de mécanismes de co-évolution par l’étude de l’évolution des profils économiques des villes françaises. L’interprétation utilisée par [schmitt2014modelisation] repose sur une entrée par la Théorie Evolutive, et consiste fondamentalement en une lecture des systèmes de villes comme entités fortement interdépendantes.

Géographie Physique

En étude des paysages, [sheeren2015coevolution] parle de co-évolution du paysage et des activités agricoles, mais ne désigne en fait pas d’effet circulaires de l’un sur l’autre. A priori, leurs résultats montrent que l’évolution des pratiques agricoles entraîne une évolution du paysage, et il n’est ainsi pas clair dans quelle mesure le cadre conceptuel de la co-évolution, mentionné sans plus de détails, est mobilisé.

Physique

Enfin, on peut noter de manière anecdotique que le terme de co-évolution a également été utilisé par la physique. L’utilisation pour des systèmes physiques peut porter à débat, selon que l’on suppose

ou non que la transmission suppose un transmission d'*information*³⁶. Dans le cas d'une transmission ontologique uniquement physique (*êtres physiques*), alors une grande partie des systèmes physiques sont évolutifs. [hopkins2008cosmological] développe un cadre cosmologique pour la co-évolution d'objets cosmiques hétérogènes dont la présence et les dynamiques sont difficilement expliquées par des théories plus classiques (certains types de galaxies, quasars, trous noirs supermassifs). [antonioni2017coevolution] étudie la co-évolution entre des propriétés de synchronisation et de coopération au sein d'un réseau d'oscillateurs de Kuramoto³⁷, montrant d'une part que le concept peut être appliqué à des objets abstraits, et d'autre part qu'un réseau de relations complexes entre variables peut être à l'origine de dynamiques présentant des causalités circulaires, c'est à dire d'une co-évolution en ce sens.

Synthèse

La plupart de ces approches rentrent dans la théorie des systèmes complexes adaptatifs développée par HOLLAND, notamment dans [holland2012signals] : il voit tout système comme une imbrication de systèmes de limites, filtrant des signaux ou des objets. Au sein d'une limite donnée, le sous-système correspondant est relativement autonome de l'extérieur, est appelé *niche écologique*, en correspondance directe avec les communautés fortement connectées au sein des réseaux trophiques ou écologiques. Ainsi, des entités interdépendantes au sein d'une niche sont dites en co-évolution. Nous reviendrons sur cette entrée lors de la construction théorique en 9.2 lorsque nous aurons développé d'autres concepts qui lui sont nécessaire.

Nous retenons de cet aperçu multidisciplinaire de la co-évolution les points fondamentaux suivants précurseurs à une définition propre de la co-évolution que nous donnerons plus loin, en conclusion de la première partie :

1. La présence de *processus d'évolution* est primaire, et leur définition se ramène presque toujours à l'existence de processus de transmission et de transformation.

³⁶ L'information est définie dans la théorie Shanonienne comme une probabilité d'occurrence d'une chaîne de caractère. [morin1976methode] montre que le concept d'information est en fait bien plus complexe, et qu'il doit être pensé conjointement à un contexte donné de génération d'un système auto-organisateur négentropique, i.e. réalisant des diminutions locales d'entropie notamment grâce à cette information. Ce type de système est nécessairement vivant. Nous prendrons ici cette vision complexe de l'information.

³⁷ Le modèle de Kuramoto s'intéresse à la synchronisation au sein de systèmes complexes, en étudiant l'évolution de phases θ_i couplée par les équations d'interaction $\dot{\theta} = \vec{\omega} + \vec{W}[\vec{\theta}] + \vec{B}$ où $\vec{\omega}$ sont les phases propres de forçage et la force de couplage entre i et j est donnée par $\vec{W}_i = \sum_j w_{ij} \sin(\theta_i - \theta_j)$ et \vec{B} du bruit.

2. La co-évolution suppose des entités ou systèmes, appartenant à des classes distinctes, dont les dynamiques évolutives sont couplées de manière circulaire causale. Les approches peuvent différer selon l'hypothèse de populations de ces entités, d'objets singuliers, ou de composantes d'un système global alors en interdépendance mutuelle sans qu'il y ait circularité directe.
3. La délimitation des systèmes ou des sous-systèmes, à la fois dans l'espace ontologique (définition des objets étudiés), mais aussi dans l'espace et le temps, ainsi que leur distribution dans ces espaces, est fondamental pour l'existence de dynamiques co-évolutives, et a priori dans un grand nombre de cas, pour leur caractérisation empirique.

3.3.3 *Nature de la Complexité et Production de Connaissances*

Les deux premiers points épistémologiques que nous venons de traiter relevaient respectivement du positionnement en lui-même, c'est à dire du cadre de lecture des processus de production de connaissance scientifique, puis de la nature des concepts considérés. Nous proposons de monter encore en généralité par rapport au premier et d'introduire un développement contribuant modestement (c'est à dire dans notre contexte) à *la Connaissance de la Connaissance*. Il s'agit d'interroger les liens entre complexité et processus de production de connaissance.

Un aspect de la production de connaissance sur des Systèmes Complexes, auquel nous nous heurtons plusieurs fois ici (voir chapitre 9), et qui semble être récurrent voire inévitable, est un certain niveau de réflexivité (et qui serait inhérent aux systèmes complexes en comparaison aux systèmes simples, comme nous le développerons plus loin). Nous entendons par là à la fois une réflexivité pratique, c'est à dire la nécessité d'élever le niveau d'abstraction, comme le besoin de reconstruire de manière endogène les disciplines dans lesquelles une réflexion cherche à se positionner comme proposé en 2.2, ou de réfléchir à la nature épistémologique de la modélisation lors de l'élaboration d'un modèle comme en B.5, mais également une réflexivité théorique en le sens que les appareils théoriques ou les concepts produits peuvent s'appliquer de manière récursive à eux-mêmes. Cette constatation pratique fait écho à des débats épistémologiques anciens questionnant la possibilité d'une connaissance objective de l'univers qui serait indépendante de notre structure cognitive, ou bien la nécessité d'une "rationalité évolutive" impliquant que notre système cognitif, produit de l'évolution, reflète les processus complexes ayant conduit à son émergence, et que toute structure de connaissance sera

par conséquent réflexive³⁸. Nous ne prétendons pas ici apporter une réponse à une question aussi vaste et vague telle quelle, mais proposons un lien potentiel entre cette réflexivité et la nature de la complexité.

Complexité et Complexités

Ce qui est entendu par complexité d'un système mène souvent à des malentendus car celle-ci peut être qualifiée selon différentes dimensions et visions. Nous distinguons dans un premier temps la complexité au sens d'émergence faible et d'autonomie entre les différents niveaux d'un système, et sur laquelle différentes positions peuvent être développées comme dans [deffuant2015visions]. Nous ne rentrerons pas dans une granularité plus fine, la vision de la complexité sociale donnant encore plus de fil à retordre au démon de Laplace, peut être par exemple comprise par une émergence plus forte (au sens d'émergence faible et forte développée précédemment en 3.1). Nous simplifions ainsi et supposons que la nature des systèmes joue un rôle secondaire dans notre reflexion, et considérons la complexité au sens d'une émergence.

D'autre part, nous distinguons deux autres "types" de complexité, la complexité computationnelle et la complexité informationnelle, qui peuvent être vues comme des mesures de complexité, mais qui ne sont pas directement équivalentes à l'émergence, puisqu'il n'existe pas de lien systématique entre les trois. On peut par exemple imaginer utiliser un modèle de simulation, pour lequel les interactions entre agents élémentaires se traduisent par un message codé au niveau supérieur : il est alors possible en exploitant les degré de liberté de minimiser la quantité d'information contenue dans le message. Les différentes langues demandent des efforts cognitifs différents et compressent différemment l'information, ayant différents niveau de complexité mesurables [febres2013complexity]. De même, des artefacts architecturaux sont le résultat d'un processus d'évolution naturelle puis culturelle et peuvent témoigner plus ou moins de cette trajectoire.

De nombreuses autres caractérisations conceptuelles ou opérationnelles de la complexité existent, et il est clair que la communauté scientifique n'a pas convergé sur une définition unique [chu2008criteria]³⁹. Nous proposons de nous concentrer sur ces trois concepts en particulier, pour lesquels les relations ne sont déjà pas évidentes.

³⁸ Nous remercions D. Pumain d'avoir pointé cette vue alternative du problème que nous allons développer par la suite

³⁹ Dans une approche en un sens réflexive, [chu2008criteria] propose de continuer d'explorer les différentes approches existantes, comme des proxys de la complexité dans le cas d'un essentialisme, ou comme des notions à part entière. La complexité devrait émerger d'elle-même de l'interaction entre ces différentes approches étudiant la complexité, d'où la réflexivité.

En effet, les liens entre ces trois types de complexité ne sont pas systématiques, et dépendent du type de système. Des liens épistémologiques peuvent néanmoins être introduits. Nous traitons ceux entre émergence et les deux autres complexités, étant donné que le lien entre complexité computationnelle et complexité informationnelle est assez bien compris et relève de problématiques de compression de l'information et de traitement du signal, ou encore de cryptographie.

Complexité computationnelle et émergence

Différents indices suggèrent une certaine nécessité de complexité computationnelle pour avoir émergence dans des systèmes complexes, tandis que réciproquement un certain nombre de systèmes complexes adaptatifs sont dotés de capacités de calcul élevées.

Un premier lien où complexité computationnelle implique émergence est suggéré par un examen algorithmique des problèmes fondamentaux de la Physique Quantique. En effet, [2014arXiv1403.7686B] démontre que la résolution de l'équation de Schrödinger avec Hamiltonien quelconque est un problème NP-difficile et NP-complet, et donc que l'acceptation de $P \neq NP$ implique une séparation qualitative entre le niveau quantique microscopique et le niveau d'observation macroscopique. Ainsi, c'est bien la complexité (ici au sens de leur calcul) des interactions au sein du système et de son environnement qui explique l'apparente réduction du paquet d'onde, ce qui rejoint l'approche de GELL-MANN par la décohérence quantique [gell1996quantum], qui explique que des probabilités ne peuvent être associées qu'aux histoires décohérentes (dans lesquelles les corrélations ont fait prendre une trajectoire au système à l'échelle macroscopique)⁴⁰. Le paradoxe du chat de Schrödinger nous apparaît ainsi comme une perspective fondamentalement réductionniste, puisqu'il suppose que la superposition d'états peut se propager à travers les niveaux successifs et qu'il n'y aurait pas émergence, au sens de constitution d'un niveau supérieur autonome. En d'autres termes, le

40 Le *Problème de la Mesure Quantique* se pose lorsqu'on considère une fonction d'onde microscopique donnant l'état d'un système pouvant être superposition de plusieurs états, et consiste en un paradoxe théorique, les mesures étant toujours déterministes alors que le système a des probabilité d'états d'une part, et le problème de la non-existence d'états macroscopiques superposés (réduction du paquet d'onde). Comme revu par [schlosshauer2005decoherence], différentes interprétations épistémologiques de la physique quantiques sont liées à différentes explications de ce paradoxe, dont celle "classique" de Copenhague qui donne à l'acte d'observation le rôle de reduction du paquet d'onde. GELL-MANN précise que cette interprétation n'est pas absurde puisque c'est bien les corrélations entre l'objet quantique et le monde qui produisent l'histoire décohérente, mais qu'elle est bien trop spécifique, et que la réduction a lieu dans l'émergence elle-même : le chat est bien mort ou vivant, mais pas les deux, avant que l'on ouvre la boîte.

travail de [2014arXiv1403.7686B] suggère que la complexité computationnelle est suffisante pour la présence d'émergence.⁴¹

Dans le sens inverse, le lien entre complexité computationnelle et émergence est mis en valeur par les questions liées à la nature de la computation [moore2011nature]. Des automates cellulaires, qui sont par ailleurs cruciaux pour la compréhension de divers systèmes complexes, ont été montrés Turing-complets⁴², comme le Jeu de la Vie [beer2004autopoiesis]⁴³. Des organismes sans système nerveux central sont capables de résoudre des problèmes décisionnels difficiles [reid2016decision]. Un algorithme à base de fourmis est montré par [Pintea2017] comme résolvant un Problème du Voyageur de Commerce Généralisé (GTSP), problème NP-difficile. Ce lien fondamental avait déjà été envisagé par TURING, puisqu'au delà de ses contributions fondamentales à l'informatique moderne, il s'était intéressé à la morphogenèse et a tenté de produire des modèles chimiques d'explication de celle-ci [turing1952chemical] (qui étaient très loin de effectivement l'expliquer - elle n'est toujours pas bien comprise aujourd'hui, voir 5.1 - mais dont les contributions conceptuelles ont été fondamentales, notamment pour la notion de réaction-diffusion). On sait par ailleurs qu'un minimum de complexité en termes d'interactions constituantes dans un cas particulier de système basé-agent (modèles de réseaux booléens), et donc d'émergences possibles, implique une borne inférieure sur la complexité computationnelle, qui devient conséquente dès que les interactions avec l'environnement sont ajoutées [tovsic2017boolean].

Complexité informationnelle et émergence

La complexité informationnelle, ou la quantité d'information contenue dans un système et la manière dont celle-ci est stockée, entretient également des liens fondamentaux avec l'émergence. L'information est équivalente à l'entropie d'un système et donc à son degré d'organisation - c'est ce qui a permis de résoudre le paradoxe apparent du

⁴¹ A priori, cette séparation effective des échelles n'implique pas que le niveau inférieur ne joue pas un rôle crucial, puisque [vattay2015quantum] prouve que les propriétés de criticalité quantiques sont typiques des molécules du vivant, sans qu'il n'y ait a priori de spécificité pour la vie dans cette détermination complexe par les échelles inférieures : [2016arXiv161102269V] a introduit une nouvelle approche liant théories quantiques et relativité générale dans laquelle il est montré que la gravité est un phénomène émergent et que la dépendance au chemin dans la déformation de l'espace de base introduit un terme supplémentaire au niveau macroscopique, qui permet d'expliquer les déviations attribuées jusqu'alors à la *matière noire*.

⁴² Un système est Turing-complet s'il est capable de calculer les mêmes fonctions qu'une machine de Turing, communément accepté comme l'ensemble du "calculable" (thèse de CHURCH). Pour mémoire, une machine de Turing est un automate fini à bande d'écriture infinie [moore2011nature].

⁴³ Il existe même un langage de programmation permettant de programmer en *Game of Life*, disponible à <https://github.com/QuestForTetris>. Sa genèse trouve son origine dans un défi posté sur *codegolf* ayant pour but la conception d'un Tetris, et a abouti à un projet collaboratif extrêmement avancé.

Démon de Maxwell qui serait capable de diminuer l'entropie d'un système isolé et donc contredire la deuxième loi de la thermodynamique : celui-ci utilise en fait l'information sur les positions et vitesses des molécules du système, et son action compense la perte d'entropie par sa captation d'information⁴⁴.

Cette notion d'accroissement local de l'entropie a été étudiée largement par CHUA sous la forme du *Local Activity Principle*, qui est introduit comme un troisième principe de la thermodynamique, permettant d'expliquer par des arguments mathématiques l'auto-organisation pour une certaine classe de systèmes complexes typiquement impliquant des équations de réaction-diffusion [[mainzer2013local](#)].

La manière dont l'information est stockée et compressée est essentielle pour la vie, puisque l'ADN est bien un système de stockage d'information, dont le rôle à différents niveaux bien loin d'être compris complètement. La complexité culturelle témoigne également d'un stockage de l'information à différents niveaux, par exemple au sein des individus mais aussi des artefacts et des institutions, et des flux d'information relevant nécessairement des deux autres types de complexité. Les flux d'information sont essentiels pour l'auto-organisation dans un système multi-agent. Les comportements collectifs de poissons ou d'oiseau sont des exemples typiques utilisés pour illustrer l'émergence et font partie des cas d'école de systèmes complexes. On commence cependant seulement à comprendre comment ces flux structurent le système, et quels sont les motifs spatiaux de transfert d'information au sein d'un *flock* par exemple : [[crosato2017informative](#)] introduit des premiers résultats empiriques avec l'entropie de transfert pour des poissons et pose les bases méthodologiques de ce type d'étude.

Production de connaissances

Nous avons à présent la matière suffisante pour en venir à la réflexivité. Il est possible de positionner la production de connaissances à l'intersection des interactions entre types de complexité développées ci-dessus. Tout d'abord, la connaissance telle que nous l'envisageons ne peut se passer d'une construction collective, et implique donc un encodage et une transmission de l'information : il s'agit à un autre niveau de toutes les problématiques liées à la communication scientifique. La production de connaissances nécessite donc cette première interaction entre complexité computationnelle et complexité informationnelle. Le lien entre complexité informationnelle et émergence est mobilisé si on considère l'établissement de connaissances comme un processus morphogénétique. Il est montré en [5.1](#) que le lien entre forme et fonction est fondamental en psychologie : nous pouvons l'in-

⁴⁴ Le démon de Maxwell est plus qu'une construction intellectuelle : [[cottet2017observing](#)] implémente un démon expérimentalement au niveau quantique.

terpréter comme un lien entre information et sens, puisque la sémantique d'un objet cognitif ne peut se passer d'une fonction. HOFSTADTER rappelle dans [hofstadter1980godel] l'importance des symboles à différents niveaux pour l'émergence d'une pensée, qui consistent à un niveau intermédiaire en des signaux. Enfin, la dernière relation entre complexité computationnelle et émergence est celle qui nous permet d'affirmer qu'on s'intéresse particulièrement à une production de connaissance sur des systèmes complexes, les deux premiers pouvant s'appliquer à tout type de connaissance.

Ainsi, toute *connaissance du complexe* embrasse non seulement toutes les complexités et leur relations dans son contenu, mais aussi dans sa nature comme nous venons de montrer. La structure de la connaissance en termes de complexité est analogue à la structure des systèmes qu'elle étudie. Nous postulons que cette correspondance structurelle implique une certaine récursivité, et donc un certain niveau de *réflexivité* (au sens de connaissance d'elle-même et de ses propres conditions).

On peut tenter d'étendre à la réflexivité en tant que réflexion sur le positionnement disciplinaire : suivant PUMAIN dans [pumain2005cumulativite], la complexité d'une approche est également liée à la diversité des points de vue nécessaire pour la construire. Pour atteindre ce nouveau type de complexité⁴⁵, qui serait une dimension supplémentaire liée à la connaissance des systèmes complexes, la réflexivité doit être au coeur de la démarche. [read2009innovation] rappelle que l'innovation a été rendue possible quand les sociétés ont été capables de produire et diffuser de l'information sur leur propre structure, c'est à dire quand elles ont pu atteindre un certain niveau de réflexivité. La *connaissance du complexe* serait donc le produit et le support de sa propre évolution grâce à la réflexivité qui a joué un rôle fondamental dans l'évolution du système cognitif : on pourrait ainsi suggérer de rassembler ces considérations, comme proposé par PUMAIN, sous une nouvelle notion épistémologique de *Rationalité Evolutive*.

Pour conclure, notons qu'étant donné la loi de la *requisite complexity*, proposée par [gershenson2015requisite] comme extension de la *requisite variety* [ashby1991requisite]⁴⁶, la *connaissance du complexe* devra nécessairement être *connaissance complexe*. Cet autre point de vue ren-

45 Pour laquelle des liens avec les types précédents apparaissent naturellement : par exemple, [gell1995quark] considère la complexité effective comme le *Contenu d'Information Algorithmique* (proche de la complexité de Kolmogorov) d'un Système Complex Adaptatif *observant un autre* Système Complex Adaptatif, ce qui donne son importance aux complexités informationnelle et computationnelle et suggère l'importance du point de vue d'observation, et par extension de la combinaison de ceux-ci - ce qui est par ailleurs à mettre en relation avec l'approche perspectiviste des sciences complexes présentée précédemment.

46 L'un des principes cruciaux de la cybernétique, la *requisite variety* postule que pour contrôler un système ayant un certain nombre d'états, le contrôleur doit avoir au moins autant d'états. GERSHENSON propose une extension conceptuelle à la complexité, qui peut être justifiée par exemple par [allen2017multiscale] qui introduit

force la nécessité de la réflexivité, puisque suivant MORIN (voir par exemple [morin1991methode] sur la production de connaissance), la *Connaissance de la Connaissance* est centrale dans l'établissement d'une pensée complexe.

Conséquences pratiques

Pour conclure cette section épistémologique, nous proposons de synthétiser l'ensemble des idées introduites sous forme de manifestations concrètes en découlant directement, et qui conditionneront fortement l'ensemble de la forme et de la sémantique de la connaissance introduite par la suite. Ces directions (que nous n'irons pas jusqu'à nommer principes car seulement à l'état d'ébauche) peuvent être regroupées en trois grandes familles : pratiques de modélisation, pratique de la Science Ouverte, et épistémologie. Sur le plan des pratiques de modélisation, dans chaque section se dégagent différents axes plus ou moins complémentaires :

- La modélisation, qui sera dans la majorité des cas équivalente à la simulation, doit être comprise comme un instruments de connaissance indirect sur des processus au sein d'un système complexe ou sur la structure de celui-ci (d'après la sous-section sur "pourquoi modéliser"), et les modèles devront nécessairement être complexes (d'après la réflexion sur les différents types de complexité) au sens qu'il capturent un phénomène d'émergence faible, tout en respectant des exigences de parcimonie.
- L'exploration des modèles est partie intégrante de l'entreprise de modélisation (voir reproductibilité), et le calcul intensif est un élément clé pour explorer efficacement les modèles de simulation (voir calcul intensif). Les méthodes d'analyse de sensibilité doivent être questionnées et étendues si besoin (comme l'illustre l'exemple de la sensibilité à l'espace).
- Comme suggéré par le positionnement perspectiviste, le couplage de modèles devra jouer un rôle crucial dans la capture de la complexité.

Pour la Science Ouverte, on peut extraire les points suivants :

- La nécessité de l'ensemble des démarches liées à la Science Ouverte pour parvenir à la construction de modèles toujours plus complexes, vers la co-construction de modèles par différentes disciplines.
- Dans ce cadre, l'ouverture complète du code source, ainsi que sa lisibilité sont cruciaux. L'explicitation complète du modèle

la *requisite variety* multi-échelle, démontrant la compatibilité avec une théorie de la complexité basée sur la théorie de l'information.

dans le compte-rendu scientifique, ainsi qu'une documentation du code auto-suffisante, sont deux aspects de celle-ci.

- La question des données ouvertes n'est pas négociable dans ce cadre. La quasi-totalité de nos traitements est basée sur des données initialement ouverte, et lorsque ce n'est pas le cas nous travaillons à un niveau agrégé auquel on peut fournir les données. Les jeux de données construits sont ouverts.
- Concernant les méthodes d'exploration interactive, qui sont un pendant de l'ouverture de la Science, nous en développons un certain nombre, mais restons limités par rapport au pré-requis idéal qui devrait rendre celles-ci totalement compatibles avec une démarche reproductible.

Enfin, sur le point épistémologique, on peut également tirer des implications "pratiques" qui seront bien évidemment plus implicites dans notre démarche, mais pas moins structurantes :

- Notre inspiration sera essentiellement interdisciplinaire et cherchera à croiser les différents points de vue.
- Les différents domaines de connaissance (notion que nous préciserons en 9.3, mais qu'on peut comprendre pour l'instant au sens des domaines théorique, empirique et de la modélisation introduits par [livet2010]) sont indissociables pour toute démarche de production scientifique, et nous les mobiliserons de manière fortement dépendante.
- Notre démarche devra comprendre un certain niveau de réflexivité.
- La construction d'une connaissance complexe ([morin1991methode]) est ni inductive ni déductive, mais constructive dans l'idée d'une morphogenèse de la connaissance : il peut par exemple être délicat d'identifier clairement des "verrous scientifiques" précis puisque cette métaphore suppose qu'il faut débloquer un problème déjà construit, et de même de faire rentrer notions, concepts, objet ou modèles dans des cadres analytiques stricts, en les catégorisant selon une classification fixe, alors que l'enjeu est de comprendre si la construction des catégories est pertinente. Le faire a posteriori relève d'une négation de la circularité et de la récursivité de la production de connaissance. L'élaboration de modes de compte-rendus rendant compte du caractère diachronique et des propriétés évolutives de celle-ci est un problème ouvert.



*

CONCLUSION DU CHAPITRE

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 extrapoler certains développements des narrations qui se doivent synthétique (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 Evolutive* 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 univoque 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 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 a largement mobilisé ce type d'étude comme par exemple [paulus2004coevolution]), 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 9.

CONCLUSION DE LA PARTIE I : UNE DÉFINITION DE LA CO-ÉVOLUTION

Cette première partie nous permet de cerner bien plus précisément notre question de recherche. En effet,

1. le premier chapitre nous a permis de dresser la diversité des processus impliqués et des échelles temporelles et spatiales concernées ;
2. le deuxième chapitre nous a donné une vue très générale des modélisations existantes et de leur contexte scientifique précis ;
3. le troisième chapitre positionne la question de manière épistémologique, apporte un éclairage multi-disciplinaire sur la co-évolution, et clarifie la complexité dans laquelle nous nous situons.

Cela nous permet d'ouvrir sur les directions à prendre par la suite pour mener à bien l'entreprise de modélisation de la co-évolution.

Définir la co-évolution

Après l'aperçu de la littérature donné en 2.1, incluant différents degrés de couplage entre les composantes des réseaux et territoires, nous sommes tout d'abord en mesure de préciser ce que nous entendrons par *modéliser la co-évolution*, en fixant une définition de la co-évolution au regard de l'aperçu multi-disciplinaire mené en 3.3.

Nous proposons l'entrée suivante pour le cas spécifiques des réseaux de transport et des territoires, qui fait écho on le rappelle au trois point essentiels (existence de processus évolutif, définition des entités ou des populations, isolation de sous-systèmes dans le temps et l'espace) dégagés en 3.3 :

- Les processus évolutifs correspondent aux transformations des composantes du système territorial aux différentes échelles : transformation sur le temps long des villes, de leur réseaux, transmission entre villes des caractéristiques socio-économiques portées par les agents microscopiques mais aussi culturelle, reproduction et transformation des agents eux-mêmes (firmes, ménages, opérateurs)⁴⁷.

⁴⁷ Cette liste s'appuie sur les hypothèses de la théorie évolutive que nous avons déjà introduite brièvement et que nous développerons à part entière en Chapitre 4. Elle ne peut être exhaustive, puisque ce qui ferait "l'ADN d'une ville" reste une question ouverte comme nous le rappelle DENISE PUMAIN dans un entretien dédié D.3.

- Au sein d'un système territorial, pourront être en co-évolution à la fois : (i) des entités précises (telle infrastructure et telles caractéristiques de tel territoire par exemple), lorsque leur influence mutuelle sera circulairement causale (à l'échelle leur correspondant); (ii) des populations d'entités, ce qui se traduira par exemple par tel type d'infrastructure et telle composante territoriale co-évoluent au niveau statistique dans une région géographique donnée; (iii) l'ensemble des composantes d'un système à petite échelle géographique lorsqu'il existe de fortes interdépendances globales. Notre vision est donc fondamentalement *multi-échelles* et articule différentes significations à différentes échelles.
- Enfin, la contrainte d'une isolation implique, en lien avec le point précédent, que la co-évolution et l'articulation des significations auront un sens s'il existe des isolations spatio-temporelle de sous-systèmes où s'effectuent les différentes co-évolutions, ce qui est en accord direct avec un vision en *Systèmes de systèmes multi-échelles*.

L'une de nos contributions en synthèse faite en 9.2 sera de formaliser cette définition au regard des résultats que nous aurons obtenus. Elle constituera jusque là notre base d'investigation.

Nous pouvons alors synthétiser les résultats fondamentaux de cette première partie dans les deux faits marquants suivants :

1. Il est légitime de parler de co-évolution des réseaux de transport et des territoires d'un point de vue théorique et thématique, et nous en donnons une définition dans ce cas particulier.
2. Celle-ci reste très peu explorée dans la littérature de modélisation urbaine, les caractéristiques des disciplines concernées et leurs interactions pouvant en être une cause.

Développons à présent les perspectives qui s'ouvrent à ce stade.

Du besoin d'une caractérisation empirique

La signification la plus large, l'interdépendance généralisée, trouve vite ses limites si les motifs ne sont pas finement caractérisés. Elle permet comme prémisses épistémologique de considérer certaines ontologies et certaines démarches de modélisation, mais permet difficilement de comprendre finement la structure et les processus d'un système. Il s'agira alors de descendre en généralité et de considérer des sous-systèmes, au sein desquels on peut s'intéresser à la co-évolution d'entités et de population. Une compréhension à ce niveau nécessite alors une caractérisation empirique fine, sans quoi notre distinction n'aurait pas de sens. Une question qui s'ouvre, et que nous devrons

traiter par la suite, est alors quelles sont les méthodes empiriques possibles pour caractériser une coévolution entre entités ou populations d'entités.

Deux pistes complémentaires

L'état de l'art fait en 2.1 ci-dessus témoigne d'une faiblesse de la littérature dans le domaine du couplage fort entre évolution des territoires et croissance des réseaux, vu la portée restreinte et la disparité des travaux revus. Les lacunes à combler sur ce point seraient donc liées à l'introduction de modèles fortement couplés dans le temps plus ou moins multi-processus et multi-échelles, pour lesquels une partie des modèles décrits en 2.1 puis en 2.3 sont précurseurs.

Les premières recherches exploratoires que nous devrons mener doivent répondre à différentes tensions conceptuelles qui découlent des conclusions que nous venons de tirer :

- permettre à la fois une approche empirique, et en particulier un méthode de caractérisation, ainsi qu'une approche de modélisation ;
- permettre la prise en compte de différentes échelles ;
- permettre l'inclusion d'ontologies pour les territoires et pour les réseaux qui ne sont pas toujours directement compatibles.

Nous choisirons pour répondre simultanément à ces différentes problématiques une stratégie originale, par double entrée thématique, dont l'introduction, la contextualisation et le développement fera l'objet de la deuxième partie.

Deuxième partie

BRIQUES ÉLÉMENTAIRES

Cette partie construit les briques élémentaires qui seront utilisées ensuite pour la synthèse. Celles-ci sont à la fois empiriques, méthodologiques, conceptuelles et sur le plan de la modélisation. Elles s'inscrivent dans deux axes complémentaires, l'un basé sur la Théorie Evolutive des Villes, l'autre sur la Morphogenèse urbaine.

INTRODUCTION DE LA PARTIE II

Il aura finalement pu le faire, ce voyage. Pas, ou très peu de villes. Quelle âme dans ces street et avenues perpendiculaires, qu'on traverse nécessairement en bagnole. Encore un plein, à croire que c'est fait exprès, pour le charme de l'odeur d'essence. Tiens ça serait amusant de regarder ce que racontent ces stations d'ailleurs, à garder en tête. Un aller-retour au pas de course au Mont Elbert, puis à Longs Peak. On sort bientôt du Colorado, faudra dire au revoir au gummy bears. Damn it, Denver est si proche, ça vaudrait la peine. Tant pis, the mountains are calling and I must go, comme dirait l'autre. Que connaît-on finalement d'un territoire en conséquence de nos découvertes si sélectives ? Une infime partie du spectre des échelles ? Une infime étendue spatiale : on ne s'invente pas une dimension supplémentaire si facilement. Peut-être au moins la prise de conscience des antagonismes, des dualités. Et la conscience d'avoir à chaque fois dû privilégier l'un des aspects. Pour faire des ponts il faut être préparé. Pour voir le monde par un regard qui en capture plusieurs, il faut déjà avoir compris, c'est à dire intégré subjectivement, les processus correspondants. Souvenir d'une des premières courses sérieuses : les arêtes de la Meije, 23h consécutives pour terminer par des hallucinations sur le chemin à prendre que les étincelles des crampons sur les éboulis ne suffisaient plus à éclairer dans la nuit qui était retombée. A fine line, ce ressenti concret du gouffre de part et d'autre qui implique le tâtonnement, s'ancre dans l'inconscient avant même d'avoir atteint le stade des hallucinations : nous parcourons à chaque instant une fine arête, qui est autant celle de l'arbitraire du road trip que celle des ponts qui résistent difficilement quand vient la crue. Sur cette arête, les points d'ancre bien évidemment solides mais aussi hétérogènes sont gages de vie : la diversité combat l'adversité.

Un paradoxe intrinsèque à nombre de démarches de production de connaissance est un besoin de consistence intrinsèque et d'une portée satisfaisante d'explication des phénomènes concernés, qui s'oppose à une inévitable réduction des dimensions explorées mais également à la fragilité des ponts qu'elle tente de former vers d'autres corpus de connaissances. L'image prise ci-dessus suggère que le tâtonnement, c'est à dire une progression pas à pas sans précipitations, ainsi que la solidité des ancrages, sont des atouts solides pour affronter ce paradoxe.

Cette partie ouvre directement les pistes de réponse thématiques pour la modélisation de la co-évolution que nous avons évoqué en conclusion de la première partie, et pose ainsi ces ancrages forts. Elle pose toutefois des bases sans entrer dans le cœur du sujet par ce souci de robustesse par entrée progressive, et construit donc les *briques élé-*

mentaires de notre démarche. Deux chapitres traitent ainsi successivement les thématiques suivantes :

1. Un premier chapitre s'intéresse à la Théorie Evolutive Urbaine est une entrée privilégiée sur les systèmes urbains d'un point de vue évolutif, et intègre en son cœur un point de vue multi-scalar de ces systèmes. Il éclaire des propriétés fondamentales des systèmes territoriaux impliquées par la théorie évolutive, en introduisant une première analyse empirique de la variabilité spatiale des interactions entre forme urbaine et forme de réseau, puis en développant une méthodologie de caractérisation statistique de la co-évolution (au sens intermédiaire de la population). Il introduit ensuite un premier modèle d'interaction entre système de ville et flux du réseau de transport, avec réseau statique.
2. Un second chapitre explore le concept de morphogenèse, qui permet une entrée conceptuelle à la caractéristique de modularité nécessaire pour avoir co-évolution. Après avoir développé une définition interdisciplinaire de la morphogenèse, il introduit un modèle de morphogenèse urbaine basé sur des processus d'agrégation-diffusion pour la densité de population, et est ensuite couplé séquentiellement à un modèle de génération de réseau.

★ ★

★

PRÉLIMINAIRES MATHÉMATIQUES

Afin de toucher l'audience la plus large possible, nous proposons de préciser dans cet intermède préliminaire les définitions de notions ou méthodes clés qui seront utilisées de manière par la suite, souvent hors d'un cadre mathématique. Ce choix permet de garder un cadre rigoureux sans rendre indigeste la lecture du manuscrit à une grande partie de son public légitime. Sauf indication contraire, les spécifications données ici feront référence lors de l'utilisation des termes correspondants.

Statistiques

Correlation

Sauf indication contraire, nous estimerons la covariance entre deux processus par estimateur de Pearson, c'est à dire si $(X_i, Y_i)_i$ est un jeu d'observations des processus X, Y , la corrélation est estimée par

$$\hat{\rho} = \frac{\hat{\text{Cov}}[X, Y]}{\sqrt{\hat{\sigma}[X] \cdot \hat{\sigma}[Y]}}$$

où la covariance est estimée par l'estimateur non biaisé $\hat{\text{Cov}}$.

Causalité de Granger

Une série temporelle multi-dimensionnelle $\vec{X}(t)$ présente une causalité de Granger si avec

$$\vec{X}(t) = A \cdot \left(\vec{X}(t-\tau) \right)_{\tau>0} + \varepsilon$$

il existe τ, i tels que $a_{i\tau} > 0$ significativement. Nous utiliserons une version faible de la causalité de Granger, c'est à dire un test sur les corrélations retardées définies par

$$\rho_\tau [X_i, X_j] = \hat{\rho} [X_i(t-\tau), X_j(t)]$$

avec τ retard ou avance. Cela nous permettra de quantifier des relations entre variables aléatoires définies dans l'espace et dans le temps.

Régression Géographique Pondérée

La Régression Géographique Pondérée est une technique d'estimation de modèles statistiques permettant de prendre en compte la non-stationnarité spatiale des processus. Si Y_i est une variable à expliquer

et X_i un jeu de variables explicatives, mesurés en des mêmes points de l'espace, on estime un modèle $Y_i = f(X_i, \vec{x}_i)$ à chaque point \vec{x}_i , en prenant en compte les observations par pondération spatiale autour du point, où les poids sont fixés par un noyau pouvant prendre plusieurs formes, par exemple un noyau exponentiel est de la forme

$$w_i(\vec{x}) = \exp(-\|\vec{x} - \vec{x}_i\|/d_0)$$

L'échelle de stationnarité spatiale supposée par le modèle est alors de l'ordre de d_0 . Celle-ci peut être ajustée par validation croisée par exemple.

Apprentissage statistique

On désignera par *Apprentissage supervisé* toute méthode d'estimation d'une relation entre variables $Y = f(X)$ où la valeur de Y est connue sur un échantillon de données. On parlera de classification si la variable est discrète. La classification non-supervisée consiste à construire Y lorsque seul X est donné. On utilisera pour classifier une technique basique qui donne de bons résultats sur des données qui n'ont pas une structure exotique : la méthode des *k-means*, répétée un nombre suffisant de fois pour prendre en compte son caractère stochastique. Le complexité du *k-means* est polynomiale en moyenne, bien que la résolution exacte du problème de partition soit NP-difficile.

Overfitting

La question de l'overfitting est particulièrement importante lors de l'estimation de modèles, puisque un nombre trop important de paramètres pourra conduire à capturer le bruit de réalisation comme structure. Lors de l'estimation de modèles statistiques, des critères d'information sont mobilisables pour quantifier le gain d'information produit par l'ajout d'un paramètre, et obtenir un compromis entre performance et parcimonie.

Le *Critère d'Information d'Akaike* (AIC) permet de quantifier le gain d'information permis par l'ajout de paramètres dans un modèle. Pour un modèle statistique qui dispose d'une Fonction de Vraisemblance (*Likelihood*), l'AIC est alors défini par

$$AIC = 2k - 2 \ln \mathcal{L}$$

si k est le nombre de paramètres du modèle et \mathcal{L} la valeur maximale de la fonction de vraisemblance. [akaike1998information] montre que cette expression correspond à une estimation du gain d'information de Kullback-Leibler. Une correction pour les petits échantillons de taille n est donnée par

$$AICc = 2 \cdot \left(k + \frac{k^2 + k}{n - k - 1} - \ln \mathcal{L} \right)$$

Un critère similaire mais dérivé dans un cadre bayésien est le *Critère d'Information Bayésien* (BIC) [**burnham2003model**], qui conduit à une pénalisation plus forte du nombre de paramètres : $BIC = \ln n \cdot k - 2 \ln \mathcal{L}$.

Ces critères sont appliqués pour la sélection de modèles en étudiant leur différences entre modèles (seules les différences ont un sens, ceux-ci étant définis à une constante près) : le “meilleur” modèle est celui ayant le critère le plus faible. Dans le cas de modèles de performance comparables, il peut être pertinent de combiner les modèles par les poids d’Akaike $w_i = \exp(-\Delta AIC/2)$.

Processus stochastiques : Stationnarité

Les propriétés de stationnarité informent sur la variabilité de la distribution d’un processus stochastique. Soit $(\vec{X}_i)_{i \in I}$ un processus stochastique multidimensionnel. Il sera dit fortement stationnaire si sa loi ne dépend pas de i , c’est à dire si $\mathbb{P}[\vec{X}_i] = \mathbb{P}[\vec{X}_{i+1}]$. La stationnarité forte implique l’égalité de tous les moments pour tout i .

Nous utiliserons une notion plus faible de la stationnarité des processus stochastiques, ou *Weak Stationarity*, qui utilise les deux premiers moments : $(\vec{X}_i)_{i \in I}$ est faiblement stationnaire si

1. $\mathbb{E}[\vec{X}_i] = \mathbb{E}[\vec{X}_0]$ pour tout i
2. $\text{Cov}[\vec{X}_i, \vec{X}_j]$ ne dépend que de $i - j$

On peut parler de stationnarité faible du premier ordre si seulement la condition sur l’espérance est vérifiée, et du second ordre si on a aussi la condition sur l’autocovariance [**zhang2014test**].

Exploration de modèles de simulation

Nous désignerons par modèle de simulation tout algorithme associant une réalisation $\mathcal{M}[\vec{x}, \vec{\alpha}]$ à des données \vec{x} étant donné des paramètres α . L’enjeu est alors de comprendre le comportement du modèle de manière empirique, en le simulant, possiblement avec plusieurs répétitions pour des mêmes paramètres si celui-ci est stochastique. Il est alors par exemple possible de calibrer le modèle, c’est à dire trouver un jeu de paramètres permettant de remplir des objectifs donnés (qui peuvent être des distances à des données observées).

Plan d’expérience par échantillonnage

Le sort de la dimension (*Dimensionality Curse*), qui correspond simplement au fait que la taille de l’espace des paramètres est exponentielle en le nombre de paramètre. Lorsque celle-ci grandit mais qu’on veut garder un aperçu du comportement d’un modèle sur des valeurs très

variées des paramètres d'entrée, on peut alors échantillonner l'espace par un nombre donné de point.

L'échantillonnage par Hypercube Latin (LHS) permet d'assurer que pour chaque dimension, l'ensemble de la plage des valeurs est couverte lorsqu'on projette les points générés sur la dimension. L'échantillonnage par suite de Sobol permet de générer des points de discrépance faible (voir [B.4](#) pour une définition précise de la discrépance, qu'il faut comprendre comme une couverture de l'espace), et est particulièrement adapté au calcul d'intégrales.

L'échantillonnage peut devenir laborieux si le modèle est très irrégulier, ou pour un objectif précis de calibration. Pour cela, il existe des algorithmes spécifiques d'exploration et de calibration pour lesquels nous pouvons donner des exemples.

Calibration par algorithme génétique

Les algorithmes génétiques sont une alternative largement utilisée en optimisation, et font plus généralement partie des méta-heuristiques de computation évolutionnaire [[rey2015plateforme](#)]. Nous utiliserons généralement pour la calibration des modèles l'algorithme standard implémenté dans OpenMole, décrit en détails par [[pumain2017evaluation](#)]. Il s'agit d'une extension stochastique de l'algorithme NSGA2 pour l'optimisation multi-objectif. Il possède les caractéristiques principales suivantes :

- étant donné une population de paramètres candidats comme solution au problème multi-objectif, le front de Pareto est déterminé comme les points non-dominés ;
- un ensemble est construit à partir de ce front en prenant en compte une contrainte de diversité ;
- une descendance est générée à partir de cet ensemble par croisements et mutations et évaluée pour sa performance ;
- l'algorithme itère sur la nouvelle population.

[[pumain2017evaluation](#)] ajoute l'objectif du nombre de réplications aux objectifs de l'algorithme, afin de prendre en compte la stochasticité et trouver un compromis entre optimalité et robustesse des solutions.

Algorithmes spécifiques

Se basant sur des algorithmes génétiques, divers algorithmes ont été proposés pour raffiner l'exploration des modèles. Mentionnons deux exemples développés dans le cadre d'OpenMole : l'algorithme PSE (*Pattern Space Exploration*) [[10.1371/journal.pone.0138212](#)] vise à découvrir l'ensemble de l'espace des sorties d'un modèle, dans l'idée

d'une recherche de l'ensemble des comportements possibles. L'algorithme *Calibration Profile* [reuillon2015] vise quant à lui à établir le caractère nécessaire d'un paramètre pour arriver à un objectif, indépendamment des autres paramètres.

★ ★

★

4

CO-EVOLUTION : UNE ENTRÉE PAR LA THÉORIE EVOLUTIVE URBAINE

L'ouverture de la première Ligne à Grande Vitesse en France entre Paris et Lyon a-t-elle eu un impact sur les dynamiques territoriales concernées ? [bonnafous1987regional] montre qu'elle en a eu à l'échelle régionale, dans des secteurs particuliers, comme par exemple le tourisme en Bourgogne. En-a-t-elle eu sur le temps long ? À quelles échelles, pour quels territoires ? Nous trouvons la question des hypothétiques *effets structurants*, que nous avons abordé en Chapitre 1 par une entrée à plusieurs échelles (micro, meso et macro), ainsi que par le développement progressif de l'idée de co-évolution. Ces caractéristiques sont en fait au cœur de la Théorie Évolutive des villes, dont nous proposons donc ici d'approfondir les implications pour notre problématique.

Après avoir rappelé en préliminaire les caractéristiques essentielles de la Théorie Evolutive, nous étudions dans une première section à l'échelle mesoscopique une manifestation simple des interactions entre territoires et réseaux, que nous capturons dans des indicateurs morphologiques pour chacun, et pour lesquels nous étudions les corrélations spatiales. Nous introduisons ensuite l'aspect dynamique en étudiant la notion de causalité spatio-temporelle dans la section 4.2. Celle-ci est essentielle d'une part d'un point de vue méthodologique par l'introduction d'une méthode originale permettant dans certains cas de mieux cerner les influences respectives entre réseaux et territoires, mais également d'un point de vue thématique concernant l'existence avérée d'une co-évolution. Les multiples configurations obtenues mises en évidence pour un modèle simple de croissance urbaine couplant fortement croissance du réseau et densité, qu'on désignera comme *régimes de causalité*, témoignent de causalités circulaires qui sont bien des marques d'une co-évolution. L'application au cas de la croissance du réseau ferroviaire et des villes en Afrique du Sud montre que cette méthode s'applique sur données empiriques et que différents régimes peuvent en être extraits. Nous explorons enfin dans une dernière section 4.3 les possibilités offertes par les modèles d'interaction issus de la théorie évolutive, à une petite échelle spatiale et longue échelle de temps, ce qui permet de démontrer l'existence d'effets de réseau de manière indirecte, sans même introduire d'aspects de co-évolution dans un premier temps. Ainsi, nous façonnons les premières briques, sur différents aspects des interactions et de la co-évolution entre réseaux et territoires, qui peuvent paraître lointain en lecture rapide, mais qui sont bien reliés en filigrane par

les questions fondamentales à laquelle la Théorie Evolutive tente de répondre.

* * *

*

Ce chapitre est composé de divers travaux. La première section reprend une partie traduite de [raimbault2017calibration] pour l'analyse morphologique, puis les résultats présentés par [raimbault2016cautious] pour l'analyse des correlations ; la deuxième section correspond à la majorité de [raimbault2017identification] pour la formulation théorique et l'illustration sur données synthétiques, puis présente les résultats de [raimbault:halshs-01584914] pour l'application. Enfin la dernière section est une traduction de [raimbault2017indirect].

THÉORIE EVOLUTIVE URBAINE

Nous avons déjà évoqué divers aspects de la Théorie Evolutive des villes, en relation à la complexité en géographie, puis à certains modèles de systèmes urbains qu'elle a produit. Une synthèse est ici nécessaire pour poser précisément le cadre dans lequel nos développements s'inscriront. Cette théorie a été introduite initialement dans [pumain1997pour] qui argumente pour une vision dynamique des systèmes de ville, au sein desquels l'auto-organisation est essentielle. Les villes sont des entités spatiales évolutives interdépendantes dont les interrelations font émerger le comportement macroscopique à l'échelle du système de villes. Le système de villes est aussi vu comme un réseau de villes, ce qui renforce sa vision en tant que système complexe. Chaque ville est elle-même un système complexe dans l'esprit de [berry1964cities], l'aspect multi-scalaire, au sens d'échelles autonomes mais ayant chacune un rôle spécifique dans les dynamiques du système, étant essentiel dans cette théorie, puisque les agents microscopiques véhiculent les processus d'évolution du système à travers des rétroactions complexes entre les échelles. Le positionnement de cette théorie au regard des Sciences des Systèmes Complexes a plus tard été confirmé [pumain2003approche].

Il a été montré que la théorie évolutive fournit une interprétation des lois d'échelle qui sont omniprésentes dans les systèmes urbains, qui découleraient de la diffusion des cycles d'innovation entre les villes [pumain2006evolutionary], qui ont par ailleurs été mis en évidence de manière empirique pour plusieurs systèmes urbains [pumain2009innovation]. La notion de résilience d'un système de villes, induit par le caractère adaptatif des ces systèmes complexes, implique que les villes sont les moteurs et les adaptateurs du changement social [pumain2010theorie]. Enfin, la dépendance au chemin est source de non-ergodicité (voir définition en 4.1) au sein de ces systèmes, rendant les interprétations "universelles" des lois d'échelle développées par les physiciens incompatibles avec la théorie évolutive [pumain2010theorie].

La Théorie Evolutive des Villes a été élaborée conjointement avec des modèles de systèmes urbains : par exemple le modèle Simpop2 introduit par [bretagnolle2006theory] est un modèle basé agent qui prend en compte des processus économiques, et simule sur de longues échelles de temps les motifs de croissance urbaine pour l'Europe et les Etats-unis [doi:10.1177/0042098010377366]. Les accomplissements les plus récents de la théorie évolutive reposent sur les productions du projet ERC GeoDiversity, présentées dans [pumain2017urban], qui incluent de progrès avancés à la fois techniques (logiciel Open-Mole¹ [reuillon2013openmole]), thématiques (connaissance issue des modèles SimpopLocal [schmitt2014modelisation] et Marius [cottineau2014evolution]) et méthodologiques (modélisation incrémentale [cottineau2015incremental]).

¹ <http://openmole.org/>

Pour une analyse épistémologique par méthode mixtes de la théorie évolutive, qui permet de renforcer cet aperçu bibliographique par une de sa genèse, en quelque sorte de sa *forme*, se référer à 9.3 qui l'utilise comme cas d'étude pour construire un cadre de connaissances. En particulier, une analyse croisée des entretiens avec DENISE PUMAIN et ROMAIN REUILLOU, révèle la fertilisation croisée entre connaissances géographiques et connaissances informatiques, permis par l'effort interdisciplinaire de développement des modèles et de leurs méthodes d'exploration.

Implications

Reprendons le cœur de la Théorie Evolutive, comme synthétisé par DENISE PUMAIN elle-même (entretien en D.3) : “[Il s’agit d’]une Théorie Géographique ayant pour ambition de rassembler la plupart des faits stylisés connus sur les villes et leur organisation dans les territoires, dans une perspective hors-équilibre et non statique, en les suivant sur de longues périodes de temps et mettant une emphase sur les facteurs structurants et les bifurcations.”

Nous pouvons ainsi considérer la complexité des systèmes de villes au sens de la Théorie Evolutive comme un macro-concept morinien [morin1976methodel], c'est à dire la combinaison complexe de multiples concepts chacun nécessaires à la construction. Les concepts suivants sont ainsi nécessaires :

- Aspect hors-équilibre des systèmes urbains. Celui-ci étant spatialisé, il implique des dynamiques spatio-temporelles complexes, et donc des propriétés de non-stationnarité pour les processus spatio-temporels associés.
- Dynamique systémique au niveau du système de villes, c'est à dire une forte interdépendance entre villes pouvant être interprété comme une co-évolution.
- Rôle central des interactions entre villes comme moteurs des processus de croissance, existence d'effets de structure sur le temps long.

Ces concepts seront ainsi explorés selon différentes perspectives dans ce chapitre, dans les sections suivantes :

1. D'un point de vue empirique, nous étudierons d'abord un exemple de propriétés de non-stationnarité de caractéristiques pour les territoires et les réseaux, ainsi que de leur interactions.
2. Nous introduisons ensuite d'un point de vue méthodologique une approche permettant de mieux comprendre les motifs d'interdépendance spatio-temporels, et donc la *co-évolution* que nous rattacherons à son sens statistique intermédiaire que nous avons donné.

3. Enfin, une approche de modélisation permet d'explorer les interactions entre villes sur le temps long, en particulier en lien avec le réseau dans le cadre de nos questionnements.

★ ★

★

4.1 CORRÉLATIONS ENTRE FORME DES TERRITOIRES ET FORME DES RÉSEAUX

Au travers des processus de relocalisation, parfois induits par les réseaux, on peut s'attendre à ce que ces derniers influencent la distribution des populations dans l'espace. Réciproquement, les caractéristiques du réseau peuvent être influencées par celle-ci. Nous proposons ici d'étudier ces liens potentiels au travers de caractérisations issues d'indicateurs synthétiques pour ces deux objets, et des corrélations entre ces indicateurs.

A l'échelle du système de ville, le caractère spatial du système urbain peut être synthétisé par les positions des villes, associées aux variables agrégées au niveau de la ville qui représentent entièrement le système. Nous nous placerons ici à l'échelle mesoscopique, à laquelle la distribution spatiale fine des activités est nécessaire pour comprendre la structure spatiale du système territorial. Nous parlerons ainsi de caractéristiques morphologiques pour la densité de population et le réseau routier.

Le choix de limites "pertinentes" pour le territoire² ou la ville est un problème relativement ouvert qui dépendra souvent de la question à laquelle on cherche à répondre : [guerois2002commune] montrent que les entités obtenues sont largement différentes si on considère une entrée par la continuité du bâti (morphologique), par les fonctions urbaines (zones d'emploi par exemple) ou par les limites administratives. Nous choisissons ici l'échelle mesoscopique d'un centre métropolitain, de l'ordre de la centaine de kilomètres, d'une part pour la cohérence du champ spatial calculé, et d'autre part parce que des échelles plus grandes deviennent moins pertinentes pour la notion de forme urbaine, tandis que des échelles plus petites contiennent un bruit trop grand.

A cette échelle, on peut supposer les caractéristiques du territoire, pour la population et le réseau, définies localement et variant de manière relativement continue dans l'espace. Ainsi, la construction de champs d'indicateurs morphologiques permettra de reconstruire de manière endogène des ensembles territoriaux par la structure spatiale émergente des indicateurs aux échelles plus petites. Par exemple, les villes émergeront naturellement au sein des espaces non urbains. L'enjeu de cette section est ainsi d'étudier les propriétés de ces indi-

² Par exemple, tenter de définir un territoire *Parisien* présenterait plusieurs facettes. Du point de vue du territoire subjectif, les Parisiens intra-muros considèrent une barrière stricte au Boulevard Périphérique, tandis que des banlieues plus ou moins proches seront vues comme parisiennes depuis la province. Le territoire fonctionnel du Métropolitain s'étend légèrement plus loin que la limite administrative de Paris, mais couvre quasiment toute l'Ile-de-France lorsqu'on y ajoute RER et Transilien. Les périmètres de gouvernance sont en train d'évoluer avec le projet de gouvernance métropolitaine (voir 1.2). Des perceptions complémentaires du territoires peuvent ainsi être multipliées.

cateurs et de leurs interactions, et donc indirectement les interactions entre territoire et réseau par l'intermédiaire de leur forme.

4.1.1 Mesures morphologiques

Morphologie Urbaine

Les manières de quantifier et qualifier la *forme urbaine* à l'échelle considérée, et par extension à toute distribution de population dans l'espace ce qu'on peut appeler *forme territoriale*, sont nombreuses.

[guerois2008built] étudient la forme des villes Européennes par l'utilisation d'une mesure simple des gradients de densité du centre vers la périphérie. Nous avons cependant besoin de mesures ayant un certain niveau d'invariance pour extraire des formes typiques. Par exemple, deux villes monocentriques devraient être classifiées comme morphologiquement proches tandis qu'une comparaison directe des distributions de population pourra donner une distance très élevée³ entre les configurations selon la position des centres.

Une autre solution pour quantifier la morphologie urbaine est l'utilisation d'indices issus de l'analyse fractale, comme par exemple appliquée systématiquement par [2016arXiv160808839C] pour classifier les formes urbaines. Le lien entre morphologie urbaine et topologie du graphe de relations sous-jacent a été suggéré dans une approche théorique par [badariotti2007conception]. D'autres indices plus originaux peuvent être proposés, comme par [lee2017morphology] qui utilise les variations de trajectoire d'itinéraires traversant une ville pour établir une classification et montrer que celle-ci est fortement corrélée aux variables socio-économiques.

Nous choisissons pour notre étude de nous référer à la littérature en morphologie urbaine qui propose des jeux d'indicateurs variés pour décrire la forme urbaine [tsai2005quantifying]. [le2009quantifier] rappelle la nécessité d'une mesure multi-dimensionnelle de la forme urbaine. Le nombre de dimensions peut par ailleurs être réduit pour obtenir une description robuste avec un petit nombre d'indicateurs indépendants [Schwarz201029]. Il faut noter que nous ne considérons ici des indicateurs sur la distribution spatiale de la densité de population uniquement, et que des considérations plus élaborées sur la forme urbaine peuvent inclure par exemple la distribution des opportunités économiques et la combinaison de ces deux champs par des mesures d'accessibilité. Pour le choix des indicateurs, nous suivons l'analyse faite par [le2015forme] où une typologie morphologique des grandes villes européennes est obtenue. La cohérence de celle-ci

³ On peut comparer des distributions spatiales par une distance euclidienne entre les matrices correspondantes, ou par des distances plus élaborées comme la distance de Monge qui résout un problème de transport minimal et donne la quantité de déplacements nécessaires pour passer d'une distribution à l'autre.

suggère la capacité du jeu d'indicateurs utilisé à capturer la forme urbaine à cette échelle.

Indicateurs

Nous donnons à présent une définition formelle des indicateurs morphologiques. Nous considérons des données de population en grille $(P_i)_{1 \leq i \leq N^2}$, écrivons $M = N^2$ le nombre de cellules, d_{ij} la distance euclidienne entre les cellules i, j , et $P = \sum_{i=1}^M P_i$ la population totale. La forme urbaine est mesurée par :

1. Pente de la loi rang-taille γ , qui exprime le degré de hiérarchie de la distribution, calculé en ajustant une loi de puissance par Moindres Carrés Ordinaires par $\ln(P_i/P_0) \sim k + \gamma \cdot \ln(\tilde{i}/i_0)$ où \tilde{i} sont les indices de la distribution triée de manière décroissante (la constante k de l'ajustement ne joue pas de rôle dans la hiérarchie). Elle est toujours négative ou nulle, et des valeurs proches de zéro signifient une distribution complètement homogène. Son utilisation sur une grille est amenée de manière originale par [le2015forme].
2. Entropie de la distribution [le2015forme], qui exprime l'uniformité de la distribution, ce qui est une façon de capturer une agrégation :

$$\mathcal{E} = \sum_{P_i \neq 0} \frac{P_i}{P} \cdot \ln \frac{P_i}{P} \quad (2)$$

$\mathcal{E} = 0$ signifie que toute la population est dans une cellule tandis que $\mathcal{E} = 0$ signifie que la population est distribuée uniformément.

3. L'auto-corrélation spatiale donnée par l'indice de Moran [tsai2005quantifying], avec des poids spatiaux simples donnés par $w_{ij} = 1/d_{ij}$:

$$I = M \cdot \frac{\sum_{i \neq j} w_{ij} (P_i - \bar{P}) \cdot (P_j - \bar{P})}{\sum_{i \neq j} w_{ij} \sum_i (P_i - \bar{P})^2}$$

Celui-ci varie théoriquement entre -1 et 1, des valeurs positives impliquent des lieux d'agrégation ("centres de densité"), des valeurs négatives des fortes variations locales, tandis que $I = 0$ correspond à des valeurs de population totalement aléatoires.

4. Distance moyenne entre individus [le2009quantifier], qui témoigne de la dispersion spatiale de la population et quantifie un degré d'acentrisme (éloignement à un modèle monocentrique) :

$$\bar{d} = \frac{1}{d_M} \cdot \sum_{i < j} \frac{P_i P_j}{P^2} \cdot d_{ij}$$

où d_M est une constante de normalisation que nous prenons comme la diagonale de l'étendue considérée dans notre cas.

Les deux premiers indices ne sont pas spatiaux, mais nécessaires pour une bonne qualification des distributions de population comme le montre la part de variance expliquée par l'ensemble des indicateurs comme nous le présenterons par la suite, et sont complétés par les deux derniers prenant en compte l'espace.

RÉSULTATS Nous calculons les mesures morphologiques données ci-dessus sur des données réelles de densité, en utilisant la grille de population de l'Union Européenne à la résolution de 100m fournie de manière ouverte par Eurostat [eurostat]⁴. Le choix de la résolution, de la portée spatiale, et de la forme de la fenêtre sur laquelle les indicateurs sont calculés, sont faits suivant les spécifications thématiques précédentes. Nous considérons des fenêtres carrées de largeur 50km. Comme une résolution trop détaillée n'est pas désirable à cause de la qualité des données, nous agrégeons les données de la grille initiale à une résolution de 500m pour avoir des fenêtres de taille $N = 100$ correspondant à 50km de côté. Pour obtenir une distribution des indicateurs relativement continue dans l'espace, nous superposons les fenêtres en posant un décalage de 10km entre chaque, ce qui induit un lissage des valeurs et permet de s'extraire des effets de bord dus à la forme. Nous avons par ailleurs testé la sensibilité à la taille de la fenêtre en calculant des échantillons avec des tailles de 30km et 100km et avons obtenu des distributions spatiales similaires, ainsi que de fortes corrélations entre les champs et leur lissage à une résolution plus fine, comme détaillé en Annexe A.5.

L'implémentation des indicateurs doit être faite avec attention, puisque les complexités computationnelles peuvent atteindre $O(N^4)$ pour l'indice de Moran par exemple : nous utilisons la convolution par Transformée de Fourier Rapide, qui est une technique permettant de calculer l'indice de Moran avec une complexité en $O(\log^2 N \cdot N^2)$ ⁵.

Nous montrons en Fig. 16 des cartes donnant les valeurs des indicateurs, pour la France seulement afin de permettre une lisibilité. La première caractéristique frappante est la diversité des motifs morphologiques au travers de l'ensemble du territoire. L'auto-correlation est relativement haute dans les zones englobant les métropoles (Paris, Lyon, Marseille par exemple), avec les environs de Paris qui se détachent clairement. Lorsqu'on s'intéresse aux autres indicateurs, il est intéressant de constater des régimes régionaux : les zones rurales ont beaucoup moins de hiérarchie dans le Sud que dans le Nord, tandis que la distance moyenne est plutôt distribuée uniformément sauf

⁴ Cette base a certains défauts de précision qui ont été reconnus [bretagnolle2016ville] mais l'agrégation à une résolution supérieure devrait permettre de diminuer d'éventuels biais.

⁵ C'est à dire ayant un temps d'exécution borné par $\log^2 N \cdot N^2$ si N est la taille des données, ce qui est un gain considérable par rapport à N^4 : pour le traitement d'une grille de côté 100, le facteur de gain asymptotique sera d'environ 10000.

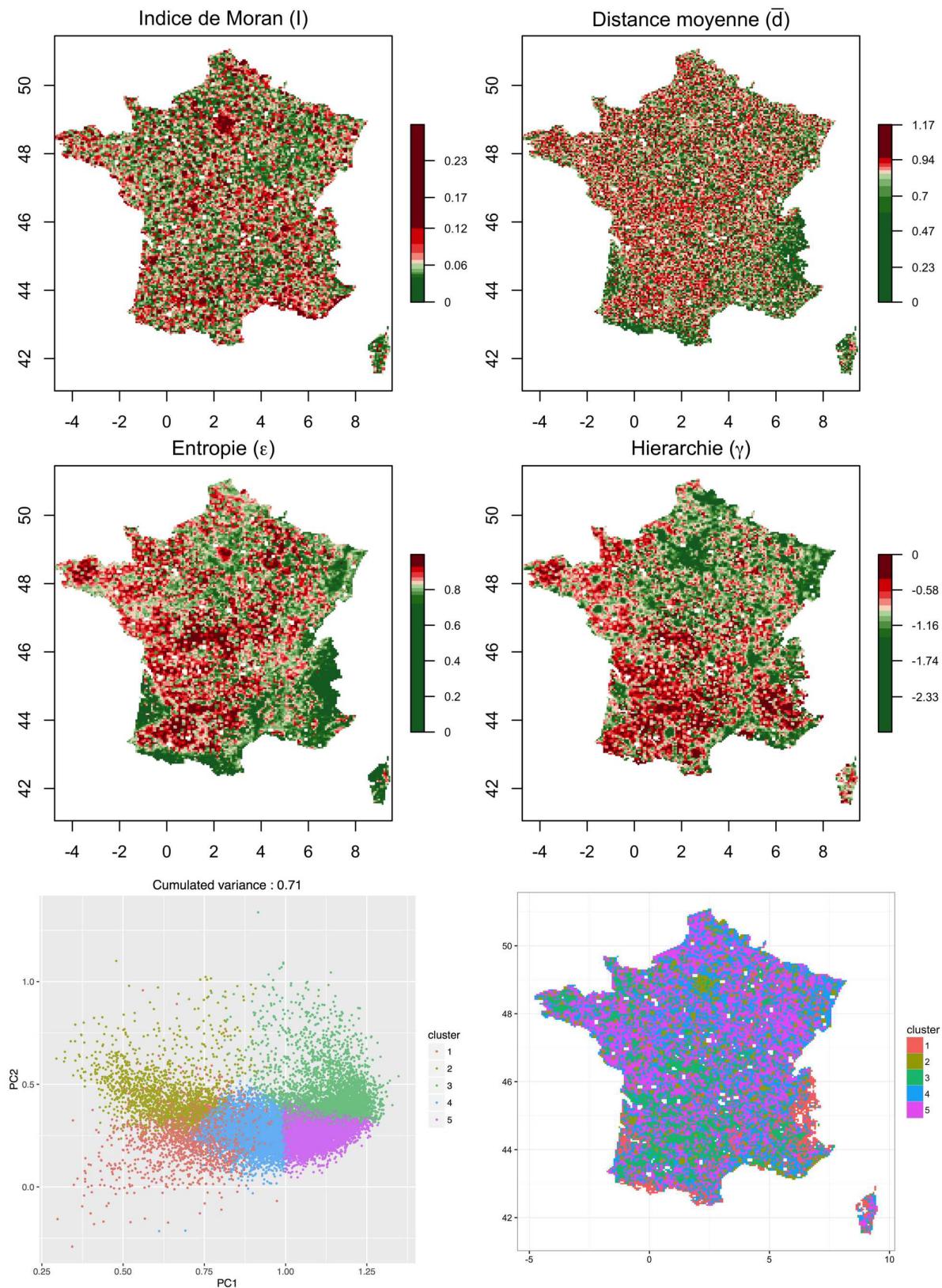


FIGURE 16 : Valeurs empiriques des indicateurs morphologiques. (Quatre cartes du haut) Distribution spatiale des indicateurs morphologiques pour la France. La détermination de l'échelle de couleur est faite par quantiles pour faciliter la lecture des cartes. (Bas gauche) Projection des valeurs morphologiques sur les deux premières composantes d'une analyse en composantes principales. La couleur donne le cluster dans une classification non supervisée (voir texte). (Bas droite) Distribution spatiale des clusters. Se référer au texte pour les détails sur la procédure d'estimation spatiale des indicateurs et sur la procédure de classification.

dans les zones montagneuses. Des régions qui présentent de fortes valeurs de l'entropie sont observées dans le centre et le Sud-ouest.

Pour avoir une meilleure compréhension des classes morphologiques existantes, nous utilisons une classification non-supervisée⁶ avec un algorithme des k-means⁷. Le nombre de clusters $k = 5$ induit une transition dans la variance inter-cluster, ce qui signifie qu'une variation de structure opère à ce nombre, que nous choisissons alors comme nombre de clusters. La séparation entre les classes est montrée en Fig. 16, panneau bas gauche, où nous représentons les mesures projetées sur les deux premières composantes d'une Analyse en Composantes Principales (expliquant 71% de la variance, ce qui est relativement conséquent). La carte des classes morphologiques confirme une opposition Nord-Sud dans le régime rural de fond (vert clair contre bleu), l'existence d'un régime de montagne (rouge) et d'un régime métropolitain (vert sombre). Une telle variété d'établissements sera l'un des objectifs du modèle en 5.2. Un calcul similaire des indicateurs morphologiques a été effectué pour la Chine en utilisant la grille de population à 1km fournie par [fu1km]. Les cartes sont disponibles en Appendice A.5.

4.1.2 Mesures de Réseau

Nous considérons d'autre part les mesures agrégées de réseau comme un moyen de caractériser les propriétés des réseaux de transport sur un territoire donné, de la même façon que les indicateurs morphologiques informent sur la structure urbaine. Nous proposons de calculer des indicateurs simples sur des étendues spatiales similaires à celles retenues pour la mesure de la morphologie, pour être en mesure d'explorer les relations entre ces mesures statiques.

L'analyse statique de réseau a été intensément documentée dans la littérature, voir par exemple [louf2014typology] pour une étude comparative des villes ou [2015arXiv151201268L] pour l'exploration de nouvelles mesures pour le réseau de rues. [2017arXiv170902939M] utilise des techniques issues de l'apprentissage profond pour établir une typologie des réseaux viaires urbains pour un grand nombre de villes dans le monde. Les enjeux derrière ce genre d'approches sont multiples : elles peuvent viser à des typologies ou caractérisations de réseaux spatiaux, à des compréhensions des processus dynamiques sous-jacents dans un but de modélisation de la morphogenèse, ou même de planification urbaine comme sont appliquées parfois les approches par *Space Syntax* [hillier1989social]. Nous nous plaçons ici

⁶ Qui on le rappelle consiste à partitionner l'espace des données selon leur structure endogène.

⁷ Vu la distribution des points qui ont une densité relativement homogène, des méthodes alternatives comme l'algorithme DBScan sont relativement équivalentes. Nous prenons ici un nombre de répétitions $b = 100$ de l'algorithme pour avoir un résultat robuste à la stochasticité.

plutôt dans les deux premières logiques. Notre contribution significative est la caractérisation du réseau routier sur de grandes étendues spatiales, couvrant l'Europe et la Chine.

Pré-traitement des données

Nous travaillons ici avec le réseau de rues, dont la structure est finement conditionnée aux configurations territoriales des densités de population. De plus, les données du réseau de routes actuel est disponible ouvertement par l'intermédiaire du projet OpenStreetMap (OSM) [[openstreetmap](#)]. Sa qualité a été étudiée pour différents pays comme l'Angleterre [[haklay2010good](#)] et la France [[girres2010quality](#)]. Il a été établi pour ces pays une qualité équivalente aux données officielles pour le réseau de rues primaire, au sens à la fois de la couverture spatiale et de la précision locale. Dans le cas de la Chine, bien que [[zheng2014assessing](#)] soulève une récente accélération de la couverture et de la précision des données OSM pour les routes, leur usage pour le calcul d'indicateurs de réseau peut être questionné à une échelle très fine. [[zhang2015density](#)] fournit une partition de la Chine en régions entre lesquelles le comportement qualitatif des données OSM varie. Nous devrons garder à l'esprit cette variabilité, et pour être assuré de la fiabilité des résultats, nous simplifierons le réseau à un niveau d'agrégation suffisant.

Le réseau constitué des segments de rue primaires est agrégé à la granularité fixe de la grille de densité pour créer un graphe. Celui-ci est ensuite simplifié pour garder uniquement la structure topologique du réseau, les indicateurs normalisés étant relativement robustes à cette opération. Celle-ci est nécessaire pour un calcul simple des indicateurs et une cohérence thématique avec la couche de densité. On garde uniquement les noeuds ayant un degré strictement supérieur ou inférieur à deux, et les liaisons correspondantes, en prenant soin d'agréger la distance géographique réelle en construisant le lien topologique correspondant. Vu l'ordre de grandeur de taille des données (pour l'Europe, la base initiale a $\simeq 44.7 \cdot 10^6$ liens, et la base finale simplifiée $\simeq 20.4 \cdot 10^6$), un algorithme spécifique parallèle est mis en place, de structure *split-merge*. Celui-ci découpe l'espace en zones qui peuvent être traitées indépendamment puis fusionnées. Il est détaillé en Appendice A.5.

Indicateurs

Nous introduisons des indicateurs pour avoir une idée large de la forme du réseau, utilisant un certain nombre d'indicateurs pour capturer le maximum de dimensions des propriétés des réseaux, plus ou moins liées à l'utilisation de ceux-ci. Ces indicateurs résument la structure mesoscopique du réseau sont calculés sur les réseau topologiques obtenus par les étapes précédentes de simplification. Notant

le réseau $N = (V, E)$, les noeuds V ont des positions spatiales $\vec{x}(v)$ et des populations $p(v)$ obtenues par agrégation de la population dans la partition de Dirichlet correspondante, les liens E ont des *distances effectives* $l(E)$ qui prennent en compte les impédances et les distances réelles (pour inclure la hiérarchie primaire du réseau). Nous utilisons alors :

- Caractéristiques du graphe, issues de la théorie des graphes, comme définies par **[haggett1970network]** : nombre de noeuds $|V|$, nombre de lien $|E|$, densité d , degré moyen $\bar{\delta}$, nombre cyclotomique μ , connectivité α , longueur moyenne des liens \bar{d}_l , population moyenne \bar{p} , coefficient de clustering moyen \bar{c} , nombre de composantes c_0 .
- Mesures liées au plus courts chemins : diamètre r , performance euclidienne v_0 (définie par **[banos2012towards]**), longueur moyenne des plus courts chemins \bar{l} .
- Mesures de centralité : celles-ci sont agrégées au niveau du réseau en prenant leur moyenne et leur niveau de hiérarchie, calculé par un ajustement des moindres carrés d'une loi rang taille, pour les mesures de centralité suivantes :
 - *Betweenness Centrality* [**crucitti2006centrality**], moyenne \bar{bw} et hiérarchie α_{bw} : étant donné la distribution de la centralité sur l'ensemble des noeuds, on prend la pente d'un ajustement rang-taille ainsi que la moyenne de la distribution.
 - *Closeness Centrality* [**crucitti2006centrality**], moyenne \bar{cl} et hiérarchie α_{cl}
 - Accessibilité [**hansen1959accessibility**], qui est dans notre cas calculée comme une *closeness* pondérée par les populations : moyenne \bar{a} et hiérarchie α_a
- Modularité [**blondel2008fast**], qui exprime la structure en communautés du réseau.

Le concept d'accessibilité est capturé ici par un indicateur de réseau, puisque son calcul implique d'attribuer des poids aux noeuds avec une population correspondante, et peut être interprété ensuite comme un temps de trajet moyen pondéré (comme nous l'avons fait en Chapitre 1). Cet indicateur est intéressant car à l'interface entre forme urbaine et forme du réseau, puisque la distribution de population sur les noeuds est prise en compte. On verra que celle-ci est fortement corrélée au même non-pondéré (corrélation de $\rho = 0.86$, estimée sur l'ensemble des points de mesure pour la Chine).

Résultats

Les indicateurs de réseau ont été calculés sur les mêmes zones que les indicateurs de forme urbaine, pour pouvoir les mettre en correspondance directe et calculer les corrélations par la suite. Nous montrons en Fig. 17 un échantillon pour la France.

Le comportement spatial des indicateurs révèle comme pour la forme urbaine des régimes locaux (urbain, rural, métropolitain), mais aussi des régimes régionaux très marqués. Ceux-ci peuvent être dus aux différentes pratiques agricoles selon les régions dans le cas du rural par exemple, impliquant une partition différente des parcelles ainsi qu'une organisation particulière de leur desserte. En taille du réseau, la Bretagne se détache nettement et rejoint les régions urbaines, témoignant de parcelles très fragmentées (et a fortiori d'un découpage foncier fragmenté également dans l'hypothèse simplificatrice d'une coincidence des parcelles et du foncier). Cela est partiellement corrélé à une faible hiérarchie dans l'accessibilité. Le Sud et l'Est du Bassin Parisien étendu se distinguent par une forte centralité d'intermédialité moyenne, en accord avec une forte hiérarchisation du réseau.

Pour la Chine, pour laquelle une sélection d'indicateurs est également donnée en A.5, on observe des variations locales et régionales encore plus marquées. Les zones urbaines fortement peuplées se détachent, correspondant à un régime bien particulier.

4.1.3 Correlations Statiques Effectives et Non-stationnarité

Corrélations spatiales

Les corrélations spatiales locales sont calculées sur des fenêtres regroupant un certain nombre d'observation, et donc de fenêtres sur lesquelles les indicateurs ont été calculés. Notons l_0 (qui vaut 10km dans les résultats précédents) la résolution des distributions des indicateurs. L'estimation des corrélations s'effectue alors sur des carrés de taille $\delta \cdot l_0$ (avec δ pouvant varier typiquement de 4 à 16). La valeur de δ influe directement sur le nombre d'observations, et donc la fiabilité de l'estimation. Nous montrons en Fig. 18 des exemples de corrélations estimées avec $\delta = 12$ dans le cas de la France. Avec 29 indicateurs, la matrice de corrélation est assez conséquente en taille, mais la dimension effective est réduite : une analyse en composante principale montre que $p = 10$ capture 60% de la variance, et la première composante capture déjà 16%, ce qui est considérable dans un espace où la dimension est de 406⁸.

8 Il s'agit de la dimension de la matrice de correlation entre 29 indicateurs, c'est à dire le nombre d'éléments de sa moitié moins sa diagonale. Si les corrélations étaient distribués aléatoirement, la première composante capturerait $1/406 = 0.2\%$ seulement, met les 10 premières 2%, puisque la variance se répartit équitablement entre des dimensions indépendantes.

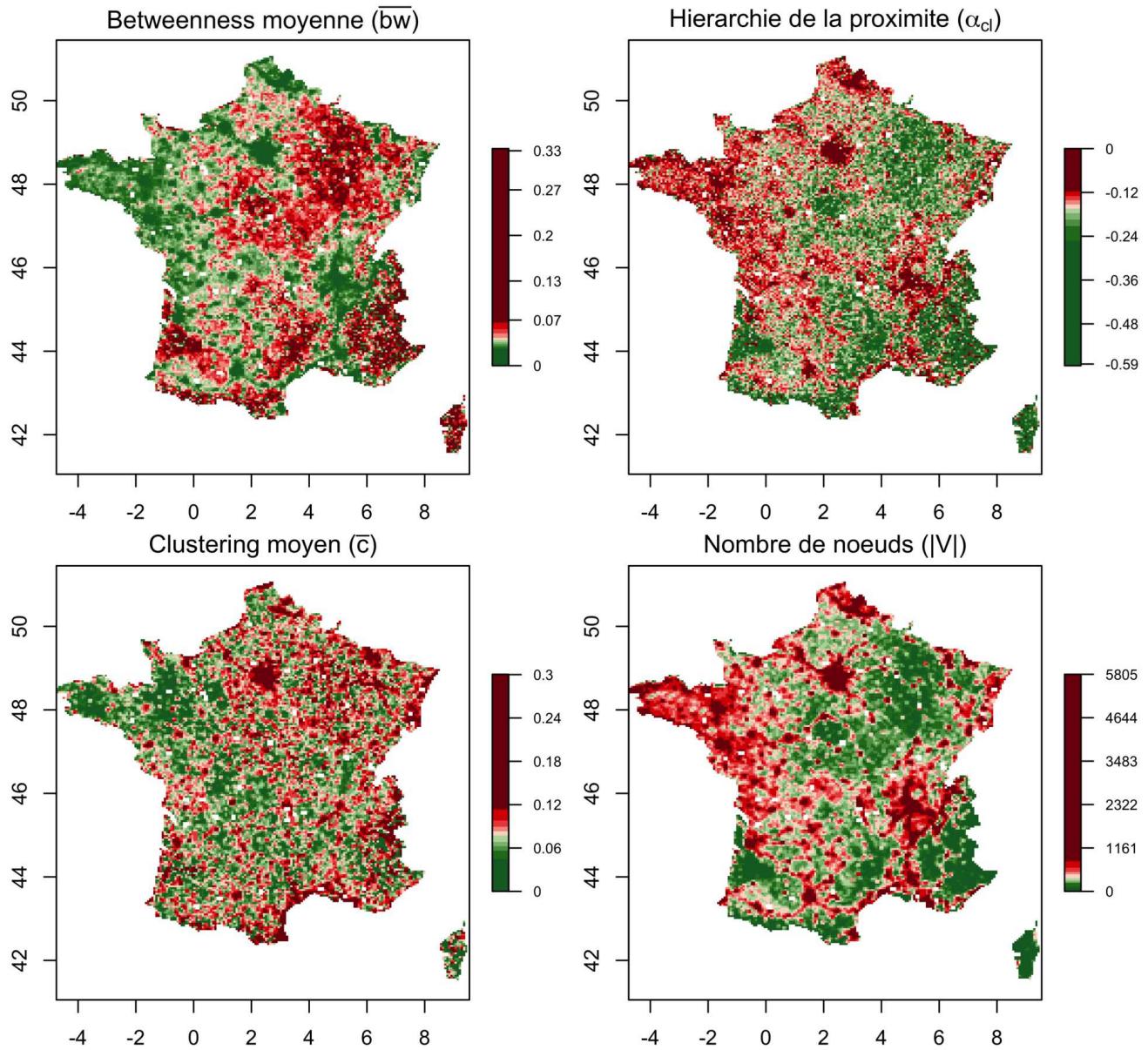


FIGURE 17 : **Distribution spatiale des indicateur de réseau.** Nous calculons les indicateurs pour la France, en correspondance avec les indicateurs morphologiques décrits précédemment. Nous donnons ici la centralité de chemin moyenne \bar{bw} , la hiérarchie de la centralité de proximité α_{cl} , le coefficient de clustering moyen \bar{c} et le nombre de noeuds $|V|$.

On peut s'intéresser aux sous-blocs morphologique, de réseau, ou les corrélations croisées, qui exprime directement un lien entre les propriétés de la forme urbaine et celles du réseau. Par exemple, la relation entre Betweenness moyenne et hiérarchie morphologique que l'on visualise permet de comprendre la processus correspondant à la correspondance des hiérarchies : une population hiérarchisée peut induire un réseau hiérarchisé ou le sens inverse, mais elle peut également induire un réseau distribué ou celui-ci peut créer une hiérarchie de population - il faut bien comprendre en terme de correspondance et non de causalité, mais cette correspondance informe sur différents régimes urbains.

Les métropoles semblent exhiber une corrélation positive dans ce cas, et des espaces ruraux une corrélation négative. Cela suggère une très grande variété de régimes d'interaction. La variation spatiale de la première composante confirme celle-ci, ce qui révèle clairement la non-stationnarité spatiale des processus d'interaction entre formes, puisque les premiers et second moments varient dans l'espace. Nous donnons en Annexe A.5 d'autres exemples de cartographie de la matrice des corrélation. La significativité statistique de la non-stationnarité spatiale peut être vérifiée de différentes façons⁹. Nous utilisons ici la méthode de [leung2000statistical] qui consiste à estimer par bootstrap la robustesse de modèles de Régression Géographique Pondérée. Ceux-ci seront développés ci-dessous, mais on obtient pour l'ensemble des modèles testés une non-stationnarité sans équivoque ($p < 10^{-3}$).

Par ailleurs, la distribution statistique des corrélations donnée en Fig. 81 en Annexe A.5 suit une loi asymétrique pour la morphologie seule, et plutôt symétrique pour le réseau et le croisement, ce qui voudrait dire que certaines zones ont des contraintes morphologiques assez fortes tandis que la forme du réseau est plutôt libre. Enfin, on constate sur les nuages de points de la même figure, croisant les valeurs des corrélations dans les différents blocs, que les configurations où les corrélations croisées sont les plus fortes correspondent à celles où les corrélations morphologiques et de réseau sont également fortes, confirmant l'imbrication des processus dans ce cas.

Variations des corrélations estimées

Nous montrons en Fig. 19 la variation de l'estimation des corrélations en fonction de la taille de la fenêtre. Plus précisément, on observe une forte variation (de plus de 0.1 pour chaque type) des correlations fonction de δ , qui se reflète dans la valeur moyenne de la matrice donnée ici. L'augmentation de δ cause pour l'ensemble un décalage dans le positif, mais également un rétrécissement de la distribution,

⁹ Il n'existe à notre connaissance pas de test générique de non-stationnarité spatiale. [zhang2014test] développe par exemple un test pour des régions rectangulaires de dimensions quelconques, mais dans le cas spécifiques des *point processes*.

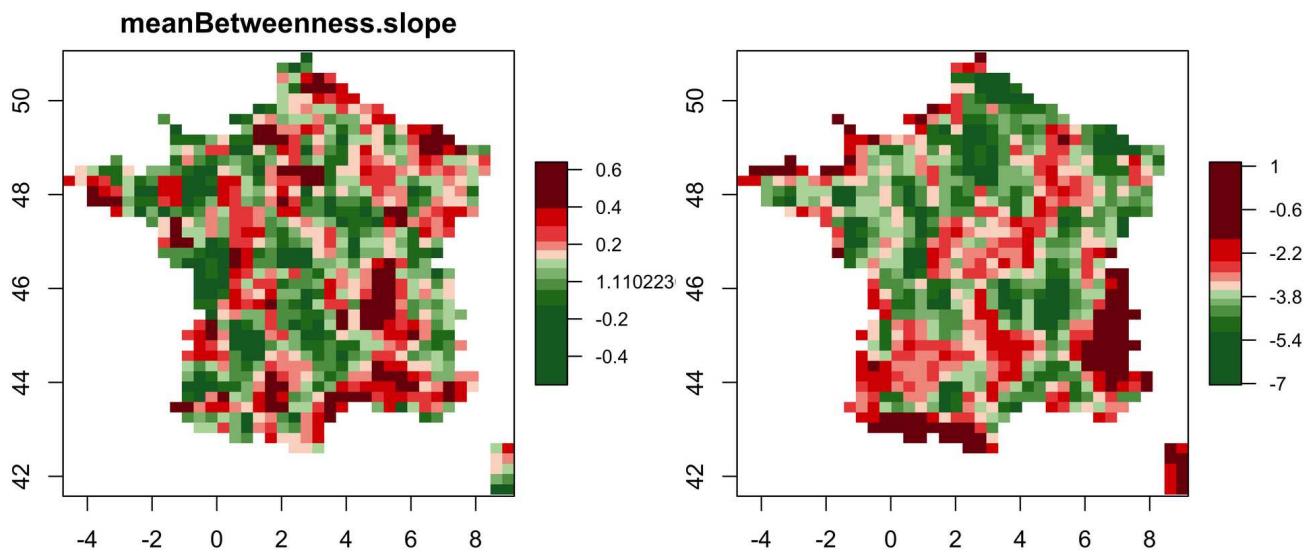


FIGURE 18 : Exemples de corrélations Spatiales. Pour la France, les cartes donnent $\rho [bw, \gamma]$ (Gauche) et la première composante de la matrice réduite (Droite).

ces deux effets se traduisant par une décroissance des corrélations absolues moyennes, qui se stabilisent approximativement pour les grandes valeurs de δ . Cette variation pourrait être révélatrice d'un comportement multi-échelle : le changement de la taille de la fenêtre ne devrait pas influer l'estimateur si un seul processus était sous-jacent, elle devrait seulement changer la robustesse de l'estimation. Le développement en Annexe A.5 illustre ce lien dans le cas de processus superposés à deux échelles, et démontre que cette structure de processus implique une variation de la corrélation estimée en fonction de δ , au moins dans les faibles valeurs, ce que nous observons ici en Fig. 19.

Par ailleurs, la variation de la taille normalisée de l'intervalle de confiance pour les corrélations, qui en théorie sous hypothèse de normalité devrait conduire $\delta \cdot |\rho_+ - \rho_-|$ à être constant, puisque les bornes varient asymptotiquement comme $1/\sqrt{N} \sim 1/\sqrt{\delta^2}$ (la démonstration est donnée en Appendice A.5), va dans la direction de cette hypothèse de processus superposés à plusieurs échelles comme proposé précédemment.

Ainsi, les processus sont à la fois non-stationnaires, et des indices poussent à laisser penser qu'ils résultent de la superposition de processus à différentes échelles¹⁰.

¹⁰ La notion de processus multi-scalaire est par ailleurs très diverse, et peut s'expliquer par exemple par la manifestation de lois d'échelles [west2017scale]. Une démarche plus proche de la notre est donnée par [Chodrow31102017] qui mesure les échelles intrinsèques aux phénomènes de ségrégation en utilisant des mesures issues de la Théorie de l'Information.

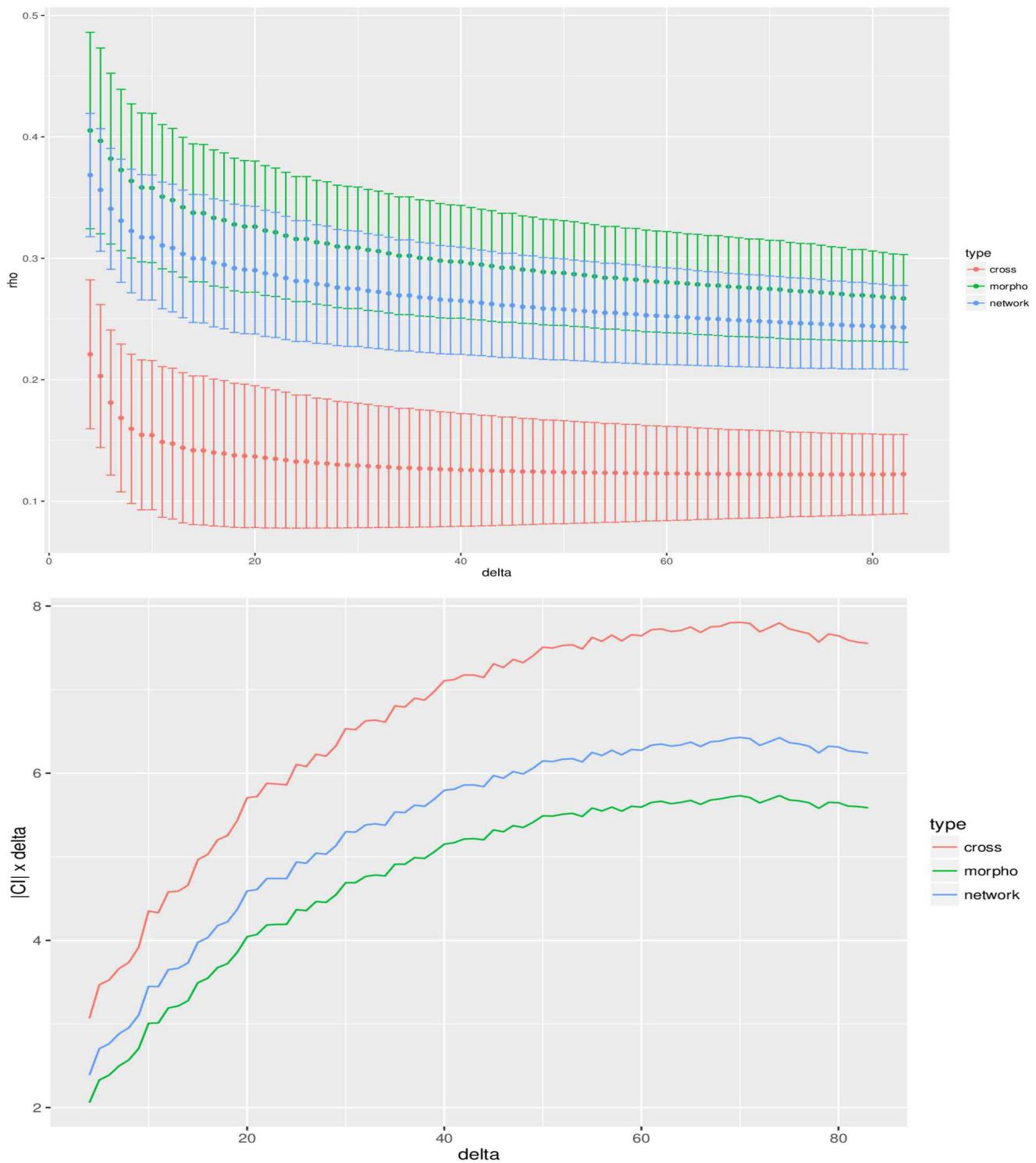


FIGURE 19 : Variation des corrélations avec l'échelle, pour les corrélations calculées sur l'Europe. (Haut) Correlations absolues moyennes et leur déviation standard, pour les différents blocs, en fonction de δ ; (Bas) Taille de l'intervalle de confiance normalisée $\delta \cdot |\rho_+ - \rho_-|$ (IC estimé par méthode de Fisher) en fonction de δ

Echelles optimales

Nous proposons d'autre part d'explorer la propriété éventuelle de multi-scalarité par extraction d'échelles endogène présentes dans les données. Une Analyse en Composantes Principales Géographique Pondérée (GWRPCA) [harris2011geographically] exploratoire suggère des poids et importances variables dans l'espace, ce qui est cohérent avec la non-stationnarité des structures de corrélation obtenue ci-dessus. Il n'y a a priori pas de raison pour que les échelles de variation des différents indicateurs soient strictement identiques. Nous proposons donc d'extraire les échelles typiques pour les relations croisées entre forme urbaine et forme de réseau.

Nous implémentons pour cela la méthode suivante : nous considérons un échantillon typique d'indicateurs (quatre pour chaque aspect, voir la liste en Table 9), et pour chaque indicateur nous formulons l'ensemble des modèles linéaires possibles en fonction des indicateurs opposés (réseau pour un indicateur morphologique, morphologique pour un indicateur de réseau), visant à capturer directement l'interaction sans contrôle sur le type de forme ou de réseau. Ces modèles sont alors ajustés par une Régression Géographique Pondérée (GWR) à portée optimale déterminée par critère d'information corrigé (AICc)¹¹. Pour chaque indicateur, on retient le modèle ayant la meilleure valeur du critère d'information. Nous ajustons les modèles sur les données de la France, avec un noyau *bisquare* et une portée adaptable en nombre de voisins.

Les résultats sont présentés en Table 9. Il est intéressant de noter dans un premier temps que l'ensemble des modèles ne comprend qu'une seule variable, suggérant des correspondances relativement directes entre topologie et morphologie. L'ensemble des indicateurs morphologiques est expliqué par la performance du réseau, c'est à dire la quantité de détour qu'il comprend. Au contraire, la topologie est expliquée par le Moran pour les centralités, et par l'entropie pour la performance et le nombre de sommets. On a ainsi une dissymétrie des relations, le réseau étant conditionné de manière plus complexe à la morphologie que la morphologie au réseau. Les ajustements sont bons ($R^2 > 0.5$) pour une majorité d'indicateurs, et les *p-values* obtenues pour l'ensemble des modèles (constante et coefficient) sont inférieures à 10^{-3} . Les échelles optimales sont quant à elles très localisées, de l'ordre de la dizaine de kilomètres, c'est à dire une plus grande variation que celle obtenue par les corrélations. Cette analyse confirme ainsi statistiquement d'une part la non-stationnarité, et d'autre part donne un point de vue complémentaire sur la question des échelles endogènes.

¹¹ En utilisant le package R GWModel [gollini2013gwmodel].

TABLE 9 : Relation croisées entre indicateurs de réseau et morphologiques. Chaque relation est ajustée par une Régression Géographique Pondérée, pour la portée optimale ajustée par AICc.

Indicateur	Modèle	Portée (km)	Ajustement (R^2)
Distance moyenne \bar{d}	$\bar{d} \sim v_0$	11.609078	0.3145114
Entropie \mathcal{E}	$\mathcal{E} \sim v_0$	8.795393	0.7511295
Moran I	$I \sim v_0$	8.795393	0.4886164
Hiérarchie γ	$\gamma \sim v_0$	8.795393	0.6840567
<i>Betwenness</i> moyenne \bar{bw}	$\bar{bw} \sim I$	12.294076	0.5776019
<i>Closeness</i> moyenne \bar{cl}	$\bar{cl} \sim I$	13.883025	0.2565047
Performance v_0	$v_0 \sim \mathcal{E}$	8.595512	0.8640510
Nombre de noeuds $ V $	$ V \sim \mathcal{E}$	8.595512	0.8825783

Développements

Nous avons montré ainsi montré empiriquement la non-stationnarité des interactions entre morphologie de la distribution des population et topologie du réseau routier. Divers développement de cette analyse sont possibles.

Des grilles de densité de population existent pour l'ensemble des régions du monde, comme par exemple celles fournies par [10.1371/journal.pone.0107042]. L'analyse peut être répétée pour d'autres régions, pour comparer les régimes de corrélations et tester si les propriétés des systèmes urbains restent les mêmes, en gardant à l'esprit les difficultés liées aux différences de qualité dans les données. On peut s'attendre à des régimes très différents pour les Etats-Unis en comparaison à l'Europe par exemple [bretagnolle2010comparer], mais la différence se devrait d'être étudiée quantitativement, pour par exemple dégager des caractéristiques différentes d'interactions entre réseaux et territoires.

La recherche d'échelles locales, c'est à dire avec une fenêtre d'estimation adaptative en taille et forme pour les corrélations, permettrait de mieux comprendre la façon dont les processus influent localement sur leur voisinage. Le critère de validation de la taille de la fenêtre resterait à déterminer : il peut s'agir comme ci-dessus de portée optimale pour des modèles explicatifs ajustés localement.

La question de l'ergodicité devrait également être explorée sur des bases dynamiques, en comparant les échelles de temps et d'espace d'évolution des processus, ou plus précisément les corrélations entre les variations dans le temps et celles dans l'espace, mais la question de l'existence de bases de données assez fines dans le temps paraît problématique. L'étude d'un lien entre la dérivée de la corrélation en fonction de la taille de la fenêtre et les dérivées des processus est

¹² Disponibles à <http://www.worldpop.org.uk/>. La variabilité potentielle de la qualité des données selon les zones doit toutefois amener une prudence dans leur utilisation.

également une piste pour obtenir des informations indirectes sur la dynamique à partir des données statiques.

Enfin, la recherche de classes de processus sur lesquels il est possible d'établir directement la relation entre corrélations spatiales et corrélations temporelles, est une direction possible de recherche. Celle-ci est hors de portée de ce présent travail, mais ouvrirait des perspectives pertinentes sur la co-évolution, puisque celle-ci implique évolution dans le temps et en isolation dans l'espace, et donc une relation complexe entre covariances spatiales et temporelles.

★ ★

★

4.2 CAUSALITÉS SPATIO-TEMPORELLES

Cette section contribue à la compréhension des processus spatio-temporels fortement couplés, en proposant une méthode générique basée sur la causalité de Granger, qui est une méthode introduite en économie pour caractériser des possibles relations causales à partir de relations de corrélations entre variables retardées. Notre méthode est validée par l'identification robuste de régimes de causalité et de leur diagramme de phase pour un modèle de morphogenèse urbaine couplant croissance du réseau et de la densité. L'application au cas réel de l'Afrique du Sud démontre des interactions qui changent dans le temps, témoins des événements historiques entre les dynamiques démographiques territoriales et la croissance du réseau.

Il existe dans la littérature un petit nombre d'exemple d'utilisation de statistiques spatiales sur les relations dynamiques entre réseaux et territoires, c'est-à-dire cherchant à exhiber des relations de causalité entre les deux. Par exemple, [levinson2008density] explique pour Londres les variations de population et de connectivité au réseau par ces mêmes variables décalées dans le temps, démontrant des effets causaux réciproques. [doi:10.1068/b39089] utilise des techniques similaires sur une région d'Italie sur des données historiques sur le temps long, mais modère les conclusions en rappelant l'importance des événements historiques sur les relations estimées. [cuthbert2005empirical] procède à des estimations économétriques des influences réciproques, et conclut que dans le cas d'étude (au Canada à une échelle régionale) le développement du réseau induit le développement de l'usage du sol, mais pas l'inverse. L'échelle de temps et d'espace devrait logiquement être responsable de cette non-circularité. [koning:hal-00962384] procède à une analyse économétrique de la relation entre existence d'une desserte TGV et variables économiques sur les unités urbaines Françaises, et conclut à un effet en propre négatif pour la desserte, après contrôle de l'endogénéité de la desserte par un modèle de sélection, et un effet significatif des caractéristiques propres des unités urbaines : par exemple, pour les unités urbaines desservies par TGV hors LGV, l'effet de la desserte est de -1% sur les emplois entre 1982 et 2006. Cette étude reste cependant limitée car non spatialisée et ne prenant en compte un décalage d'une unité de temps seulement. [MANC:MANC1073] montre sur le temps long un lien de causalité entre stock d'infrastructure et croissance économique sur un panel mondial, mais que ces effets sont atténués localement par des sous ou sur-investissements : dans ce cas, des effets macroéconomiques sont révélés.

4.2.1 Causalités Spatio-temporelles

L'étude des processus spatio-temporels fortement couplés implique la prise en compte d'intrications entre ceux-ci généralement difficiles à isoler. Essence même des approches par la complexité, ces interactions qui sont à l'origine du comportement émergent d'un système font sens comme objet d'étude en lui-même, et une séparation des processus paraît alors contradictoire avec une vision intégrée du système. Dans le cas des systèmes territoriaux, l'exemple des interactions entre réseaux de transport et territoires est une bonne instantiation de ce phénomène, comme le montre le débat sur les effets structurant développé en 1. Le débat est toujours d'actualité puisque la question se pose toujours par exemple pour la construction de lignes à grande vitesse [[crozethalshs01094554](#)]. La réalité des processus territoriaux est en fait bien plus compliquée qu'une simple relation causale entre la mise en place d'une infrastructure et les retombées sur le développement local, mais correspond au contraire à une *co-évolution* complexe [[bretagnolle00459720](#)]. Sur le temps long et à petite échelle, certains effets de renforcement des dynamiques dans les systèmes de villes par l'insertion dans les réseaux, ont été mis en valeur par l'application de la Théorie Evolutive des Villes [[espacegeo2014effets](#)], montrant que la mise en évidence de régularités est toutefois possible dans certains cas par une compréhension plus globale du système. A une autre échelle, toujours concernant les relations entre réseaux et territoires, on peut citer les liens entre pratiques de mobilité, également urbain et localisation des ressources dans un cadre métropolitain qui s'avèrent tout autant complexes : [[cerqueira2017inegalites](#)] montre par exemple une forte correspondance entre conditionnement des pratiques de mobilité par l'accessibilité et classe socio-professionnelle. Ce type de problématique est bien sûr présent dans d'autres domaines : en économie, l'exemple des liens entre innovation, impacts locaux de la connaissance et agrégation des agents économiques est une illustration typiques de processus économiques spatio-temporels présentant des causalités circulaires difficiles à démêler [[audretsch1996r](#)]. Des méthodes spécifiques sont introduites, comme l'utilisation d'instruments statistiques comme par [[aghion2015innovation](#)] dans lequel l'origine géographique des membres du Bureau du Congrès américain attribuant les subventions locales est une bonne variable instrumentale pour lier caractère innovant et inégalités des plus hauts salaires, et permet de montrer que la corrélation significative entre les deux est en fait une causalité de l'innovation sur les inégalités¹³.

¹³ Cet exemple est important sur le plan méthodologique, mais pas seulement puisqu'il se lie en filigrane au thème de la diffusion de l'innovation qui est crucial dans la Théorie Évolutive.

Causalité en géographie

Le couplage fort spatio-temporel implique généralement l'introduction de la notion de causalité, à laquelle la géographie s'est toujours intéressée : [loi1985étude] montre que les questions fondamentales que se pose la géographie théorique récente (isolation des objects, lien entre espace et structures causales, etc.) étaient déjà présentes dans la géographie classique de VIDAL. [claval1985causalite] critique d'ailleurs les nouveaux déterminismes ayant émergé, notamment celui proposé par certains tenants de l'analyse systémique¹⁴ : dans ses débuts, cette approche héritait de la cybernétique et donc d'une vision réductionniste impliquant un déterminisme même dans une formulation probabiliste. CLAVAL note que des travaux contemporains à son écriture (l'école de Prigogine et la Théorie des Catastrophes de Thom) devraient permettre de capturer la complexité qui fait la particularité des décisions humaines. Ce point de vue a anticipé les développements antérieurs, puisque comme le rappelle [pumain2003approche], le glissement de l'analyse des systèmes à l'auto-organisation puis à la complexité a été long et progressif, et ces travaux ont été fondamentaux pour le permettre. FRANÇOIS DURAND-DASTÈS résume cette situation plus récemment dans [durand2003geographes], en appuyant l'importance des bifurcations et de la dépendance au chemin lors des instants initiaux de la constitution du système qu'il désigne par *systèmogenèse*¹⁵. Ce type de dynamique complexe implique généralement une co-évolution des composantes du système, qu'on peut interpréter comme des causalités circulaires entre processus : la question de pouvoir les identifier est donc cruciale au regard de la notion de causalité pour la géographie complexe contemporaine. Cette vision d'une causalité complexe [morin1976méthode] peut être aussi mise en perspective avec le concept de *causalité cumulative* en économie [skott1995cumulative], qui insiste sur le rôle de la dépendance au chemin et la possibilité pour de petites perturbations de causer des effets conséquents par rétroaction négative : il est alors impossible de séparer les effets des causes dans les perturbations infinitésimales.

Identification de causalités

L'identification opérationnelle des causalités peut prendre des formes très diverses, dans différents domaines. Celle-ci dépendra des définitions utilisées, de la même manière que les méthodes à disposition pour lesquelles nous pouvons donner quelques illustrations, en essayant de s'intéresser à des champs divers pour mettre en valeur les différents enjeux et possibilités méthodologiques. [goudet2017learning] utilise des réseaux de neurones pour inférer des relations de causalité

¹⁴ Voir [chamussy1984dynamique] pour un exemple de modèle à but de planification se plaçant dans ce courant.

¹⁵ Cette notion peut être rapprochée de celle de *morphogenèse* que nous approfondissons en Chapitre 5.

entre variables au sens des probabilités conditionnelles. [**liu2011discovering**] propose la détection de relations spatio-temporelles entre perturbations des flux de trafic, introduisant une définition particulière de la causalité basée sur une correspondance de points extrêmes. Les algorithmes associés sont toutefois spécifiques et difficilement applicables à des types de systèmes différents. L'utilisation des corrélations spatio-temporelles a été démontrée comme ayant dans certains cas un fort pouvoir prédictif pour les flots de traffic [**min2011real**]. Également dans le domaine des transports et de l'usage du sol, [**xie2009streetcars**] applique une analyse par causalité de Granger, qu'on pourra interpréter comme une corrélation retardée, pour montrer dans un cas particulier que la croissance du réseau induit le développement urbain et est elle-même tirée par des externalités comme les habitudes de mobilité.

Les neurosciences ont développé de nombreuses méthodes répondant à des problématiques similaires. [**luo2013spatio**] définit une causalité de Granger généralisée prenant en compte la non-stationnarité et s'appliquant à des régions abstraites issues d'imagerie fonctionnelle. Ce genre de méthode est également développée en Vision par Ordinateur, comme l'illustre [**ke2007spatio**] qui exploite les corrélations spatio-temporelles de formes et de flux dans des successions d'images pour classifier et reconnaître des actions. Les applications peuvent être très concrètes comme la compression de fichiers vidéos par extrapolation des vecteurs de mouvement [**chalidabhongse1997fast**]. Dans l'ensemble de ces cas, l'étude des correlations spatio-temporelles rejoint les notions faibles de causalité vues précédemment.

Nous cherchons ici à explorer la possibilité d'une méthode analogue pour des données spatio-temporelles présentant a priori des causalités circulaires complexes, et donc de tenter l'exercice d'équilibriste de concilier un certain niveau de simplicité et de caractère opérationnel à une prise en compte de la complexité. Nous introduisons ainsi une méthode d'analyse des corrélations spatio-temporelles similaire à une causalité de Granger estimée dans le temps et l'espace, dont la robustesse est démontrée systématiquement par l'application à un modèle de simulation complexe de morphogenèse urbaine et par l'isolation de régimes de causalités distincts dans l'espace des phases du modèle. Notre contribution inclut également l'application à un cas d'étude empirique, ce qui la positionne à l'interface des domaines de la méthodologie, de la modélisation et de l'empirique.

La suite de cette section est organisée de la façon suivante : le cadre générique de la méthode proposée est décrit. Nous l'appliquons ensuite à un jeu de données synthétiques afin de la valider partiellement et de tester ses potentialités, ce qui permet de l'appliquer ensuite au système urbain sud-Africain sur le temps long. Nous discutons finalement la proximité avec d'autres méthodes existantes et des développements possibles.

Méthode

Nous formalisons ici de manière générique la méthode, basée sur un test similaire à la causalité de Granger¹⁶, pour tenter d'identifier des relations causales dans des systèmes spatiaux. Soit $X_j(\vec{x}, t)$ des processus aléatoires spatiaux unidimensionnels, se réalisant dans le temps et l'espace. On se donne un ensemble d'unités spatiales fondamentales (u_i) qui peuvent être par exemple les cellules d'un d'une image raster ou un pavage quelconque de l'espace géographique. On suppose l'existence de fonctions $\Phi_{i,j}$ permettant de faire correspondre les réalisations de chaque composante aux unités spatiales, possiblement par une première agrégation locale. Une réalisation d'un système est donnée par un ensemble de trajectoires pour chaque processus $x_{i,j,t}$, et on pourra noter un ensemble de réalisations $x_{i,j,t}^{(k)}$ (accessibles dans le cas d'un modèle de simulation par exemple, ou par hypothèse de comparabilité de sous-systèmes territoriaux dans des cas réels). On suppose disposer d'un estimateur de corrélation $\hat{\rho}$ s'exerçant dans le temps, l'espace et les répétitions, c'est-à-dire que la covariance est estimée par

$$\hat{\text{Cov}}[X, Y] = \hat{\mathbb{E}}_{i,t,k}[XY] - \hat{\mathbb{E}}_{i,t,k}[X]\hat{\mathbb{E}}_{i,t,k}[Y]$$

Il est important de noter ici l'hypothèse de stationnarité spatiale et temporelle, qui peut toutefois aisément être relaxée dans le cas d'une stationnarité locale. D'autre part, l'autocorrelation spatiale n'est pas explicitement incluse, mais est prise en compte soit par l'agrégation initiale si l'échelle caractéristique des unités est plus grande que celle des effets de voisinage, soit par un estimateur spatial adéquat (statistiques spatiales pondérées de type *GWR*¹⁷ [brunsdon1998geographically] par exemple). Cela nous permet de définir la corrélation retardée entre les composantes X_{j_1} et X_{j_2} pour le délai τ par

$$\rho_\tau[X_{j_1}, X_{j_2}] = \hat{\rho}\left[x_{i,j_1,t-\tau}^{(k)}, x_{i,j_2,t}^{(k)}\right] \quad (3)$$

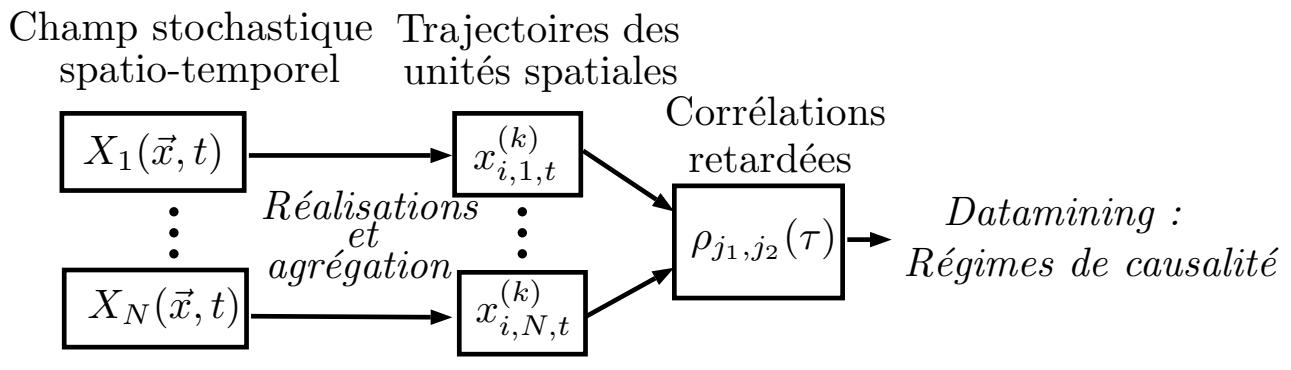
La corrélation retardée n'est pas directement symétrique, mais on a de manière évidente $\rho_\tau[X_{j_1}, X_{j_2}] = \rho_{-\tau}[X_{j_2}, X_{j_1}]$. On applique alors cette mesure de manière simple : si $\text{argmax}_\tau \rho_\tau[X_{j_1}, X_{j_2}]$ ou $\text{argmin}_\tau \rho_\tau[X_{j_1}, X_{j_2}]$ sont "clairement définis" (les deux pouvant l'être simultanément), leur signe donnera alors le sens de la causalité entre les composantes j_1 et j_2 et leur valeur absolue le retard de propagation.

¹⁶ On rappelle que la causalité de Granger correspond à l'existence d'une relation significative entre les composantes retardées dans le temps d'un vecteur et celui-ci.

¹⁷ On rappelle que la Régression Géographique Pondérée consiste à estimer des modèles statistiques à différents endroits de l'espace, en pondérant les informations par la distance, c'est à dire en d'autres termes de prendre en compte la non-stationnarité spatiale.

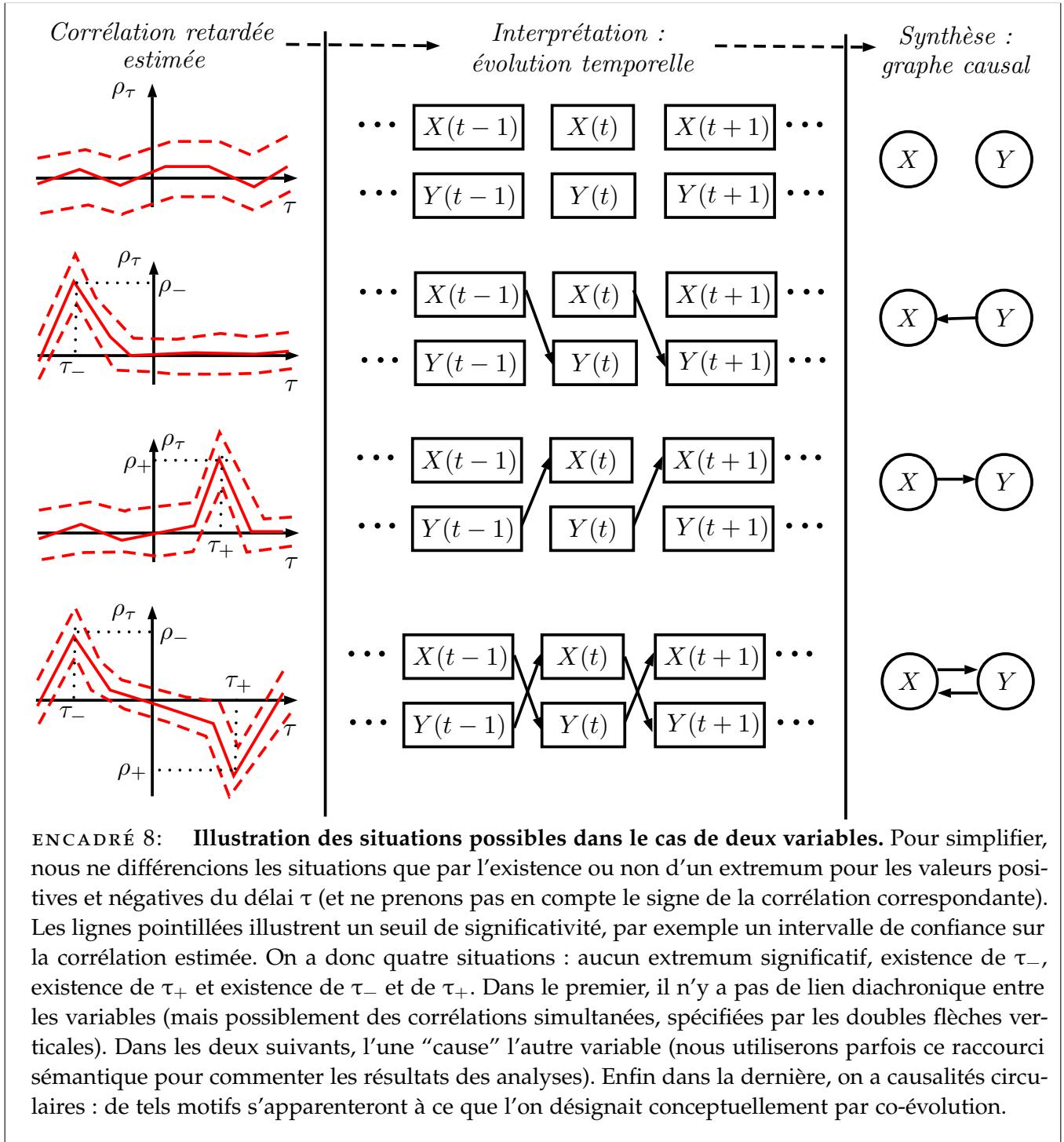
Les critères de significativité dépendront du cas d'application et de l'estimateur utilisé. Ils peuvent prendre en compte différents aspects de la robustesse de l'estimation. Par exemple, un filtrage sur la significativité du test statistique (test de Fisher dans le cas d'un estimateur de Pearson) permet de s'assurer d'isoler des relations qui sont statistiquement significatives. On peut aussi vouloir s'assurer de la significativité d'une corrélation minimale, et regarder la position des bornes d'un intervalle de confiance à un niveau donné. Enfin, on peut aussi fixer un seuil exogène θ sur $|\rho_\tau|$ pour forcer un certain degré de corrélation.

Pour résumer la structure de la méthode et l'enchaînement des traitements effectués, nous proposons le schéma dans l'encadré 7 ci-dessous. La méthode que nous proposons n'est pas nouvelle dans les éléments utilisés, mais l'enchaînement des différentes étapes est originale.



ENCADRÉ 7: Structure de la méthodologie. Nous partons d'un champ stochastique dans le temps et l'espace $X_j(\vec{x}, t)$. Un certain nombre de ses réalisations sont capturées, et mesurées sur des unités spatiales. Nous obtenons des trajectoires k par unité i dans le temps t , notées $x_{i,j,t}^{(k)}$, sur lesquelles la matrice des corrélations retardées $\rho_{j_1,j_2}(\tau)$ est estimée. La fouille de données sur celles-ci permet d'établir différents *régimes de causalité*.

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 8 synthétise de manière stylisée des situations idéales 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 $\tau > 0$. Les quatre possibilités sont illustrées et on représente les interactions entre variable sous forme graphique, dans le temps et de manière synthétique.



ENCADRÉ 8: Illustration des situations possibles dans le cas de deux variables. Pour simplifier, nous ne différencions les situations que par l'existence ou non d'un extremum pour les valeurs positives et négatives du délai τ (et ne prenons pas en compte le signe de la corrélation correspondante). Les lignes pointillées illustrent un seuil de significativité, par exemple un intervalle de confiance sur la corrélation estimée. On a donc quatre situations : aucun extremum significatif, existence de τ_- , existence de τ_+ et existence de τ_- et de τ_+ . Dans le premier, il n'y a pas de lien diachronique entre les variables (mais possiblement des corrélations simultanées, spécifiées par les doubles flèches verticales). Dans les deux suivants, l'une "cause" l'autre variable (nous utiliserons parfois ce raccourci sémantique pour commenter les résultats des analyses). Enfin dans la dernière, on a causalités circulaires : de tels motifs s'apparenteront à ce que l'on désignait conceptuellement par co-évolution.

Emergence et mesure de la co-évolution ?

Prenons également un court instant pour clarifier le statut épistématique 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 macroscopique¹⁸. 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.

Est-il alors possible de répondre de manière équivoque à la question “*s'il y a co-évolution dans un cas particulier*”¹⁹? 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 éparses, le problème du MAUP, etc.) seront bien toujours présents, et que la question ci-dessus qui y est naturellement soustraite 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 Données Synthétiques

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.

Séries temporelles auto-régressives

Illustrons les motifs qui peuvent être attendus, notamment ceux stylisés donnés précédemment en Encadré 8, 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, 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

¹⁸ 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

¹⁹ A laquelle nous ajoutons : pour ces composantes, sur cette portée spatiale et temporelle et sur ces échelles spatiale et temporelle.

$\tau \neq \tau_0$ et $\mathbf{A}(\tau_0) = \begin{pmatrix} 0 & a \\ a & 0 \end{pmatrix}$ pour $-1 < a < 1$, le calcul des corrélations théoriques est immédiat (voir Appendice A.6), et on obtient, en notant $\mathbf{X} = (X, Y)$, pour $\tau > 0$

$$\rho[X(t), Y(t-\tau)] = \begin{cases} a^{2k+1} \text{ si } \tau = (2k+1)\tau_0 \\ 0 \text{ sinon} \end{cases}$$

L'expression est la même pour $\tau < 0$ en échangeant X et Y . Ainsi, on contrôle la corrélation retardée au retard voulu et à ses multiples impairs. 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 8.

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-supervisée²⁰ sur les séries temporelles $[\rho(\tau)]_{a_1, a_2}$. Nous montrons en Fig. 20 les profils typique obtenu en correspondance avec leur position dans l'espace des paramètres (a_1, a_2) . On trouve exactement les neuf profils stylisés possibles, en correspondance avec les valeurs relatives des paramètres comme attendu. A partir de profil 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.

Modèle de croissance urbaine

Cette méthode doit être testée et validée plus en profondeur, ce que nous faisons à nouveau sur des données synthétiques, méthode qui permet une connaissance plus fine des comportements des modèles [raimbault2016generation]. En écho à l'exemple des relations entre réseaux de transport et territoires qui a permis d'introduire notre problématique précédemment, nous proposons de générer des configurations urbaines stylisées dans lesquelles réseau et densité s'influencent mutuellement, et pour lesquelles les causalités ne sont pas évidents *a priori* étant donné les paramètres du modèle génératif.

[raimbault2014hybrid] décrit et explore un modèle simple de morphogenèse urbaine²¹ (modèle RBD) répondant parfaitement à ces contraintes.

²⁰ Par algorithme des *k-means* avec $k = 9$ et $b_c = 1000$ répétitions.

²¹ Nous n'explorons pas ici le concept de morphogenèse, qui fera l'objet du Chapitre 5, mais utilisons ce modèle comme producteur de données synthétiques.

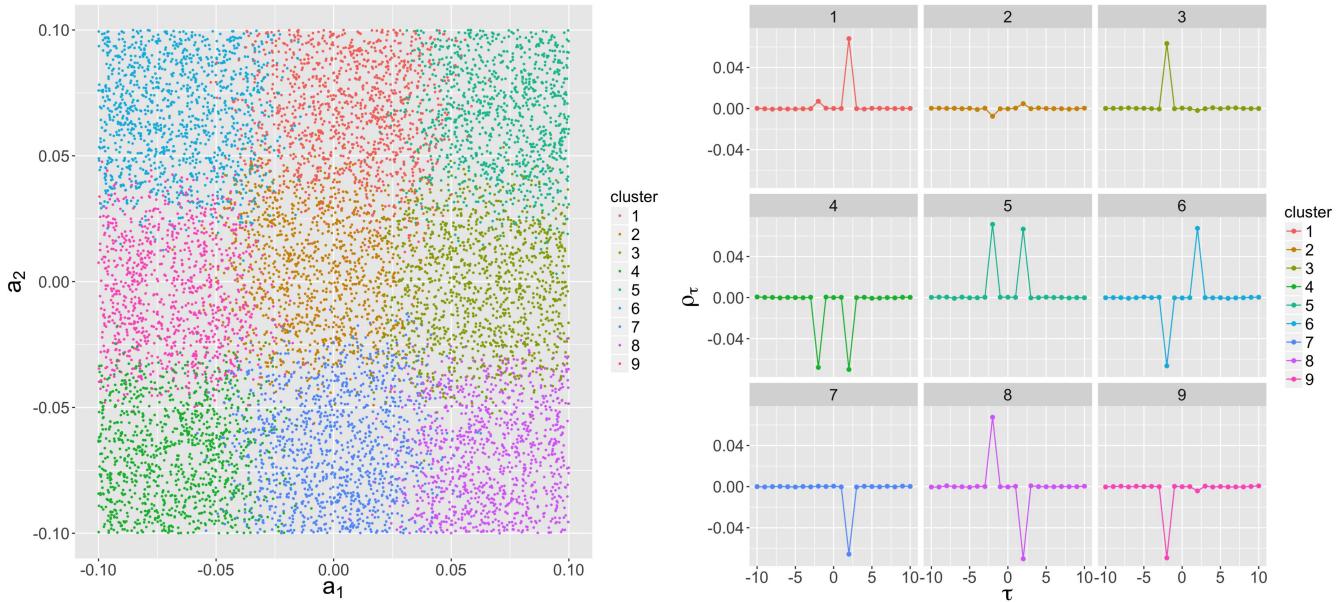


FIGURE 20 : Estimation des régimes de corrélation dans le cas de séries temporelles auto-régressives linéaires.

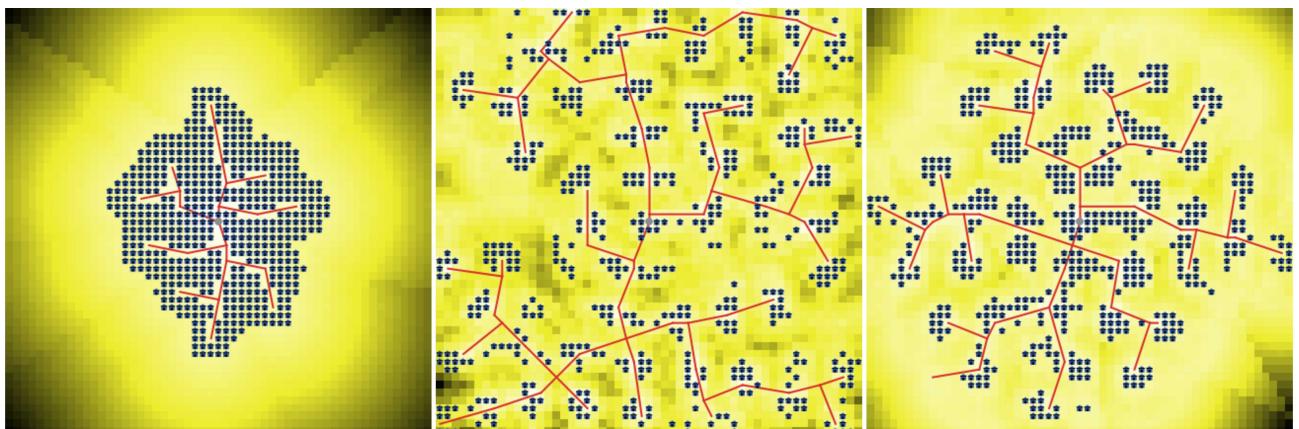
Résultats de la méthode de classification des régimes pour des processus AR simples. Nous simulons $b = 10000$ séries temporelles de longueur $t_f = 10000$, avec des coefficients aléatoires $(a_1, a_2) \in [-0.1, 0.1]$ et un délai $\tau_0 = 2$. (*Gauche*) Valeur des coefficients (a_1, a_2) , la couleur donnant le cluster obtenu; (*Droite*) Trajectoires de centroïdes correspondants. Nous retrouvons les profils stylisés attendus, qui correspondent aux valeurs relatives des paramètres : par exemple, le cluster 1 est pour a_1 faible et a_2 fort, et correspond bien à une situation où ρ_+ existe c'est à dire une configuration $X \rightarrow Y$, et le signe de ρ_+ correspond à $a_2 > 0$,

Ce modèle est décrit en détails pour la configuration dans laquelle nous l'utilisons en Encadré 9. Les variables explicatives de la croissance urbaine, les processus d'extension du réseau et le couplage entre densité urbaine et réseau ne sont pas trop complexes. Cependant, hormis dans des cas extrêmes (par exemple lorsque la distance au centre détermine la valeur foncière uniquement, le réseau dépendra de manière causale de la densité, ou lorsque la distance au réseau seule compte, la causalité sera inversée), les régimes mixtes ne présentent pas de causalités évidentes : c'est donc un parfait cas pour tester si la méthode est capable d'en détecter. Les données synthétiques nous permettent de contrôler la cohérence dans les cas où la relation est attendue.

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 est caractérisée par les variables x_d (densité dans un rayon fixé $r = 5$), x_r (distance à la route la plus proche) et x_c (distance au centre via le réseau). Ces variables permettent de calculer une valeur de potentiel pour chaque cellule $U_i = \sum w_k x_k(i)$, où les w_k sont des paramètres du modèle permettant d'influencer les formes urbaines produites. 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).

ENCADRÉ 9: Description du modèle RBD.

Nous utilisons une implémentation adaptée²² du modèle initial, permettant de capturer les valeurs des variables étudiées pour chaque patch et à chaque pas de temps et de calculer les corrélations retardées entre variables au sein du modèle. Nous explorons une grille de l'espace des paramètres du modèle RBD, faisant varier les paramètres de poids de la densité w_d , de la distance au centre w_c et de la distance au réseau w_r (voir Encadré 9 de description du modèle), dans $[0; 1]$ avec un pas de 0.1. Les autres paramètres sont fixés à leur valeurs par défaut données par [raimbault2014hybrid]. Pour chaque valeur des paramètres, nous procédons à $N = 100$ répétitions ce qui est suffisant pour une bonne convergence des indicateurs. Les explorations sont effectuées via le logiciel OpenMole [reuillon2013openmole], le grand nombre de simulations (1,330,000) nécessitant l'utilisation d'une grille de calcul²³.

Nous calculons sur l'ensemble des cellules les corrélations retardées par estimateur de Pearson non biaisé entre les variations des variables suivantes²⁴ : densité locale, distance au centre et distance au réseau. Il s'agit des variables explicatives pour la dynamique du modèle, et donc celles sur lesquelles on peut identifier des relations dynamiques entre caractéristiques territoriales locales.

La Fig. 21 montre le comportement de ρ_τ pour chaque couple de variables (non dirigé, τ prenant des valeurs négatives et positives), pour les combinaisons des valeurs extrêmes des paramètres. Nous donnons également l'interprétation sous forme de graphe de relations entre variables et une illustration de configuration urbaine générée pour les valeurs de paramètres correspondantes. On peut voir déjà différents régimes émerger : par exemple,

Nous voyons un certain nombre de régimes émerger, et tirer les interprétations synthétiques suivantes :

- Le lien négatif $D \rightarrow R$ résulte du mécanisme simple d'extension du réseau : un accroissement de la densité conduit à une diminution de la distance au réseau par la construction d'une nouvelle route. Certaines configurations inhibent ce lien, par l'interaction complexe avec les autres variables (par exemple $(0, 1, 0)$).
- Des graphes de relations élaborés peuvent émerger : $(1, 0, 1)$ conduit par exemple à une relation circulaire entre distance au réseau et densité, et une causalité de ces deux variables sur la distance au centre.
- Des comportements non attendus a priori émergent, comme par exemple la relation circulaire entre distance au réseau et

²² disponible sur le dépôt ouvert du projet à

<https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Simple/ModelCA>

²³ Les résultats de simulation sont disponibles à <http://dx.doi.org/10.7910/DVN/KGHZZB>.

²⁴ Calculer les corrélations sur les variables directement n'a pas de sens puisque leur valeur n'en a pas en absolu.

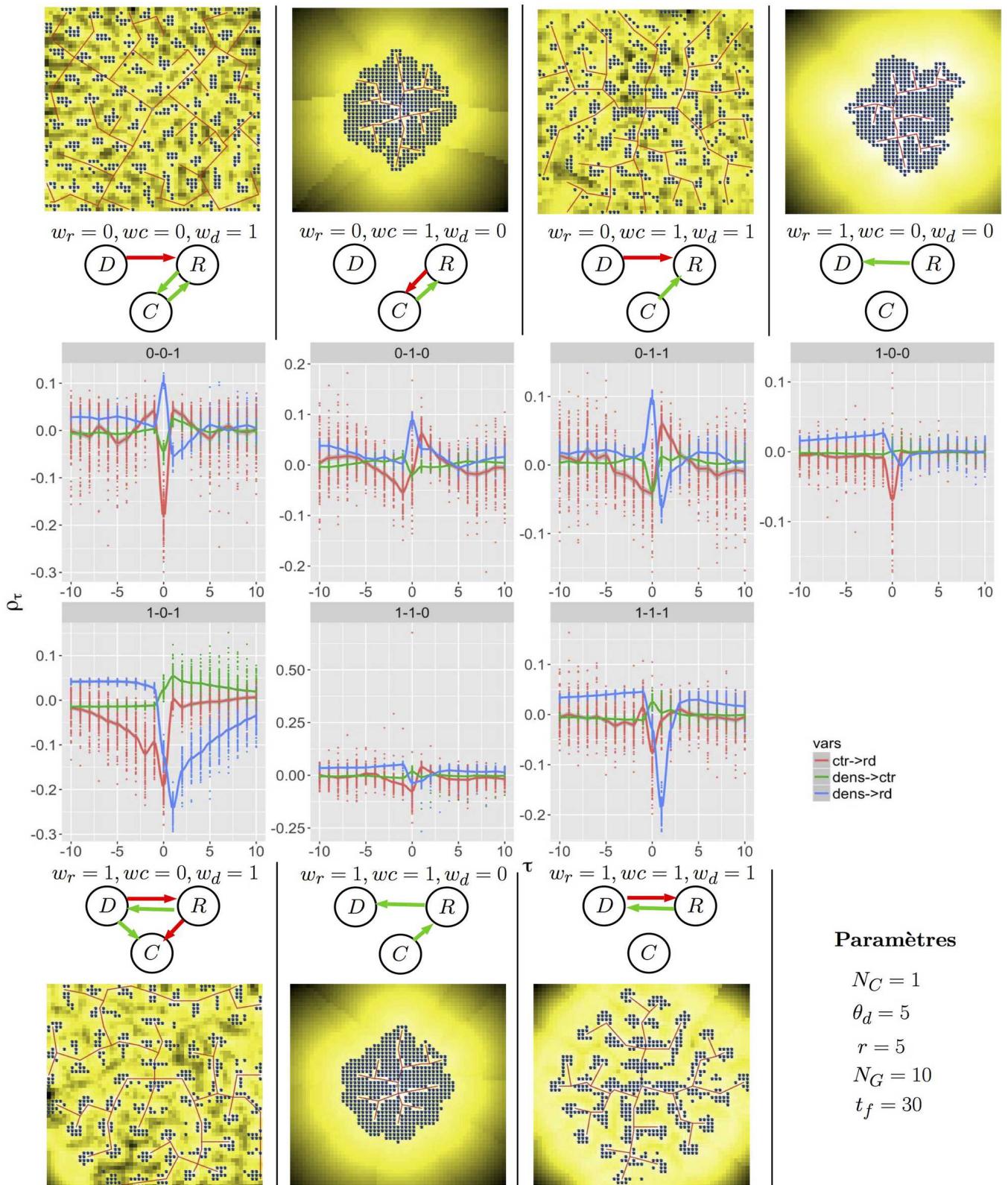


FIGURE 21 : Correlations retardées dans le modèle RDB. Corrélations retardées, pour chaque combinaison des valeurs extrêmes pour l'ensemble des paramètres w_r, w_c, w_d . Chaque graphe donne la corrélation ρ_τ en fonction du retard τ . Les différentes couleurs correspondent à chaque couple de variables : distance au centre (ctr), densité (dens) et distance au réseau (rd). Les points correspondent aux corrélations individuelles pour chaque répétition du modèle (estimateurs sur i et t), tandis que les courbes donnent l'estimateur complet sur l'ensemble des répétitions également. Pour chaque graphe, nous donnons en correspondance directe une configuration finale, et l'interprétation de terme de motifs de causalité sous forme d'un graphe entre les variables. La couleur des liens dirigés donne le signe de la relation (rouge pour corrélation négative, verte pour corrélation positive).

distance au centre pour $(0,0,1)$ où seule la densité joue; au contraire le modèle fait émerger une relation qui correspond au mécanisme microscopique quand $w_r = 1$ seulement.

Régimes de causalité

Nous démontrons à présent qu'il est possible d'établir une typologie endogène des comportements des corrélations retardées. Afin d'étudier ces comportements de manière systématique, nous proposons d'identifier des régimes de manière endogène, en procédant à un apprentissage non-supervisé. Nous appliquons comme précédemment une classification des *k-means*, robuste à la stochasticité (5000 répétitions), avec les points caractéristiques (*features*) suivants : pour chaque couple de variable, $\text{argmax}_\tau \rho_\tau$ et $\text{argmin}_\tau \rho_\tau$ si la valeur correspondante est telle que $\frac{\rho_\tau - \bar{\rho}_\tau}{|\bar{\rho}_\tau|} > \theta$ avec θ paramètre de seuil, 0 sinon. L'inclusion des *features* (variables caractéristiques) supplémentaires des valeurs de ρ_τ n'influence pas significativement les résultats, celles-ci n'ont pas été prises en compte pour réduire la dimension. Le choix du nombre de clusters k est en général épineux dans ce genre de problème [hamerly2003learning], dans notre cas le système possède une structure qui lève l'ambiguité : les courbes de la proportion de variance inter-cluster et de sa dérivée (voir Fig. 82 en Annexe A.6), en fonction de k pour différentes valeurs de θ , présentent une transition pour $\theta = 2$, ce qui donne pour cette courbe une rupture à $k = 5$. Un examen visuel des clusters dans un plan principal confirme la bonne qualité de la classification pour ces valeurs. Une classe correspond alors à un *régime de causalité*, dont nous pouvons représenter le diagramme de phase en fonction des paramètres du modèle, ainsi que les trajectoires des centres des clusters (calculées comme barycentre dans l'espace complet initial) en Fig. 22.

Interprétation

Nous proposons finalement d'interpréter les régimes obtenus, représentés en Fig. 22. Le comportement obtenu est particulièrement intéressant : les régions du diagramme de phase selon les paramètres correspondant aux régimes sont clairement délimitées et connexes. Par exemple, on observe l'émergence d'un régime (numéroté 1) où la densité cause fortement la distance au réseau de manière négative (au sens de l'existence d'un ρ_+ négatif), mais la distance au centre cause la distance au réseau, régime dont l'étendue maximale dans le plan (w_d, w_c) est pour une valeur intermédiaire $w_r = 0.7$. Ainsi, pour maximiser l'impact du réseau sur la densité, il ne faut pas maximiser le poids correspondant mais prendre une valeur intermédiaire, ce qui peut paraître contre-intuitif en premier abord : cela illustre l'intérêt de la méthode dans le cas de relations circulaires difficiles à démêler a priori. Le régime 2, où la distance au réseau influence la

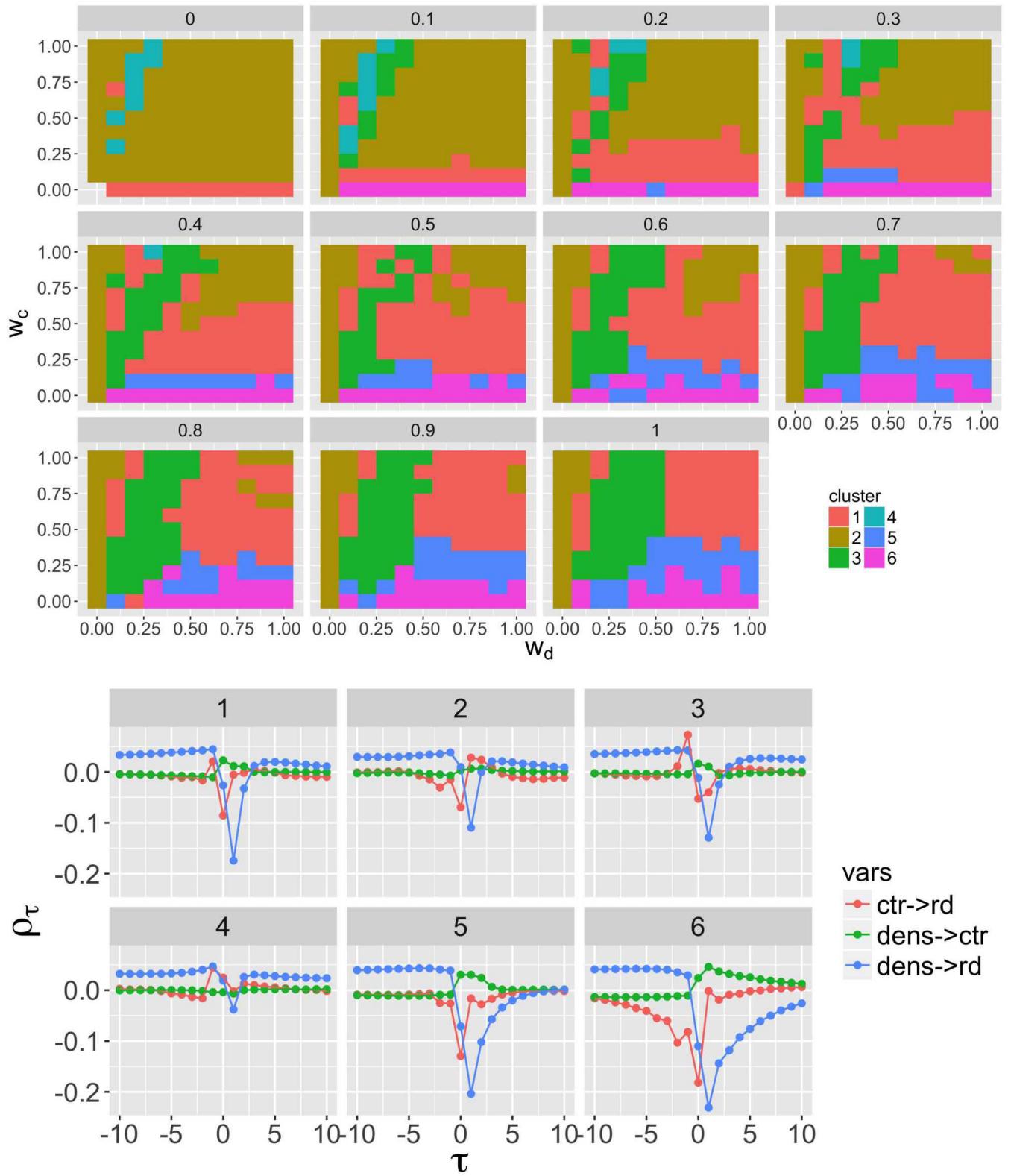


FIGURE 22 : Régimes de causalité identifiés par classification non supervisée. (Haut) Diagramme de phase des régimes (cluster) dans l'espace des paramètres (w_d, w_c, w_r), w_r variant entre les différents sous-diagrammes de (w_d, w_c). (Bas) Trajectoires correspondantes des centroïdes, en termes de profils de corrélations retardées ρ_τ , pour l'ensemble des couples de variables (couleur) et pour chaque cluster.

densité de la même manière, mais la relation entre distance au centre et route est inversée, est prédominant dans les faibles w_r : ainsi, l'atténuation du rôle de la distance à la route conduit le centre à inverser sa relation avec la distance à la route. Le régime 6, extrême quant à sa position dans l'espace des paramètres car dans un voisinage de $w_c = 0$, correspond à une situation isolée dans laquelle la distance au centre n'importe pas comme variable explicative, et on observe une causalité de la densité sur la distance au centre (ρ_+ pour $D \rightarrow R$) et de la distance à la route sur la distance au centre (ρ_- pour $C \rightarrow R$), c'est à dire que c'est aspect est totalement dominé par les autres.

Cette application sur données synthétiques démontre ainsi d'une part la robustesse de la méthode vu la cohérence des régimes obtenus, et constitue aussi une qualification beaucoup plus précise des comportements du modèle que celle réalisée dans l'article initial. Dans ce cas précis, il peut s'agir d'un instrument de connaissance des relations entre réseaux et territoires en lui-même, permettant le test d'hypothèses ou la comparaison de processus dans le modèle stylisé.

4.2.3 Relations Réseaux-territoires en Afrique du Sud

Nous démontrons à présent les potentialités de notre méthode sur des données géo-historiques sur le temps long, pour le cas du réseau ferré en Afrique du Sud au cours du 20ème siècle. En faisant l'hypothèse que les territoires et les réseaux réagissent différemment aux événements historiques, les motifs de causalité devraient informer sur leur relations sur le temps long.

Contexte

Les réseaux de transport peuvent être utilisés comme un puissant outil de contrôle des populations, avec des effets encore plus significatifs lorsque ceux-ci perturbent les relations avec les territoires. Le cas de l'Afrique du Sud est une illustration pertinente, puisque [baffi:tel-01389347] montre que lors de l'apartheid la planification du réseau ferré était utilisée comme un outil de ségrégation raciale par l'établissements de motifs de mobilité et d'accessibilité fortement contraints. En particulier, il est montré qualitativement que les dynamiques entre réseaux et territoires ont profondément changé à la fin de l'apartheid, transformant un outil de ségrégation planifiée (une forme de réseau optimisée pour minimiser une accessibilité non désirée) en un outil d'intégration grâce à des changements récents dans la topologie du réseau. Nous étudions ici les potentielles propriétés *structurelles* de ce processus historique, en se concentrant sur les motifs dynamiques des interactions entre le réseau ferré et la croissance des villes. Plus précisément, nous essayons d'établir si les politiques de planification ségréguantes ont effectivement modifié la trajectoire

du système couplé, ce qui correspondrait à des impacts plus larges et profonds que leurs effets immédiats.

Données

Nous utilisons une base de données complète couvrant l'ensemble du réseau ferré Sud-Africain de 1880 à 2000 avec les dates d'ouverture et de fermeture pour chaque station et liaison, couplée à une base de données pour les villes s'étendant de 1911 à 1991 pour laquelle des ontologies consistantes pour les aires urbaines ont été assurées. Ces bases de données sont décrites par [baffi:tel-01389347], mais ne sont pas ouvertes. Pour respecter notre exigence d'ouverture, nous ne mettons ainsi à disposition que les données agrégées utilisées dans l'analyse.

Mesures de réseau

Une analyse préliminaire consiste à regarder l'évolution dynamique des mesures de réseau, celles-ci pouvant témoigner de ruptures dans les propriétés structurelles du réseau et donc de mutations historiques profondes. L'évolution de certaines propriétés du réseau, comme les distributions de la centralité ou de l'accessibilité, peut témoigner de l'existence d'une planification les ayant influencées. Nous montrons en Figure 23 l'évolution des mesures de réseau dans le temps²⁵, correspondant aux mesures les plus basiques de celles définies en 4.1. La centralité de proximité, que nous définissons comme le temps moyen de trajet vers les autres noeuds, présente un comportement intéressant. En effet, la taille du réseau et les valeurs moyennes des centralités présentent un comportement concordant, qui correspond à l'expansion initiale du réseau. Par contre, la tendance de la hiérarchie de la centralité de proximité à se réduire est soudainement rompue à la date correspondant à l'officialisation des politiques ségrégatives en 1951, alors que taille et forme géométrique globale du réseau, traduite par l'efficience, restent constants. Ainsi, dans le meilleur des cas la planification après cette date est une coïncidence avec la variation de cette propriété. Dans le pire des cas elle est en effet responsable de cette rupture de tendance, c'est à dire a eu les effets escomptés sur l'accessibilité, dans le but d'empêcher la diminution de la ségrégation, puisque plus la hiérarchie est faible plus le réseau est égalitaire²⁶.

²⁵ Globalement, [baffi:tel-01389347] (p. 154) montre que le réseau ferré sud-africain s'est développé en arborescence et que les politiques de ségrégation ont figé sa structure, l'empêchant de mailler le territoire.

²⁶ La politique de ségrégation est interprétée par [baffi:tel-01389347] (p. 189) comme une intention de "connecter sans connectivité", conjointement avec les migrations forcées de la population ségrégée dans des zones spécifiques à l'écart des fonctions urbaines, appelées *bantustans*. Le réseau a alors été spécifiquement développé pour relier ceux-ci aux zones de production sans les connecter efficacement aux centres urbains.

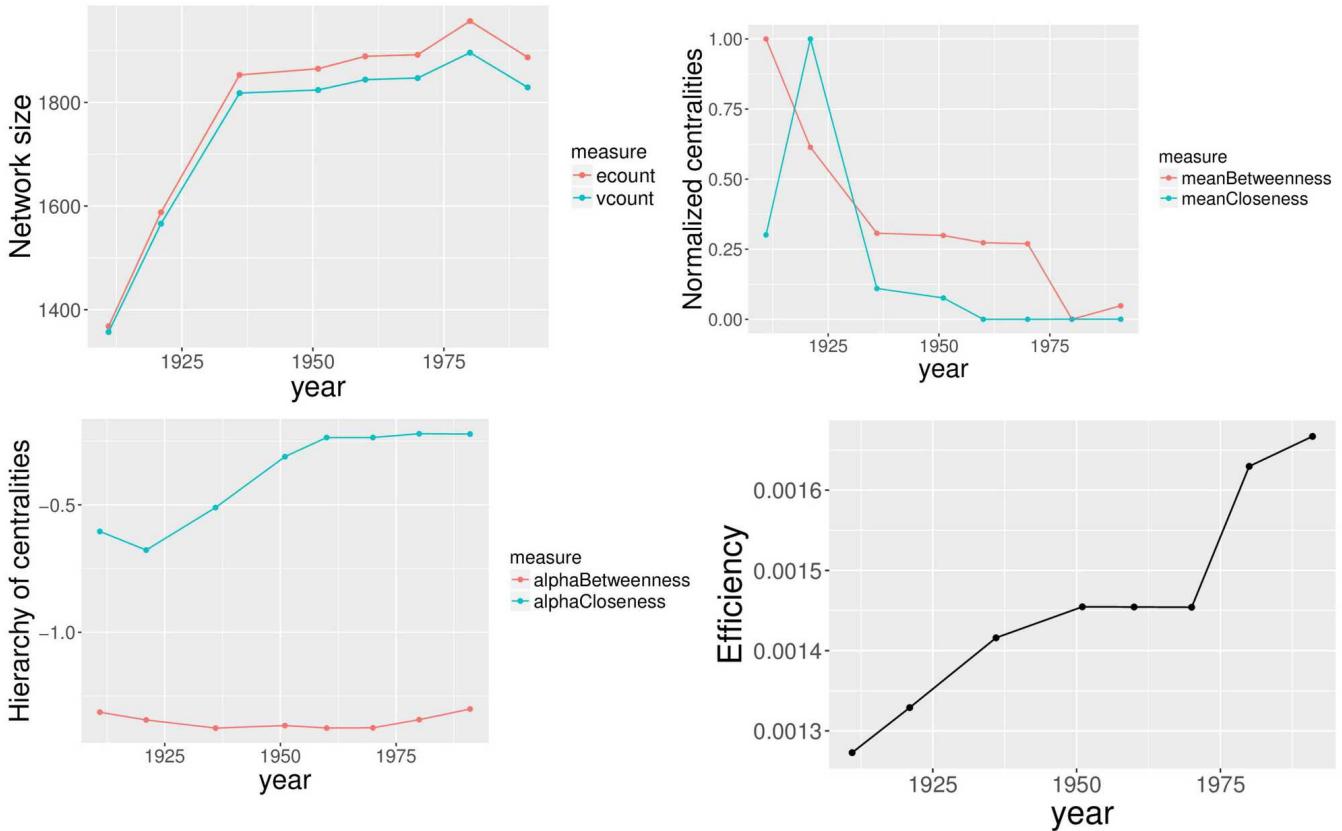


FIGURE 23 : Evolution des mesures de réseau. On calcule pour l'ensemble des dates les mesures basiques de réseau : taille, centralités résumées par leur hiérarchie et leur moyenne, efficience. Les centralités sont normalisées pour comparaison de leur variation respective ($\max \bar{bw} = 0.07$, $\max \bar{cl} = 1.5e-4$).

Motifs de causalité

Nous examinons à présent les interactions dynamiques entre le réseau ferré et la croissance urbaine. Pour cela, nous appliquons la méthode développée dans la première partie, qui consiste à l'étude des causalités de Granger, au sens large des corrélations entre les variables retardées, estimées entre les taux de croissance des villes et les différentiels d'accessibilité dus à la croissance du réseau, pour toutes les villes ou aires urbaines ayant une connection au réseau. Nous testons à la fois l'accessibilité en terme de distance et pondérée par la population à l'origine et aux deux extrémités. Si P_i sont les populations, d_{ij} la matrice de distance dans le réseau, l'accessibilité de i sera donnée par $Z_i = w_i \sum_j w_j \exp(-d_{ij}/d_0)$ où d_0 est le paramètre de décroissance et les poids w_i sont $1/N$ ou $P_i / \sum_j P_j$ selon la modalité. Nous faisons varier les valeurs de d_0 pour prendre en compte les relations à différentes échelles spatiales. De plus les corrélations retardées sont estimées sur des fenêtres temporelles de taille variable T_W , pour tester différentes échelles de stationnarité temporelles potentielles.

Les résultats des estimations sont montrés en Figure 24. Nous obtenons des résultats significatifs avec l'accessibilité non-pondérée seulement, que nous montrons ici²⁷. Le meilleur compromis pour la fenêtre temporelle apparaît être une trentaine d'année, si on cherche à avoir à la fois un bon nombre de corrélations significatives (définies par $p < 0.1$ pour un test de Fisher) et un niveau moyen élevé de corrélation absolue sur l'ensemble des retards et des paramètres de décroissance. Nous interprétons cette valeur comme approximativement l'échelle de stationnarité du système. De plus, le nombre de corrélations significatives présente clairement une transition de phase dans ses valeurs intermédiaires en fonction de d_0 (Fig. 83), ce qui devrait correspondre au passage entre l'échelle spatiale des aires urbaines et celle du pays, et donne ainsi l'échelle locale de stationnarité spatiale, autour de $d_0 = 500\text{km}$.

Quand on examine le comportement des corrélations retardées pour la distance, on observe des motifs de causalité assez clairs, puisque le sens de la causalité de Granger s'inverse autour de 1950, celle-ci étant à chaque fois marquée par des corrélations allant jusqu'à 0.5 pour certaines valeurs du paramètre de décroissance. On passe ainsi d'une accessibilité causant la croissance de la population avec un délai de 10 à 20 ans avant l'apartheid (1948), à l'opposé, c'est à dire une population induisant les changements d'accessibilité après l'apartheid (avec un délai de 20 ans). Ce résultat est en cohérence avec les relocalisations de population et la conception du réseau en accord avec celles-ci. Nous interprétons ce phénomène comme une *ségrégation structurelle*, c'est à dire un impact significatif des politiques de planification sur les dynamiques des interactions entre les réseaux et les territoires. En effet, on peut interpréter le premier régime comme un effet direct du transport sur les motifs de migration dans un contexte de liberté, en opposition au second régime qui correspondrait à un contrôle de la population et d'une adaptation du réseau en fonction. Ainsi, l'évènement historique a eu un effet au second ordre sur les relations dynamiques. Ces motifs rejoignent les conclusions empiriques obtenues par [baffi:tel-01389347] sur le sujet de l'apartheid, qui montre par exemple un fort effet des mesures sur les déplacements forcés de population, ainsi qu'une baisse de l'accessibilité pour les zones cibles de la ségrégation.

Développements possibles

Une première extension pourra consister en une étude similaire avec des variables socio-économiques plus précises, pour quantifier par

²⁷ Nous donnons en Annexe A.6, Fig. 83 les profils de corrélations retardées estimées pour les accessibilités pondérées à l'origine et à l'origine et à la destination. L'auto-corrélation domine a priori l'accessibilité pondérée : en effet, on a pour les deux variables pondérées des valeurs positives pour les faibles valeurs de d_0 uniquement, les autres n'étant pas significatives.

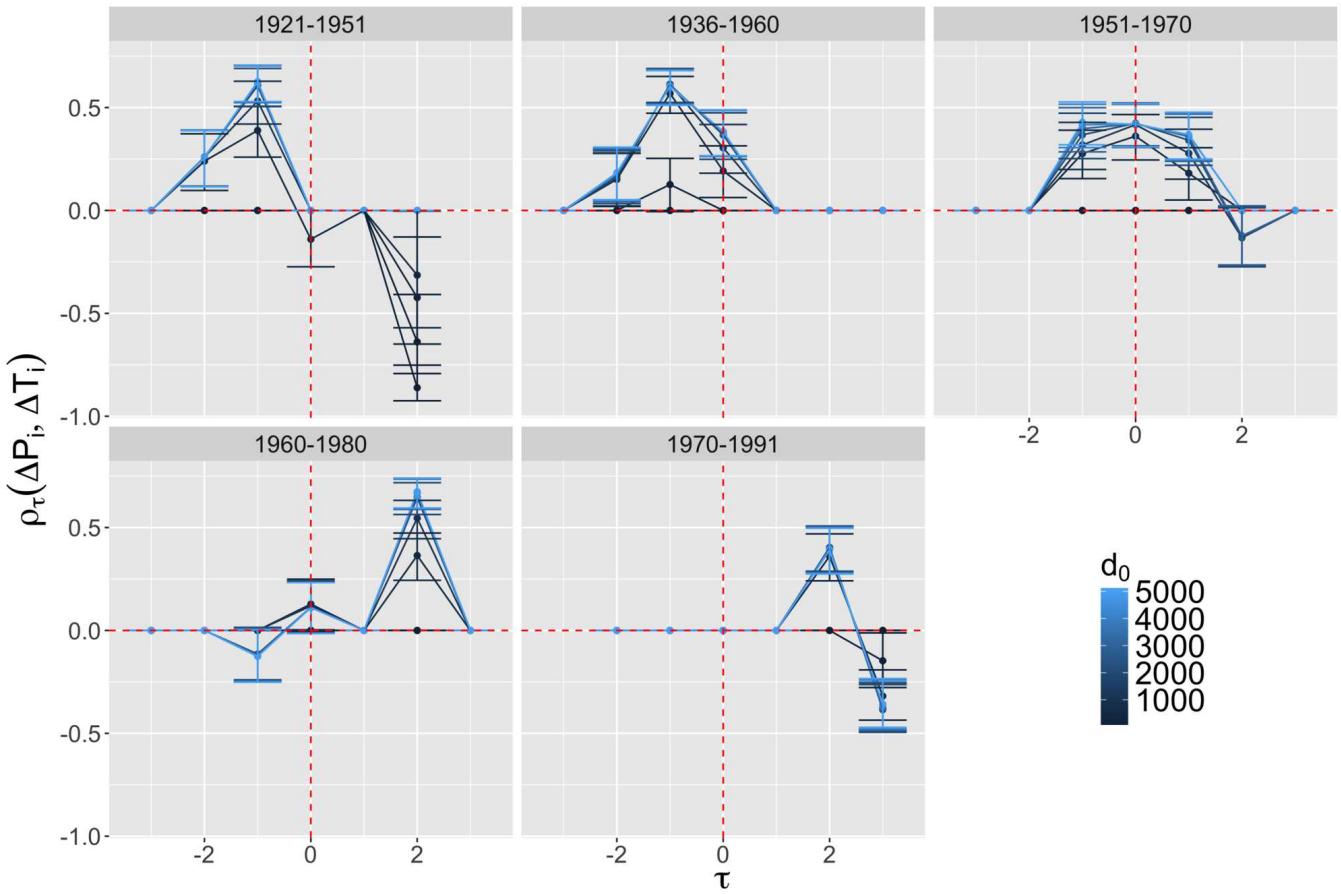


FIGURE 24 : Corrélations retardées. Corrélations retardées en fonction du délai τ , pour la fenêtre temporelle $T_W = 3$, sur les différentes périodes successives (colonnes), et pour d_0 variable (couleur). Pour interpréter, on observe un maximum de la corrélation retardée qui se décale dans le temps passant d'un retard négatif à un retard positif, signifiant une inversion du sens de la causalité.

exemple directement les motifs de ségrégation. D'autre part, des variables qualitatives liées aux évènements historiques pourraient faire office de variable d'instrumentation. La méthode des variables instrumentales [angrist1996identification] est utilisée pour identifier des relations causales entre variables, d'une façon complémentaire à celle que nous avons mis en place. On pourrait chercher à rendre nos conclusions plus robustes, notamment vérifier si les corrélations ne sont pas fortuites, par l'application de cette approche, qui serait cependant difficile à réaliser vu la rareté des données dans notre cas, un bon instrument n'étant par ailleurs pas évident dans ce cas.

★ ★

*

4.3 MODÈLE DE CROISSANCE MACROSCOPIQUE

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 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.

Nous décrivons ainsi un modèle spatial simple de croissance urbaine pour les systèmes de villes à l'échelle macroscopique, qui combine les interactions directes entre les villes et un effet indirect des flux du réseau physique comme moteurs de la croissance de population. Le modèle est paramétré sur les données de population pour le système de villes français entre 1831 et 1999, dont la forte non-stationnarité des motifs de corrélation suggère d'appliquer le modèle sur des fenêtres temporelles locales. Les calibrations correspondantes du modèle par l'utilisation d'algorithmes génétiques fournit l'évolution des processus d'interaction et des effets de réseau dans le temps. De plus, l'amélioration de l'ajustement par l'ajout du module de réseau apparaît comme effectif lorsqu'on contrôle pour les paramètres supplémentaires, ce qui confirme la capacité du modèle à révéler des effets de réseau dans le système de villes.

Dans cette section, nous visons à explorer plus en détail l'hypothèse, centrale à la Théorie Evolutive des Villes de PUMAIN, selon laquelle les interactions spatiales sont des moteurs significatifs de leur croissance. Plus précisément, nous considérons à la fois les interactions abstraites et les interaction avec les flux portés par les réseaux physiques, principalement les réseaux de transport. Nous étendons les modèles existants de manière correspondante. Les apports de cette section consistent en deux points : (i) nous montrons que des modèles d'interaction très basiques basés uniquement sur la population peuvent être ajustés aux données empiriques et que les valeurs ajustées des paramètres sont directement interprétables ; et (ii) nous introduisons une nouvelle méthodologie pour quantifier l'overfitting dans les modèles de simulation, comme une extension de Critères d'Information pour les modèles statistiques, qui appliquée à nos modèles calibrés confirme que l'amélioration du fit n'est pas due seulement aux paramètres supplémentaires et donc artificielle, mais que le modèle étendu capture effectivement plus d'information sur les processus du système. Cela révèlera des effets de réseaux de manière indirecte. Nous précisons d'abord le contexte thématique dans lequel le modèle se placera et revoyons les approches de modélisation de la croissance urbaine basées sur les interactions spatiales.

4.3.1 Modéliser la croissance urbaine par les interactions spatiales

Modèles de croissance urbaine

Replaçons la présente démarche dans un contexte plus global de modélisation de la croissance urbaine²⁸, notion plus générale que notre problématique précise. Elle nous permet ici toutefois de construire un premier modèle du point de vue de la Théorie Evolutive. Une bonne connaissance de la façon dont les villes se différencient, interagissent et croissent est ainsi un sujet pertinent à la fois pour les applications en termes de politiques et d'un point de vue théorique. [pumain2009innovation] suggère que les villes sont l'incubateur du changement social, leur destin étant étroitement lié à celui des sociétés, et donc comme nous l'avons développé en Chapitre 1, de leurs territoires et de leurs réseaux. Diverses disciplines ont étudié des modèles de croissance urbaine avec différents objectifs et prenant en compte des aspects variés. Par exemple, l'économie est toujours prudente à inclure les interactions spatiales dans les modèles [krugman1998space], le prenant en compte de façon très simplifiée même en économie géographique, mais produisant des modèles extrêmement détaillés en termes de processus de marché. Au contraire, la géographie se concentre plus sur les spécificités territoriales et les interactions dans l'espace mais produira des conclusions générales avec plus de difficulté [marchionni2004geographical]. L'exemple de ces deux disciplines montre comment il est difficile de créer des ponts, comme il a fallu des efforts exceptionnels pour effectuer des traductions de l'une à l'autre (comme P. HALL le fit avec le travail de VON THÜNEN [taylor2016polymath]), et ainsi comment il est loin d'être évident de capturer la complexité des systèmes urbains de manière intégrée.

Le modèle le plus simple pour expliquer²⁹ la croissance urbaine, le modèle de Gibrat, qui suppose des taux de croissance aléatoires, a été montré par [gabaix1999zipf] produisant asymptotiquement la loi rang-taille (loi de Zipf) attendue pour les systèmes de ville et qui est considérée comme l'un des faits stylisés les plus réguliers, au moins dans sa formulation généralisée sous forme de loi d'échelle [nitsch2005zipf]. Expliquer les lois d'échelles urbaines est étroitement lié à la compréhension de la croissance urbaine, comme [bettencourt2008large] suggère que celles-ci reflètent des processus universels sous-jacents et que les propriétés individuelles des villes s'expliquent par un passage à l'échelle. Cette approche reflète cependant peu les relations com-

²⁸ Qu'on comprend ici comme l'évolution dans le temps des villes, vue typiquement par l'évolution de leur population ou des activités économiques, et de leur distribution spatiale.

²⁹ C'est à dire essayant de traduire et d'inclure ses processus fondamentaux et de reproduire ses faits stylisés.

plexes entre agents économiques pour lesquelles [storper2009rethinking] se positionnent.

Par l'utilisation d'une reconstruction par le bas des zones de population contigüe via des données microscopiques dynamiques de population, [rozenfeld2008laws] montrent en effet que des déviations aux tailles attendues par la loi rang-taille existent systématiquement, celles-ci étant sous-estimées, ce qui est probablement un effet des interactions spatiales entre les aires urbaines. Les approches par la complexité sont de bon candidats pour intégrer celles-ci dans les modèles. [andersson2006complex] introduit par exemple un modèle d'économie urbaine comme un réseau complexe de relations en croissance. La Théorie Evolutive Urbaine, introduite par [pumain1997pour], se concentre sur les villes comme des entités en co-évolution et produit des explications pour la croissance au niveau du système de villes. [pumain2006evolutionary] montrent que les lois d'échelles pourraient être dues à la différentiation fonctionnelle et la diffusion de l'innovation entre les villes. Le positionnement au regard de l'universalité des lois est plus modéré que les théories du *Scaling* comme celle de [west2017scaling], puisque [pumain2012urban] souligne que l'ergodicité peut difficilement être prise pour acquise dans le cadre des systèmes complexes territoriaux. Un aspect crucial de ce paradigme est l'importance des interactions entre agents constituants du système, généralement les villes, qui produisent les motifs émergents à l'échelle supérieure. [pumain2013theoretical] a investigué les avantages des modèles basés-agents comparé à des systèmes d'équations plus classiques, et cet aspect méthodologique est en accord avec le positionnement théorique, comme cela permet de prendre en compte l'hétérogénéité des interactions possibles, les particularités géographiques, et de traduire naturellement l'émergence entre les niveaux et rendre compte de motifs multi-échelles.

Croissance Urbaine et Interactions Spatiales

Dans un premier temps, nous devons préciser que nous considérons seulement les modèles à l'échelle macroscopiques, ne considérant pas les nombreuses approches très riches à l'échelle mesoscopique, qui incluent par exemple les modèles à automates cellulaires, les modèles de morphogenèse urbaine que nous aborderons en Chapitre 5 ou les modèles de changement d'usage du sol. Nous excluons aussi naturellement les modèles économiques qui n'incluent pas explicitement les interactions spatiales. Un certain nombre de modèles de croissance urbaine à l'échelle macroscopique ont insisté sur le rôle de l'espace et des interactions spatiales, que nous allons illustrer par la suite. [bretagnolle2000long] a proposé une extension spatiale du modèle de Gibrat. Le modèle d'interaction basé sur la gravité que [anders1992systeme] utilise pour appliquer les concepts de la Synergétique aux villes est également proche de cette idée de crois-

sance urbaine interdépendante, contenue physiquement dans le phénomène de migration entre les villes. Une extension plus raffinée avec des cycles économiques et des vagues d'innovation a été développé par [favar2011gibrat], fournissant une version du cœur des modèles Simpop [pumain2012multi] en termes de systèmes dynamiques.

La famille des modèles Simpop a été développée en symbiose avec la Théorie Évolutive des villes, avec la caractéristique principale de modèles basés sur les agents prenant en compte l'interaction spatiale. Les modèles ont été progressivement raffiné, et spécifiés pour divers cas d'étude. Nous pouvons donner un aperçu chronologique d'un échantillon de ceux-ci. Cette famille de modèles a commencé avec un modèle stylisé basé sur les interactions économiques entre les villes comme agents, qui produit des motifs de hiérarchie à l'échelle du système [sanderson1997simpop]. Plus tard, le modèle Simpop2, toujours basé sur l'interaction en fonction de la distance pour les échanges commerciaux, incluant les vagues successives d'innovation, a dévoilé des différences structurelles entre le système de villes Européen et le système aux Etats-Unis [bretagnolle2010comparer]. Le modèle SimpopLocal [pumain2017simpoplocal] est utilisé pour montré l'émergence des motifs initiaux d'établissement humains. Enfin, le dernier modèle similaire en date, le modèle Marius [cottineau2014evolution] couple la croissance de la population et économique avec les interactions entre les ville, permettant de reproduire assez fidèlement les trajectoires réelles (au sens de l'erreur carré moyenne dans le temps sur l'ensemble des populations) sur l'ancienne Union Soviétique après calibration avec multi-modélisation des processus.

Croissance Urbaine et Réseaux de Transports

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 revu en Chapitre 2.

Dans des hypothèses similaires aux modèles précédemment revus, l'inclusion des réseaux de transports a été rarement poursuivie, contrairement à l'échelle mesoscopique à laquelle les relations entre réseaux et territoires ont été largement étudiées par les modèles Luti par exemple [chang2006models]. Les modèles de croissance de réseau [xie2009modeling], prolifiques en économie et physique, ne peuvent pas être utilisés pour expliquer la croissance urbaine.

[bigotte2010integrated] étudie un modèle d'optimisation pour la conception du réseau combinant les effets de la hiérarchie urbaine et de la hiérarchie du réseau de transport. [baptiste1999interactions] a modélisé l'intrication dynamique entre la capacité des liens du réseau et la croissance des villes sur un sous-ensemble du système de villes français. Le modèle SimpopNet [schmitt2014modelisation] va un pas plus loin dans la modélisation de la co-évolution entre les villes et les réseaux de transport, puisqu'il permet que de nouveaux liens soit

créés dans le temps. Ces exemples montrent la difficulté de coupler ces deux aspects des systèmes urbains dans les modèles de croissance, et nous prendrons en compte pour cette raison les effets de réseau d'une manière simplifiée comme nous le détaillerons par la suite.

La suite de cette section est organisée de la manière suivante : nous introduisons notre modèle macroscopique et le décrivons de manière formelle ; puis nous donnons les résultats obtenus par l'exploration et la calibration du modèle sur les données pour les villes françaises, plus particulièrement la révélation d'effets de réseaux influençant de manière significative les processus de croissance, grâce à une nouvelle méthodologie spécifiquement introduite. Nous discutons finalement les implications de ces résultats.

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 mènerons par la suite.

4.3.2 *Modèle et Résultats*

Description du modèle

HYPOTHÈSES Une confusion peut régner lorsqu'on s'intéresse aux modèles stochastiques et déterministes de croissance urbaine. Dans quelle mesure un modèle proposé est-il "complexe" et la simulation de la stochasticité nécessaire ? Concernant le modèle de Gibrat et la plupart de ses extensions, les hypothèses d'indépendance et la linéarité produisent un comportement totalement prédictable, ce qui ne les rend pas complexes au sens d'exhiber une émergence faible [**bedau2002downward**]. En particulier, la distribution complète des modèles de croissance aléatoire peut être déterminée analytiquement à tout instant [**gabaix1999zipf**], et dans le cas de l'étude du premier moment seulement, une simple relation de récurrence évite de procéder à toute simulation de Monte-Carlo. Sous ces hypothèses, il est raisonnable de travailler avec un modèle déterministe, comme il est fait par exemple pour le modèle Marius [**cottineau2014evolution**]³⁰. Nous travaillerons sous cette hypothèse, capturant la complexité par la non-linéarité. Nous travaillons sur des systèmes territoriaux simples supposés comme des systèmes de villes régionaux, dans lesquels les villes sont les entités de base. Nous avons ainsi de l'ordre d'une centaine de villes, et les territoires intermédiaires ne sont pas pris en compte. L'échelle de temps correspond à l'échelle caractéristique associée à cette échelle spatiale, i.e. autour d'un ou deux siècles. Les interactions spatiales sont capturées par des interactions de type gravitaire, cette formulation ayant l'avantage de la simplicité et de capturer la première loi de Tobler, c'est à

³⁰ Le modèle Marius introduit par [**cottineau2014evolution**] pour l'évolution des villes en ex-Union soviétique, lie les variables économiques et de population à l'échelle de villes en prenant en compte leurs interactions. Le modèle a été conçu dans une perspective de modélisation incrémentale et divers processus peuvent être considérés.

dire que la force d'interaction décroît avec la distance. Ce choix n'a a priori pas une influence fondamentale sur le comportement du modèle, puisque d'autres approches différentes du paradigme gravitaire comme le modèle de radiation, introduites plus récemment, ont des performances similaires à cette échelle [masucci2013gravity].

DESCRIPTION Nous considérons une extension déterministe du modèle de Gibrat, ce qui est équivalent à considérer seulement les espérances des populations dans le temps et ne plus simuler des trajectoires aléatoires. Soit $\vec{P}(t) = (P_i(t))_{1 \leq i \leq n}$ le vecteur des populations des villes dans le temps. Sous les hypothèses d'indépendance de Gibrat³¹, nous avons $\text{Cov}[P_i(t), P_j(t)] = 0$, c'est à dire que les villes ne s'influencent pas mutuellement dans leur processus de croissance. Une version étendue linéaire du modèle de Gibrat s'écrirait alors

$$\vec{P}(t+1) = \mathbf{R} \cdot \vec{P}(t)$$

où \mathbf{R} est une matrice aléatoire indépendante de taux de croissance (l'identité à un scalaire près dans le cas original). Cela conduit directement grâce à l'hypothèse d'indépendance que $\mathbb{E}[\vec{P}(t+1)] = \mathbb{E}[\mathbf{R}] \cdot \mathbb{E}[\vec{P}(t)]$.

Nous généralisons cette relation linéaire à une relation non-linéaire qui permet d'être plus cohérent avec les interprétations du modèle et plus flexible. Notant $\vec{\mu}(t) = \mathbb{E}[\vec{P}(t)]$, nous généraliserons cette relation avec une fonction donnée f et un pas de temps quelconque Δt , sous la forme

$$\vec{\mu}(t + \Delta t) = \Delta t \cdot f(\vec{\mu}(t))$$

Il faut noter que dans ce cas, les versions stochastiques et déterministes ne sont plus équivalentes, précisément à cause de la non-linéarité, mais nous gardons une version déterministe pour rester simple. La spécification des taux de croissance interdépendants est donnée par

$$f(\vec{\mu}) = (1 + r_0) \cdot \mathbf{Id} \cdot \vec{\mu} + \mathbf{G}(\vec{\mu}) \cdot \vec{1} + \vec{N}(\vec{\mu}) \quad (4)$$

où $\vec{1}$ est le vecteur colonne unité, et $\mathbf{G} = G_{ij} = w_G \cdot \frac{V_{ij}}{\langle V_{ij} \rangle}$ est le terme d'interaction directe, de telle façon que le potentiel d'interaction V_{ij} suit une expression de type gravitaire donnée par, avec d_{ij} distance entre i et j (distance euclidienne ou distance de réseau),

$$V_{ij} = \left(\frac{\mu_i \mu_j}{(\sum_k \mu_k)^2} \right)^{\gamma_G} \cdot \exp(-d_{ij}/d_G) \quad (5)$$

³¹ On rappelle que le modèle de Gibrat suppose des taux de croissance aléatoires indépendants : $P_i(t+1) = r \cdot P_i(t)$.

où γ_G est un exposant de hiérarchie des interactions par rapport aux populations, d_G un paramètre de décroissance donnant la distance typique d'interaction, et w_G est le poids relatif des interactions directes.

Le dernier terme capture un effet de réseau : \vec{N} est donné par $N_i = w_N \cdot \frac{W_i}{\langle W_i \rangle}$ où le potentiel du flux de réseau W_i suit

$$W_i = \sum_{k < l} \left(\frac{\mu_k \mu_l}{\left(\sum_j \mu_j \right)^2} \right)^{\gamma_N} \cdot \exp(-d_{kl,i}/d_N) \quad (6)$$

où $d_{kl,i}$ est la distance de la ville i au plus court chemin entre k, l calculé dans l'espace géographique, qui peut être par un réseau de transport ou dans un champ d'impédance dans l'espace euclidien. Les paramètres γ_N , d_N et w_N sont analogues à ceux pour l'interaction directe. Les sept paramètres du modèle sont détaillés ci-dessous et récapitulés en Table 10. Précisons d'abord la logique de cette formulation.

Le premier terme de l'équation est le modèle de Gibrat seul, qui est obtenu en fixant les poids $w_G = w_N = 0$. La deuxième composante capture les interdépendances directes entre les villes, sous la forme d'un potentiel gravitaire séparable comme celui utilisé dans [sanderson1992systeme]. La logique du troisième terme, qui a pour but de capturer l'effet de réseau en exprimant une rétroaction des flux du réseau entre les villes k, l sur la ville i . Intuitivement, un flux démographique et économique transitant physiquement par une ville ou dans son voisinage est attendu d'avoir une influence sur son développement (par des arrêts intermédiaires e.g.), cet effet étant bien sûr dépendant du mode de transport puisqu'une ligne à grande vitesse avec peu d'arrêts ignorera la majorité des territoires traversés. Notons que nous n'utilisons pas exactement les flux gravitaires dans le terme de réseau, puisqu'il n'y a pas de décroissance des interactions générant les flux avec la distance, mais une décroissance de l'effet du flux en fonction de la distance au réseau. Cela est équivalent à supposer une utilisation du réseau sur de très longues portées en moyenne dans le temps, puisque le terme d'atténuation tend vers 1 si le paramètre de décroissance tend vers l'infini, ce qui est ainsi complémentaire au premier terme de gravité.

ESPACE DES PARAMÈTRES Nous donnons en Table 10 la description des paramètres du modèle, détaillant les processus associés et les bornes des paramètres. Les interactions directes et les effets au second ordre des flux du réseau ont tous deux la même structure, c'est à dire le caractère séparable de l'effet de la distance et de l'influence des populations, un paramètre de décroissance exponentielle et un paramètre de hiérarchie exprimant l'inégalité des contributions

selon les tailles relatives des villes : plus l'exposant est grand, plus les contributions des petites villes seront négligeables au regard des grandes villes. Le paramètre de décroissance de la distance peut s'interpréter comme une distance caractéristique d'atténuation des interactions³². Finalement, nous ne considérerons que des poids positifs, pour suivre les observations empiriques comme détaillé ci-dessous. Les valeurs numériques pour les poids seront données normalisées par le nombre de villes impliquées dans le processus, i.e. $w'_G = w_G/n$ et $w'_N = w_N/(n(n-1)/2)$.

TABLE 10 : **Espace des paramètres du modèle d'interaction.** Nous donnons le nom du paramètre et sa notation, le processus du modèle qui lui est associé, une interprétation possible, et son domaine typique de variation.

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	r_0	Effet des flus	Portée de l'effet	$[0, +\infty]$

Données

Le modèle est construit pour être hybride, car nous proposons de l'étudier sur une semi-paramétrisation par données empiriques. Il pourrait être possible de l'étudier comme un modèle complètement jouet, la configuration initiale et l'environnement physique étant construits comme données synthétiques. Nous visons cependant à révéler des faits stylisés sur des données réelles plutôt que sur le comportement du modèle en lui-même, et initialisons ainsi le modèle à partir des données que nous décrivons à présent.

DONNÉES DE POPULATION Nous travaillons avec la base de données historique Pumain-INED pour les villes françaises [[pumain1986fichier](#)], qui donne les populations des Aires Urbaines (définition de l'INSEE) à des intervalles de temps de 5 ans, de 1831 à 1999 (31 observations

³² Il est possible d'établir formellement son expression exacte. Fixons une fraction arbitraire α et des portées spatiales typiques pour un système urbain local d_L et pour un système urbain à longue portée d_R , considérons une ville i et deux voisines j, j' de population égale $\mu_j = \mu_{j'}$, à des distances respectives d_L et d_R de i . Si on veut répondre à la question à quelle différence de distance est équivalent une atténuation de α du potentiel d'interaction avec i , nous obtenons $d_L - d_R = -d_G \cdot \ln \alpha$. Pour cela, d_G est exactement le coefficient de proportionnalité répondant à ce questionnement intuitif.

temporelles). La version la plus récente de la base de données intègre les aires urbaines, permettant de les suivre sur de longues périodes de temps, suivant l'ontologie de BRETAGNOLLE pour les villes sur le temps long [bretagnolle:tel-00459720], qui construit une définition fonctionnelle des villes comme entités dont les limites évoluent dans le temps. Pour simplifier, nous travaillons avec les 50 plus grandes villes en 1999³³. Nous isolons de plus des périodes de longueur similaires excluant les guerres, obtenant 9 périodes³⁴ de 20 ans sur lesquelles l'ajustement du modèle non-stationnaire dans le temps sera exécuté.

FLUX PHYSIQUES Comme rappelé précédemment, cet exercice de modélisation se concentre sur l'exploration du rôle des flux physiques, quelle que soit la forme effective du réseau. Nous choisissons pour cette raison de ne pas utiliser de vraies données de réseau qui sont de plus difficiles à obtenir à différentes périodes de temps, et nous supposons que les flux physiques prennent le plus court chemin géographique prenant en compte la pente du terrain. Cela évite des absurdités géographiques comme des villes difficilement accessibles ayant un taux de croissance surestimé. Utilisant le Modèle d'Elevation Numérique de l'IGN à la résolution 1km, nous calculons les plus courts chemins de manière standard [collischon:2000:direction], par la construction d'un champ d'impédance de la forme

$$Z = \left(1 + \frac{\alpha}{\alpha_0}\right)^{n_0}$$

où Z est l'impédance des liens du réseau de la grille de 1km dans laquelle chaque cellule est connectée à ses huit voisins. α est la pente du terrain calculée avec la différence d'altitude entre les deux cellules. Nous prenons des valeurs des paramètres fixes $\alpha_0 = 3$ (correspondant approximativement à la valeur réelle d'une pente de 5%) et $n_0 = 3$ ce qui donne des chemins plus réalistes que des valeurs significativement plus petites ou plus grandes³⁵.

³³ Ce choix n'ayant que peu d'influence sur la majorité des trajectoires des villes puisque les petites villes influent peu dans les processus d'interaction. Il peut en avoir sur l'ajustement des villes ajoutées, mais notre objectif n'étant pas de reproduire fidèlement l'ensemble des trajectoires mais de comprendre le rôle du réseau, nous fixons ce seuil.

³⁴ Qui sont précisément : 1831-1851, 1841-1861, 1851-1872, 1881-1901, 1891-1911, 1921-1936, 1946-1968, 1962-1982, 1975-1999.

³⁵ Plus précisément, on a validé "à dire d'expert", en inspectant visuellement les chemins entre quelques destinations typiques (incluant Paris-Lyon, Lyon-Marseille, Lyon-Bordeaux par exemple), pour $\alpha_0 = 2, 3, 4$. Pour $\alpha_0 = 4$, le chemin est généralement trop rectiligne et coupe à travers des relief évités par les grands axes; pour $\alpha_0 = 2$ le chemin est trop sinueux au contraire. On a testé $n_0 = 2, 3$, le deuxième étant également plus crédible. Une calibration précise de ces paramètres nécessiterait l'ajustement par rapport au réseau autoroutier par exemple, mais est hors de portée de cet exercice ici.

Indicateurs de performance

Nous travaillons sur un modèle explicatif plutôt qu'un modèle exploratoire. Pour cette raison, les indicateurs pour évaluer les sorties du modèle ne sont pas directement liés aux propriétés intrinsèques des trajectoires ou des états finaux obtenus, mais plutôt à une distance au phénomène que l'on cherche à expliquer, i.e. les données. Etant donné des populations réelles $p_i(t)$ (réalisations historiques de $P_i(t)$) et les espérances simulées $\mu_i(t)$ obtenues par $\vec{\mu}(t_0) = \vec{p}(t_0)$ sur une période de longueur T , on peut évaluer deux aspects complémentaires de la performance du modèle :

- Performance globale du modèle, donnée par le logarithme de l'erreur carrée moyenne dans l'espace et le temps

$$\varepsilon_G = \ln \left(\frac{1}{T} \sum_t \frac{1}{n} \sum_i (p_i(t) - \mu_i(t))^2 \right)$$

- La performance locale moyenne, donnée par l'erreur carrée moyenne des logarithmes, comme proposé par [pumain2017incremental]

$$\varepsilon_L = \frac{1}{T} \sum_t \frac{1}{n} \sum_i (\ln p_i(t) - \ln \mu_i(t))^2$$

Les deux sont en fait complémentaires, puisqu'utiliser seulement ε_G comme il est généralement fait se concentrera seulement sur les plus grandes villes et donnera des résultats mitigés sur les villes de taille moyennes et les petites villes (pour la France seul Paris aura une estimation raisonnable comme il domine fortement les autres aires urbaines et villes). ε_L permet pour cela de prendre en compte la performance du modèle sur l'ensemble des villes simulées par le modèle.

Résultats

FAITS STYLISTÉS Des faits stylisés typiques peuvent être extraits d'une telle base de données, comme il a été déjà largement été exploré dans la littérature [guerin1990150]. Nous retrouvons les meilleurs ajustements de distributions log-normales des taux de croissance à toutes les dates en comparaison à des distributions normales, et également le fait que les taux de croissance sont essentiellement positifs, sur les villes que nous considérons et enlevant les guerres.

Un aspect intéressant à examiner en relation avec nos considérations sur les interactions spatiales sont les corrélations entre les séries temporelles, et plus particulièrement leur variation en fonction de la distance. Nous considérons des fenêtres temporelles de 50 ans se superposant pour avoir assez d'observations temporelles, finissant respectivement en (1881, 1906, 1931, 1962, 1999) et estimons sur chacune, pour chaque couple de villes (i, j), la corrélation entre les log-returns,

que nous définissons par $\Delta X_i = X_i(t) - X_i(t-1)$ et $X_i(t) = \ln\left(\frac{P_i(t)}{P_i(t_0)}\right)$, est donnée par $\hat{\rho}_{ij} = \rho[\Delta X_i, \Delta X_j]$ avec un estimateur de Pearson classique. Cette méthode permet de révéler des interactions dynamiques sans être biaisée par les tailles [mantegna1999introduction].

Nous montrons en Figure 25 les courbes de corrélations lissées en fonction de la distance, pour chaque période temporelle. Tout d'abord, les fortes différentes entre chaque confirme la non-stationnarité des taux de croissance sur l'ensemble de la période temporelle, et justifie l'utilisation d'ajustements locaux dans le temps pour le modèle. Nous pouvons aussi interpréter ces motifs en termes d'événements historiques pour le système de villes et le réseau de transport. La dynamique du système commence par une corrélation plate en 1881, autour de 0.2, qui pourrait être fortuite à cause de croissance similaire simultanée pour toutes les villes. Elle reste ensuite plate mais tend vers 0, témoignant de fortes différentiations dans les motifs de croissance entre 1856 et 1906. Après 1931, l'effet de la distance est clair avec des courbes décroissantes, commençant entre 0.4 et 0.5. Nous postulons que cette évolution doit être partiellement liée à l'évolution du réseau de transport : en considérant le réseau ferré par exemple [thevenin2013mapping], le développement initial global a pu encourager des interactions à longue portée rendant ainsi les courbes de corrélation plates, tandis que sa maturation dans le temps a conduit au retour d'interactions plus classiques décroissant rapidement avec la distance.

EXPLORATION DU MODÈLE La pré-traitement des données, le traitement des résultats et le profilage des modèles sont implémentés en R. Pour des raisons de performance et une intégration plus facile dans le logiciel OpenMole pour l'exploration de modèles [reuillon2013openmole], une version scala a également été développée. La question du compromis entre performance d'implémentation et inter-opérabilité est un problème typique de ce genre de modèle, puisque des explorations et calibrations totalement aveugles peuvent être trompeuses pour les directions de recherches futures ou les interprétations thématiques. Une implémentation NetLogo, permettant l'exploration interactive et la visualisation dynamique, a également été développée pour cette raison. Le code source des modèles, les données brutes nettoyées, les données de simulation, et les résultats utilisés ici sont disponibles sur le dépôt ouvert du projet³⁶. Nous montrons en Fig. 26 un exemple de sortie du modèle. Les couleurs des villes donnent l'écart

³⁶ A <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/NetworkNecessity/InteractionGibrat>. Les trois versions du modèle sont destinées à être reprises, réutilisées et étendues, et chaque implémentation a son utilité propre : la version R permet une intégration directe avec des scripts d'analyse de données, la version scala peut être utilisée comme plugin OpenMole, et la version NetLogo permet une exploration interactive et la visualisation directe des trajectoires.

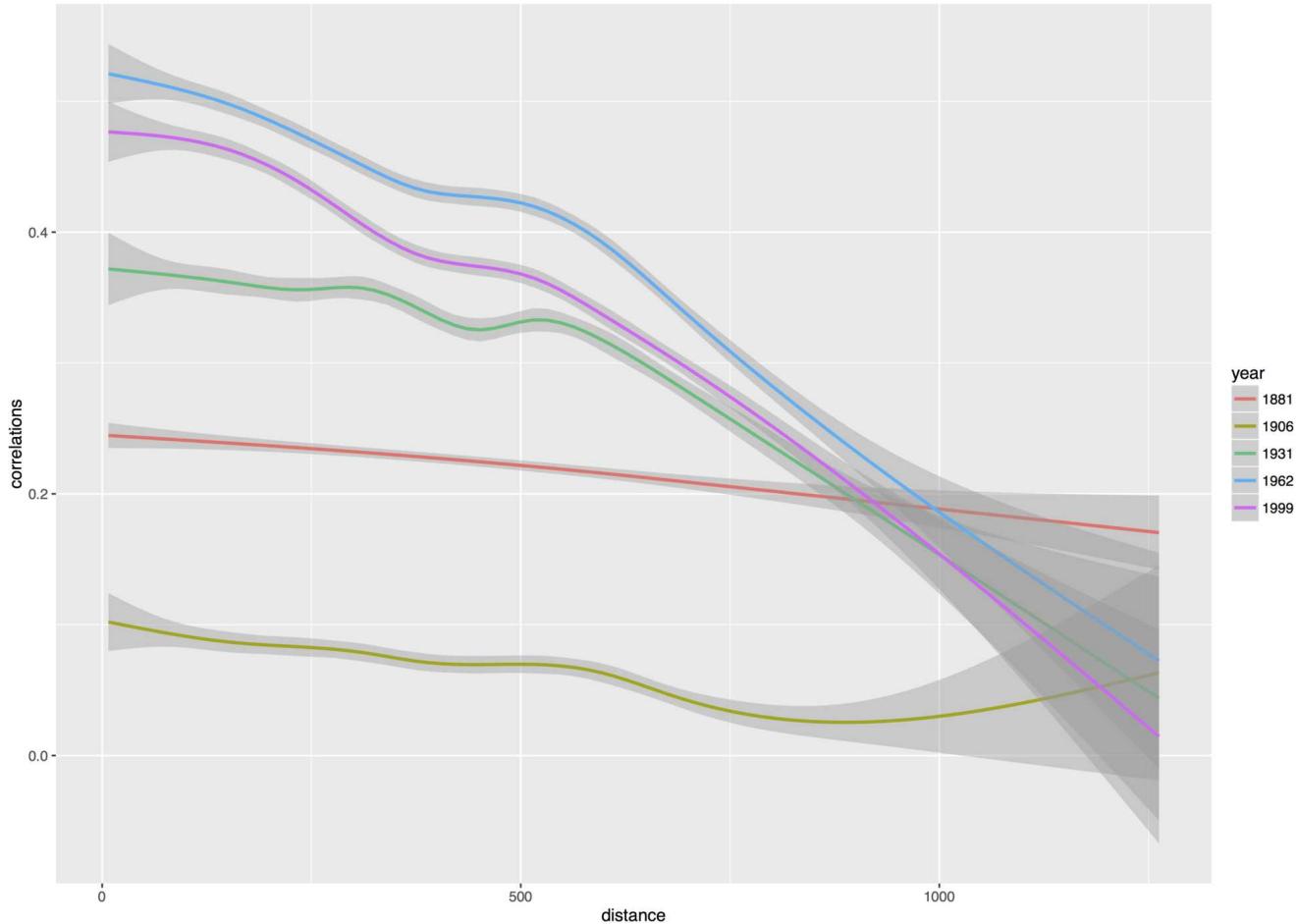


FIGURE 25 : Corrélations entre séries temporelles en fonction de la distance. Les lignes pleines correspondent aux corrélations lissées, calculées entre chaque paire des log-returns normalisés des séries temporelles de population, sur des périodes successives données par la couleur de la courbe.

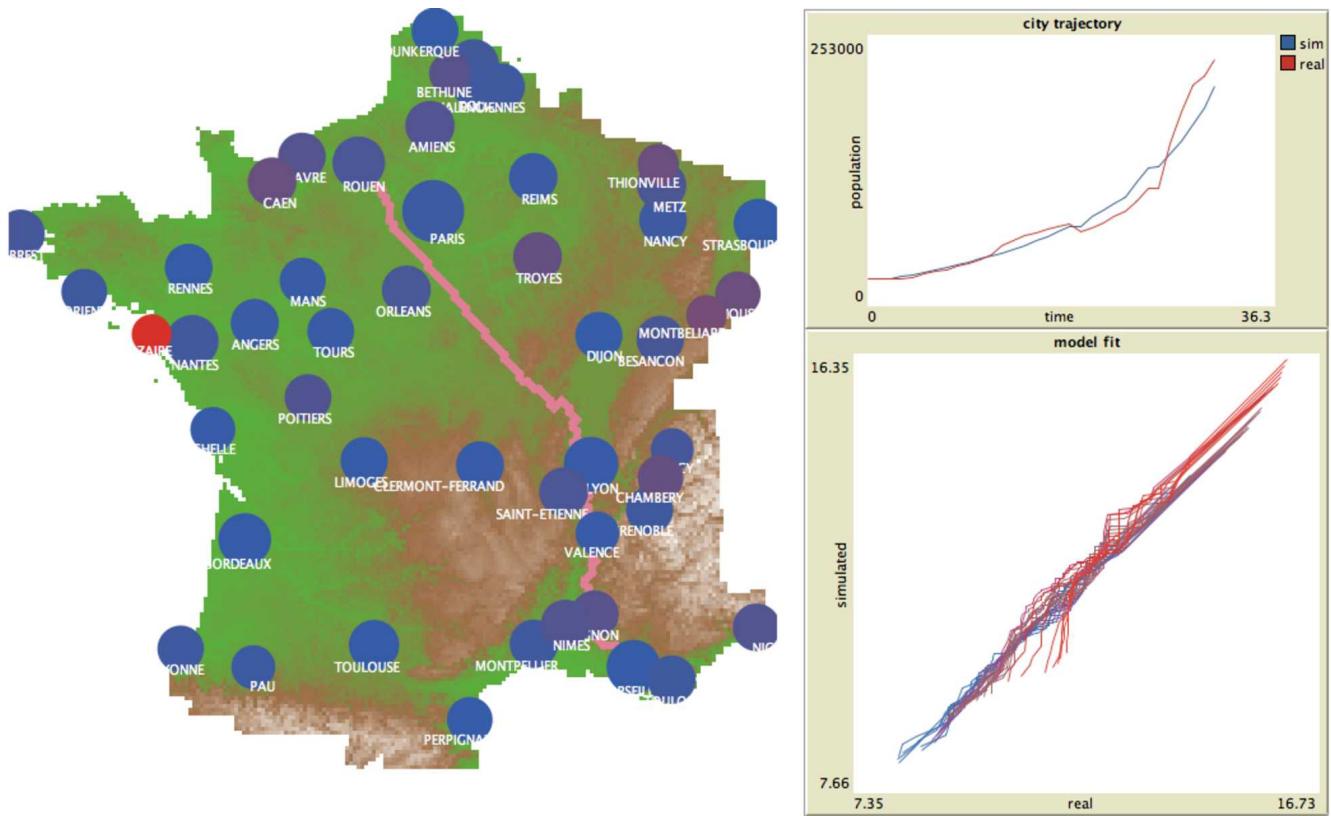


FIGURE 26 : Exemple de sortie du modèle. L’interface graphique permet d’explorer de manière interactive sur quelles villes les changements s’opèrent après un changement de paramètres, ce qui est nécessaire pour interpréter les résultats bruts de calibration. La carte permet de visualiser les erreurs d’ajustement par ville (couleur) et leur population (taille). Nous illustrons en rose le plus court chemin géographique entre Rouen et Avignon. Le graphe supérieur permet de suivre dans le temps la trajectoire d’une ville donnée, en comparant la population simulée à la population réelle. Le graphe inférieur trace à chaque date l’ensemble des données simulées en fonction des données réelles : plus la courbe est proche de la diagonale plus l’ajustement est bon.

à l’observation au niveau de la ville et leur taille la population. Les valeurs extrêmes peuvent ainsi être aisément repérées (comme Saint-Nazaire ayant le pire fit dans l’exemple montré) et des possibles effets régionaux identifiés. Nous illustrons en rose un exemple de plus court chemin géographique, de Rouen à Marseille, qui correspond raisonnablement au plus court chemin effectif actuel par autoroute. Le graphe du haut montre la trajectoire dans le temps pour une ville donnée, tandis que celui du bas donne la qualité globale de l’ajustement dans le temps, en traçant les données simulées en fonction des données réelles. Plus la courbe est proche de la diagonale, meilleur est l’ajustement.

Les premières explorations du modèle, en parcourant simplement des grilles fixées de l’espace des paramètres, suggèrent déjà la présence d’effets de réseau, au sens de flux physiques ayant effectivement une influence sur la reproduction des taux de croissance ob-

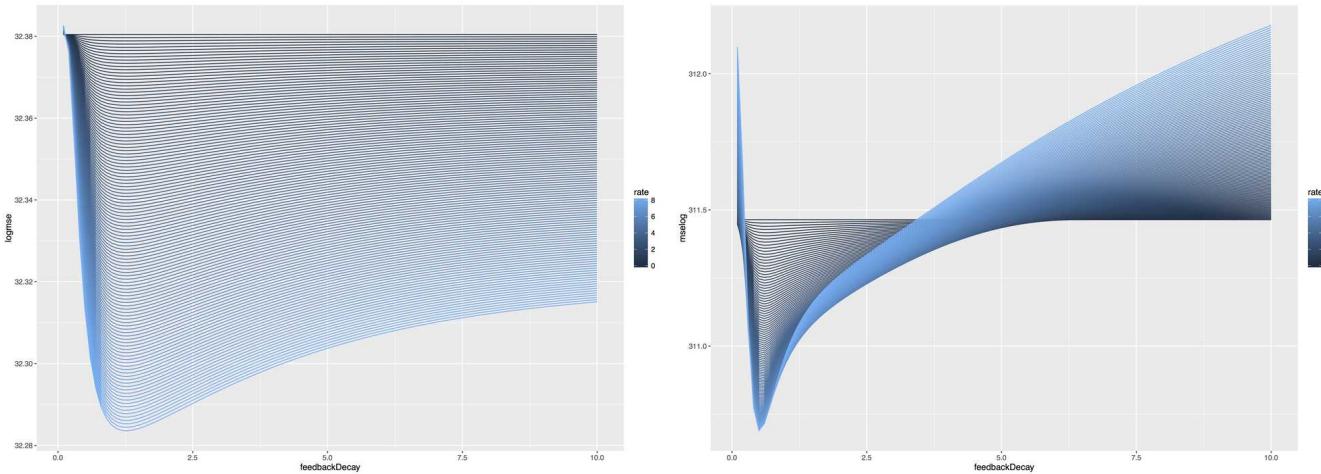


FIGURE 27 : Effets de réseau révélés par l'exploration du modèle. Le graphe de gauche donne ε_G comme fonction de d_N pour w_N/r_0 variant (rate), à effet de gravité fixe et $\gamma_N = 3$. Le graphe de droite est similaire pour ε_L . Partant d'un modèle gravitaire pur (courbe horizontale pour w_N), la prise en compte progressive du réseau augmente les performances au regard des deux objectifs, dans une plage restreinte pour d_N . Les valeurs de d_N donnant les minima correspondent à la distance typique de l'effet du réseau.

servés. Nous montrons en Fig. 27 une configuration dans laquelle c'est le cas. À paramètres de gravité et taux de croissance fixés, nous étudions les variations des paramètres w_N , d_N et γ_N et la réponse correspondante de ε_G et ε_L . A des valeurs fixes de γ_N , on observe un comportement similaire des indicateurs quand w_N et d_N varient. L'existence d'un minimum pour les deux comme fonction de d_N , qui devient plus marqué quand w_N augmente, montre que l'introduction du terme de rétroaction du réseau améliore les fits locaux et globaux en comparaison du modèle de gravité seul, i.e. que les processus associés ont un pouvoir explicatif pour les motifs de croissance.

CALIBRATION DU MODÈLE DE GRAVITÉ Nous utilisons à présent le modèle pour extraire de l'information de manière indirecte sur les processus dans le temps. En effet sous l'hypothèse de non-stationnarité, l'évolution temporelle des paramètres ajustés localement montre l'évolution de l'aspect des processus correspondant. Dans une première expérience, nous fixons $w_N = 0$ et calibrions le modèle avec quatre paramètres sur les neuf périodes temporelles successives de 20 ans. Le problème d'optimisation associé à la calibration du modèle ne présente pas de caractéristiques le rendant agréable à résoudre (expression fermée d'une fonction de likelihood, convexité ou caractère creux du problème d'optimisation), nous devons nous reposer sur des techniques alternatives pour le résoudre. Une exploration de grille par force brute est rapidement limitée par le sort de la dimension. Les méthodes classiques [batty1972calibration] comme une descente du gradient échouent à cause de la forme assez compliquée du pay-

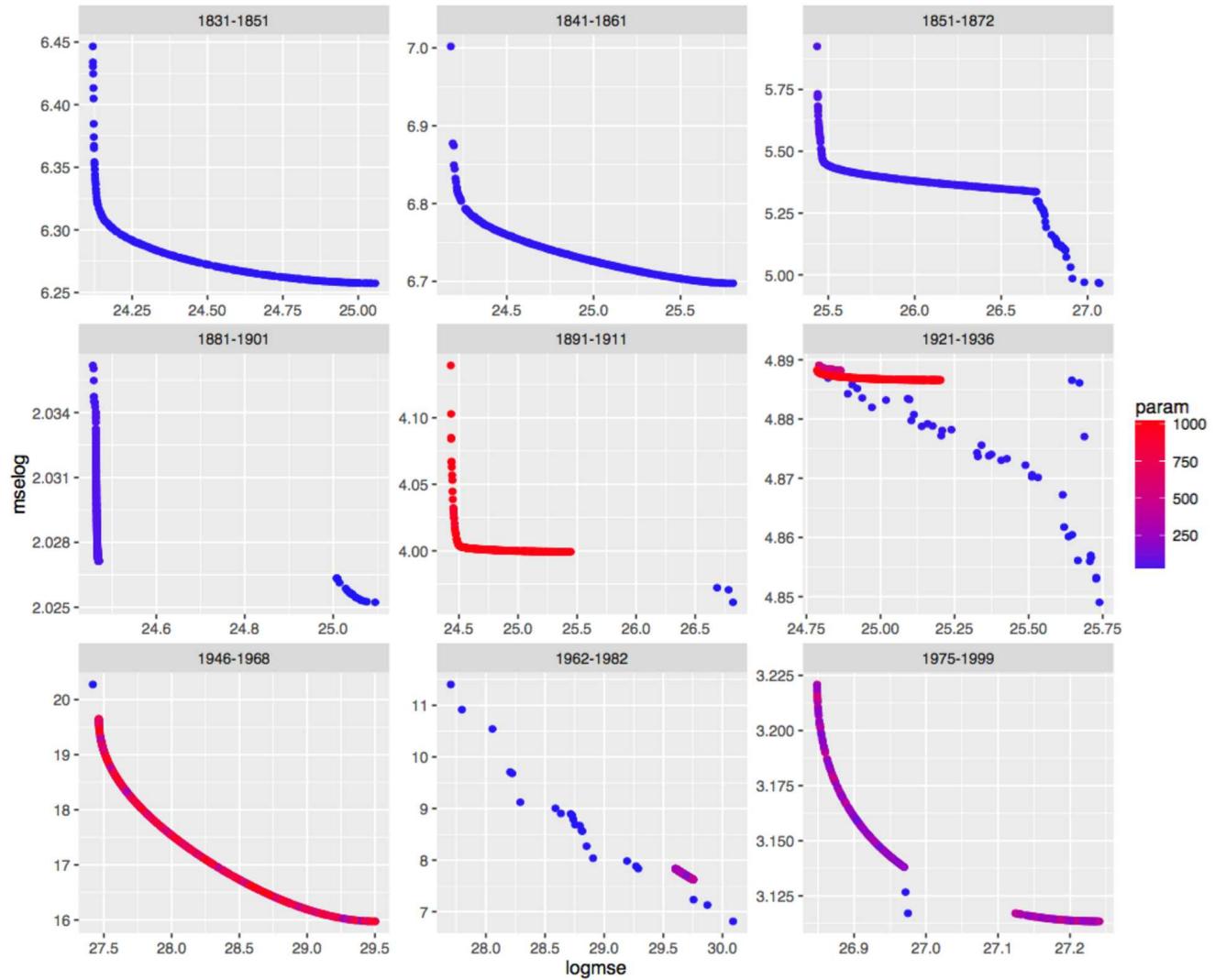


FIGURE 28 : Calibration du modèle de gravité. Fronts de Pareto sur les périodes successives. La couleur donne la valeur du paramètre de décroissance de la distance.

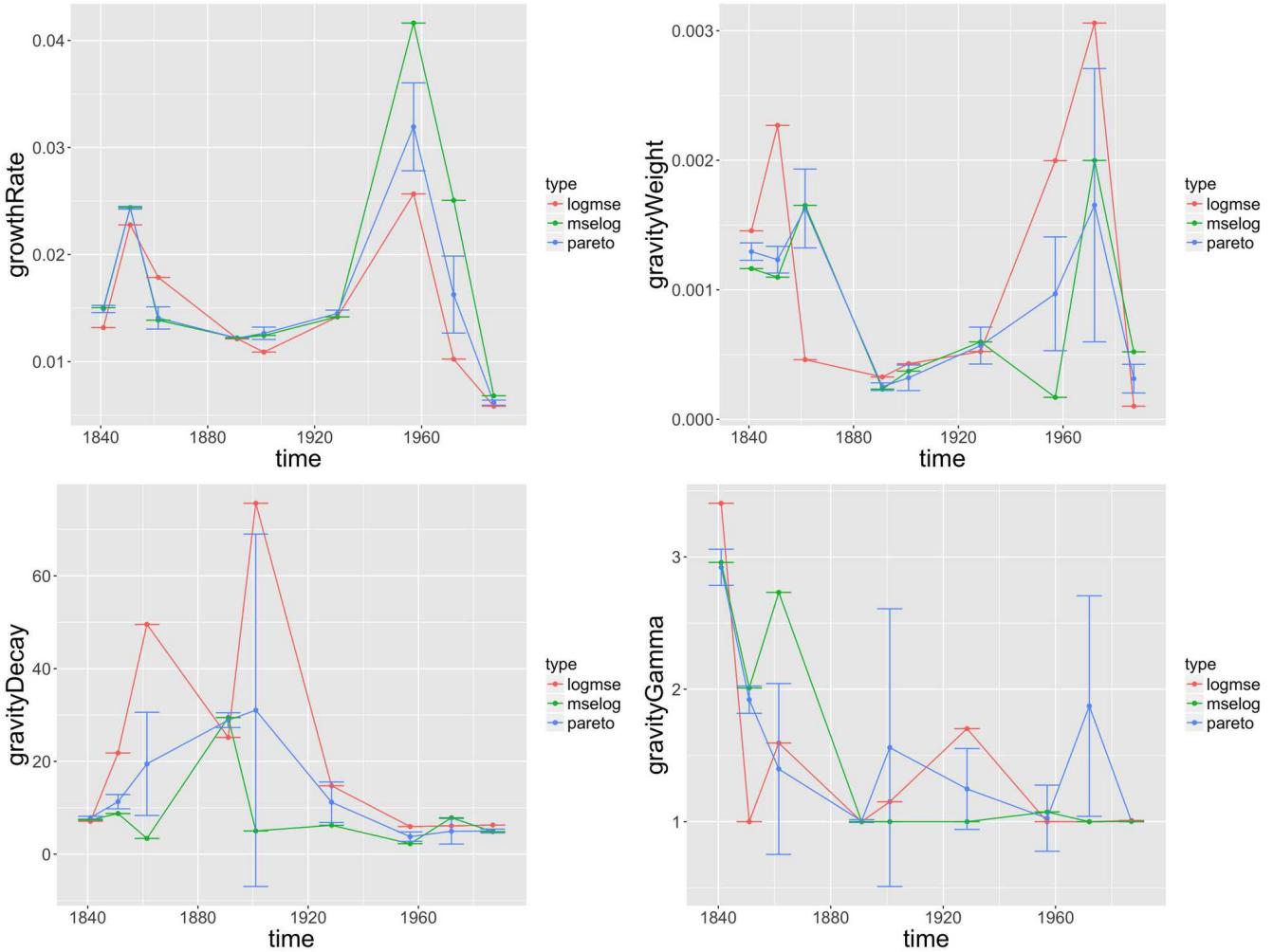


FIGURE 29 : Valeurs des paramètres calibrés pour le modèle de gravité seul. Chaque graphe donne les valeur ajustées dans le temps pour chaque paramètre. Les courbes rouge et verte correspondent aux points optimaux pour ε_G (respectivement ε_L), tandis que la courbe bleue donne la valeur moyenne sur l'ensemble du front de Pareto avec la déviation standard. Selon qu'on s'intéresse à la performance sur les petites villes (mselog), les grandes villes (logmse) ou des compromis (pareto), on obtiendra une évolution des paramètres pouvant différer. Des grandes tendances se dégagent, comme le double pic du taux de croissance endogène (growthRate) qui se retrouve dans le poids des interactions, et une hiérarchie décroissante des interactions.

sage d'optimisation. La calibration par Algorithme Génétique (GA) est une solution efficace pour trouver des solutions approximatives en un temps raisonnable. OpenMole inclut une collection de telles méta-heuristiques pour différents buts : [schmitt2014half] démontre les potentialités de ces méthodes pour calibrer les modèles de simulation. Dans notre cas, cela permet de plus de procéder à une optimisation bi-objectif sur $(\varepsilon_G, \varepsilon_L)$. Nous utilisons le GA steady state standard fournit par OpenMole, distribué sur 25 îles, avec une population de 200 individus et 100 générations³⁷.

Nous montrons en Fig. 28 les résultats de la calibration sur les périodes successives, en représentant la population finale dans l'espace des indicateurs. Comme attendu, des fronts de Pareto correspondant à des compromis entre les deux objectifs opposés sont la règle. Cela signifie que le modèle ne peut pas être précis à la fois globalement et localement, et qu'une solution intermédiaire doit être trouvée. Cela peut être dû au fait que la portée d'interaction change avec la taille de la ville (i.e. que les termes dans le potentiel ne sont plus séparables), que nous gardons comme un développement potentiel du modèle. La forme des fronts de Pareto révèle un paysage d'optimisation chaotique, puisque pour certaines périodes comme 1921-1936 ou 1962-1982 les fronts ne sont pas réguliers et épars. Le changement dans les formes traduit également différents régimes dynamiques selon les périodes : pour 1881-1901, la forme quasi-verticale suivi par un front isolé à de fortes valeurs de ε_G révèle un comportement quasi-binaire du modèle dans les régimes optimaux, au sens où améliorer ε_L sous la limite n'est possible uniquement à travers un saut qualitatif à un fort coût pour ε_G .

Les valeurs prises par d_G pour les périodes 1892-1911 et 1921-1936 montrent que les grandes villes ont des portées d'interaction plus grandes, puisqu'une valeur plus grande donne des meilleures valeurs pour ε_G . Nous montrons en Fig. 29 les valeurs des paramètres ajustés dans le temps, par leur moyenne sur le front de Pareto et pour les deux meilleures solutions à objectif simple. Tout d'abord, les deux motifs en pic pour r_0 correspondent globalement au comportement observé sur les taux de croissance moyens. L'évolution de w_G a une forme similaire mais décalée de 20 ans : cela peut être interprété comme une répercussion de la croissance endogène sur les motifs d'interaction les années suivantes, ce qui est cohérent avec une interprétation des processus d'interaction en termes de migration. Les valeurs de d_G , avec une augmentation jusqu'en 1900 suivie d'une décroissance progressive, est cohérent avec le comportement des corrélations empiriques commenté précédemment : les deux premières fenêtres de 50 ans ont des portées d'interaction plus grandes ce qui correspond à des courbes de corrélations plates. Enfin, le niveau de hié-

³⁷ Voir [pumain2017urban] pour la présentation la plus récente des méthodes de calibration de modèles urbains par algorithme génétique fournit par OpenMole.

rarchie γ_G a été régulièrement décroissant, ce qu'on peut lire comme une atténuation du pouvoir des grandes villes, qui peut être comprise en termes de la décentralisation progressive en France qui a été encouragée par l'administration³⁸.

EFFETS DE RÉSEAU Nous nous intéressons à présent à la calibration du modèle complet sur des périodes successives, dans le but d'interpréter les paramètres liés aux flux de réseau et obtenir des informations sur les effets de réseau. La calibration complète est faite de manière similaire avec les sept paramètres libres. Nous montrons en Fig. 30 les valeurs ajustées dans le temps pour certains de ces paramètres. Le comportement du taux de croissance et du poids de la gravité relatif au taux de croissance, qui est similaire au modèle de gravité seul, confirme que les effets de réseau sont bien au second ordre et que la croissance endogène et les interactions directes sont les facteurs principaux. Les effets de réseaux sont cependant loin d'être négligeables, puisqu'ils améliorent l'ajustement comme montré précédemment lors de l'exploration du modèle, capturant ainsi des processus de second ordre. L'évolution de d_N , correspondant à la portée sur laquelle le réseau influence le territoire qu'il traverse, montre un minimum en 1921-1936 pour se stabiliser à nouveau plus tard, mais à une valeur plus basse que les valeurs du passé. Cela pourrait correspondre à l'effet tunnel, puisque les transports à grande vitesse ne s'arrêtent peu sur les territoires qu'ils traversent. En effet, l'évolution du réseau ferré a témoigné une forte décroissance des lignes locales à une date similaire au minimum, et plus tard l'émergence de lignes à grande vitesse spécifiques, ce qui expliquerait cette valeur finale plus basse. La hiérarchie des flux a été légèrement décroissante comme pour la gravité, mais est extrêmement haute. Cela signifie que seuls les flux entre les grandes villes ont un effet significatif. Ainsi, le modèle donne une information indirecte sur les processus liés aux effets de réseau.

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.

³⁸ Sachant que la réalité est forcément plus complexe et qu'une telle tendance peut être tirée aussi par une inscription plus globale dans un changement de nature des structures urbaines, dont témoigne par exemple l'émergence des MCR (voir 1.2).

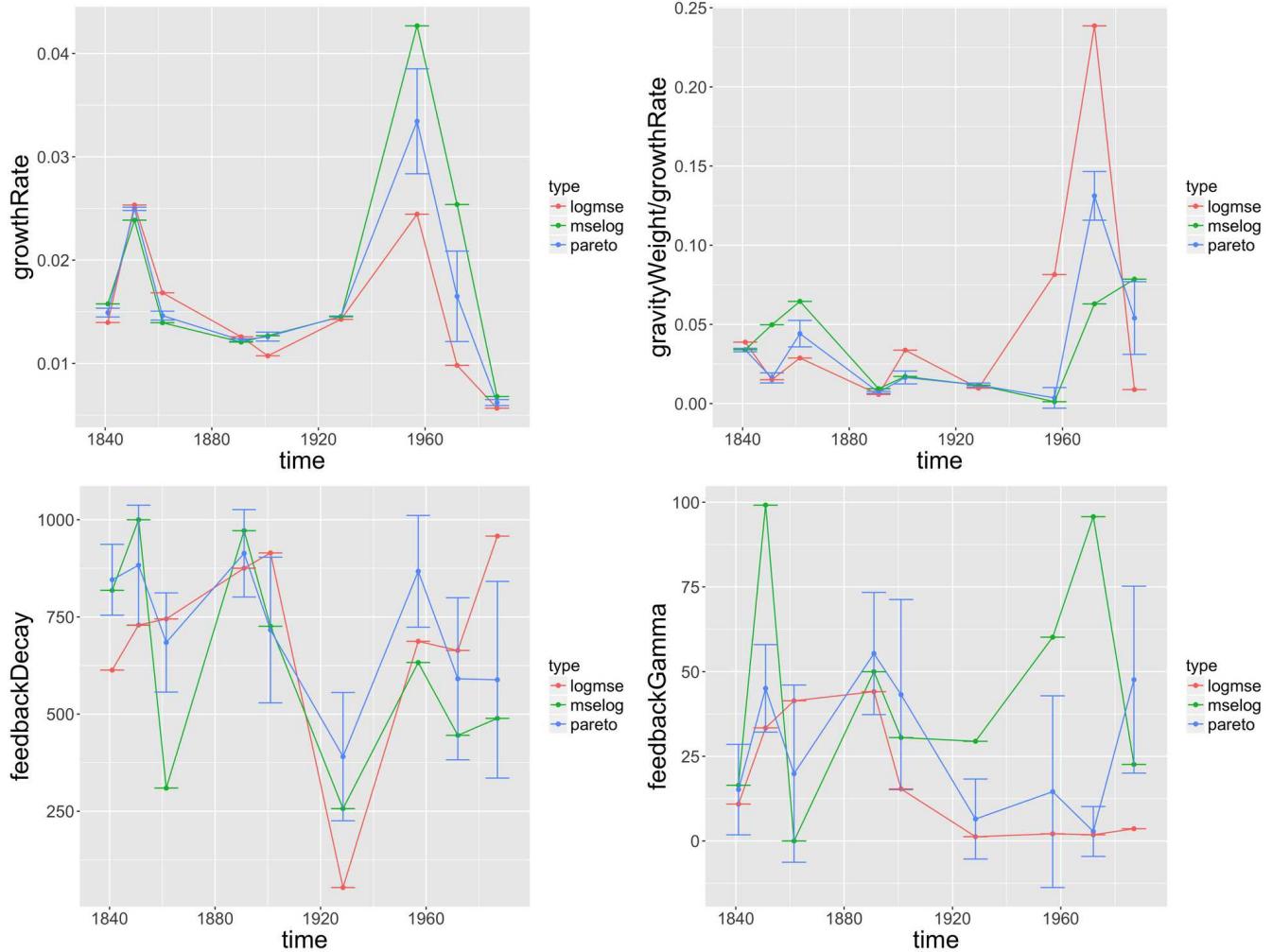


FIGURE 30 : Paramètres ajustés pour le modèle complet. Nous donnons les valeurs de r_0 , w_G/r_0 , d_N et γ_N dans le temps, pour les points optimaux pour les objectifs simples (courbes rouge et verte) et moyen sur le front de Pareto (bleu).

ESTIMER LE COMPROMIS ENTRE PUISSANCE D'AJUSTEMENT ET NOMBRE DE PARAMÈTRES Nous visons dans cette dernière expérience à quantifier la “performance” du modèle, prenant en compte ses capacités prédictives, mais aussi sa structure. Plus précisément, nous voulons traiter la question de la surestimation (*overfitting*), qui a été reconnue depuis un certain temps en Apprentissage Statistique par exemple [[dietterich1995overfitting](#)], mais pour lequel il manque des méthodes applicables aux modèles de simulation³⁹. Nous avons besoin d'introduire un outil qui confirme que l'amélioration de l'ajustement n'est pas uniquement artificiellement due aux paramètres supplémentaires. Cette étape est importante à ce stade d'introduction de modèles préliminaires : il s'agit en effet de construire des briques pertinentes mais également relativement simples.

Le critère d'information d'Akaike (AIC) fournit pour les modèles statistiques pour lesquels une fonction de vraisemblance est disponible le gain d'information entre deux modèles [[akaike1998information](#)], corrigeant l'amélioration du fit par le nombre de paramètres. Des méthodes similaires incluent le critère d'information Bayesien (BIC), qui repose sur des hypothèses légèrement différentes et corrige différemment. [[biernacki2000assessing](#)] propose une likelihood intégrée comme une généralisation de ces critères pour la classification non-supervisée. [[2017arXiv170108673P](#)] montre que dans le cas de la sélection du nombre d'états pour des Modèles de Markov Cachés, les cas réels induisent trop d'embûches pour que les méthodes standard fonctionnement de manière robuste, et suggèrent une sélection pragmatique basée sur leurs résultats et le jugement d'expert. Dans notre cas, le problème est qu'il n'est même pas possible de les définir.

La méthode que nous proposons est basée sur l'idée intuitive d'approcher les modèles de simulation par des modèles statistiques et d'utiliser l'AIC correspondant sous certaines conditions de validité. [[2017arXiv170609773B](#)] utilise une astuce similaire de considérer les modèles comme des boîtes noires et de les approcher pour gagner de l'information, dans leur cas pour extraire une structure interprétable sous forme d'arbres de décision. Soit (X, Y) les données initiales et les observations des réalisations. Nous considérons les modèles computationnels comme des fonctions $(X, \alpha_k) \mapsto M_{\alpha_k}^{(k)}(X)$ faisant correspondre les valeurs des données à une variable aléatoire. Ce qui est vu comme données et comme paramètres est dans une certaine mesure arbitraire mais est séparé dans la formulation puisque les dimensions correspondantes auront des rôles différents. Nous supposons

³⁹ Cette question est en fait au cœur de la compréhension de la complexité : il s'agit de trouver une dimension qui capture la plus grande partie des dynamiques du système à un niveau donné, ce qui revient à déterminer un niveau de simplification qui capture les dynamiques agrégées, et donc l'émergence. Il pourrait d'ailleurs exister un lien entre cette question et le problème de la détermination de l'*embedding dimension* pour les séries temporelles, qui revient à trouver une dimension effective de l'espace des phases d'un système.

que les modèles ont été ajustés aux données au sens où une heuristique a été utilisée pour trouver une solution optimale approximative $\alpha_k^* = \operatorname{argmin}_{\alpha_k} \|M_{\alpha_k}^{(k)}(X) - Y\|$, et nous écrivons $\varepsilon_k = \|M_{\alpha_k}^{(k)}(X) - Y\|^2$ l'erreur carrée moyenne correspondante. Pour chaque modèle computationnel optimisé, un modèle statistique $S^{(k)}$ avec le même degré de liberté peut être ajusté sur un ensemble de réalisations : $M_{\alpha_k^*}^{(k)}(X) = S^{(k)}(X)$, avec une erreur $s_k = \|M_{\alpha_k^*}^{(k)}(X) - S^{(k)}(X)\|^2$. Si les modèles statistiques sont de bonnes approximations des modèles en comparaison de la distance des modèles à la réalité, c'est à dire si $s_k \ll \varepsilon_k$, alors le gain d'information entre les deux devrait majoritairement capturer le gain d'information entre les modèles de simulation. Nous définissons ainsi une mesure d'AIC *empirique* entre deux modèles de simulation par

$$I(M^{(1)}, M^{(2)}) = \Delta AIC [S^{(1)}, S^{(2)}] \quad (7)$$

En pratique, nous calibrons le modèle de gravité seul et le modèle complet sur la période temporelle complète, et choisissons deux solutions intermédiaires donnant $M^{(1)}$ à $r_0 = 0.0133, d_G = 4.02e12, w_G = 1.28e-4, \gamma_G = 3.82$ avec $\varepsilon_G = 31.2375, \varepsilon_L = 302.89$ et le modèle complet $M^{(2)}$ à $r_0 = 0.0128, d_G = 8.43e14, w_G = 1.230e-4, \gamma_G = 3.81, w_N = 0.60, d_N = 7.47e14, \gamma_N = 1.15$ avec $\varepsilon_G = 31.2366, \varepsilon_L = 302.93$. Il n'est pas clair dans quelle mesure la méthode empirique est sensible au type de modèle statistique utilisé, nous utilisons pour cela un certain nombre pour la robustesse, à chaque fois avec les nombre de paramètres correspondants (4 pour le premier et 7 pour le second modèle) : un modèle polynomial de la forme $a_0 + \sum_{i>0} a_i X^i$, une mixture de logarithme et polynôme comme $a_0 + a_1 \ln X + \sum_{i>1} a_i X^i$ et un polynôme généralisé avec des exposants réels qui ont été optimisés pour la performance du modèle par utilisation d'un algorithme génétique $a_0 + \sum_{i>0} a_i X^{\alpha_i}$. Nous ajustons les modèles statistiques en utilisant les années successives comme des réalisations différentes. Les résultats pour chaque sont donnés en Table 11. Nous donnons les valeurs de s_k/ε_k et le ΔAIC . Nous donnons aussi le ΔBIC pour vérifier la robustesse au regard du critère d'information utilisé. Nous trouvons une valeur positive pour 5 critères sur 6, ce qui signifie que le gain d'information est effectivement positif. Le gain décroît quand la performance du modèle statistique augmente, et seul le BIC pour le modèle optimisé échoue à montrer une amélioration. L'hypothèse des erreurs négligeables est toujours vérifiée puisque le taux est toujours autour de 1%. Cette approche est bien sûr préliminaire et des développements supplémentaires seraient nécessaires pour un test plus systématique et une justification plus robuste de la méthode. Cela suggère cependant que l'amélioration de fit dans le modèle de simulation sont effectifs, et que le modèle révèle par cela des effets de réseau.

TABLE 11 : Valeurs de l'AIC empirique.

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 Vers des modèles co-évolutifs

Rappelons le positionnement de l'étude que nous venons de mener par rapport à notre objectif général. Notre compréhension des effets de réseau reste ici assez limitée puisque (i) nous ne considérons pas une infrastructure réelle mais uniquement des flux abstraits, et (ii) nous ne prenons pas en compte la possible évolution du réseau, due aux progrès techniques [**bretagnolle200olong**] et à la croissance de l'infrastructure dans le temps. Un développement intéressant sera d'abord l'application du modèle sur des données réelles de réseau, en utilisant les matrices de distance réelles dans le temps, calculées e.g. avec le réseau ferré utilisé par [**thevenin2013mapping**]. Ensuite, permettre au réseau d'évoluer de manière dynamique dans le temps, comme fonction des flux, produira un modèle de co-évolution entre les villes et les réseaux de transport pour un système de villes, qui a été prouvée empiriquement par [**bretagnolle:tel-00459720**]. Ce type de modèle est très rare, et [**schmitt2014modelisation**] fournit avec SimpopNet l'un des exemples. Il est montré dans la section 2.2 que la séparation des disciplines pourrait être à l'origine de l'absence relative de tels types de modèles dans la littérature. En effet, cela impliquerait d'inclure des processus hétérogènes comme des règles économiques pour régir la croissance du réseau, qui sont assez loin de l'approche prise. Cela permettrait cependant d'investiguer dans quelle mesure le raffinement de la structure spatiale du réseau et des dynamiques de réseau peut améliorer l'explication des dynamiques des systèmes urbains. La pertinence d'un tel développement est confirmée par les approches empiriques, comme [**dupuy1996cities**] qui montre le rôle de la position des villes dans le réseau autoroutier Européen pour leur relations respectives et leur compétitivité.

Nous avons introduit un modèle spatial de croissance pour un système de villes à l'échelle macroscopique, incluant des effets de réseau au second ordre avec la croissance endogène et les interaction directes comme moteurs de la croissance. Le modèle est initialisé sur les données réelles du système de villes français entre 1831 et 1999. La calibration du modèle dans le temps fournit des interprétations pour l'évolution des processus d'interaction dans le système de villes. Nous montrons de plus que le modèle révèle effectivement des effets de réseau en contrôlant l'overfitting. Ce travail ouvre la voie pour des

modèles plus compliqués avec des réseaux dynamiques, qui capturentraient la co-évolution entre les réseaux de transport et les territoires, qui seront développés au Chapitre 6.

★ ★

★

CONCLUSION DU CHAPITRE

La notion de co-évolution, qui était jusqu'ici relativement conceptuelle, apparaît sous de multiples angles nouveaux complémentaires. Ce chapitre permet d'éclairer son rôle au sein de la Théorie Evolutive. Celle-ci sera également centrale pour la construction théorique que nous élaborerons en 9.2. 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 pu permettre 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. La question de la non-stationnarité et de la non-ergodicité dans les systèmes urbains est cruciale mais très peu comprise, et nous l'avons à peine effleurée. Dans notre cas, l'aspect le plus important de celle-ci pour la construction des modèles est son implication pour les échelles considérées, et les hypothèses d'équilibre ou de stochasticité correspondantes. On y reviendra par un point de vue différent en Chapitre 5.

* * *

*

MORPHOGENÈSE URBAINE

L'importance des relations spatiales et de la mise en réseau est bien établie en géographie, comme le prouve par exemple la "première loi de la géographie" de TOBLER [tobler2004first]¹. 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 à l'idée de sous-système relativement indépendant. Il serait alors possible d'isoler certaines règles locales régissant un sous-système, étant par ailleurs 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. Reprenons un exemple concret de terrain déjà évoqué en 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 différentes 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 défective ont pu subir des répercussions politiques et économiques bien plus conséquentes. D'autre part, 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. 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. On supposera à la fois une certaine forme de stationnarité locale ("régime urbain local") mais aussi une certaine indépendance avec l'extérieur. Nous pouvons également noter que dans ce cadre, son auto-organisation locale impliquera nécessairement des relations fortes entre forme et fonction, de par la distribution spatiale des fonctions urbaines mais aussi car *la forme fait la fonction* dans certains cas de figure, au sens des motifs d'utilisation entièrement conditionnés à cette forme. Le type de raisonnement que nous avons esquissé mobilise les éléments essentiels propres à l'idée

¹ "Tout intéragit avec tout, mais deux choses plus proches auront plus de chances d'intéragir."

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. La morphogenèse, qui a été importée de la biologie vers de nombreux champs, a dans chaque cas ouvert des voies pour l'étude des systèmes complexes propres à ce champ selon un point particulier. Il est important de noter que le monument qu'est la Théorie des Catastrophes de RENÉ THOM introduit une façon originale de comprendre la différentiation qualitative et donc la morphogenèse. Cette théorie a toujours un potentiel d'application considérable aux problèmes qui nous concernent, comme l'a suggéré DURAND-DASTÈS [[durand2003geographes](#)] en évoquant la systémogenèse, que nous développerons en ouverture. 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 l'ensemble des formes d'établissements humains en Europe, en utilisant les résultats de [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 module de morphogenèse de réseau dans la section [5.3](#), afin d'établir un espace possible des correlations 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 [[antelope2016interdisciplinary](#)]; la deuxième section est traduite de [[raimbault2017calibration](#)]; et enfin la troisième section a été écrite pour les Actes des Journées de Rochebrune 2016 [[raimbault2016generation](#)].

5.1 UNE APPROCHE INTERDISCIPLINAIRE DE LA MORPHOGENÈSE

Une première étape essentielle est la clarification de ce qui est entendu par le terme de morphogenèse. La démarche prise ici est voulue *interdisciplinaire*, au sens où elle se pose comme objectif de construire une connaissance synthétique à partir des disciplines abordées², et se veut donc à la fois large (disciplines couvertes), profonde (profondeur dans chaque discipline³) et synthétique (intégration en résultant).

Initialement introduit en biologie, son transfert à d'autres champs s'est accompagné d'une déformation des concepts associés. Nous adaptons et traduisons ici le texte de [antelope2016interdisciplinary] qui propose une entrée interdisciplinaire sur la morphogenèse. Brique essentielle de nos constructions, il est en effet crucial de lui donner une armature rigoureuse et claire. Nous prenons le parti d'une vision croisée, dans l'idée d'un perspectivisme appliqué comme introduit en section 3.3, pour obtenir des concepts aussi génériques et larges que possible.

La notion de morphogenèse semble jouer un rôle important dans l'étude d'une large gamme de systèmes complexes. Si le concept a été introduit initialement en embryologie pour désigner la croissance des organismes, il a été rapidement utilisé dans différentes disciplines, e.g. l'urbanisme, la géomorphologie, et même la psychologie. Toutefois, l'utilisation du concept semble généralement floue et avoir une définition spécifique à chaque champ pour chacune de ses utilisations. Nous menons dans cette section une étude épistémologique, commençant par une revue interdisciplinaire large puis en extrayant les notions essentielles liées à la morphogenèse dans chaque champ. Cela permet de construire un meta-cadre général consistant pour la morphogenèse. Des applications peuvent inclure une application concrète du cadre sur des cas particuliers pour opérer un transfert interdisciplinaire de concepts, et des analyses quantitatives de texte pour renforcer ces résultats qualitatifs.

CONTEXTE Durant chaque période historique, l'avancée technologique principale a été utilisée comme une métaphore pour expliquer d'autres phénomènes de la nature. D'abord, la nature a été mécanique, puis électrique, et à présent computationnelle. Ici, nous suggérons qu'une métaphore alternative peut permettre de mieux étudier les propriétés d'un système, et ainsi comprendre comment le concept de morphogenèse qui a trouvé son origine en biologie du développement, peut être utilisé pour d'autres types de systèmes.

² Nous nous différentions ici de l'effort concernant la co-évolution mené en 3.3, qui proposait une revue multidisciplinaire mais construisait une définition propre aux systèmes territoriaux. Le travail ici est issu d'une collaboration interdisciplinaire et s'inscrit dans une logique plus intégrative

³ Bien sûr pas au sens d'un compte rendu des fronts de connaissance actuels, mais d'une présentation non vulgarisée.

La morphogenèse est une métaphore très puissante qui est bien distincte des trois précédentes qui ont été très populaires dans l'histoire. Contrairement aux explications mécaniques, électriques ou computationnelles de la nature, la morphogenèse n'est pas un processus conçu par l'homme. La morphogenèse met l'accent sur le rôle du changement et de la croissance, plutôt qu'un état statique. Comme [thompson1942growth] mentionnait déjà, "l'histoire naturelle traite de l'éphémère et les accidents, pas par des choses éternelles ou universelles". Le but de notre exercice est de répondre à trois questions : (i) comment la morphogenèse est définie dans différents champs ; (ii) existe-t-il des champs qui utilisent des approches et concepts incluant la notion de morphogenèse mais sans utiliser le terme ; (iii) dans quelle mesure les approches étudiant la morphogenèse peuvent-elles être transférées entre les champs ? Un effort similaire a été mené par [bourgine2010morphogenesis] mais consiste plus en une collection de points de vue de sujets liés à la morphogenèse plutôt qu'une reconstruction épistémologique de la notion comme nous proposons de faire. De plus, les exemples sur ce sujet sont loin d'être épuisés et notre revue est pour cela complémentaire.

La suite de cette section est organisée de la façon suivante : nous produisons d'abord une revue autonome de la notion de morphogenèse pour différents champs, s'étendant de la biologie aux sciences sociales, la psychologie et les sciences territoriales. Une synthèse est ensuite faite et un cadre aussi général que possible proposé. Nous discutons finalement des développements futurs et des applications potentielles de cette analyse épistémologique.

5.1.1 *Revues*

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.

Biologie du Développement

En biologie du développement, la morphogenèse fait référence aux mécanismes conduisant un organisme à acquérir sa forme et différentes unités fonctionnelles, en partant d'une unique cellule. De manière générale, ces mécanismes doivent être fiables pour garantir une issue similaire pour chaque individu. Cela suppose que les cellules connaissent leur position par rapport à un cadre de référence afin de se différencier, c'est à dire prendre une fonction particulière, ou pour décider si elles doivent se diviser ou non, ce qui est une étape cruciale lors de la croissance. Nous décrivons par la suite les modèles qui ont été appliqués en biologie du développement.

MÉCANISMES DE RÉACTION-DIFFUSION Le terme de réaction-diffusion avait été utilisé par ALAN TURING dans son article séminal de 1952 [turing1952chemical], pour décrire l'émergence de motifs dans un anneau théorique de cellules. Bien que ce travail soit aujourd'hui reconnu comme l'une des contributions les plus fondamentales dans le champ de la formation de motifs, il a fallu des années pour qu'il trouve une reconnaissance comme modèle effectif pour les systèmes biologiques. [gierer1972theory] a plus tard suggéré d'utiliser des modèles similaires pour expliquer la polarité intracellulaire, qui correspond à la capacité d'une cellule à différencier des zones dans son intérieur. Ces réseaux de réaction-diffusion sont un exemple de l'émergence de motifs à partir d'un état homogène, parmi d'autres comme la coloration ou la segmentation. Ces motifs à grande échelle sont générés par l'interaction entre un petit nombre d'espèces chimiques, chacune suivant une diffusion, une production et une dégradation. Il est ainsi possible d'utiliser des systèmes d'équations aux dérivées partielles, pour lesquelles certains paramètres généreront des motifs stables à partir de conditions initiales homogènes, où les perturbations aléatoires sont amplifiées par le système. Des motifs complexes peuvent être produits à partir d'un nombre très restreint d'espèces [kondo2010reaction]. L'une des réactions capables de produire des motifs stables les plus étudiées comporte deux types de molécules, un activateur et un répresseur. La différence dans le taux de diffusion entre les deux molécules est responsable de l'amplification du bruit dans le système [gierer1972theory]. Le système à l'origine d'une coloration le plus étudié sont les réactions responsables des rayures jaunes et noires du poisson zèbre [nakamasu2009interactions]. L'émergence de la polarité cellulaire est expliquée chez certaines levures par un mécanisme similaire [goryachev2008dynamics]. Des exemples impliquant des fonctions comme la segmentation du corps de *Drosophila melanogaster* impliquent des réseaux d'espèces chimiques bien plus complexes pour assurer la robustesse de l'émergence de ces fonctions.

LE MODÈLE FRENCH FLAG De façon similaire, le modèle French Flag a été conçu initialement pour expliquer la différentiation des cellules de manière régulière [Wolpert1969]. Le modèle suppose un gradient de concentration d'une protéine, généralement appelée le morphogen, auquel les cellules d'un tissu réagiront différemment selon leur niveau (d'où les rayures du drapeau). Un tel gradient doit être produit par une diffusion, à partir d'une source, complété par un mécanisme de stabilisation impliquant un puits ou une dégradation locale dans le tissu (mécanismes qui sont passés en revue par [Rogers2011]). Le gradient peut ensuite être utilisé localement de manière linéaire (l'expression d'un gène variant de manière linéaire par exemple) ou par seuils grâce à des boucles de retroaction locales. D'après [Wolpert2011], aucun de ces systèmes n'est parfaitement bien

compris, mais les preuves empiriques de leur existence sont claires à une granularité assez grande. Les expériences nécessaires pour leur vérification exacte sont en effet très difficiles et encore hors de portée pour la plupart.

FORCES INTER-CELLULAIRES Les réagencements cellulaires sont souvent conduits par des forces physiques intracellulaires [Heisenberg2013], qui sont ensuite transmises entre cellules, par des jonctions intercellulaires modulables. Ce phénomène peut conduire à un comportement quasi-fluidique lorsqu'un stress extérieur est appliqué pour une certaine durée. A de plus petites échelles temporelles, les cellules gardent cependant un comportement élastique et gardent leur forme lorsque aucune force extérieure n'est appliquée. Pour que le tissu change de forme, ont lieu des divisions, morts, extrusions ou intercalages de cellules [Guillot2013]. Un exemple de dynamique de tissu bien étudiée est présent chez *Drosophila melanogaster* également. Dans ce cas, des cellules formant initialement une couche plate deviennent un long sillon en contractant la membrane cellulaire d'un côté [Lecuit2007].

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 [tero2010rules]⁴.

Artificial Life

La notion de *Programmable Self-Assembly* semble être en *Artificial Life*⁵ très proche du concept biologique de morphogenèse : [crosato2014self] note dans une large revue que "le meilleur exemple de *Programmable Self-Assembly* dans la nature est probablement l'organisation des cellules en organismes multi-cellulaires, qui est encodée par l'ADN". Une approche importante dans ce champ est le concept de *Morphogenetic Engineering* introduit par DOURSAT, qui se concentre sur la conception de systèmes complexes par le bas. Une revue du champ est faite dans [doursat2013review]. Une distinction essentielle entre auto-organisation et morphogenèse qui y est introduite est la présence d'une *architecture*, au sens d'une structure macroscopique bien discernable ayant des propriétés fonctionnelles (mais que nous ne considérerons pas nécessairement télééconomique [monod1970hasard] pour garder un certain niveau de généralité). Un exemple d'une nuée hétérogène de particules, produisant des architectures complexes, est

⁴ Voir aussi une initiative récente de biologiste 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/>.

⁵ Au sens du domaine scientifique propre, animé par exemple par la *International Society for Artificial Life* (voir <http://alife.org/>).

décrit dans [doursat2008programmable]. Les processus d'interactions locales (correspondant en biologie aux forces physiques locales) et l'information de position par la propagation du gradient sont tous deux intégrés dans le modèle et permettent l'émergence par le bas de motifs complexes. [sayama2009swarm] développe des modèles similaires en y ajoutant la possibilité d'évolution des espèces de particules, dirigée de manière interactive par le modélisateur, ce qui permet de effectivement orienter l'architecture émergente. Dans quelle mesure ces systèmes artificiels sont proches de systèmes vivants est une question ouverte : [Schmickl_2016] exhibent des règles de mouvement similaires qui conduisent à l'émergence de structures aux propriétés de reproduction, avec différentes fonctions dans un écosystème propre, qu'ils qualifient de "imitant la vie"⁶. L'ajout d'un environnement avec ses propriétés propres influence fortement les dynamiques morphogénétiques, comme montré par [cussat2012synthesis] qui combine une couche de réaction chimique avec une couche hydrodynamique dans laquelle cette première prend place. L'application de ces méthodes à des questions concrètes d'ingénierie commence à être développée avec des résultats prometteurs : [Aage:2017aa] utilise un modèle de morphogenèse pour la conception de la structure interne d'une aile d'avion et obtient des gains de masse allant jusqu'à 5%, et une structure finale très proche de formes aux fonctions similaires dans la nature comme une aile d'oiseau. Dans ce dernier cas, le biomimétisme est émergent, les processus de morphogenèse faisant le lien.

Sciences Territoriales

Le concept est utilisé dans de nombreuses disciplines s'intéressant aux territoires et à l'environnement bâti : géographie, planification et design urbains, urbanisme, architecture. Il ne semble pas exister de vue unifiée ni de théorie entre les champs ni dans chaque champ lui-même.

ENVIRONNEMENT BÂTI L'architecture et l'urbanisme sont des disciplines étudiant les établissements humains et l'environnement bâti à des échelles généralement grandes⁷. La théorie du Métabolisme Urbain de OLSEN [olsen1982urban] relie la morphogenèse de la ville à son métabolisme et à l'écologie urbaine. La ville est vue comme une organisme vivant avec différentes échelles de temps d'évolution (les cycles de vie). L'étude de la Morphologie Urbaine [moudon1997urban], qui s'intéresse aux processus morphogénétiques, est présenté comme un champ émergent en lui-même, à l'interface de la géographie, l'ar-

⁶ Nous développerons plus en détails plus loin les concepts nécessaires pour creuser cette affirmation.

⁷ nous n'incluons pas l'aménagement du territoire, mais considérons bien le contexte de projets urbains qui ne dépassent jamais l'échelle métropolitaine

chitecture et la planification urbaine : cette vision appuie sur le rôle crucial de la forme dans ce genre de processus. [burke1972dublin] étudie la croissance d'une ville particulière (Dublin) durant une période temporelle donnée, et attribue l'évolution de la morphologie urbaine aux *agents morphogénétiques*, i.e. les habitants et les promoteurs. A une autre échelle, en architecture, un bâtiment peut être vu comme le résultat de processus microscopiques ayant un sens propre, et un style architectural particulier peut être interprété par l'utilisation d'une grammaire générative de formes [ceccarini2001essai]. Cette méthodologie se rapproche du travail de C. ALEXANDER, un architecte ayant produit une théorie des processus de design [mehaffy2007notes], inspirée de l'informatique et de la biologie et liée par certains aspects à la complexité. La notion de morphogenèse est dans ce cas cependant assez floue, puisqu'elle se rapporte au processus de la génération de forme en général, de la même façon que [whitehand1999urban] étudie les changements concrets dans la forme des maisons comme un témoin de la morphogenèse urbaine, montrant par exemple que les quartiers de plus forte densité étaient plus susceptibles à la contagion des adaptations mineures par les habitants. DOLLENS fait référence à l'autopoïèse [dollens2014alan]⁸, impliquant un cas particulier de morphogenèse, pour défendre l'influence de TURING sur la pensée contemporaine en design, et pour proposer une approche plus organique de l'architecture. [desmarais1992premisses] soutient que les structures humaines sont porteuses d'une morphologie abstraite, et que celle-ci est générée par des processus porteurs de sens, rejoignant la conception de [ceccarini2001essai]. Cela fait écho aux usages de la morphogenèse en psychologie comme nous verrons plus loin : l'élaboration de la forme concrète va alors de pair avec le processus cognitif qui est lui-même une morphogenèse. [levy2005formes] soulève la difficulté d'une définition propre du terme de forme urbaine, et propose de le revisiter en liant la production de la forme à celle du sens dans l'ensemble de la dynamique du système. Ce positionnement rejoint partiellement celui que nous prendrons plus loin pour définir la morphogenèse.

MODÉLISATION URBAINE La littérature de modélisation de la croissance urbaine se réfère souvent au processus de croissance comme morphogenèse quand l'échelle impliquée permet de révéler des motifs de forme. Un exemple de l'émergence de fonctions urbaines qualitativement différenciées, basé sur le modèle d'Alonso-Mills-Muth, est proposé dans [bonin2012modele]. [bonin2014modelisation] étend également ce modèle standard d'Economie Urbaine et fait directement référence aux *morphogens*, substance qui diffusent et réagissent dans

⁸ Le concept d'*autopoïèse* sur lequel nous reviendrons plus en détail par la suite, consiste principalement en la capacité d'un système de s'auto-entretenir comme réseau de processus autonome.

la formulation initiale de la morphogenèse par TURING. Ce modèle prend en compte la population, les prix immobiliers, les surfaces de logement, les emplois et des aménités endogènes, et simule l'évolution de leur distribution spatiale et de la forme urbaine résultante.

[makse1998modeling] étudie un modèle de croissance urbaine impliquant la forme urbaine locale. Dans ce cas les corrélations spatiales locales induisent la structure urbaine quand les villes gagnent de nouveaux habitants. Des modèles plus hétérogènes impliquent un couplage entre les composantes urbaines et les réseaux de transport. [achibet2014model] décrit un modèle de co-évolution entre réseau de rues et la structure des blocs urbains. A une plus petite échelle et impliquant des fonctions plus abstraites, [raimbault2014hybrid] couple croissance urbaine et croissance de réseau, incluant une rétroaction locale de la forme par une contrainte de densité et une rétroaction globale de la position par la centralité de réseau et l'accessibilité aux aménités. Ces deux mécanismes sont analogues aux interactions locales et à la diffusion du flux d'information global en biologie.

ARCHÉOLOGIE La morphogenèse des établissements humains du passé, vue du point de vue de la Théorie des Catastrophes de THOM, est introduite dans [renfrew1978trajectory]. Des changements soudains (changement qualitatifs, ou changements de régime) se sont produits à toute époque et peuvent être vus comme des bifurcations durant le processus de morphogenèse. Une autre manière simplifiée de le comprendre est d'interpréter la transition comme un changement des meta-paramètres d'une dynamique stationnaire.

Sciences Sociales et Psychologie

La morphogenèse a été occasionnellement utilisée comme une métaphore efficace pour comprendre différents processus en sciences sociales et dans divers champs de la psychologie. En psychologie du développement par exemple, l'influence des processus d'apprentissage culturel sur le comportement sont une bonne illustration [hart_held_2013]. Pour la psychologie clinique, des analogies sont utilisées pour l'auto-organisation des relations avec le Moi et l'Autre, ainsi que pour les dynamiques impliquant l'émergence créative, qui doit être encouragée pour une psychothérapie "aboutie" [piers_self-organizing_2007]. D'autres part, en neurosciences, la structure du cerveau en elle-même et la mise en place des réseaux de neurones est typiquement l'issue de processus morphogénétiques [_issues_2013]. En psychologie sociale, la co-évolution de l'individu et de la société peut également être vu par ce prisme [archer_margaret_1999]. La théorie de RENÉ THOM que nous détaillerons plus loin a certainement joué un rôle dans l'utilisation de ce concept en psychologie [de_luca_pacione_processes_2016]. Toutefois, au delà d'une unité systématique au travers de ces différents champs, les usages sont plutôt discontinus, et on pourrait sup-

poser que l'utilité du concept de morphogenèse réside plutôt dans sa portée épistémologique. Celle-ci consisterait dans une perception partagée du pouvoir descriptif de la morphogenèse pour mieux comprendre l'émergence de la structure des divers phénomènes.

Histoire de la notion

L'étude de la morphogenèse a démarré avec l'embryologie juste avant les années 30. Il s'agit environ de la même période à laquelle les mouvements cellulaires de bactéries ont été découverts [abercrombie1977concepts]. Les statistiques issues de Google Books donne le premier usage du mot dans un livre en 1871. L'usage montre ensuite un pic d'utilisation entre 1907 et 1909, pour continuer d'augmenter jusqu'en 1990 avant de décroître progressivement.

Mise en 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.

EPISTÉMOLOGIE La morphogenèse peut aussi être utilisée pour étudier la science elle-même : par exemple [gilbert2003morphogenesis] étudie l'évolution de la biologie évolutionnaire du développement par la métaphore de la morphogenèse. Il voit les idées scientifiques comme des agents en interaction, desquels émergent de nouveaux phénotypes par des processus de différentiation, qui sont désignés comme la morphogenèse du champ.

UNE APPROCHE MATHÉMATIQUE René Thom a développé dans *Stabilité Structurelle et Morphogenèse* [thom1974stabilite] une théorie de la dynamique des systèmes, la théorie des catastrophes, qui étudie en profondeur l'impact de la structure topologique des variétés de l'espace des phases sur les dynamiques du système. Soit M une variété différentielle, dans laquelle l'état du système (m, \dot{m}) est embarqué. On suppose l'existence d'un ensemble fermé K appelé *Ensemble de Catastrophe*. Le type topologique de K est en fait déterminé de manière endogène par la dynamique du système (dans les cas simples, il réfère au types "classiques" d'attracteurs/points fixes que l'on connaît habituellement : points et cycles limites). Quand m traverse K , le système rencontre un changement *qualitatif* dans sa forme, ce qui constitue la base de la *morphogenèse*. Cette théorie abstraite de la morphogenèse est indépendante de la nature du système étudié, sa contribution principale étant de classifier les catastrophes locales qui surviennent lors de la morphogenèse. La différentiation et la richesse des motifs ont ainsi une explication géométrique à travers les types

topologiques des catastrophes. THOM note qu'à cette époque, l'étude de la forme a majoritairement été ciblée par la biologie, mais que de nombreuses applications pourraient être développées en physique et géomorphologie par exemple. Il formule l'hypothèse que parce que cela implique des discontinuités et de l'auto-organisation, à laquelle les mathématiciens étaient réticents, que cela n'a pas été appliqué facilement à divers champs. Nous pouvons lier cela à l'émergence des approches complexes, avec des paradigmes de la complexité qui se sont progressivement répandus dans diverses disciplines, et l'étude de la morphogenèse semble avoir suivi.

Les mathématiques, peu mentionnées dans notre revue, sont toutefois concernées à la fois comme outil mais comme discipline à part entière, les constructions mathématiques obtenues à partir des questions liées à la morphogenèse sont des sujets de recherche à part entière. Comme l'a récemment rappelé CEDRIC VILLANI [villani2017chauvesouris], *"la morphogenèse est une discipline pas très bien identifiée ayant toujours un certain nombre de mystère, à l'intersection entre les mathématiques, la chimie et la biologie, (...) où des modèles mathématiques jouent un rôle pour faire émerger les structures"*.

AUTOPOIÈSE ET MORPHOGENÈSE

AUTOPOIÈSE Le concept d'autopoïèse, qui provient initialement de la biologie, est lié à la morphogenèse de manière intriquée. Dans le cas des systèmes psychologiques, il fournit une interprétation dépendante de l'observateur de la cognition et de la conscience. Celle-ci a eu des impacts en psychologie et sociologie, comme certaines théories des systèmes [gershenson_requisite_2014]. Les systèmes sociaux et psychiques sont alors compris comme des systèmes fortement couplés, comme le témoigne le langage qui est un phénomène social profondément ancré dans les manifestations cognitives [seidl_luhmanns_2004]. Ces approches rejoignent également les visions du sujet comme dynamique et récursif [pichon_riviere_processus_2004]. L'interpénétration du social et du psychologique trouvent echo chez l'anthropologie psychanalytique de FREUD qui appuie les relations entre les symptômes neurotiques et les phénomènes socio-culturels [freud_totem_1989].

Dans son approche biologique, il exprime la capacité d'un système à s'auto-reproduire. Une caractérisation rudimentaire est l'existence d'une frontière semi-perméable produite par le système et la capacité à reproduire ses composants. Une définition plus générale est proposée par [bourgine2004autopoiesis]⁹. La notion de processus dynamique est une notion clé, et pourrait être liée à la théo-

⁹ *"Un système autopoïétique est un réseau de processus qui produit les composants permettant de reproduire le réseau, et qui régule également les conditions au bord nécessaire pour son existence continue en tant que réseau"*.

rie de la morphogenèse de THOM. Ils introduisent de plus une définition de la cognition (déclenchement d'actions en fonction d'entrées sensorielles pour assurer la viabilité), et d'un organisme vivant comme autopoïétique et cognitif, les deux notions étant bien distinctes [bitbol_{autopoiesis_2004}]. Dans ce cadre par exemple, l'arbotron [jun2005formation] est cognitif mais pas autopoïétique. Un exemple de lien entre autopoïèse et morphogenèse est montré dans [niizato2010model], où un type d'organisme Physarum doit jouer à la fois sur la mobilité des cellules et sur l'évolution de la forme pour être capable de collecter la nourriture nécessaire à sa survie. A cette étape, nous pouvons déjà postuler une inclusion stricte des systèmes autopoïétiques, aux systèmes morphogénétiques, aux systèmes auto-organisés.

CO-ÉVOLUTION La morphogenèse pouvant être transposée aux écosystèmes ou aux sociétés, dont les composantes sont en co-évolution dans ce cas, la présence d'une co-évolution pourrait être liée à la morphogenèse, comme une autre façon de voir le système. La symbiose en biologie peut mener à des causalités très fortes dans l'évolution de l'organisme (co-évolution) : ce phénomène a été désigné comme *symbiogenesis*. La symbiose induit un changement dans les motifs morphogénétiques des organismes symbiotiques comme montré pour différentes espèces par [chapman1998morphogenesis]. D'où un lien potentiellement fort entre morphogenèse et co-évolution : dans ce cas la morphogenèse est utilisée pour désigner plus des trajectoires évolutionnaires de motifs morphogénétiques, i.e. sur une échelle de temps différente.

DEFINITION ET FRONTIÈRES DU SYSTÈME 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 fois. La théorie des systèmes complexes adaptatifs de [holland2012signals] 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 l'ouverture et de la fermeture [morin1976methode].

5.1.2 *Synthèse*

Notions clés

Nous listons à présent les concepts importants découlant de cette revue, et dont une vision synthétique doit émerger. Chacun peut être

dépendant du domaine, et les conceptions sous-jacentes peuvent varier d'un champ à l'autre.

- **Auto-organisation** : la morphogenèse implique auto-organisation mais le contraire n'est pas nécessairement vrai (les deux concepts étant parfois confondus comme en géomorphologie [[cholley1950morphologie](#)]), certains aspects sont spécifiques à la morphogenèse, comme la présence de fonctions résultant de la forme.
- **Motifs et Forme** : "l'émergence de formes" semble être commun à toutes les approches de la morphogenèse.
- **Embryogenèse / modélisation des tissus** en biologie, les processus typiques de la morphogenèse sont généralement observés au stades initiaux de la vie, durant l'embryogenèse, incluant la formation initiale des tissus.
- **Apostosis** la morphogenèse est souvent liée à la vie (voir la section sur l'autopoïèse), mais aussi à la mort : la mort programmée de cellules, l'apoptose, peut dans certains cas faire partie de processus morphogénétiques.
- **Qualitatif vs Quantitatif** Les bifurcations qualitatives sont un concept fondamental pour la morphogenèse : e.g. la différentiation des organes en biologie ; l'émergence de fonctions urbaines différencierées.
- **Symétrie** Des ruptures de symétrie se produisent, majoritairement dans les étapes initiales, mais aussi à tous les stades de la morphogenèse.
- **Unité et Echelle** : les systèmes sont-ils conçus par le haut ou par le bas, auto-organisés, ou présentant une architecture ? Les deux ne sont pas nécessairement incompatibles, les unités fondamentales et les échelles jouant un rôle crucial dans la définition de la morphogenèse. Les systèmes semblables à des fractales, comme les coraux (tissus collaboratifs) ou les villes, mais aussi le sujet et la société peuvent être étudiés du point de vue des processus morphogénétiques à différents niveaux.
- **Frontières** : les frontières sont un aspect crucial pour l'étude des Systèmes Complexes Adaptatifs (voir par exemple l'approche de HOLLAND par *Signals and Boundaries* [[holland2012signals](#)]). La morphogenèse peut impliquer des frontières claires (d'un embryon e.g.) mais pas nécessairement (organismes sociaux, villes pour lesquelles la définition des frontières est toujours une question ouverte [[2015arXiv150707878C](#)]).
- **Relation entre forme et fonction** : les relations causales entre forme et fonction sont au centre de l'architecture émergente.

Processus communs et divergences

DES INTERACTIONS LOCALES AUX FLUX GLOBAUX D'INFORMATION Les imbrications des relations entre agents, soit par des effets de voisinage comme des interactions mécaniques et la diffusion, ou par des interactions de réseaux comme le signalement, et la retroaction d'un flux d'information global (i.e. une causation descendante du niveau supérieur) apparaît être commun à la majorité des utilisations de la morphogenèse. Cela souligne la nature fondamentalement multi-niveaux des processus morphogénétiques et le rôle central de l'émergence.

DE L'AUTO-ORGANISATION À LA MORPHOGENÈSE : LA NOTION D'ARCHITECTURE La plupart des systèmes étudiés semblent avoir la particularité de présenter une architecture, ce qui permettrait de faire la distinction entre auto-organisation et morphogenèse. Cette idée vient du champ du *morphogenetic engineering*, qui peut être vu comme un sous-champ de l'intelligence artificielle. Ce point peut être une divergence pour certains champs, comme par exemple en géographie physique où la "morphogenèse" de motifs d'érosion est une auto-organisation en notre sens. La notion d'architecture peut être difficile à définir. Une façon d'y parvenir est de considérer les fonctions des niveaux macroscopiques du système : l'émergence d'une fonction à un niveau supérieur implique une architecture, qui est *le lien entre la forme et la fonction*. Ici ce dernier concept prend tout son sens et son importance au regard de la morphogenèse.

Proposition d'un cadre meta-épistémologique

CADRE Nous proposons une imbrication hiérarchique des concepts, qui peut être vue comme un cadre meta-épistémologique, puisque les définitions sont construites de la synthèse des diverses disciplines évoquées ici, et que leur application dans chaque discipline particulière fournit un cadre épistémologique. Les concepts sont organisés de la façon suivante :

$$\text{Auto-organisation} \supsetneq \text{Morphogenèse} \supsetneq \text{Autopoïèse} \supsetneq \text{Vie} \quad (8)$$

chacun ayant une définition générique, élaborée de la synthèse des disciplines. L'inclusion stricte signifie qu'un concept implique l'autre mais qu'ils sont différents. L'ensemble des concepts est nécessaire pour bien situer la morphogenèse.

Definition : Auto-organisation. Un système est dit auto-organisé s'il exhibe une émergence faible [bedau2002downward].

Définition : Morphogenèse. Un système auto-organisé est le produit de processus morphogénétiques s'il présente une architecture émergente, au sens de relations causales circulaires entre forme et fonction à différents niveaux.

La *forme* est comprise comme *propriétés topologiques ou géométriques* d'un système ou de l'une de ses parties, tandis que la *fonction* est son rôle au sein des chaînes de processus, dans une perspective *téléonomique*¹⁰.

Définition : Autopoïèse et Vie. Nous prenons la définition de BOURGINE pour l'autopoïèse [bourgine2004autopoiesis], qui étend celle de BITBOL [bitbol_autopoiesis_2004], qui définit également la vie comme autopoïèse avec cognition.

La frontière entre auto-organisation et morphogenèse est l'existence de liens causaux entre forme et fonction, qui peut être définie comme une *architecture* [doursat2013review], qui sera généralement émergente. Nous observons que la complexité du système augmente avec la profondeur de la notion, ce qui peut être traduit de façon simplifiée par :

- La force de l'émergence [bedau2002downward] diminue avec la profondeur, au sens que le nombre d'échelles autonomes, ainsi que le nombre de propriétés aux pouvoir causaux irréductibles, augmentent.
- Le nombre de bifurcations augmente [thom1974stabilite], i.e. la dépendance au chemin augmente.

Ce deux propriétés peuvent être interprétées comme *l'une des caractérisations de la complexité* (voir 3.3).

APPLICATION Une spécification ontologique [livet2010ontology], i.e. la définition des entités à laquelle la notion s'applique, fournit une application à un champ donné, chaque champ développant ses propres propriétés et niveaux d'inclusion entre les concepts. Il n'existe a priori pas de raison pour une correspondance directe ou une équivalence entre les concepts projetés, ainsi le transfert de connaissances entre les domaines doit rester sujet à caution.

¹⁰ Au sens donné par MONOD dans [monod1970hasard], c'est à dire participant à répondre à un projet, à un but donné. Les êtres vivants sont téléonomiques au sens que l'ensemble de leur fonctions visent à finalement reproduire leur ADN. Une vision non *téléologique* de l'univers postule que celui-ci n'a pas de projet, et que la plupart des objets physiques ne rentrent pas dans cette catégorie. L'ensemble des autres cas d'étude que nous avons revu dans notre construction sont téléonomiques à différents niveaux : les systèmes territoriaux sont aménagés selon des logiques d'acteurs qui répondent à des projets; les systèmes de robots en *morphogenetic engineering* répondent à un besoin; les idées ou pensées participent à l'écosystème de l'esprit. Nous postulons ainsi cette nécessité téléonomique de la fonction pour avoir morphogenèse, position qui peut être discutée, comme en géomorphologie le réseau de rivières sera supposé avoir la fonction de drainer l'eau de pluie. Dans tous les cas une dichotomie claire entre morphogenèse en notre sens et auto-organisation ne pourra être distinctement établie, et un continuum correspond plus sûrement à la réalité (de la même manière que BEAU imagine un continuum entre émergence faible et émergence forte). En effet, dans une vision perspectiviste (voir 3.3), l'observateur joue un rôle essentiel dans la définition d'une fonction : le Jeu de la Vie utilisé comme ordinateur (par ses propriétés de Turing-complétude) sera morphogénétique, tandis qu'il sera auto-organisé s'il est simulé sans raison, rejoignant l'absurdité de la définition d'un *objet sans sujet* soulevée par MORIN dans [morin1976methode].

5.1.3 Discussion

VERS UNE CONSTRUCTION SYSTÉMATIQUE Ce travail repose pour l'instant sur une revue large mais non *systématique*, au sens de la méthodologie utilisée en évaluation thérapeutique par exemple, et où elle joue un rôle aussi important que les études primaires, une nouvelle connaissance étant créée par la comparaison systématique des résultats et la meta-analyse. Cela impliquerait dans notre cas une approche itérative, en utilisant de manière couplée les différents outils et méthodes développés en 2.2 :

- Une revue systématique aveugle, sans aucun a priori des champs concernés ou des moyens d'exprimer la notion.
- Extraction des champs principaux ; extraction des synonymes et notions proches (comme il a été fait ici avec l'autopoïèse et la *self-assembly* par exemple) ; si besoin itération de la première revue générale.
- Revue systématique spécifique à chaque champ, puisque chaque a ses propres bases bibliographiques, moyens spécifiques de communiquer, etc.
- Confrontation de chaque notion depuis un champ vers les autres

L'objectif dans notre cas serait d'enrichir, comme nous l'avons déjà fait de manière préliminaire, mais systématiquement, le concept de morphogenèse urbaine.

EPISTEMOLOGIE QUANTITATIVE Notre position peut également être renforcée par des approches quantitatives à l'analyse de la littérature. Avec la fouille de texte, l'extraction de mots-clés et de concepts à partir des résumés (ou même des textes complets) est possible, et devrait permettre de confronter notre analyse qualitative à la réalité empirique, en répondant à des questions telles que : un concept est-il central, ou quel concept est utilisé de la même façon dans la plupart des disciplines. [chavalarias2013phylomemetic] par exemple reconstruisent des champs scientifiques par le bas par une analyse textuelle, et étudie leur lignée et dynamique dans le temps. Une autre approche peut être la construction itérative des concepts, par une revue systématique algorithmique comme celle faite par [raimbault2015models].

Application potentielles

TRANSFERT DE CONNAISSANCES Les applications concrètes de ce cadre incluent un transfert potentiel de connaissance entre champs. Comme les systèmes biologiques inspirant l'architecture en *morphogenetic engineering*, ou comme l'usage des modèles gravitaires inspirés

par la physique a eu des applications riches en géographie, nous postulons que les tentatives de déclinaison du cadre dans des disciplines spécifiques peuvent favoriser des analogies ou d'autre modèles qui auraient été difficiles à formuler autrement.

★ ★

★

L'exploration du concept de morphogenèse réalisée dans la section précédente permet de guider la conception de modèles de croissance urbaine. Des modèles qui se baseront sur ce concept devront avoir les propriétés suivantes :

1. Rôle crucial de la *forme*, et donc inclusion à la fois d'une définition et d'une mesure de la forme, mais également rôle de celle-ci dans les ontologies.
2. Couplage fort de la forme avec la fonction. Dans un premier temps, la fonction ne sera pas explicite dans l'ontologie mais bien présente dans les processus abstraits.
3. Autonomie des sous-systèmes, c'est à dire présence d'un certain niveau de modularité dans le système global. Cette propriété nous guide à la fois dans l'échelle de modélisation, que nous prendrons "moyenne", ou mesoscopique, ainsi que dans la recherche de modèles simples, c'est à dire parcimonieux dans les processus pris en compte et dans le nombre de paramètres.

La stratégie que nous suivons pour intégrer ces propriétés dans des modèles de morphogenèse qui doivent nous conduire à des modèles de co-évolution entre réseau de transport et territoire, est progressive : progression en largeur des ontologies (en nombre d'aspects pris en compte) et progression en complexité, que nous interpréterons ici comme force du couplage. Les deux sections suivantes présentent ainsi d'abord un modèle de morphogenèse à visée minimaliste pour la densité de population uniquement, puis le couplage faible (séquentiel) de celui-ci avec un modèle de génération du réseau routier. Le couplage fort et l'explicitation des fonctions par le réseau, fournissant les bases d'un modèle de co-évolution, feront l'objet du Chapitre 7.

* * *

*

5.2 MORPHOGENÈSE URBAINE PAR AGRÉGATION-DIFFUSION

Nous étudions donc ici un modèle stochastique de croissance urbain générant des distributions spatiales de densité de population à une échelle intermédiaire mesoscopique. Le modèle se base sur le jeu antagoniste entre les deux processus abstraits opposés de l'agrégation (attachement préférentiel) et de la diffusion (étalement urbain). En utilisant des indicateurs pour quantifier précisément la forme urbaine, le modèle est d'abord validé statistiquement puis exploré intensivement pour comprendre son comportement complexe sur son espace de paramètres. Ayant calculé précédemment les mesures morphologiques réelles sur des aires locales de taille 50km couvrant l'ensemble de l'Union Européenne, nous les utilisons pour montrer que le modèle peut reproduire la plupart des morphologies urbaines existantes en Europe. Cela implique que la dimension morphologique des processus de croissance urbaine à cette échelle est capturée de manière suffisante par les deux processus abstraits d'agrégation et de diffusion.

5.2.1 Contexte

Croissance Urbaine

L'étude de la croissance urbaine, et plus particulièrement sa quantification, est plus que jamais un enjeu crucial dans un contexte où la majorité de la population mondiale vit dans des villes dont l'expansion a des impacts environnementaux significatifs [**seto2012global**] et qui doivent pour cela assurer une soutenabilité et une résilience au changement climatique accrues. La compréhension des moteurs de la croissance urbaine devrait conduire à l'élaboration de politiques mieux intégrées. Il s'agit cependant d'une question loin d'être résolue dans les diverses disciplines concernées : les systèmes urbains sont des systèmes socio-techniques complexes qui peuvent être étudiés d'une grande variété de points de vue. BATTY défend en ce sens la construction d'une science dédiée définie par ses objets d'étude plus que par les méthodes utilisées [**batty2013new**], ce qui devrait permettre des couplages plus faciles entre approches et donc des modèles de croissance urbaine prenant en compte des processus hétérogènes. Les processus qu'un modèle peut prendre en compte sont également liés au choix de l'échelle d'étude. A une échelle macroscopique, les modèles de croissance pour les systèmes de villes sont majoritairement le sujet de l'économie et de la géographie. [**gabaix1999zipf**], reprenant les idées de [**gibrat1931inegalites**], montre qu'en première approximation, le modèle de GIBRAT postulant des taux de croissance aléatoires ne dépendant pas de la taille des villes, produit la bien connue loi de Zipf, ou loi rang-taille, qui est un fait stylisé typique témoignant d'une hiérarchie dans les systèmes de villes. Il a cepen-

dant été démontré empiriquement que des déviations systématiques à cette loi existent [**rozenfeld2008laws**], et que les interactions spatiales pourraient en être responsables. Les modèles intégrant les interactions spatiales incluent par exemple [**bretagnolle200olong**] qui introduit un modèle de croissance dans lequel ces interactions, qui sont fonction de la distance et de la géographie, jouent un rôle significatif dans les taux de croissance. Plus récemment, [**favarro2011gibrat**] a étendu ce modèle en prenant en compte les vagues d'innovation entre les villes comme facteur d'influence. Les relations entre espace, croissance économique et croissance de la population sont étudiées par le modèle Marius [**cottineau2014evolution**] pour le cas de l'ex-Union Soviétique, pour lequel la performance du modèle est démontrée améliorée par rapport aux modèles sans interactions.

Automates cellulaires

A de plus grandes échelles, qui peuvent être comprises comme microscopiques ou mesoscopiques selon la résolution et l'étendue spatiale des modèles, les agents des modèles diffèrent fondamentalement. L'espace est généralement pris en compte de manière plus fine, par les effets de voisinage par exemple. [**andersson2002urban**] décrit un modèle de croissance urbaine basé sur le microscopique, dans le but de remplacer des mécanismes physiques non interprétables par des mécanismes d'agents, incluant des forces d'interaction et des choix de mobilité. Les corrélations locales sont utilisées par [**makse1998modeling**], qui développe le modèle introduit dans [**makse1995modelling**], pour modular les motifs de croissance pour qu'ils ressemblent à des configurations réelles. Le monde des modèles de croissance urbaine à automates cellulaires (CA) [**batty1994cells**] offre aussi de nombreux exemples. [**GEAN:GEAN940**] introduit un cadre générique pour les CA avec usage du sol multiple, basé sur des règles d'évolution locales. Un modèle avec des états plus simples (occupé ou non) mais prenant en compte des contraintes globales est étudié par [**ward2000stochastically**]. Le modèle Sleuth, introduit initialement par [**clarke1998loose**] pour la zone de la Baie de San Francisco, et pour lequel un aperçu des diverses applications est donné dans [**clarke2007decade**], a été calibré sur des régions tout autour du monde, fournissant des mesures comparatives au travers des paramètres calibrés.

Morphogenèse urbaine

Enfin, assez proches des modèles CA, mais au coeur de nos préoccupations ici, sont les modèles de Morphogenèse Urbaine, qui visent à simuler la croissance de la forme urbaine à partir de règles autonomes. Nous en avons déjà revu un certain nombre en [5.1](#), et proposons maintenant de les situer par rapport aux modèles précédent.

[frankhauser1998fractal] suggère que la nature fractale des villes est en relation étroite avec l'émergence de la forme urbaine à partir des interactions socio-économiques microscopiques, à savoir la morphogenèse urbaine. [courtat2011mathematics] développe un modèle de morphogenèse pour les routes urbaines seules, avec des règles de croissance basées sur des considérations géométriques. Celles-ci sont montrées suffisantes pour produire un large spectre de motifs analogues à des existants. De manière similaire, [raimbault2014hybrid] couple un CA avec un réseau évolutif pour reproduire des formes urbaines stylisées, de villes monocentriques concentrées à des banlieues étalées. Le modèle Diffusion-Limited-Aggregation, venant de la physique, et qui a été appliqué aux villes en premier par [batty1991generating], peut aussi être vu comme un modèle de morphogenèse. Ce type de modèles, qui peuvent parfois être classifiés comme CA, ont généralement la particularité d'être parcimonieux dans leur structure. Des modèles similaires ont également été étudiés en biologie pour la diffusion de population par exemple [bosch1990velocity].

La particularité de ces modèles, en comparaison aux automates cellulaires, est la rôle crucial de la forme dans leur règles d'évolution, et pour certains de la fonction, comme [bonin2012modele]. Nous nous placerons ici dans cette même logique de règles basées sur la forme (dans un premier temps) et la fonction (en Chapitre 7) pour construire des modèles d'interaction entre territoires et réseaux.

Objectif

Nous étudions dans cette section un modèle de morphogenèse, à l'échelle mesoscopique, dont le but est d'être performant pour la reproduction de motifs existants, sous contrainte de simplicité dans ses règles et variables. La question sous-jacente est l'exploration de la performance de mécanismes simples pour reproduire des formes urbaines complexes. Nous considérons des processus abstraits, précisément l'agrégation et la diffusion, comme facteurs potentiellement explicatifs de la croissance urbaine, basés sur la densité de population seule, qui seront détaillés ci-dessous. Un aspect important que nous utilisons est la mesure quantitative de la forme urbaine, basée sur une combinaison d'indicateurs morphologiques, pour quantifier et comparer les sorties de modèle et les formes urbaines réelles. Notre contribution est significative sur plusieurs points : (i) le calcul des caractéristiques morphologiques réelles sur une étendue spatiale conséquente (Union Européenne complète) ; (ii) nous apprenons le comportement du modèle par une exploration conséquente de l'espace des paramètres ; (iii) nous montrons par la calibration que le modèle est capable de reproduire la majorité des formes urbaines existantes en Europe, et que ces processus abstraits sont suffisants pour expliquer la forme urbaine seule. La suite de cette section est organisée de la façon suivante : nous décrivons d'abord formellement le modèle. Nous

étudions ensuite le comportement du modèle par une exploration de l'espace des paramètres et par une approche semi-analytique d'un cas simplifié, puis nous décrivons les résultats de la calibration du modèle.

5.2.2 Modèle et Résultats

Modèle de croissance urbaine

DESCRIPTION Notre modèle est basé sur des idées largement acceptées de processus d'agrégation-diffusion pour les processus urbains. La combinaison de forces d'attraction avec celles de répulsion, dues par exemple à la congestion, fournit déjà une issue complexe qui a été montrée représentative des processus de croissance urbaine sous certaines hypothèses simplificatrices. Un modèle capturant ces processus a été introduit dans [**batty2006hierarchy**], comme une variation cellulaire du modèle de *Diffusion-limited Aggregation* (DLA) [**batty1991generating**]. En effet, la tension entre les mécanismes antagonistes d'agrégation et d'étalement peut être un processus important pour la morphogenèse urbaine. Par exemple, [**fujita1996economics**] oppose les forces centrifuges aux forces centripètes dans une vision d'équilibre des systèmes urbains spatiaux, ce qui peut facilement être transféré aux systèmes hors équilibre dans le cadre de la complexité auto-organisée : une structure urbaine est un système *far-from-equilibrium* qui a été conduit à ce point par ces forces opposées. Par exemple, des forces concrètes de dispersion sont la congestion ou la recherche de faible densité par les habitants, tandis que des forces d'agrégation peuvent être la présence d'aménités, de lieux d'intérêts, de possibilités accrues d'interactions sociales.

Les deux processus contradictoires de concentration urbaine et d'étalement urbain sont capturés par le modèle, ce qui permet de reproduire avec une bonne précision un grand nombre de morphologies existantes. Nous pouvons supposer que des mécanismes d'agrégation comme l'attachement préférentiel sont des bons candidats pour expliquer la croissance urbaine. En effet, il a été montré que le modèle de Simon, pour lequel l'attachement préférentiel est le principal mécanisme, génère des *power-law* qui sont typiques des systèmes urbains (lois d'échelles par exemple) [**2016arXiv160806313S**]. La question de l'échelle à laquelle il est possible et pertinent de définir et d'essayer de simuler la croissance urbaine est relativement ouverte, et dépendra en fait de quels problèmes sont considérés. Travailant dans un cadre typique de la morphogenèse, les processus considérés sont locaux et notre modèle doit avoir une résolution au niveau microscopique. Nous voulons cependant quantifier la forme sur des unités urbaines cohérentes, et travaillerons ainsi sur des étendues spatiales d'ordre 50~100km. Nous résumons ces deux aspects en posant que le modèle est à l'échelle *mesoscopique*.

FORMALISATION Nous formalisons à présent le modèle et ses paramètres. Le monde du modèle est une grille carrée de côté N , dans lequel chaque cellule est caractérisée par sa population $(P_i(t))_{1 \leq i \leq N^2}$. Nous considérons la grille initialement vide, i.e. $P_i(0) = 0$, mais le modèle peut être facilement généralisé à n'importe quelle distribution initiale de population. La distribution de population est mise à jour de façon itérative. A chaque pas de temps,

1. La population totale est augmentée par un nombre fixe N_G (taux de croissance). Chaque unité de population est attribuée indépendamment à une cellule suivant un attachement préférentiel tel que

$$\mathbb{P}[P_i(t+1) = P_i(t) + 1 | P(t+1) = P(t) + 1] = \frac{(P_i(t)/P(t))^\alpha}{\sum (P_j(t)/P(t))^\alpha} \quad (9)$$

L'attribution est tirée de manière uniforme si toutes les populations sont égales à 0.

2. Une fraction β de la population est diffusée au voisinage de chaque cellule (les 8 plus proches voisins recevant chacun la même fraction de la population diffusée). Cette opération est répétée n_d fois.

Le modèle s'arrête quand la population totale atteint un paramètre fixé P_m . Pour éviter les effets de bord comme des ondes de diffusion se réfléchissant, les cellules du bord diffusent la proportion qu'elles devraient hors du monde, ce qui implique que la population totale à l'instant t est strictement plus petite que $N_G \cdot t$.

Nous résumons les paramètres du modèle dans la Table 12, donnant les processus associés et les bornes des valeurs utilisées dans les simulations. La population totale de la zone P_m est exogène, au sens qu'elle est supposée dépendre de processus de croissance à l'échelle macroscopique sur le temps long. Le taux de croissance N_G capture à la fois la croissance endogène et la balance migratoire dans la zone. Le taux d'agrégation α fixe la différence d'attractivité entre cellules, qui peut être interprétée comme un coefficient abstrait d'attraction suivant une loi d'échelle de la population. Enfin, les deux paramètres de diffusion sont complémentaires puisque diffuser avec force $n_d \cdot \beta$ est différent de diffuser n_d fois avec force β , le dernier cas donnant des configurations plus plates.

MESURE DE LA FORME URBAINE Comme le modèle se base uniquement sur la densité, nous proposons de quantifier ses sorties par la morphologie spatiale, i.e. les propriétés de la distribution spatiale de la densité. A l'échelle choisie, on s'attend à ce qu'elle traduise diverse propriétés fonctionnelles de l'environnement urbain. Le contexte et la définition des indicateurs a déjà été donnée en section 4.1.

TABLE 12 : Résumé des paramètres du modèle de morphogenèse. Nous donnons les processus correspondants à chaque paramètre et le domaine typique de variation dans la configuration que nous utilisons.

Paramètre	Notation	Processus	Domaine
Population totale	P_m	Croissance macroscopique	[1e4, 1e6]
Taux de croissance	N_G	Croissance mesoscopique	[500, 30000]
Force d'agrégation	α	Agrégation	[0.1, 4]
Force de diffusion	β	Diffusion	[0, 0.5]
Nombre de diffusions	n_d	Diffusion	{1, ..., 5}

Données réelles

Nous travaillons sur les valeurs des indicateurs calculées en section 4.1 pour l'Europe, sur les fenêtres de côté 50km avec résolution de 100 cellules. Nous posons donc pour la suite $N = 100$ pour les simulations du modèle.

Génération de structures urbaines

IMPLÉMENTATION Le modèle est implémenté à la fois en NetLogo [[wilensky1999netlogo](#)] pour des raisons d'exploration et de visualisation, et en Scala pour des raisons de performance et d'intégration plus aisée dans OpenMole [[reuillon2013openmole](#)], qui permet un accès transparent aux environnements de calcul haute performance. Le calcul des valeurs des indicateurs sur les données géographiques est fait en R avec le package raster [[hijmans2015geographic](#)]. Le code source et les résultats sont disponibles sur le dépôt ouvert du projet¹¹. Les données des valeurs réelles des indicateurs et des résultats de simulation sont disponibles sur Dataverse¹². Nous avons dans le cadre de l'implémentation Scala implémenté la convolution de distribution en deux dimensions par Transformée de Fourier rapide, permettant de transformer une complexité $O(N^4)$ en $O(N^2 \log^2 N)$ ¹³, puis implémenté les indicateurs qui ont pu être intégrés à une extension NetLogo dédiée (celle ci est détaillée en E.1.4).

¹¹ à <https://github.com/JusteRaimbault/Density>

¹² à <http://dx.doi.org/10.7910/DVN/WSUSBA>

¹³ On rappelle qu'une mesure de complexité d'un algorithme correspond à l'évaluation du temps nécessaire à la résolution d'un problème en fonction de la taille des données, notée N . Un ordre de grandeur asymptotique est noté $O(f(N))$. Ainsi, un passage d'un ordre puissance quatre à un ordre quasi puissance deux est significatif pour le temps de calcul, rendant quasi instantané un calcul prenant une dizaine de secondes pour nos tailles de grilles. La transformée de Fourier rapide utilise une décomposition creuse pour calculer la transformée de Fourier discrète en $O(N \log N)$ au lieu de $O(N^2)$. Le morphisme de la transformée du produit vers la convolution, c'est à dire $\mathcal{F}[f * g] = \mathcal{F}[f] \cdot \mathcal{F}[g]$, permet de transférer ce gain au calcul d'une convolution.

FORMES GÉNÉRÉES Le modèle a un nombre relativement faible de paramètres mais est capable de générer une grande variété de formes, qui s'étendent au delà des formes existantes, comme illustré en Fig. 31. Plus particulièrement, sa nature dynamique permet par la combinaison des paramètres P_m et N_G de choisir entre des configurations qui peuvent être non stationnaires ou semi stationnaires, tandis que l'interaction entre α et β module l'étalement et le caractère compact des formes. Nous simulons le modèle pour des valeurs de paramètres variant dans les bornes données en Table 12, pour une taille de monde $N = 100$.

La Fig. 31 montre des exemples de la variété des formes urbaines générées pour différentes valeurs des paramètres, avec les interprétations correspondantes. Parmi les quatre formes très différentes, certaines peuvent être obtenues avec la variation d'un seul paramètre seulement : passer d'une zone péri-urbaine à une zone rurale implique une agrégation accrue au même niveau de diffusion. Il faut noter que le modèle est basé sur la densité, et que le paramètre P_m/N_G est celui qui influence réellement la dynamique : les valeurs de P_m ne correspondent dans certains cas pas directement aux interprétations qui en sont faites (pour le rural en particulier) qui sont faites sur les densités. Une homothétie garde la forme des établissements et résout ce problème. Ces exemples montrent la potentialité du modèle à produire des formes diverses. Nous devons ensuite étudier systématiquement sa stochasticité et explorer son espace des paramètres.

Comportement du modèle

Dans l'étude d'un tel model computationnel de simulation, le manque de traçabilité analytique doit être compensé par une connaissance extensive du comportement du modèle dans l'espace des paramètres [banos2013pour]. Ce type d'approche est typique de ce que ARTHUR nomme le *tournant computationnel dans la science moderne* [arthur2015complexity] : la connaissance est moins extraite de résolutions analytiques exactes que par des expériences de calcul intensif, même pour des modèles "simples" comme celui que nous étudions.

CONVERGENCE Dans un premier temps il est important d'assurer la convergence du modèle et son comportement au regard de la stochasticité. Nous simulons le modèle pour une grille creuse de l'espace des paramètres contenant 81 points, avec 100 répétitions à chaque point. Les histogrammes correspondants sont montrés en Appendice A.8. Les indicateurs présentent de bonnes propriétés de convergence : la plupart des indicateurs sont aisément discernables de manière statistique entre les points de paramètres : par exemple l'indice de Moran, parmi les plus dispersé, a une étendue de 0 à 0.1 entre paramètres mais une variabilité maximale de 0.01 entre réplications.

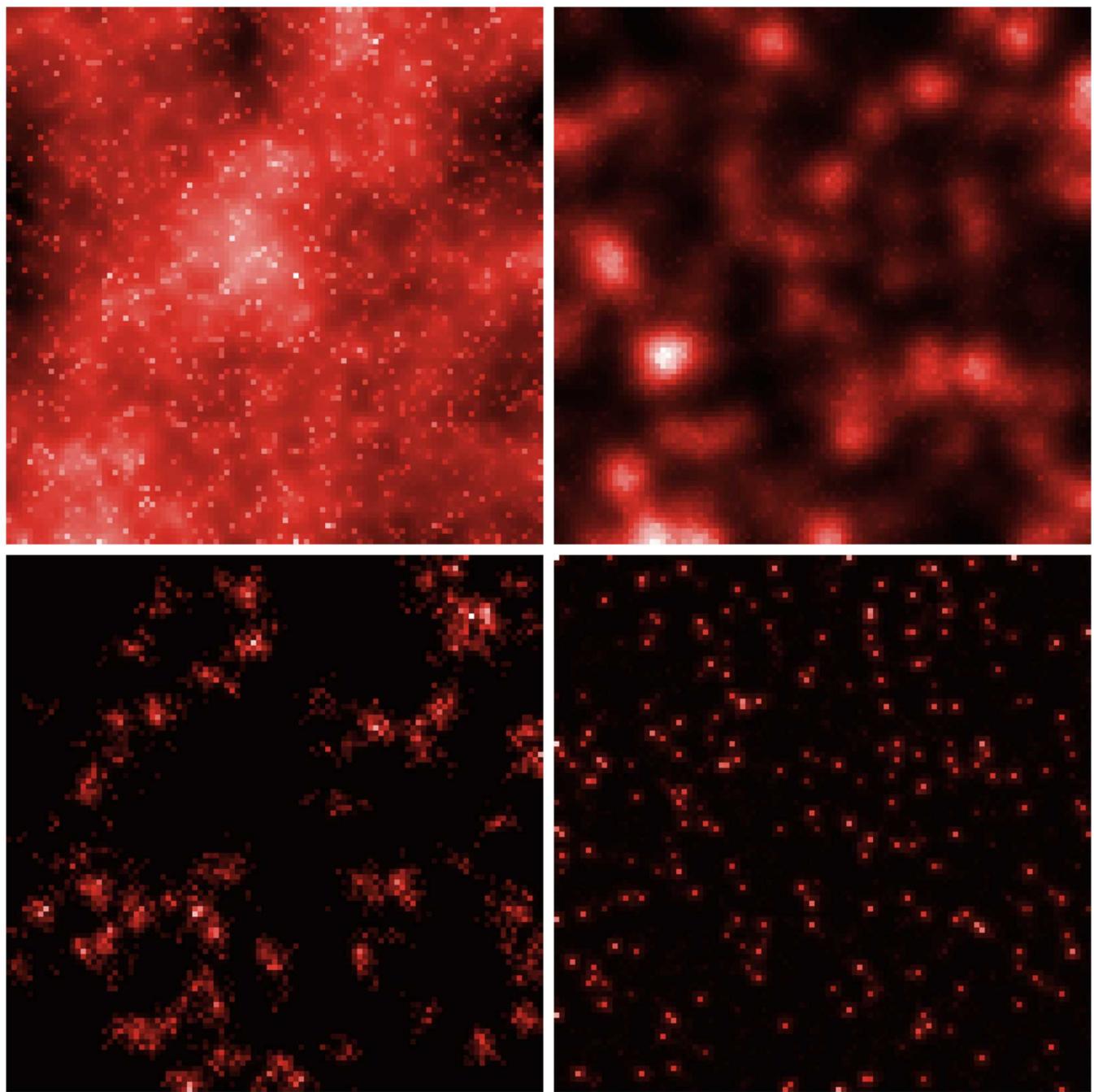


FIGURE 31 : Exemple de la variété de formes territoriales générées. (*Haut Gauche*) Configuration urbaine très diffuse, $\alpha = 0.4, \beta = 0.05, n_d = 2, N_G = 76, P_m = 75620$; (*Haut Droite*) Configuration polycentrique urbaine semi-stationnaire, $\alpha = 1.4, \beta = 0.047, n_d = 2, N_G = 274, P_m = 53977$; (*Bas Gauche*) Etablissements intermédiaires (périurbain ou zone rurale densément peuplée), $\alpha = 0.4, \beta = 0.006, n_d = 1, N_G = 25, P_m = 4400$; (*Bas Droite*) Zone rurale, $\alpha = 1.6, \beta = 0.006, n_d = 1, N_G = 268, P_m = 76376$.

Nous utilisons cette expérience pour établir un nombre raisonnable de répétitions nécessaires pour des expériences plus volumineuses. Pour chaque point, nous estimons le ratio de Sharpe pour chaque indicateur, i.e. sa moyenne normalisée par la déviation standard. L'indicateur le plus variable est l'indice de Moran avec un Sharpe minimal de 0.93, mais pour lequel le premier quartile est à 6.89. Les autres indicateurs ont tous des valeurs minimales très hautes, toutes au-dessus de 2. Cela signifie que des intervalles de confiance large comme $1.5 \cdot \sigma$ sont suffisants pour différencier entre deux configurations différentes. Dans le cas d'une distribution Gaussienne, nous savons que la taille de l'intervalle de confiance à 95% autour de la moyenne est donné par $2 \cdot \sigma \cdot 1.96/\sqrt{n}$, ce qui donne $1.26 \cdot \sigma$ pour $n = 10$. Nous utilisons pour cela ce nombre de répétitions pour chaque point de paramètres par la suite, ce qui est largement suffisant pour avoir des différences entre les moyennes qui sont statistiquement significatives comme montré précédemment. Par la suite, lorsque nous considérons les valeurs des indicateurs pour le modèle simulé, nous considérons la moyenne d'ensemble sur ces répétitions stochastiques.

EXPLORATION DE L'ESPACE DES PARAMÈTRES Nous échantillonons l'espace des paramètres en utilisant un *Latin Hypercube Sampling*, les paramètres variant¹⁴ dans $\alpha \in [0, 1, 4]$, $\beta \in [0, 0.5]$, $n_d \in \{1, \dots, 5\}$, $N_G \in [500, 30000]$, $P_m \in [1e4, 1e6]$. Ce type de criblage est un bon compromis pour avoir un échantillonnage raisonnable sans être soumis au sort de la dimension dans des capacités de calcul normales. Nous échantillonons autour de 80000 points, avec 10 répétitions chacun. Rappelons le protocole suivi ici pour obtenir le comportement d'un modèle de simulation, qui est à placer dans celui plus général présenté en 3.1 :

- échantillonnage des points de paramètres ;
- simulation du modèle pour chaque point de paramètre, répétée 10 fois ;
- calcul pour chaque execution du modèle des indicateurs de forme ;
- agrégation pour chaque point de paramètre par calcul des moyennes sur les répétitions¹⁵.

¹⁴ Comme nous l'expliquons, les valeurs relatives de P_m et N_G jouent principalement sur les formes obtenues, et nous fixons donc P_m pour obtenir des territoires contenant au maximum 1 million d'habitants, ce qui est une forte densité mais pas extrême (pour comparaison, la métropole parisienne concentre autour de 8 millions sur une zone de taille équivalente). Les valeurs de N_G varient considérablement pour couvrir un grand nombre de régimes dynamiques possibles. Les valeurs de α et β ont été obtenues par expérimentations successives.

¹⁵ Vu la forme des distributions obtenues pour 100 répétitions, présentées en A.8, l'utilisation de la moyenne ou de la médiane donnent des résultats équivalents.

Des graphes complets du comportement du modèle en fonction des paramètres sont donnés en A.8. Nous montrons en Fig. 32 des comportements particulièrement intéressants pour la pente γ et la distance \bar{d} . Tout d'abord, le comportement qualitatif général en fonction de la force d'agrégation, c'est à dire que des valeurs faibles de α donnent des configurations moins hiérarchiques et plus étalées, confirme le comportement attendu intuitivement. L'effet de la force de diffusion β est plus difficile à cerner : l'effet est inversé pour la pente entre des haut et bas taux de croissance mais pas pour la distance, qui elle présente une inversion quand α varie. Dans le cas où N_G est faible, une diffusion faible crée des configurations plus étalées quand l'agrégation est basse, mais moins étalées quand l'agrégation est forte. De plus, tous les indicateurs présentent une transition plus ou moins abrupte autour de $\alpha \simeq 1.5$. La pente se stabilise au-dessus de certaines valeurs, ce qui veut dire que la hiérarchie ne peut pas être forcée plus et dépend alors de la valeur de la diffusion, au moins pour les faibles N_G (colonne de droite). En général, des valeurs fortes pour P_m/N_G augmentent les effets de la diffusion ce à quoi on pouvait s'attendre. L'existence d'un minimum pour la pente à $n_d = 1, P_m/N_G \in [13, 26]$ et les valeurs faibles de β est inattendue et témoigne d'une interaction complexe entre agrégation et diffusion. L'émergence de ce régime "optimal" est associé avec un décalage des points de transition dans les autres cas : par exemple une diffusion plus faible implique une transition commençant à des valeurs plus faible de α pour la distance. Cette exploration confirme qu'un comportement complexe, au sens de formes émergentes qui ne peuvent être prédites, est présent dans le modèle : il n'est pas possible de donner en avance la forme finale étant donné un jeu de paramètres, sans se référer à l'exploration complète dont nous avons donné un aperçu ici.

Analyse semi-analytique

Notre modèle peut être compris comme un type de modèle de réaction-diffusion, qui ont été utilisés largement dans d'autres champs comme la biologie comme nous l'avons résumé en 5.1. Une autre façon de formuler les modèles typiques de ces approches est d'utiliser des Equations aux Dérivées Partielles (EDP). Dans le cas d'un modèle de croissance de firmes, généralisation du modèle de Simon avec forme quelconque de la fonction d'attachement, [2017arXiv171007580R] montre qu'une EDP et sa solution générale peuvent être dérivées. Notre cas est plus délicat par l'ajout du processus de diffusion. Nous proposons d'éclairer des comportements des dynamiques de temps long en les étudiant sur un cas simplifié. Nous considérons le système en une dimension, tel que $x \in [0; 1]$ avec $1/\delta x$ cellule de taille δx . Un pas de temps est donné par δt . Chaque cellule est caractérisée par sa population comme une variable aléatoire dépendant de la position x et du

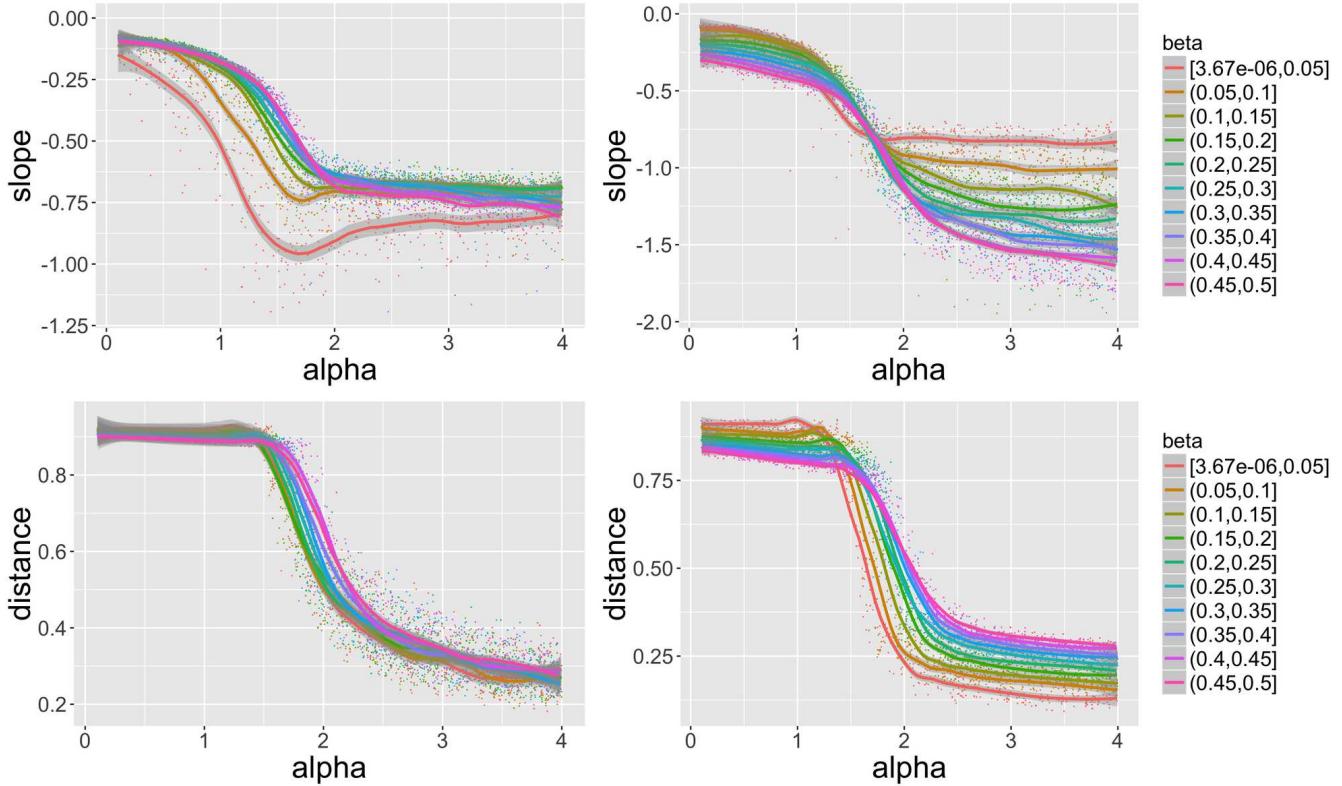


FIGURE 32 : Comportement des indicateurs. Pente γ (ligne du haut) et distance moyenne \bar{d} (ligne du bas) comme fonction de α , pour différentes valeurs de β données par la couleur des courbes, pour des valeurs particulières $n_d = 1, P_m/N_G \in [13, 26]$ (colonne de gauche) et $n_d = 4, P_m/N_G \in [41, 78]$ (colonne de droite). On observe dans chaque cas une transition en fonction de α , dont les propriétés sont influencées par les autres paramètres. Pour les faibles valeurs de P_m/N_G et de β émerge une non-monotone contre intuitive.

temps t , que nous notons $P(x, t)$. Nous travaillons sur les espérances $p(x, t) = \mathbb{E}[P(x, t)]$, et supposons que $n_d = 1$. Comme développé en Appendice A.8, on peut montrer que ce processus simplifié obéit à l'EDP suivante :

$$\delta t \cdot \frac{\partial p}{\partial t} = \frac{N_G \cdot p^\alpha}{P_\alpha(t)} + \frac{\alpha \beta (\alpha - 1) \delta x^2}{2} \cdot \frac{N_G \cdot p^{\alpha-2}}{P_\alpha(t)} \cdot \left(\frac{\partial p}{\partial x} \right)^2 + \frac{\beta \delta x^2}{2} \cdot \frac{\partial^2 p}{\partial x^2} \cdot \left[1 + \alpha \frac{N_G p^{\alpha-1}}{P_\alpha(t)} \right] \quad (10)$$

où $P_\alpha(t) = \int_x p(x, t)^\alpha dx$. Cette équation non-linéaire ne peut pas être résolue analytiquement, la présence de termes intégraux la mettant hors des méthodes standard, et la résolution numérique doit être utilisée [tadmor2012review].

Il est important de noter que le modèle simplifié peut être exprimé comme une EDP analogue aux équations de réaction-diffusion, comme celle partiellement résolue pour un modèle plus simple dans [bosch1999velocity]. Nous montrons en A.8 qu'à cause des conditions au bord, la densité (au sens de la proportion de population) converge vers une solution stationnaire sur le temps long, en passant par des états intermédiaires pour lesquels la solution est partiellement stabilisée, au sens où sa vitesse d'évolution devient relativement lente. Ces états "semi-stationnaires" sont ceux utilisés en deux dimensions avec les états dynamiques. Cette étude confirme que la variété des formes obtenues par le modèle est permise à la fois par l'interactions entre l'agrégation et la diffusion puisque l'équation les couple, mais aussi par les valeurs de P_m/N_G qui permet de fixer le niveau de convergence. En effet, la sensibilité de la solution stationnaire aux paramètres est très faible en comparaison de la forme du monde (en écho à notre étude sur la sensibilité aux conditions spatiales initiales en 3.1), et utiliser le modèle en mode stationnaire n'aurait aucun sens dans notre cas.

Enfin, nous utilisons ce cas simplifié pour démontrer l'importance des bifurcations dans la dynamique du modèle. Plus précisément, nous montrons que la dépendance au chemin est cruciale pour la forme finale. Comme illustré en Fig. 33, l'utilisation d'une condition initiale rendant les choix ambigus, correspondant à 5 cellules équidistantes et de population égale, produit des trajectoires très différentes, puisqu'en général l'un des lieux finira par dominer les autres, mais est complètement aléatoire, témoignant de bifurcations cruciales dans le système aux instants initiaux. Cet aspect est typiquement attendu dans les systèmes urbains, puisque des caractéristiques très précises feront partie des déterminants de la localisation au instants initiaux de la genèse du système : l'existence d'une ressource très locale, ou l'avantage stratégique du site (défensif, de franchissement) déterminera sur des temps très longs la forme locale de la densité. Cet aspect confirme l'importance d'indicateurs morphologiques robustes décrits précédemment.

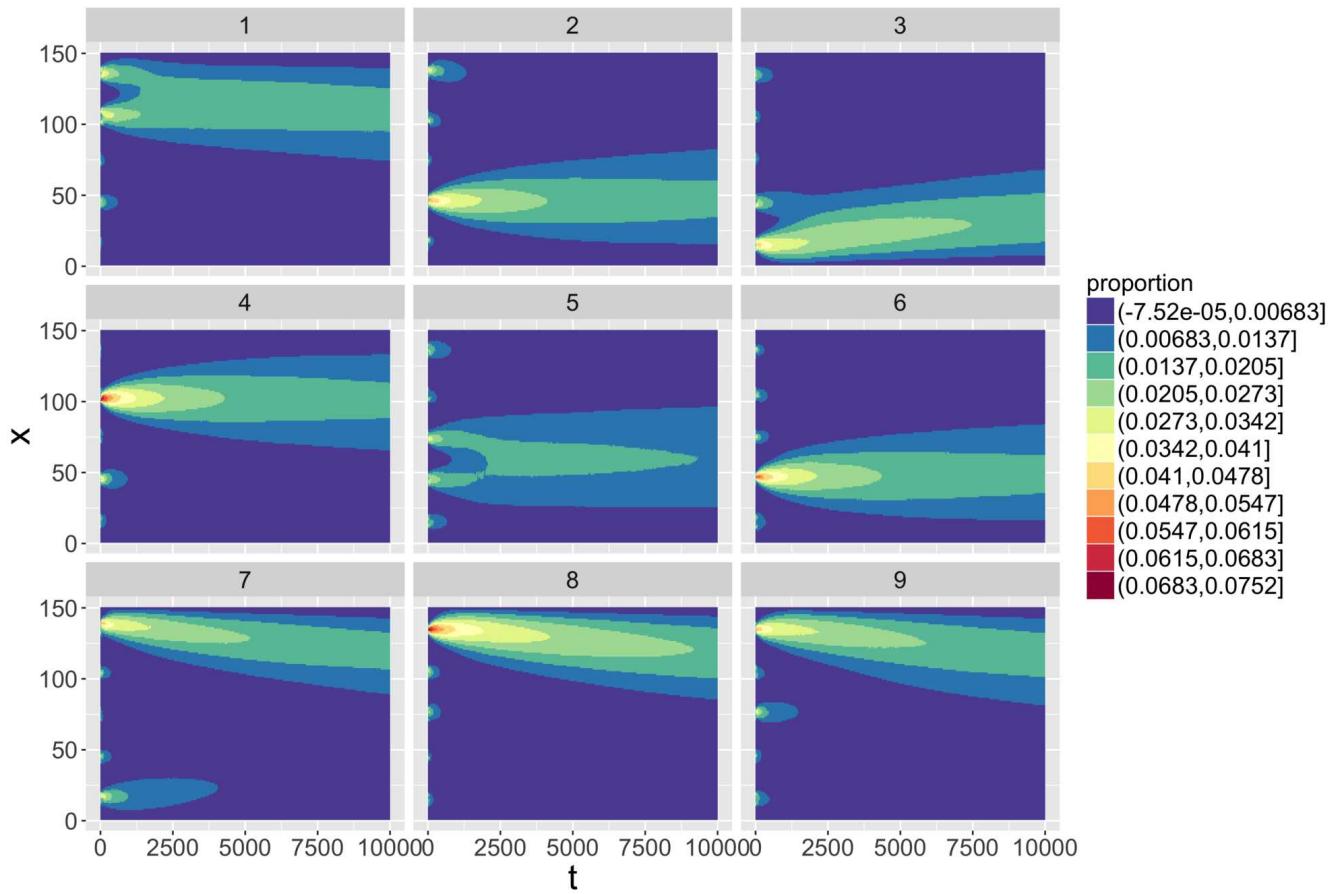


FIGURE 33 : Aléatoire et accidents figés. Nous montrons 9 réalisations aléatoires du système à une dimension avec des conditions initiales identiques, c'est à dire 5 cellules équidistantes peuplées également à l'instant initial. Les paramètres sont $\alpha = 1.4$, $\beta = 0.1$, $N_G = 10$. Chaque graphe, qui correspond à une répétition indépendante, montre le temps t en abscisse contre l'espace x en ordonnée, le niveau de couleur donnant la proportion de population dans chaque cellule à chaque instant.

Calibration du modèle

Nous traitons finalement la calibration du modèle, qui est faite sur les objectifs morphologiques. Comme une calibration pour chaque cellule réelle est hors de portée en terme de calcul, nous utilisons l'exploration précédente du modèle et superposons le nuage de points avec les valeurs réelles des indicateurs. Les scatterplots pour chaque couple d'indicateurs, pour les configurations simulées et les réelles, sont donnés en A.8. Nous constatons que le nuage de points réels est en majorité contenu dans le simulé, qui s'étend sur des zones significativement plus grandes. Cela signifie que pour une grande majorité des configurations réelles, il existe des valeurs des paramètres qui produisent en moyenne exactement la même configuration telle que résumée par les indicateurs de forme. Les plus grands écarts sont pour l'indicateur de distance, le modèle échouant à produire

des configurations avec une valeur élevée de la distance, un Moran faible et une hiérarchie intermédiaire. Cela peut par exemple correspondre à des configurations polycentriques avec de nombreux centres conséquents.

Nous considérons une contrainte de calibration plus faible, en procédant à une analyse en composantes principales sur les valeurs normalisées des indicateurs morphologiques pour les configurations synthétiques et réelles, et ne considérons que les deux premières composantes seulement. Celles-ci représentent 85% de la variance cumulée. Les nuages de points projeté sur ces dimensions est montré en Fig. 34. La majorité du nuage réel tombe dans le simulé dans cette configuration simplifiée. Nous illustrons des points particuliers avec des configurations réelles et leur contrepartie simulée : par exemple Bucarest, Roumanie, correspond à une configuration monocentrique semi-stationnaire, avec une forte agrégation mais aussi diffusion et un taux de croissance plutôt bas. Les autres exemples montrent des zones moins peuplées en Espagne et en Finlande. A partir des graphes montrant l'influence des paramètres, on peut montrer que la plupart des situations réelles tombent dans la région avec des valeurs intermédiaires pour α mais β assez variable. Cela est cohérent avec le fait que les exposants de lois d'échelles urbaines ont une plage de variation plutôt étroite (entre 0.8 et 1.3 généralement [pumain2006evolutionary]) comparée à celle que nous avons permis dans les simulations, tandis que les processus de diffusion peuvent être bien plus divers.

Ainsi, nous avons montré que le modèle est capable de reproduire la majorité des configuration de densité en Europe, malgré sa relative simplicité. Cela confirme qu'en terme de forme urbaine, la plupart des facteurs à cette échelle peuvent être traduits dans ces processus abstraits d'agrégation et de diffusion. Cela implique également que les fonctions urbaines, qui pourraient être quantifiées par des indicateurs similaires sur leur distribution spatiale, ne jouent que peu de rôle dans la localisation des populations, ou alors que celles-ci sont fortement corrélées à la distribution de la population (et sont en fait prises en compte de manière abstraite dans la fonction d'agrégation).

5.2.3 *Discussion*

RAFFINEMENT DE LA CALIBRATION ET DU MODÈLE Des développements futurs sur ce modèle simple peuvent consister à l'extraction de l'espace des paramètres exact couvrant l'ensemble des situations réelles et fournir une interprétation de sa forme, en particulier par les corrélations entre les paramètres et les expressions des fonctions de bordure. Son volume dans différentes directions devrait de plus donner l'importance relative des paramètres. Concernant l'espace faisable pour le modèle de simulation en lui-même, nous avons testé un algorithme d'exploration ciblée, qui donne des résultats pro-

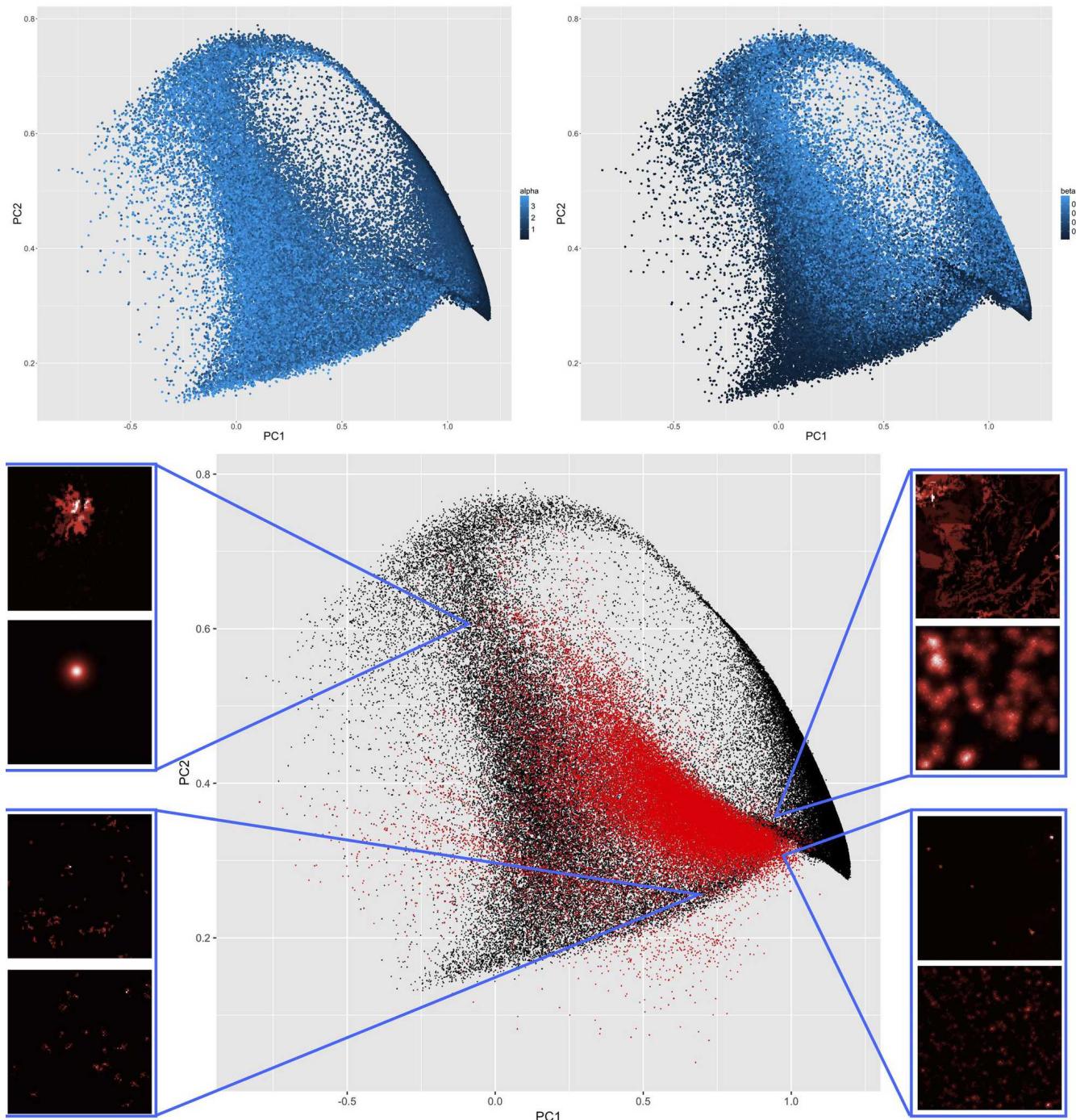


FIGURE 34 : Calibration du modèle. (*Haut*) Configurations simulées dans le plan des deux premières composantes principales, le niveau de couleur donnant l'influence de α (gauche) et de β (droite); (*Bas*) Points simulés dans le même espace (en noir) avec les configurations réelles (en rouge). Autour du graphique sont montrés des exemples typiques de configurations réelles et leur contrepartie simulée dans différentes régions de l'espace, le premier étant le réel et le second le simulé dans chaque cas : haut gauche coordonnées 25.7361,44.69989 - Romania, Bucharest - paramètres $\alpha = 3.87, \beta = 0.432, N_G = 1273, nd = 4, P_m = 63024$; Haut droite coordonnées -2.561874,41.30203 - Spain, Castilla et Leon, Soria - paramètres $\alpha = 1, \beta = 0.166, N_G = 100, nd = 1, P_m = 10017$; Bas gauche coordonnées 27.16068,65.889 - Finland, Lapland - paramètres $\alpha = 0.4, \beta = 0.006, N_G = 25, nd = 1, P_m = 849$; Bas droite coordonnées -2.607152,39.74274 - Spain, Castilla-La Mancha, Cuenca - paramètres $\alpha = 1.14, \beta = 0.108, N_G = 637, nd = 1, P_m = 13235$.

metteurs. Plus précisément, l'algorithme PSE [10.1371/journal.pone.0138212] qui est implémenté dans OpenMole, a pour but de déterminer toutes les sorties possibles d'un modèle de simulation, c'est à dire échantillonne son espace de sortie plutôt que d'entrée. Nous obtenons des résultats intéressants comme montré en Fig. 35 : nous trouvons que la borne inférieure dans le plan Moran-entropie, confirmée par l'algorithme, exhibe une loi d'échelle de manière inattendue (puisque il est impossible a priori de déterminer cet espace non-faisable avec seule les formules d'indicateurs, celui-ci étant témoin de la réalité de structures urbaines même simulées). Cela voudrait dire qu'à un niveau fixé d'auto-corrélation, qu'on pourrait vouloir atteindre pour des raisons de soutenabilité par exemple (optimalité par co-localisation), impose un désordre minimal dans la configuration des activités. D'autres relations entre indicateurs et comme fonction des paramètres peut être l'objet de développements futurs similaires. La possibilité d'une calibration dynamique du modèle, i.e. essayer de reproduire des configurations à des dates successives, est conditionnée à la disponibilité des données de population à cette résolution dans le temps.

Nous avons visé à utiliser des processus abstraits plutôt que d'avoir un modèle hautement réaliste. La modification de certains mécanismes est possible pour avoir un modèle plus proche de la réalité des processus microscopiques : par exemple plafonner la densité de population locale, ou stopper la diffusion à une distance donnée du centre s'il est bien défini. Il est cependant loin d'être clair si ceux-ci produiraient une telle variété de formes et pourraient être calibrés de la même façon, puisqu'être précis localement n'implique pas d'être précis au niveau mesoscopique pour les indicateurs morphologiques. Permettre aux paramètres de varier localement, i.e. être non-stationnaires dans l'espace, ou ajouter de l'aleatoire au processus de diffusion, sont également des raffinements potentiels du modèle.

En conclusion, nous avons produit un modèle spatial de morphogenèse urbaine à l'échelle mesoscopique, dont la calibration permet de reproduire n'importe quelle configuration urbaine Européenne en terme de morphologie. Nous démontrons que les processus abstraits d'agrégation et diffusion sont suffisants pour capturer la dimension morphologique des processus de croissance urbaine à cette échelle. Cela a des implications par exemple en terme de politiques basées sur la forme urbaine comme l'efficacité énergétique, mais aussi signifie que les questions hors de ce cadre doivent être traitées à d'autres échelles ou par d'autres dimensions des systèmes urbains.

* * *

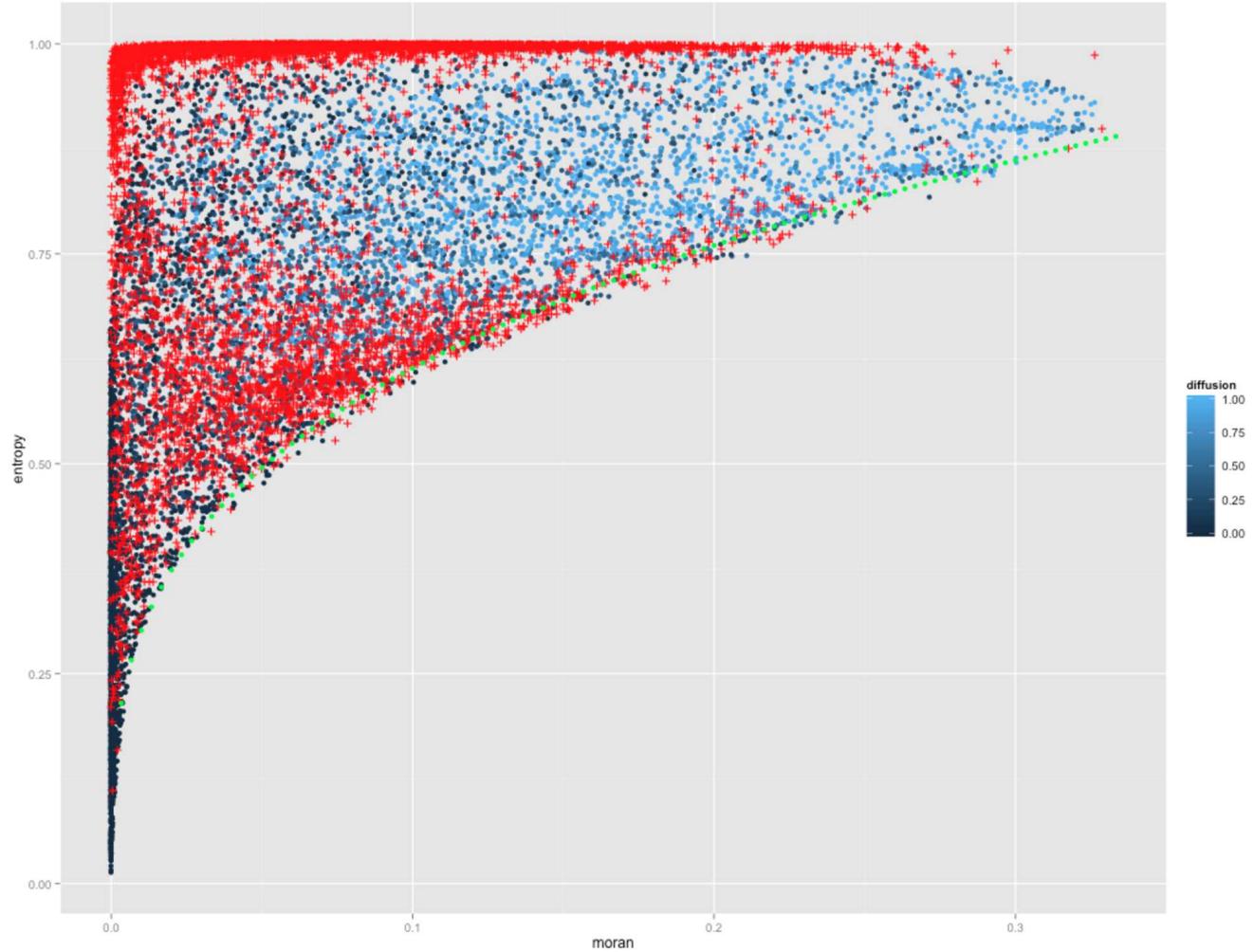


FIGURE 35 : Exploration par PSE. Scatterplot de l'entropie \mathcal{E} en fonction de l'indice de Moran I. Les points bleus, dont le niveau de couleur donne la valeur du paramètre de diffusion β , sont obtenus par criblage LHS, tandis que les points rouges sont obtenus par l'algorithme PSE. La ligne pointillée verte donne la borne inférieure faisable (qui peut être qualifiée telle quelle grâce aux propriétés de PSE), qu'il n'est pas possible d'obtenir de manière simple analytiquement. Sa forme en loi d'échelle suggère des directions de recherche, par exemple par la détermination de bornes morphologiques pour des distributions de population, et les relations fonctionnelles entre ces bornes.

5.3 GÉNÉRATION DE CONFIGURATIONS TERRITORIALES CORRÉLÉES

Cette section vise à explorer un couplage séquentiel (ou couplage simple) du modèle de génération de densité précédent avec une heuristique de croissance de réseau. Nous explorons par là un espace faisable de corrélations entre les mesures de réseau et les mesures morphologiques. Rappelons dans un premier temps les enjeux de la notion de données synthétiques et du rôle des structures de corrélation dans celles-ci.

5.3.1 *Données Géographiques corrélées de Densité et de Réseau*

L'une des inspirations et applications de la présente démarche est la génération de données synthétiques, par exemple pour alimenter les analyses de sensibilité à la configuration spatiale présentées en section 3.1. En géographie, l'utilisation de données synthétiques est plus généralement axée vers l'utilisation de populations synthétiques au sein de modèles basés agents, comme par exemple des modèles de mobilité, des modèles *LUTI* [pritchard2009advances]. On peut également citer des méthodes d'analyse spatiale qui s'en rapprochent : par exemple, l'extrapolation d'un champ spatial continu à partir d'un échantillon discret, par une estimation par noyaux par exemple, peut être compris comme la génération d'un jeu de données synthétiques, même si ce n'est pas le point de vue initial. Par exemple, dans le cas de la Regression Géographique Pondérée [brunsdon1998geographically], les noyaux de tailles variables construisent des champs abstraits représentant un potentiel généré par des variables explicites définies en des points précis de l'espace.

Dans le domaine de la modélisation en géographie quantitative, dans le cas de *modèles jouets* ou de modèles hybrides, une configuration initiale cohérente est souvent essentielle : un ensemble de configurations initiales possibles est alors un jeu de données synthétiques sur lesquelles le modèle est testé : le premier modèle Simpop [sanderson1997simpop], pionnier d'une famille de modèles par la suite configurés sur des données observées, pourrait rentrer dans ce cadre mais était lancé sur une spatialisation synthétique unique. De même, il a été souligné la difficulté de générer une configuration initiale pour une infrastructure de transport dans le cas du modèle SimpopNet [schmitt2014modelisation], alors qu'il s'agit d'un point essentiel dans la connaissance du comportement du modèle.

Il a récemment été proposé de contrôler systématiquement les effets de la configuration spatiale sur le comportement de modèles de simulation spatialisés [cottineau2015revisiting], et comme nous l'avons développé en 3.1, cette méthodologie pouvant être interprétée comme un contrôle par données statistiques spatiales. L'enjeu est

de pouvoir alors distinguer effets propres dus à la dynamique intrinsèque du modèle, d'effet particuliers dus à la structure géographique du cas d'application. Celui-ci est crucial pour la validation des conclusions issues des pratiques de modélisation et simulation en géographie quantitative. En effet, comme nous l'avons revu en 3.1, la plupart des expériences de modélisation explorent systématiquement l'influence des paramètres, mais pas des configurations spatiales initiales, alors que celles-ci peuvent avoir une influence plus grande que les paramètres.

5.3.2 Modèle et Résultats

Formalisation

Dans notre cas, nous proposons de générer des systèmes de villes représentés par une densité spatiale de population $d(\vec{x})$ et la donnée d'un réseau de transport $n(\vec{x})$, représenté de façon simplifiée, pour lesquels on serait capable de contrôler les correlations entre mesures morphologiques de la densité urbaine et caractéristiques du réseau. La question de l'interaction entre territoire et réseaux de transport est un sujet d'étude classique [offner1996reseaux], mais toujours majoritairement ouvert, extrêmement complexe et difficile à quantifier [offner1993effets]. Une modélisation dynamique des processus impliqués devrait apporter des connaissances sur ces interactions ([bretagnolle:tel-00459720], p. 162-163). Dans ce cadre, nous développons un couplage *simple* (c'est à dire sans boucle de rétroaction) entre un modèle de morphogenèse urbaine et un modèle de génération de réseau.

MODÈLE DE DENSITÉ Les modèle de densité est celui décrit et exploré dans la section précédente. Nous l'utilisons pour la génération conditionnelle du réseau.

MODÈLE DE RÉSEAU D'autre part, on est capable de générer par un modèle N un réseau de transport planaire à une échelle équivalente, étant donné une distribution de densité. La génération du réseau étant conditionnée à la donnée de la densité, les estimateurs des indicateurs de réseau seront conditionnels d'une part, et d'autre part les formes urbaines et du réseau devraient nécessairement être corrélées, les processus n'étant pas indépendants. La nature et la modularité de ces correlations selon la variation des paramètres des modèles restent à déterminer par l'exploration du modèle couplé.

Concernant le choix de l'heuristique de génération d'un réseau d'infrastructure, nous avons revu en 2.1 de nombreux modèles le permettant. D'autre part, nous comparerons différents modèles de manière opérationnelle en 7.1. Le but ici étant de démontrer la faisabilité du couplage au sein d'un modèle de morphogenèse ainsi que d'explo-

rer l'espace faisable des corrélations, nous proposons une heuristique unique, qui s'inspire du modèle de [schmitt2014modelisation], tout en le simplifiant par suppression du caractère stochastique. Cette heuristique est détaillée ci-dessous.

La procédure de génération heuristique de réseau est la suivante :

1. Un nombre fixé N_c de centres qui seront les premiers noeuds du réseau est distribué selon la distribution de densité, suivant une loi similaire à celle d'agrégation, i.e. la probabilité d'être distribué sur une case est $\frac{(P_i/P)^\alpha}{\sum(P_i/P)^\alpha}$. La population est ensuite répartie selon les zones de Voronoi des centres, un centre cumulant la population des cases dans son emprise.
2. Les centres sont connectés de façon déterministe par percolation entre plus proches clusters : tant que le réseau n'est pas connexe, les deux composantes connexes les plus proches au sens de la distance minimale entre chacun de leurs sommets sont connectées par le lien réalisant cette distance. On obtient alors un réseau arborescent.
3. Le réseau est alors modulé par ruptures de potentiels afin de se rapprocher de formes réelles. Plus précisément, un potentiel d'interaction gravitaire généralisé entre deux centres i et j est défini par

$$V_{ij}(d) = \left[(1 - k_h) + k_h \cdot \left(\frac{P_i P_j}{P^2} \right)^{\gamma_G} \right] \cdot \exp \left(-\frac{d}{d_G(1 + d/d_0)} \right)$$

où d peut être la distance euclidienne $d_{ij} = d(i, j)$ ou la distance par le réseau $d_N(i, j)$, $k_h \in [0, 1]$ un poids permettant de changer le rôle des populations dans le potentiel, γ_G régissant la forme de la hiérarchie selon les valeurs des populations, d_G distance caractéristique de décroissance et d_0 paramètre de forme. Cette forme de potentiel suppose d'une part que l'atténuation de l'interaction due à la distance est indépendante de la force de l'interaction due aux poids (hypothèse standard des modèles gravitaires) ; d'autre part qu'un terme constant du à la distance peut prendre plus ou moins de poids (pondération par k_h) ; et enfin que la fonction de distance prend comme paramètre une distance caractéristique, mais aussi un paramètre de forme, permettant par exemple de contrôler la décroissance sur les faibles distances.

4. Un nombre $K \cdot N_L$ de nouveaux liens potentiels est pris comme les couples ayant le plus grand potentiel pour la distance euclidienne ($K = 5$ est fixé).

5. Parmi les liens potentiels, N_L sont effectivement réalisés, qui sont ceux ayant le plus faible rapport $V_{ij}(d_N)/V_{ij}(d_{ij})$: à cette étape seul l'écart entre distance euclidienne et distance par le réseau compte, ce rapport ne dépendant plus des populations et étant croissant en d_N à d_{ij} fixé.
6. Le réseau est planarisé par création de noeuds aux intersections induites par les nouveaux liens (avec l'ancien réseau ou entre nouveaux liens)¹⁶.

Notons que la construction du modèle de génération est heuristique, et que d'autres types de modèles comme un réseau biologique auto-généré [**tero2010rules**], une génération par optimisation locale de contraintes géométriques [**barthelemy2008modeling**] ou un modèle de percolation plus complexe que celui utilisé, peuvent le remplacer, et permettraient la création de boucles dans le réseau. Ainsi, dans le cadre d'une architecture modulaire où le choix entre différentes implémentations d'une brique fonctionnelle peut être vue comme méta-paramètre [**cottineau2015incremental**], on pourrait choisir la fonction de génération adaptée à un besoin donné (par exemple proximité à des données réelles, contraintes sur les relations entre indicateurs de sortie, variété de formes générées, etc.).

ESPACE DES PARAMÈTRES L'espace des paramètres du modèle couplé¹⁷ est constitué des paramètres de génération de densité $\vec{\alpha}_D = (P_m/N_G, \alpha, \beta, n_d)$ (voir section 5.2 ; on s'intéresse pour simplifier au rapport entre population et taux de croissance, i.e. le nombre d'étapes nécessaires pour générer, et on fixe la population totale) et des paramètres de génération de réseau $\vec{\alpha}_N = (N_C, k_h, \gamma_G, d_G, d_0)$. On notera $\vec{\alpha} = (\vec{\alpha}_D, \vec{\alpha}_N)$.

INDICATEURS On quantifie la forme urbaine et la forme du réseau, dans le but de moduler la corrélation entre ces indicateurs. La forme est définie par un vecteur $\vec{M} = (r, \bar{d}, \varepsilon, a)$ donnant auto-corrélation spatiale (indice de Moran), distance moyenne, entropie, hiérarchie (voir [**le2015forme**] pour une définition précise de ces indicateurs). Les mesures de la forme du réseau $\vec{G} = (\bar{c}, \bar{l}, \bar{s}, \delta)$ sont, avec le réseau noté (V, E) ,

- Centralité moyenne \bar{c} , définie comme la moyenne de la *betweenness-centrality* (normalisée dans $[0, 1]$) sur l'ensemble des liens.

¹⁶ Notre modèle diffère également de [**schmitt2014modelisation**] sur ce point car nous simplifions et ne supposons pas différents niveaux de hiérarchie entre les liens.

¹⁷ Le couplage faible permet de limiter le nombre total de paramètres puisqu'un couplage fort incluant des boucles de retroaction comprendrait nécessairement des paramètres supplémentaires pour régler la forme et l'intensité de celles-ci. Pour espérer le diminuer, il faudrait concevoir un modèle intégré, ce qui est différent d'un couplage fort dans le sens où il n'est pas possible de figer l'un des sous-systèmes pour obtenir un modèle de l'autre correspondant au modèle non-couplé.

- Longueur moyenne des chemins \bar{l} définie par

$$\frac{1}{d_m} \frac{2}{|V| \cdot (|V|-1)} \sum_{i < j} d_N(i, j)$$

avec d_m distance de normalisation prise ici comme la diagonale du monde $d_m = \sqrt{2N}$.

- Vitesse moyenne [**banos2012towards**], qui correspond à la performance du réseau par rapport au trajet à vol d'oiseau, définie par $\bar{s} = \frac{2}{|V| \cdot (|V|-1)} \sum_{i < j} \frac{d_{ij}}{d_N(i, j)}$.
- Diamètre du réseau $\delta = \max_{ij} d_N(i, j)$

Nous n'avons à ce stade pas d'indicateur de "performance" du processus de génération de réseau, c'est à dire visant à reproduire des motifs typiques ou optimisant certains critères. Ceux-ci viendront plus tard en 7.1 lorsqu'on calibrera des modèles similaires sur des données réelles. Nous considérons les exemples montrés en 37 comme des éléments de l'espace faisable, la question de savoir si les formes de réseau correspondent à des réalités ou des faits stylisés donnés sera également l'objet de cette calibration.

COVARIANCE ET CORRELATION On utilisera la matrice de covariance croisée $\text{Cov}[\vec{M}, \vec{G}]$ entre densité et réseau, estimée sur un jeu de n réalisations à paramètres fixés $(\vec{M}[D(\vec{\alpha})], \vec{G}[N(\vec{\alpha})])_{1 \leq i \leq n}$ par l'estimateur standard non-biaisé. On s'intéressera à la corrélation de Pearson qui y est associée, estimée de la même façon.

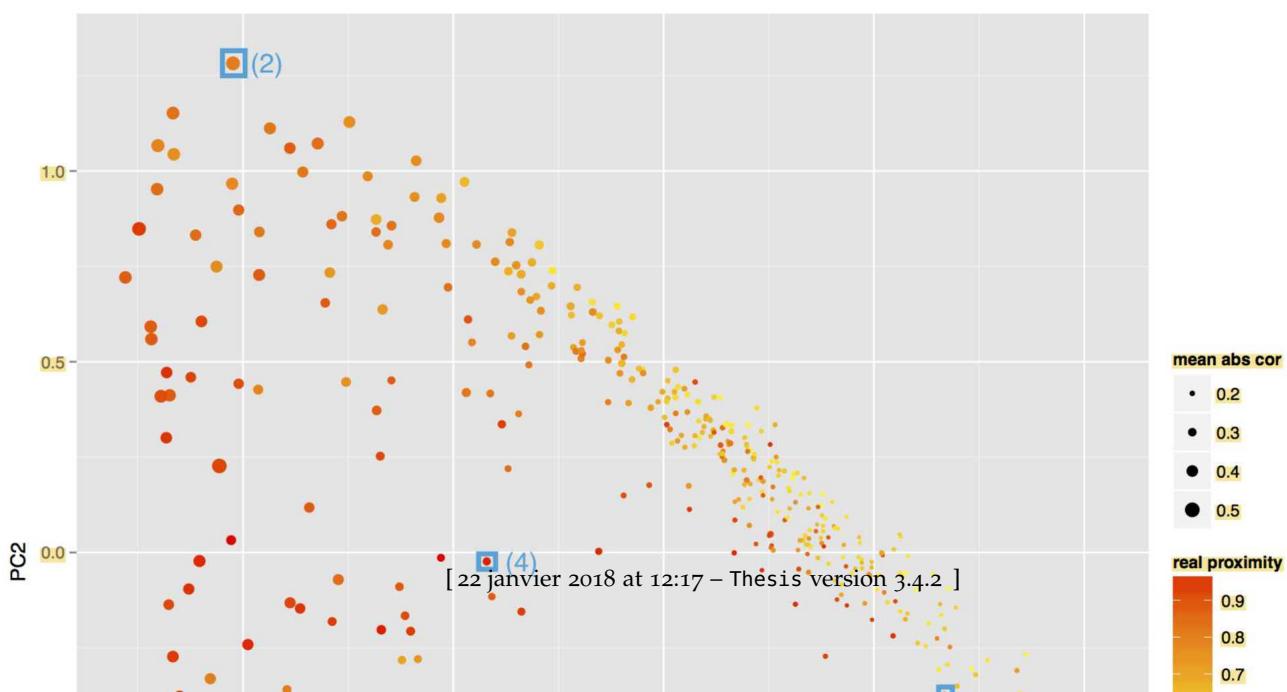
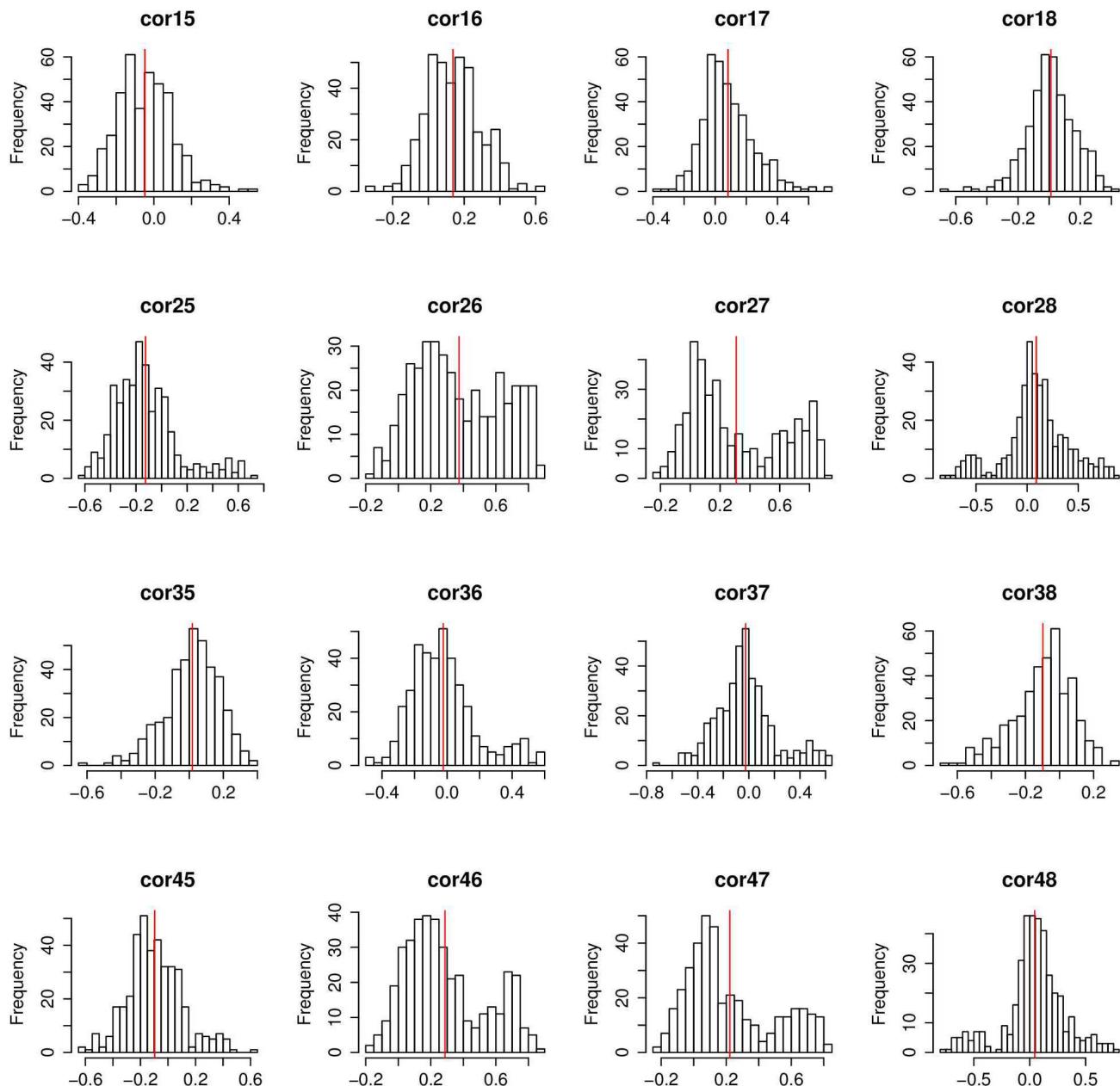
Implementation

Le couplage des modèles génératifs est effectué à la fois au niveau formel et au niveau opérationnel, c'est à dire qu'on fait interagir des implémentations indépendantes. Pour cela, le logiciel OpenMole [**reuillon2013openmole**] utilisé pour l'exploration intensive, offre le cadre idéal de par son langage modulaire permettant de construire des *workflows* par composition de tâches à loisir et de les brancher sur divers plans d'expérience et sorties. Pour des raisons opérationnelles, le modèle de densité est implémenté en langage *scala* comme un plugin d'OpenMole, tandis que la génération de réseau est implantée en langage basé-agent NetLogo [**wilensky1999netlogo**], ce qui facilite l'exploration interactive et construction heuristique interactive. Le code source est disponible pour reproductibilité sur le dépôt du projet¹⁸.

Résultats

L'étude du modèle de densité seul est développée dans la section précédente. Pour rappel, il est notamment calibré sur les données de

¹⁸ à l'adresse <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Synthetic>



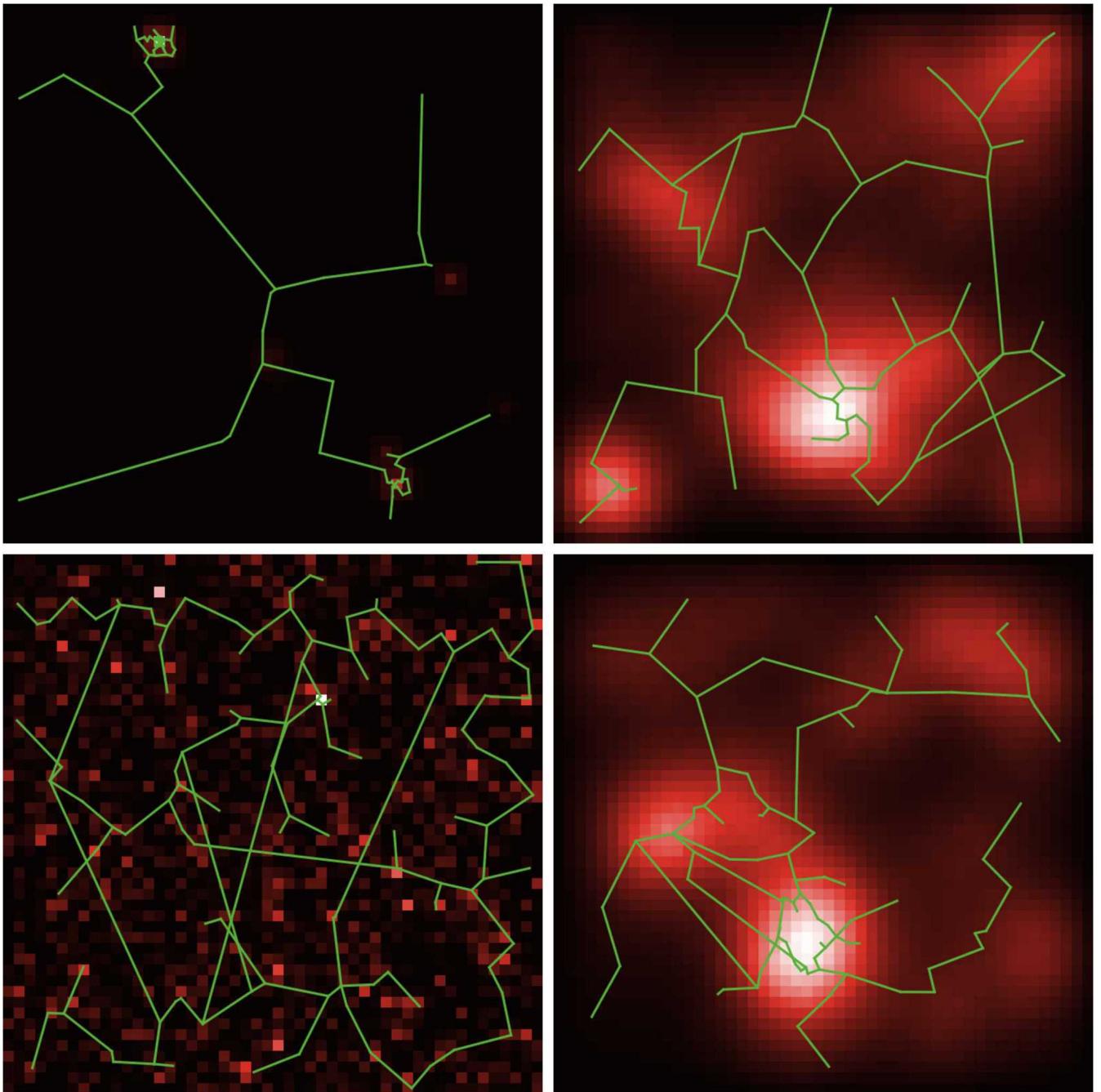


FIGURE 37 : Configurations obtenues pour les paramètres donnant les quatre points mis en évidence en 36 (d), dans l'ordre de gauche à droite et de haut en bas. Nous retrouvons des configurations de villes polycentriques (2 et 4), des établissements ruraux diffus (3) et une zone de densité agrégée faible (1). Se reporter à l'appendice A.9 pour les valeurs exhaustives des paramètres, indicateurs, et corrélations correspondantes. Par exemple \bar{d} est fortement corrélé à $\bar{l}, \bar{s} (\simeq 0.8)$ dans (1), mais pas dans (3) même si les deux correspondent à des environnements ruraux; dans le cas urbain nous observons également une forte variabilité : $\rho[\bar{d}, \bar{c}] \simeq 0.34$ pour (4) mais $\simeq -0.41$ pour (2), ce qui est expliqué par un rôle plus fort de la hiérarchie de gravité dans (2) $\gamma_G = 3.9, k_h = 0.7$ (pour (4), $\gamma_G = 1.07, k_h = 0.25$), tandis que les paramètres de densité sont similaires.

la grille européenne de densité, sur des zones de 50km de côté et de résolution 500m pour lesquelles les valeurs réelles des indicateurs ont été calculées pour l'ensemble de l'Europe. D'autre part, une exploration brutale du modèle permet d'estimer l'ensemble des sorties possibles dans des bornes raisonnables pour les paramètres (grossièrement $\alpha \in [0.5, 2]$, $N_G \in [500, 3000]$, $P_m \in [10^4, 10^5]$, $\beta \in [0, 0.2]$, $n_d \in \{1, \dots, 4\}$). La réduction à un plan de l'espace des objectifs par une Analyse en Composantes Principales (variance expliquée à deux composantes $\simeq 80\%$) permet d'isoler un nuage de points de sorties recouvrant assez fidèlement le nuage des points réels, ce qui veut dire que le modèle est capable de reproduire morphologiquement l'ensemble des configurations existantes.

Dans le but d'illustrer la méthode de génération de données synthétiques, l'exploration a été orientée vers l'étude des correlations. Etant donné la grande dimension relative de l'espace des paramètres, une exploration par grille exhaustive est impossible. On utilise un plan d'expérience par criblage (hypercube latin), avec les bornes indiquées ci-dessus pour $\vec{\alpha}_D$ et pour $\vec{\alpha}_N$, on a $N_C \in [50, 120]$, $d_G \in [1, 100]$, $d_0 \in [0.1, 10]$, $k_h \in [0, 1]$, $\gamma_G \in [0.1, 4]$, $N_L \in [4, 20]$. Concernant le nombre de réplications du modèle pour chaque valeur des paramètres, moins de 50 sont nécessaires pour obtenir sur les indicateurs des intervalles de confiance à 95% de taille inférieure aux déviations standard. Pour les correlations, une centaine donne des IC (obtenus par méthode de Fisher) de taille moyenne 0.4, on fixe donc $n = 80$ pour l'expérience. La figure 36 donne le détail des résultats de l'exploration. On retiendra les résultats marquants suivants au regard de la génération de données synthétiques corrélées :

- les distributions empiriques des coefficients de correlations entre indicateurs de forme et indicateurs de réseaux ne sont pas simples, pouvant être bimodales (par exemple $\rho_{46} = \rho[r, \bar{l}]$ entre l'index de Moran et le chemin moyen).
- On arrive à générer un assez haut niveau de correlation pour l'ensemble des indicateurs, la correlation absolue maximale variant entre 0.6 et 0.9; l'amplitude varie quant à elle entre 0.9 et 1.6, ce qui permet un large spectre de valeurs. L'espace couvert dans un plan principal a une étendue certaine mais n'est pas uniforme : on ne peut pas moduler à loisir n'importe quel coefficients, ceux-ci étant liés par les processus de génération sous-jacent. Une étude plus fine aux ordres suivants (corrélation des correlations) serait nécessaire pour cerner exactement la latitude dans la génération.
- les points les plus corrélés en moyenne sont également ceux les plus proches des données réelles, ce qui confirme l'intuition d'une forte interdépendance en réalité.

- Des exemples concrets pris sur des points particuliers distants dans le plan principal montre que des configurations de densité proches peuvent présenter des profils de correlations très différents.

5.3.3 *Discussion*

Développements

Il est possible de raffiner cette étude en étendant la méthode de contrôle des correlations. La connaissance très fine du comportement de N (distribution statistique sur une grille fine de l'espace des paramètres) conditionnée à D devrait permettre de déterminer exhaustivement $N^{<-1>}|D$ et avoir plus de latitude dans la génération des correlations. On pourra également appliquer des algorithmes spécifiques d'exploration pour essayer d'atteindre des configurations exceptionnelles réalisant un niveau de corrélation voulu, ou au moins pour découvrir l'espace des correlations atteignables par la méthode de génération [[10.1371/journal.pone.0138212](#)].

Applications Directes

Le couplage que nous venons de présenter s'est arrêté à la génération des données synthétiques. Nous proposons des pistes d'application directe qui donneront un aperçu de l'éventail des possibilités.

- La calibration de la composante de génération de réseau, à densité donnée, sur des données réelles de réseau de transport (typiquement routier vu les formes heuristiques obtenues, il devrait par exemple être aisément d'utiliser les données ouvertes d'OpenStreetMap qui sont de qualité raisonnable pour l'Europe, du moins pour la France [[girres2010quality](#)] et pour lesquelles nous avons déjà simplifié le réseau et calculé les indicateurs en [4.1](#)). Il y a toutefois des ajustements à faire sur le modèle pour supprimer les effets de bord dus à sa structure, par exemple en le faisant générer sur une surface étendue pour ne garder qu'une zone centrale sur laquelle la calibration aurait lieu) permettrait en théorie d'isoler un jeu de paramètres représentant fidèlement des situations existantes à la fois pour la forme urbaine et la forme du réseau. Il serait alors possible de dériver une "corrélation théorique" pour celles-ci, étant donné qu'une corrélation empirique n'est en théorie pas calculable puisqu'une seule instance des processus stochastiques est observée. Vu la non-ergodicité des systèmes urbains [[pumain2012urban](#)], il y a de fortes chances pour que ces processus soient différents d'une zone géographique à l'autre (ou selon un autre point de vue qu'ils soient dans un autre état des meta-paramètres, dans un autre régime) et que leur interprétation en tant que réalisations

d'un même processus stochastique n'ait aucun sens, entraînant l'impossibilité du calcul des covariations, sauf sous des hypothèses simplifiées comme nous l'avons fait en 4.1. Il s'agit alors de supposer une stationnarité locale, c'est à dire des processus dominants se manifestant selon des paramètres variables selon les régions de l'espace. En attribuant un jeu de données synthétiques similaire à une situation donnée, on serait capable de calculer une sorte de *correlation intrinsèque* propre à la situation, qui émerge en fait en réalité des interdépendances temporelles des composantes. Connaitre celle-ci renseigne alors sur ces interdépendances, et donc sur les relations entre réseaux et territoires.

- Comme déjà évoqué, la plupart des modèles de simulation nécessitent un état initial, généré artificiellement à partir du moment où la paramétrisation n'est pas effectuée totalement à partir de données réelles. Une analyse de sensibilité avancée du modèle implique alors un contrôle sur les paramètres de génération du jeu de données synthétique, vu comme méta-paramètre du modèle [cottineau2015revisiting]. Dans le cas d'une analyse statistique des sorties du modèle, on est alors capable d'effectuer un contrôle statistique au second ordre.
- On a étudié des processus stochastiques dans le premier exemple, au sens de séries temporelles aléatoires, alors que le temps ne jouait pas de rôle dans le second. On peut suggérer un couplage fort entre les deux composantes du modèle (ou la construction d'un modèle intégré) et observer les indicateurs et correlations à différents pas de temps de la génération. Dans le cas d'une dynamique, de par les rétroactions, on a nécessairement des effets de propagation et donc l'existence d'interdépendances décalées dans l'espace et le temps [pigazzi1980interurban], étendant le domaine d'étude vers une meilleure compréhension des corrélations dynamiques.

Généralisation

Nous nous sommes limités au contrôle des premier et second moments des données générées, mais il est possible d'imaginer une généralisation théorique permettant le contrôle des moments à un ordre arbitraire. Toutefois, la difficulté de génération dans un cas concret complexe, comme le montre l'exemple géographique, questionne la possibilité de contrôle aux ordres supérieurs tout en gardant un modèle à la structure cohérente et au nombre de paramètres relativement faible. Par contre, l'étude de structures de dépendances non-linéaires comme celles utilisées dans [chicheportiche2013nested] est une piste de développement intéressante.

On a ainsi proposé une méthode abstraite de génération de données synthétiques corrélées à un niveau contrôlé. Son implémentation partielle dans deux domaines très différents montre sa flexibilité et l'éventail des applications potentielles. De manière générale, il est essentiel de généraliser de telles pratiques de validation systématique de modèles par étude statistique, en particulier pour les modèles agents pour lesquels la question de la validation reste encore relativement ouverte.

* * *

*

CONCLUSION DU CHAPITRE

Une question générale relativement ouverte concernant les systèmes urbains et celle du *lien entre forme et fonction*. Si dans certains cas et à certaines échelles, celui-ci est aisément extricable, il ne semble pas exister de règle générale ni de théorie répondant à ce problème fondamental. Les futures villes intelligentes seront-elles capables de totalement déconnecter la forme de la fonction comme le suppose [batty2017age]?

Si on se place à l'échelle d'un système de ville ou d'une méga-région urbaine, pour lesquels la forme se manifestera dans les positions relatives à la fois géographique, mais aussi selon des réseaux multi-couches, des villes selon leur spécialisations, ou dans la localisation fine des différents types d'activité dans la région et les liens formés par le réseau de transport, on peut supposer au contraire que les nouvelles formes urbaines seront liées de manière toujours plus intriquées et complexes avec leurs fonctions, à différentes échelles et selon différentes dimensions. La notion de morphogenèse, que nous avons définie et explorée partiellement, semble être bonne candidate pour lier forme et fonction puisque cette hypothèse fait partie intégrante de sa définition construite en 5.1. Un modèle simple comme celui étudié en 5.2 intègre ce paradigme sans pouvoir offrir d'interprétation possible puisque les fonctions sont implicites dans les processus considérés. En couplant le modèle au réseau de transport comme fait en 5.3, on introduit explicitement des notions de fonctions puisque par exemple l'accessibilité se met à jouer un rôle, mais aussi parce que le réseau est une fonction en lui-même. Ces paradigmes seront utilisés par la suite pour modéliser la co-évolution dans une perspective correspondante en 7.2, c'est à dire à l'échelle mesoscopique avec les mêmes hypothèses de processus autonomes et de sous-système bien défini. On poussera la reflexion du rôle des fonctions et d'une forme urbaine multi-dimensionnelle dans l'étude du modèle Lutecia en 7.3, qui intégrera la gouvernance du système de transport et les relations entre actifs et emplois dans une région métropolitaine.

★ ★

*

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 :

1. 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.
2. 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 forme 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.

Faisons à ce stade un bilan conceptuel de notre construction progressive.

Définition conceptuelle

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 causalité circulaires d'un point de vue statistique ; (iii) interdépendances systémiques globales.

Une caractérisation opérationnelle

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 variable. Dans le cas de causalité réciproques entre deux populations d'entités, nous aurons bien co-évolution au deuxième sens. Nous avons donc ainsi une caractérisation empirique et opérationnelle de la co-évolution.

L'approche morphogénétique

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 [holland2012signals]. 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.

Vers un approche de modélisation

En se raccrochant au triptyque des domaines de connaissance concepts-empirique-modèles [livet2010], nous pouvons considérer être armés 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 en profondeur.

Troisième partie

SYNTHÈSE

A partir des fondations et des briques constitutives, cette partie introduit la construction de modèles de co-évolution pour les réseaux de transport et les territoires, à deux échelles ayant chacune leur paradigme propre.

INTRODUCTION DE LA PARTIE III

Les rationnelles meso-macro font écho à Gibrat-Simon.

Ontologies : dans le macro, villes fixes, pas de nouvelles ville, mais nouveaux liens de réseau. Meso : tout évolue.

C : à ce stade, expliquer lien entre les différents modèles : utiliser appendice unified framework urban growth

ECHELLES ET PROCESSUS Partant des hypothèses tirées des enseignements empiriques et théoriques, on postulera *a priori* que certaines échelles privilégient certains processus, par exemple que la forme urbaine aura une influence au niveaux micro et mesoscopiques, tandis que les motifs émergeant des flux agrégés entre villes au sein d'un système se manifesteront au niveau macroscopique. Toutefois la distinction entre échelles n'est pas toujours si claire et certains processus tels la centralité ou l'accessibilité sont de bons candidats pour jouer un rôle à plusieurs échelles¹⁹ : il s'agira par la modélisation d'également tester ce postulat, par comparaison des processus nécessaires et/ou suffisants dans les familles de modèles à différentes échelles que nous allons mettre en place, en gardant à l'esprit des possibles développements vers des modèles multi-scalaires dans lesquels ces processus intermédiaires joueraient alors un rôle crucial.

We expect to produce *models of coevolution*, with the emphasis on processes of coevolution, to directly confront the theory. They will be necessary a flexible family because of the variety of scales and concrete cases we can include and we already began to explore in preliminary studies. Processes already studied can serve either as a thematic bases for a reuse as building bricks in a multi-modeling context, or as methodological tools such as synthetic data generator for synthetic control. Finally, we mean by operational models hybrid models, in the sense of semi-parametrized or semi-calibrated on real datasets or on precise stylized facts extracted from these same datasets. This point is a requirement to obtain a thematic feedback on geographical processes and on theory.

C : faire le même tableau pour les modèles existants : vue plus large de l'ensemble des processus, pour chacun de ces modèles et de nos modèles, lister tous les processus potentiels ; faire une typologie ensuite. Q : typologie différente d'une pure empirique ? à creuser, et peut être intéressant dans le cadre du knowledge framework, comme illustration coevol connaissances.

C : justifier ici pourquoi pas modèle très fin sur processus eco par exemple (//Levinson) : prix à payer pour être across scales, disciplines et avoir vraiment de la coevol ? pour ces premières étapes oui, à justifier

C : (Florent) expliciter la différence avec ce que tu as fait jusque là

¹⁹ on entend ici par "jouer un rôle" avoir une autonomie propre à l'échelle correspondante, c'est à dire qu'ils émergent *faiblement* des niveaux inférieurs.

Processus	Analyse Empirique	Echelles	Type <i>find a typology of processes</i>	Modèle
Attachement préférentiel		Croissance Urbaine		
Diffusion/Etalement		Forme Urbaine		
Accessibilité		Réseau / Ville		
Gouvernance des Transports				
Flux direct				
Flux indirect/Effet tunnel <i>c'est le même processus, vu sous un angle différent : l'effet tunnel est l'absence de nw feedback</i>				
Centralité de proximité (accessibilité : généralisation)				
Centralité de Chemin (correspond aux flux indirect : différents niveaux de généralité / sous-processus-sous-classif?)				
Proximité au réseau				
Distance au centre (similar to agrégation?)			RBD	

TABLE 13 :

6

CO-ÉVOLUTION À L'ECHELLE MACROSCOPIQUE

Les dynamiques des systèmes territoriaux, qui nous le rappelons impliquent dans notre cadre dynamiques couplées des territoires et des réseaux, à l'échelle macroscopique peuvent être partiellement appréhendées au moyen d'une approche par les interactions, comme montré au Chapitre 4. Pour rappeler les idées sous-jacentes de manière synthétique, en echo au point de vue par la morphogenèse développé en Chapitre 5 qui au contraire se concentre sur les règles autonomes au sein des sous-systèmes à une échelle intermédiaire , le principe dans cette ontologie est de raffiner le rôle des interactions en capturant les variations propres dans des processus abstraits endogènes simples. Le pouvoir explicatif est alors différent de celui des modèles économiques classiques et concerne d'autres types de processus, basés sur les interactions à des échelles d'espace plus grandes et des échelles de temps plus longues. Le rôle des réseaux de transports dans ce cadre est crucial, comme suggéré par les résultats préliminaires obtenus précédemment. Dans quelle mesure la construction du lien ferroviaire par le tunnel sous la Manche a-t-elle pu conforter le pouvoir économique de Londres ou renforcer ses interactions avec ses proches voisins Européens , et dans quelle mesure les événements politiques récents peuvent-ils conduire à une modification des trajectoires économiques puis par conséquent à une modification des motifs de transports par une rétroaction de la demande? D'une façon similaire, les projets de lignes à grande vitesse sur la côte Est des Etats-Unis et dans le corridor Californien sont-ils une conséquence attendue des dynamiques régionales ou un choix de gouvernance plus compliqué à cerner, et s'ils sont effectivement réalisés ,dans quelle mesure influenceront-ils les trajectoires du système de ville? Nous avons déjà étudié des questions analogues dans le cas de l'Afrique du Sud de manière empirique en 4.2, et nous proposons dans ce chapitre d'éclairer celles-ci à un plus grand niveau de généralité, en introduisant les processus de co-evolution dans les modèles d'interactions déjà développés. Pour donner une idée de la nature des enseignements qu'il est possible de tirer d'une telle approche, nous commençons en 6.1 par une exploration systématique du modèle SimpopNet, approche la plus avancée en termes de modélisation de la co-evolution au sein des systèmes de villes ,comme établi au Chapitre 2. Cela permet également d'introduire les indicateurs adaptés pour la compréhension des trajectoires des systèmes de villes en termes de dynamiques co-évolutives . Nous décrivons ensuite en 6.2 le modèle générique de co-évolution, qui est testé sur des

C (FL) : cela sera pour le chap8 (wrapup)

C (FL) : sens

C (AB) : suppr. en particulier, nous avons montré qu'il est possible de capturer

C (AB) : des interactions? préciser

C (FL) : ton cadre conceptuel n'est pas très coévolution : rôle transport → territoire

C (FL) : quel rapport? de plus échelle non evoluee jusqu'à présent

C (FL) : formulation maladroite, pas assez systémique

C (FL) : donc ce n'est plus le territoire? harmoniser

C (FL) : formulation à harmoniser

données synthétiques à deux niveaux de détail pour la représentation du réseau, puis sur le système de villes français.

★ ★

★

Ce chapitre est inédit pour sa première section. La deuxième section reprend les résultats de [] pour les données synthétiques, et va paraître prochainement comme [].

6.1 MODÈLES EXISTANTS

Nous proposons dans un premier temps d'introduire les modèles de co-évolution à l'échelle macroscopique en explorant les résultats produits par des modèles existants, ce qui permettra également d'introduire les méthodes et indicateurs nécessaires à l'exploration du modèle, ainsi que d'appréhender les questionnements typiques liés à ce type de modèles. En particulier, nous procéderons à une exploration systématique du modèle SimpopNet [schmitt2014modelisation], à notre connaissance l'une des rares initiatives pour modéliser la co-evolution au sein d'un système de villes.

6.1.1 Contexte

Quelle gain de connaissances obtenues peut s'observer, de la description conceptuelle ou thématique d'un modèle, à sa formalisation mathématique, son implémentation, son exploration systématique, jusqu'à son exploration approfondie à l'aide de meta-heuristiques spécifiques ? Notre postulat, qui découle à la fois de notre positionnement (voir chapitre 3 sur la simulation) et d'expériences dont les modèles déroulés précédemment font partie, est que celui-ci est important, mais surtout de nature *qualitative*, c'est à dire que la nature même des connaissances subit des transitions abruptes lors de l'avancée de la démarche dans ce continuum. Le modèle SimpopNet introduit par [schmitt2014modelisation], qui est à notre connaissance l'unique modèle de co-évolution dans une perspective de la théorie évolutive des villes, est un exemple d'une telle démarche préliminaire qui nécessite d'être creusée, par exemple par l'exploration systématique.

Description du Modèle

Nous reformulons brièvement le modèle, suivant les notations de la formulation du modèle d'interaction en 4.3, un certain nombre de paramètres et de processus se recoupant. Les villes croissent suivant la spécification de l'équation 4, avec $r_0 = 0$, $w_G = \lambda^\beta \cdot N$ et $V_{ij} = \mu_j/d_{ij}^\beta$. Le potentiel d'interaction ne dépend pas de la population de la ville d'origine, et le choix d'une fonction puissance permet de combiner un paramètre de décroissance λ à un paramètre de forme β . Le réseau croît à chaque pas de temps par rupture topologique : un couple de villes est choisi, la première selon les populations avec une hiérarchie γ_N (c'est à dire avec une probabilité proportionnelle à $\mu_i^{\gamma_N}$) et la seconde selon les potentiels d'interaction $\mu_i\mu_j/d_{ij}^\beta$ avec la même hiérarchie γ_N , puis un lien est créé si le réseau n'est pas assez efficace, i.e. $d_{ij}/d_{ij}^{(N)} > \theta_N$. Les liens créés à une date t ont une vitesse

C (AB) : reformuler
C (FL) : c'est une question trop générique pour du chap 6
C (AB) : _

C (AB) : me semble impropre fondé (?) comme terme (?) → de quoi ? des modèles, de leur comportement, de leur résultats ?

C (FL) : apport de ce paragraphe ?
C (AB) : reformuler

C (AB) : la rappeler
C (FL) : rappeler l'interprétation thématique des notations

C (AB) : est-ce le meilleur terme ?
C (FL) : sens ?

C (AB) : expliciter ce principe
C (FL) : sens efficace ?

$v(t)$, qui dépendra des technologies de transport courantes. La création de nouvelles intersection pour produire un graphe planaire n'est effectuée que pour les liens de vitesse semblable. Pour étudier une version stylisée du modèle, nous considérons une configuration simplifiée telle que $v(t > 0) = v_0$ et $v(0) = 1$ (le modèle initial considère trois valeurs de la vitesse correspondant aux réalités des technologies de transport entre 1830 et 2000).

C (FL) : alors présente peut être seulement la version simplifiée

C (FL) : pourquoi faire cette simplification ?

C (FL) : sens ?

C (FL) : sens ?

Perspectives

Certains choix de modélisation ne sont pas en cohérence directe avec l'application qui en est faite : par exemple, une telle précision dans la paramétrisation des dates et des vitesses (dates historiques de 1800 à 2000 et vitesse correspondant approximativement aux technologiques de transport) en fait un modèle hybride, et devrait correspondre à une application sur une configuration spatiale réelle. Dans une configuration stylisée, ces paramètres n'ont de sens que si l'on connaît le comportement des dynamiques simulées, et en particulier le rôle de la configuration spatiale, c'est à dire la séparation entre effets structurels et effets conjoncturels. D'autre part, l'utilisation du modèle d'interaction sans le terme de Gibrat endogène serait difficilement adaptable pour une application du modèle sur données réelles vu les valeurs obtenues dans les études précédentes des modèles d'interaction, mais est bien cohérent dans un modèle stylisé, afin de comprendre les processus d'interaction de manière isolée, comme nous le ferons plus loin (mais en gardant à l'esprit que cette connaissance ne reflète pas nécessairement le comportement couplé, l'interaction entre les processus pouvant faire émerger de nouveaux comportements). La formulation du potentiel, donnée ci-dessus, en $(\lambda/d_{ij})^\beta$ implique que λ capture à la fois le poids et la décroissance, mais permet moins de liberté que la spécification que nous avons utilisé précédemment, et ne permet pas une interprétation en terme de flux limite¹.

Enfin, les règles permettant des valeurs variables de $v(t)$ et le mécanisme de non-planarité du réseau², permet l'introduction d'un effet tunnel, qui on le rappelle est la non-interaction d'un infrastructure traversant un territoire avec celui-ci. Celui-ci est cependant exogène puisque spécifié explicitement dans les règles du modèle, contrairement au modèle d'interaction avec rétroaction des flux, dans lequel les variations de w_N et d_N doivent capturer un effet tunnel endogène. L'introduction d'indicateurs spécifiques pour le mesurer serait une piste intéressante de développement, mais nous nous contenterons

¹ Le paramètre de poids dans notre modèle en 4.3 donne en fait la valeur du flux lorsque l'atténuation par la distance tend vers l'infini et pour l'ensemble de la population.

² Lorsqu'un nouveau lien est construit, celui-ci ne forme des intersections qu'avec les liens de vitesse similaire.

de regarder ici la hiérarchie des centralités qui en est déjà un bon indicateur³.

6.1.2 Méthode

Configuration spatiale

Un aspect important de la compréhension des processus de co-évolution impliqués dans ce modèle est le rôle de la configuration spatiale initiale dans les motifs émergents observés. Nous appliquons pour cela la méthodologie développée en 3.1, permettant d'étendre l'analyse de sensibilité d'un modèle à des méta-paramètres spatiaux⁴.

GÉNÉRATION DE CONFIGURATION SYNTHÉTIQUE Une système de villes synthétique, respectant au premier ordre de manière visuelle les critères de l'état initial du modèle de base, est construit de la façon suivante (voir l'Appendice B.3 pour la notion de données synthétiques, calibrées au premier et second ordre). Un nombre fixé de villes N est réparti uniformément dans l'espace conditionnellement à une distance minimale entre chaque, et leur population est attribuée suivant une loi rang-taille dont les paramètres P_m et α peuvent être ajustés (la distribution du modèle initial correspond à $\alpha \simeq 0.68$ avec $R^2 = 0.98$). Un squelette de réseau est créé par un algorithme de connexification , qui connecte les villes deux à deux par plus proche voisin , puis itérativement sélectionne un cluster aléatoirement et le connecte perpendiculairement au lien le plus proche hors du cluster. Le réseau est ensuite étoffé par la création de raccourcis locaux, par répétitions n_s fois de la sélection aléatoire d'une ville selon les populations, et sa connexion à un voisin dans un rayon r_s sous conditions de degré maximal d_s . Le réseau final est ensuite planarisé. Cette procédure crée des réseaux correspondant visuellement à l'initialisation du modèle, sachant qu'une instance du réseau ne permet pas de déterminer les distributions de paramètres topologiques sur lesquels une calibration plus fine pourrait être opérée.

C (AB) : expliciter

C (FL) : ?

C (FL) : c'est flou : mesure de distance? espace euclidien? pourquoi faire cela?

C (AB) : préciser critères visuels

Indicateurs

Un aspect crucial de l'étude des modèles de simulation est la définition d'indicateurs pertinents, surtout dans le cas de modèles synthétiques où il n'est pas possible de produire des sorties directement liées aux données par exemple. Des faits stylisés très généraux,

³ En effet, une distribution très hiérarchique des accessibilités signifie qu'il existe un petit nombre de villes très accessibles et un grand nombre peu accessibles. Si les villes importantes couvrent raisonnablement l'espace, alors leurs liens ignorent nécessairement les villes peu accessibles survolées, sinon la distribution serait moins hiérarchique.

⁴ Nous rappelons que dans notre cas qu'un méta-paramètre est un paramètre permettant de générer une configuration initiale en amont du modèle.

comme vouloir produire une hiérarchie urbaine ou une hiérarchie de réseau, sont relativement limités. De plus, la hiérarchie est produite mécaniquement par la majorité des modèles incluant des processus d'agrégation. Il faut donc des indicateurs plus élaborés pour comprendre les dynamiques du système. Ces indicateurs doivent notamment apporter des éléments de réponse aux questions suivantes :

- types de systèmes de villes sont produites par le modèle ;
- changement dans le temps dans l'organisation du système de ville ;
- profils typiques de trajectoires ;
- capacité à "produire de la co-évolution".

Pour se concentrer sur la capacité du modèle à produire des trajectoires à la fois diverses et complexes, et par exemple sa capacité à produire des bifurcations qui se traduirait par inversions de rang, ainsi que sa capacité à capturer différents aspects des dynamiques co-évolutives, nous proposons un jeu d'indicateurs, incluant par exemple des mesures de corrélation retardée en écho aux régimes de causalité exhibés en 4.2, ou une mesure de corrélation en fonction de la distance, pour comprendre le rôle des interactions spatiales dans les couplages de trajectoires. Etant donné une variable $X_i(t)$ définie sur chacune des villes et dans le temps (qui pourra être la population ou des mesures de centralité par exemple), nous définissons les indicateurs :

- Indicateurs caractérisant la distribution : hiérarchie (pente du rang-taille) $\alpha(t)$, entropie $\varepsilon(t)$, statistiques descriptives $E[\hat{X}](t)$, $\hat{\sigma}(t)$, de la distribution de X_i dans le temps
- Corrélation de rang entre l'instant initial et l'instant final, qui traduit la quantité de changement dans la hiérarchie lors de l'évolution du système : $\rho[X_i(t=0), X_i(t=t_f)]$
- Diversité des trajectoires $D[X_i]$, qui capture la diversité de profil des séries temporelles, avec $\tilde{X}_i(t) \in [0; 1]$ les trajectoires mises à l'échelle individuellement,

$$\frac{2}{N \cdot (N-1)} \sum_{i < j} \left(\frac{1}{T} \int_t (\tilde{X}_i(t) - \tilde{X}_j(t))^2 \right)^{\frac{1}{2}}$$

C (FL) : confusion possible avec forme urbaine et puis qu'est ce que cela signifie ?

C (AB) : expliciter cette variance → cf méthode alternative Jasss

- Changements de direction des trajectoires $C[X_i]$, que nous prenons comme le nombre de points d'inflection. Dans le cadre de ce type de modèle, qui produit majoritairement des trajectoires monotones, cet indicateur témoigne dans une certaine mesure d'une "complexité" des trajectoires.

- Corrélations en fonction de la distance, pour comprendre la manière dont l'effet de la distance est traduit au niveau macroscopique. Le profil de cette fonction, au regard des valeurs des paramètres de distance d'interaction inclus dans le modèle, traduira la tendance du modèle à faire émerger tel ou tel niveau d'interaction.

$$\hat{\rho}_d [(X(\vec{x}_1, Y(\vec{x}_2)) \mid \|\vec{x}_1 - \vec{x}_2\| \sim d]$$

- Corrélations retardées entre les variations, pour identifier des motifs de causalité entre les variables X et Y. Les motifs $\hat{\rho}_\tau$ pour l'ensemble des variables, et pour τ retard ou anticipation, sont à lire dans le sens des régimes potentiels, explorés en 4.2.

$$\hat{\rho}_\tau [\Delta X(t - \tau), \Delta Y(t)]$$

Ces indicateurs sont utilisés pour les populations $\mu_i(t)$, les centralités de proximité $c_i(t) = \frac{1}{N-1} \sum_{i \neq j} \frac{1}{d_{ij}(t)}$ qui capturent la position dans le système urbain, et les accessibilités $Z_i = \frac{1}{\sum_k \mu_k} \sum_{i \neq j} P_j \exp(-d_{ij}(t)/d_G)$ qui capturent l'insertion dans le système urbain.

Nous introduisons de plus divers indicateurs de topologie du réseau, pour comprendre les formes finales produites par l'heuristique : diamètre, longueur moyenne de chemin, centralité de chemin moyenne et niveau de hiérarchie, performance moyenne, longueur totale, comme ils ont été définis en 4.1.

6.1.3 Résultats

Plan d'expérience

Etant donné une configuration spatiale initiale (c'est à dire une valeur des meta-paramètres), nous établissons le comportement des indicateurs par l'exploration d'une grille de l'espace des paramètres. Le nombre de paramètres étant restreint et l'objectif étant un premier aperçu du comportement du modèle, nous ne faisons pas appel à des méthodes d'exploration plus élaborées. Les paramètres sont $(d_G, \gamma_G, \gamma_N, \theta_N, v_0)$ et les méta-paramètres $(N_S, \alpha_S, d_S, n_S)$. Nous explorons une grille de 16 configurations des meta-paramètres, 324 configurations de paramètres, et 30 réplications aléatoires, ce qui correspond à 155,520 simulations. Celles-ci sont exécutées sur grille de calcul par l'intermédiaire d'OpenMole⁵.

C (AB) : pas clair

C (FL) : non compréhensible

C (FL) : ce n'est pas une attente suffisante : cette exploration est nécessairement au service d'un questionnement plus vaste

C (FL) : tu ne dis pas comment tu as fait ces choix (cad dans quelle perspective)

⁵ Les résultats de simulation sont disponibles à <http://dx.doi.org/10.7910/DVN/RW8S36>.

Convergence

Le modèle étant stochastique, il est important de contrôler la convergence des indicateurs, qui sera plus ou moins facile selon leur variabilité. Pour quantifier la variabilité d'un indicateur X par rapport à la stochasticité, nous utilisons une mesure similaire à celle de 5.2, donnée par $v[X] = \mathbb{E}[\hat{X}] / \sigma[\hat{X}]$ avec les estimateurs basiques pour l'espérance et l'écart-type. Sur l'ensemble des réplications, on obtient sur l'ensemble des indicateurs donnés précédemment, une médiane pour le ratio $v[X]$ estimé au sein des réplications, estimée sur toutes les valeurs des paramètres, qui prend une valeur minimale de 3.94, pour la moyenne de l'accessibilité à l'instant final, ce qui témoigne d'une faible variabilité stochastique. On peut de plus utiliser cette valeur pour estimer le niveau de convergence : elle correspond à un intervalle de confiance à 95% autour de la moyenne de taille relative 0.18 (sous hypothèse de distribution normale de la moyenne), c'est à dire une bonne convergence. Cet aspect est essentiel pour la robustesse des résultats.

Sensibilité à l'espace

C (AB) : pas vraiment exploitez

La table 14 donne les valeurs de \tilde{d} pour 16 configurations des métaparamètres⁶, par rapport à une configuration de référence arbitraire (première colonne). La hiérarchie au sein du système de villes initial est le plus fort déterminant, puisque l'ensemble des configurations avec $\alpha_S = 1.5$ donne des valeurs supérieures à 1.7, ce qui témoigne d'une très forte sensibilité relative à cette hiérarchie. Ensuite, le nombre de villes joue un rôle secondaire non-négligeable, donnant les plus forts effets de l'espace. Ainsi, il est crucial de garder à l'esprit ce rôle de la configuration initiale lors de l'analyse des diagrammes de phase. Pour rester dans l'esprit du modèle initialement proposé, nous commenterons toutefois un diagramme de phase pour une configuration spatiale donnée. L'étude du modèle étendu avec intégration des meta-paramètres auquel il est sensible comme paramètres à part entière est hors de portée de cette analyse auxiliaire.

C (AB) : pas clair

C (FL) : tu peux parler de ça de façon plus simple

Motifs

C (FL) : pourquoi ces choix?

La Fig. 38 rend compte du comportement du modèle selon les divers indicateurs donnés ci-dessus. Nous commentons une configuration spatiale particulière qui correspond à un système peu hiérarchisé avec un réseau n'ayant que des raccourcis locaux, donnée par les métaparamètres $N_S = 80$, $\alpha_S = 0.5$, $d_S = 10$, $n_S = 30$. Les graphes complets sont disponibles en Appendice A.10.

⁶ La définition de la mesure relative de sensibilité, donnée en 3.1, est pour deux diagrammes de phase f_1, f_2 et d distance euclidienne, $\tilde{d} = 2d(f_1, f_2) / (\text{Var}[f_1] + \text{Var}[f_2])$.

TABLE 14 : Sensibilité à l'espace du modèle SimpopNet. Chaque colonne correspond à une instance du diagramme de phase, pour laquelle les méta-paramètres sont donnés, ainsi que la distance relative à un diagramme de référence arbitraire. En entrée on a les méta-paramètres N_S , α_S , d_S , n_S et en résultat des simulation la distance \tilde{d} .

N_S	40	40	40	40	40	40	40	40	80	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	1.5
d_S	5	5	10	10	5	5	10	10	5	5	10	5	5	10	10	10
n_S	10	30	10	30	10	30	10	30	10	30	10	30	10	30	10	30
\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	2.21

Les valeurs prises par l'entropie pour les centralités (premier panel de la Fig. 38), en fonction du temps, pour $\gamma_N = 2.5$ et $v_0 = 110$, exhibent différents régimes en fonction de d_G et γ_G . Une faible hiérarchie conduit à une entropie se stabilisant dans le temps, correspondant à une certaine uniformisation des distances. Au contraire, une forte hiérarchie produit un régime avec un minimum, puis une augmentation des disparités dans le temps.

Cette variété de comportements se retrouve avec la corrélation de rang ρ_R , que nous montrons ici pour la variable de population, en fonction de d_G . Celle-ci est peu sensible à θ_N et γ_N , mais varie fortement en fonction de d_G et γ_G : des interactions à plus longue distance induisent systématiquement un plus grand nombre d'inversions de hiérarchie (qui est la notion capturée par cet indicateur), et celles-ci peuvent avoir lieu quand la hiérarchie du potentiel gravitaire est faible. En résumé, l'augmentation de la portée des interactions complexifie la trajectoire du système de villes, tandis que l'augmentation de leur hiérarchie la simplifiera.

Le comportement des indicateurs de corrélation est montré en Fig. 39. Concernant l'effet de la distance sur les corrélations entre variables, c'est à dire l'évolution de ρ_d , il est intéressant de noter que l'augmentation de d_G diminue systématiquement les niveaux de correlation, ce qui correspond à la complexification mise en valeur précédemment. Comme attendu, $\rho_d [d]$ décroît en fonction de la distance, et des valeurs non-nulles pour la correlation entre population et centralité pour une forte hiérarchie γ_G , ce qui montre que les régimes d'adaptation simultanée sont rares dans ce modèle.

De même, les régimes de causalité au sens de 4.2, sont relativement pauvres : la population est systématiquement causée par la centralité, mais il n'existe pas de régime où l'on observe le contraire. On est dans la logique d'un effet de renforcement de la hiérarchie par la centralité, mais pas dans une configuration avec causalités circulaires, et donc pas dans une co-évolution à proprement parler comme nous l'avons définie au sens statistique. Cette exploration brève nous permet d'affirmer que ce modèle capture des trajectoires urbaines d'une certaine

C (FL) : expliciter plus clairement ce que cela signifie

C (FL) : B

C (FL) : ?

C (FL) : alors est ce que la hiérarchie est un input ou bien un output dans ce modèle ?

C (AB) : Aerer

C (AB) : faire des ponts avec le thématique

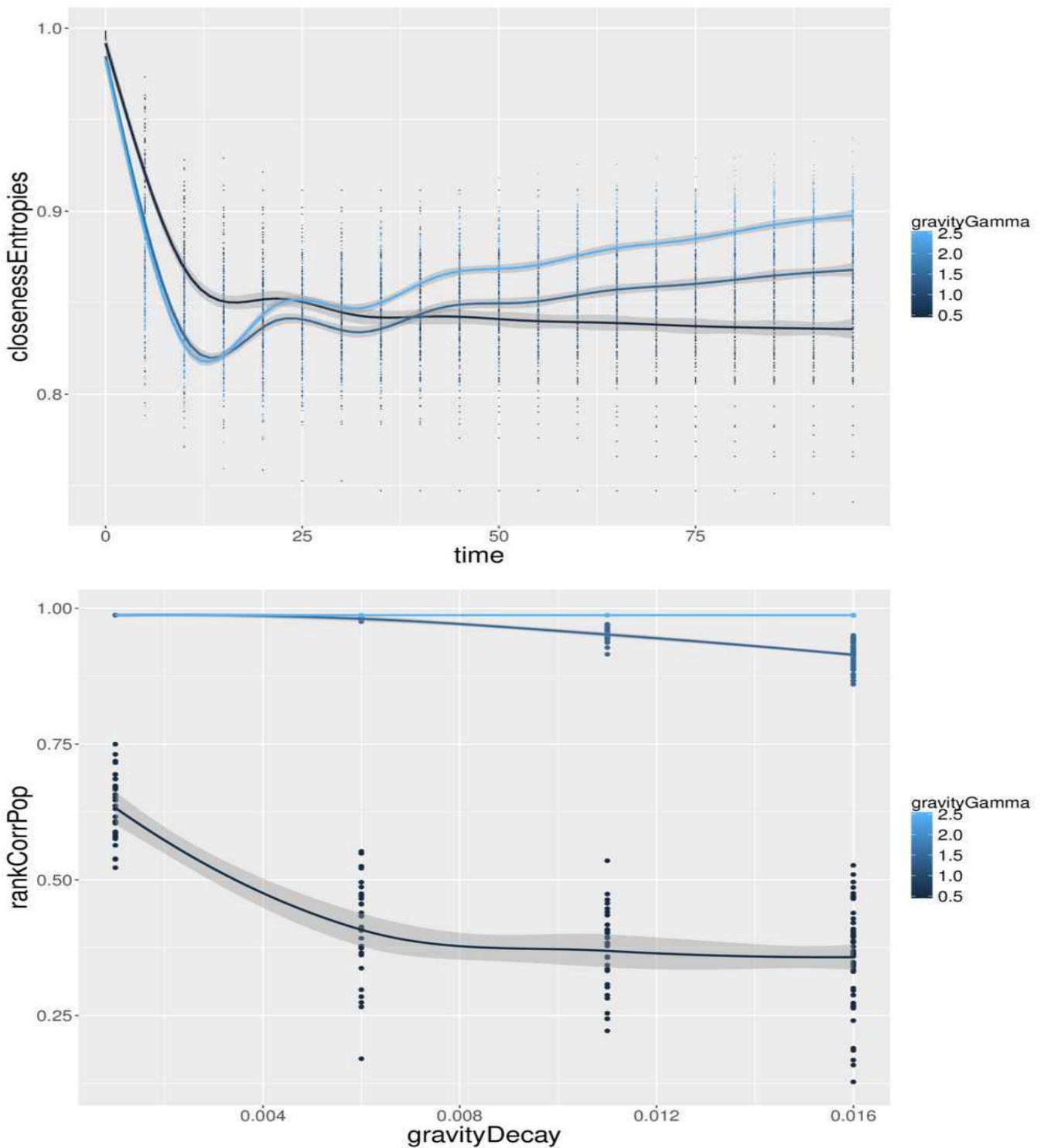


FIGURE 38 : Comportement du modèle pour la configuration spatiale $N_S = 80$, $\alpha_S = 0.5$, $d_S = 10$, $n_S = 30$. (Haut) Trajectoires temporelles de l'entropie des centralités de proximité, pour $\gamma_N = 2.5$, $v_0 = 110$, $d_G = 0.016$, $\theta_N = 11$, en fonction de γ_G (couleur); (Bas) Corrélation de rang pour la population, en fonction de d_G et de γ_G (couleur), pour $\theta_N = 11$, $\gamma_N = 2.5$; (bas Gauche) Corrélations en fonction de la distance, pour les couples de variables (couleur), pour $\gamma_N = 2.5$, $\theta_N = 21$, $v_0 = 10$, et pour d_G (colonnes) et γ_G (lignes) variables; (Bas Droite) Corrélations retardées pour les mêmes paramètres.

complexité, mais qu'il ne reproduit pas des régimes de co-évolution mais de co-adaptation. Par la suite, nous explorerons dans le même esprit une extension co-évolutive du modèle d'interaction développé en 4.3, et chercherons à établir dans quelle mesure il est capable de capturer des dynamiques co-évolutives.

★ ★

★

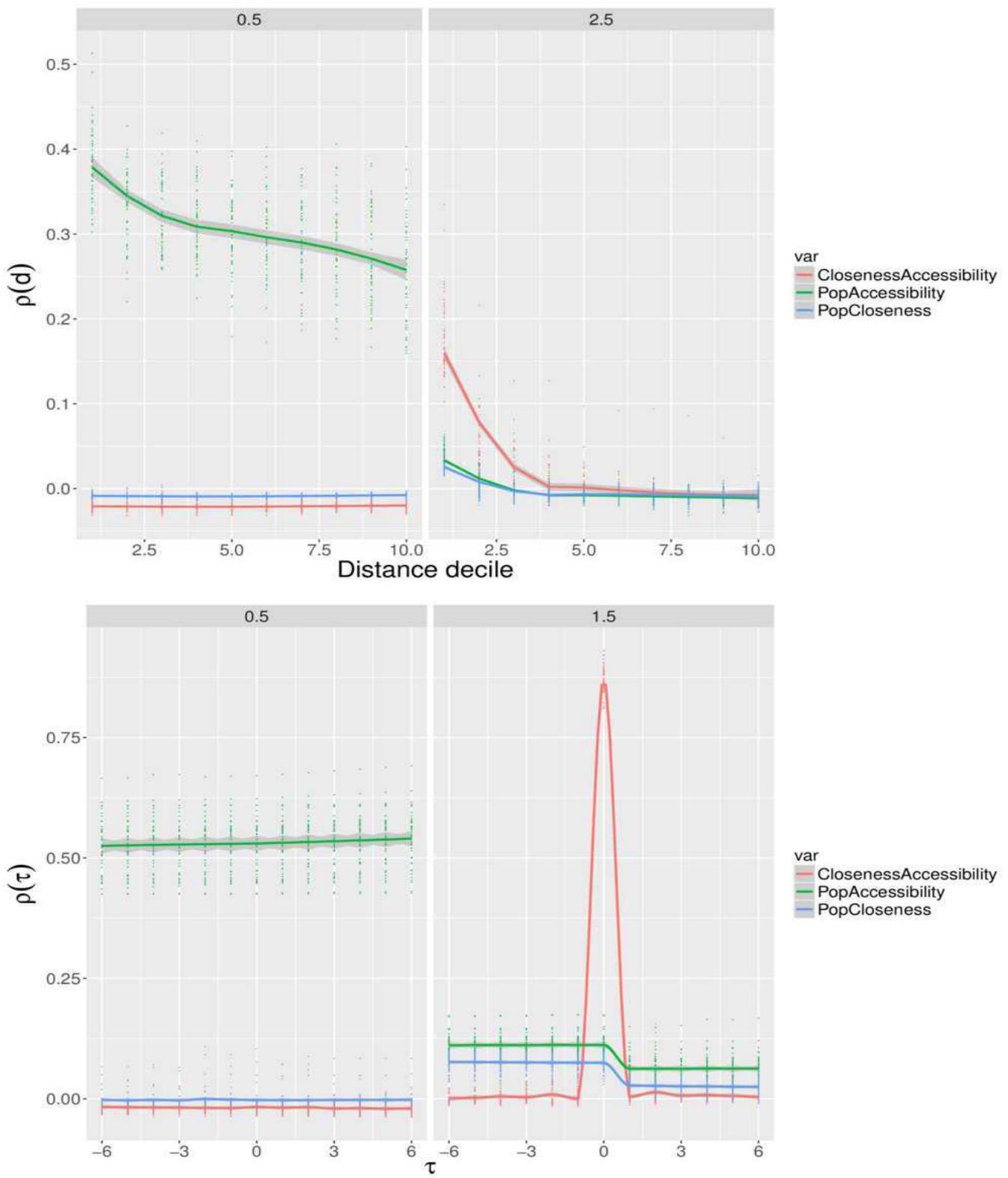


FIGURE 39 : Corrélations dans le modèle pour la configuration spatiale $N_S = 80$, $\alpha_S = 0.5$, $d_S = 10$, $n_S = 30$. (Haut) Corrélations en fonction de la distance, pour les couples de variables (couleur), pour $\gamma_N = 2.5$, $\theta_N = 21$, $v_0 = 10$, et pour d_G (colonnes) et γ_G (lignes) variables ; (Bas) Corrélations retardées pour les mêmes paramètres.

6.2 EXTENSION DYNAMIQUE DU MODÈLE D'INTERACTION

Nous pouvons à présent faire la synthèse d'une part des paradigmes d'intégration d'un système de ville et du réseau de transport, effectué de manière statique pour le comportement du réseau dans le modèle d'interaction développé et exploré en section 4.3, d'autre part des indicateurs pour la compréhension d'un modèle de co-évolution pour un système de villes, expérimentés dans la section précédente 6.1, ainsi que de manière indicative pour comparaison des comportements obtenus pour le modèle SimpopNet. Cette synthèse consiste en une première formulation d'un *modèle de co-évolution macroscopique pour les systèmes de villes*, qui est un élément clé pour apporter un éclaircissement partiel de notre problématique générale.

C (FL) : reprendre tous ces concepts : modèle, indicateur etc sont ici trop mis au même niveau

C (AB) : reformuler

6.2.1 Modèle macroscopique de co-évolution

Hypothèses et choix de modélisation

Cette première approche se place dans une logique d'extension directe du modèle d'interactions au sein d'un système de villes présenté en chapitre 4, c'est à dire à une échelle macroscopique et avec une ontologie typique aux systèmes de villes. Toujours dans un choix de simplicité, nous restons ici à une description unidimensionnelle des villes par leur population. Concernant la croissance du réseau, nous proposons de nous placer également à un niveau relativement agrégé et simplifié, en permettant de tester des heuristiques de croissance à différents niveaux d'abstraction. Dans une logique de multimodélisation, le modèle peut prendre en compte divers processus comme les interactions directes entre les villes, les interactions intermédiaires par le réseau, la rétroaction des flux de réseau et une croissance du réseau induite par la demande. Les éléments empiriques mis en valeur pour le réseau ferré français par [thevenin2013mapping] suggèrent l'existence de retroactions de l'utilisation du réseau, ou des flux le traversant, sur sa persistence et son développement, dont les propriétés ont évolué dans le temps : une première phase de développement fort correspondrait à une réponse à des demandes même faibles, tandis que le renforcement des liens principaux et la disparition des liens faibles observés plus tard s'interpréterait comme une rétroaction dont le signe est fixé par seuil.

C (AB) : _

C (FL) : sens ?

C (FL) : ?

C (AB) : ?

C (AB) : pas super clair

C (AB) : pas terrible, p 272 N = effectif, ce qui est plus logique

Formulation Générique

Le système urbain est caractérisé par les populations $\mu_i(t)$ et le réseau $G(t)$ auquel on peut associer une matrice de distance $d_{ij}^N(t)$. Les flux entre villes ϕ_{ij} suivent les expressions données en 4.3 avec la distance réseau. De la même manière, la variation des populations suit les

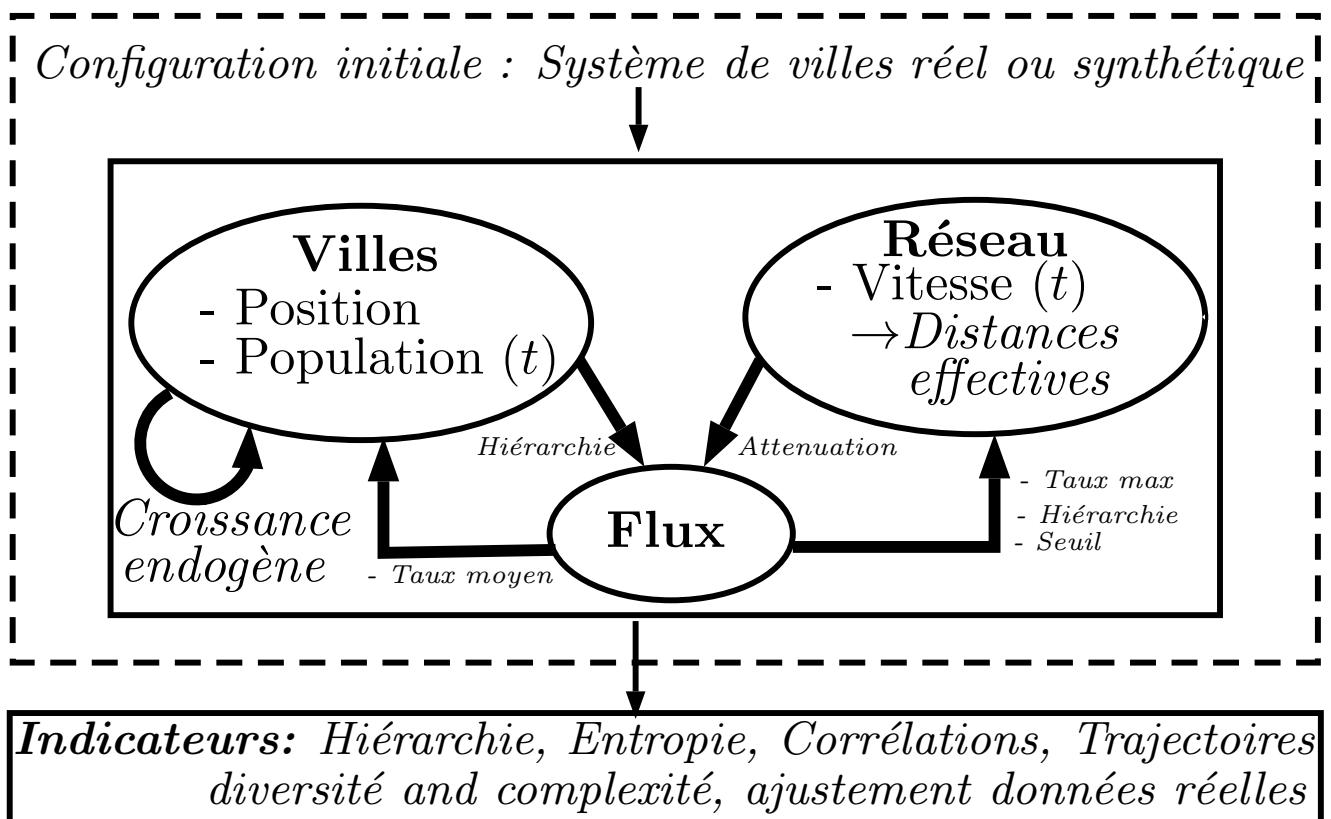


FIGURE 40 : **Représentation abstraite du modèle.** Les ovales correspondent aux éléments ontologiques principaux (Villes, Réseau, Flux), tandis que les flèches traduisent des processus et les paramètres associés sont indiqués. Le modèle est décrit dans son écosystème plus large d'initialisation et d'indicateurs de sortie.

spécifications du modèle de base. La Fig. 40 exprime le modèle sous forme schématisée.

CROISSANCE DU RÉSEAU Concernant le réseau, nous faisons l'hypothèse que celui-ci évolue suivant $\mathbf{N}(t+1) = F(\mathbf{N}(t), \phi_{ij}(t))$, de telle façon qu'une assignation des flux dans le réseau ainsi qu'un variation locale de ses éléments est possible. Nous proposons dans un premier temps de nous intéresser aux motifs liés à la distance uniquement, et de spécifier une relation sur un réseau abstrait uniquement par $d_{ij}^N(t+1) = F(d_{ij}^N(t), \phi_{ij}(t))$. Dans cette logique, nous restons dans un modèle d'interaction à l'échelle macroscopique uniquement (puisque qu'un spatialisation précise du réseau implique la prise en compte d'une échelle plus fine).

C (FL) : la différence entre les deux formules est subtiles, expliciter

C (AB) : développer

Suivant l'heuristique de rétroaction par seuil, étant donné un flux ϕ dans un lien, on suppose que sa distance effective est mise à jour par :

$$d(t+1) = d(t) \cdot \left(1 + g_{\max} \cdot \left[\frac{1 - \left(\frac{\phi}{\phi_0} \right)^{\gamma_s}}{1 + \left(\frac{\phi}{\phi_0} \right)^{\gamma_s}} \right] \right)$$

C (AB) : donner une intuition + references

[tero2007mathematical]

avec γ_s un paramètre de hiérarchie, ϕ_0 le paramètre de seuil et g_{\max} le taux de croissance maximal à chaque étape.

Implémentation

Le couplage du modèle d'interaction à une explicitation du réseau plus fine (par exemple encodage de l'ensemble de la structure du réseau) rend plus difficile l'intégration complète dans un plugin OpenMole comme c'était le cas pour le modèle étudié en 4.3, nécessitant une implémentation ad hoc. L'utilisation d'un workflow comme médiateur pour le couplage est une solution intéressante mais réaliste uniquement dans le cas d'un couplage faible. L'un des défis que devra relever la bibliothèque de métamodélisation en cours de développement autour d'OpenMole, serait la possibilité de coupler fortement (par exemple au sens de dynamiquement dans l'évolution de la simulation) des composantes hétérogènes de manière transparente, permettant de tirer parti des avantages de différents langages ou d'implémentation déjà existantes.

Nous optons pour ce modèle pour une implémentation complète en NetLogo pour une simplicité de couplage des composantes. Une attention particulière est portée à la dualité de la représentation du réseau, à la fois sous forme de matrice de distance et sous forme physique, pour permettre facilement l'extension à des heuristiques de réseau physique.

6.2.2 Application à des Données Synthétiques

Le modèle est d'abord testé et exploré sur des systèmes de villes synthétiques, afin de comprendre certaines de ses propriétés intrinsèques. Dans ce cas, nous considérons le modèle avec réseau abstrait comme spécifié ci-dessus, c'est à dire sans explicitation spatiale du réseau et avec les règles d'évolution agissant directement sur d_{ij}^N selon les spécifications données précédemment.

Données Synthétiques

Un système de villes synthétiques est généré, en suivant l'heuristique utilisée dans la section précédente : (i) des villes en nombre N_S sont placées aléatoirement dans le plan euclidien ; (ii) les populations sont attribuées aux villes selon une loi de puissance inverse, avec un paramètre de hiérarchie α_S et de telle façon que la plus grande ville ait une population P_{\max} , c'est-à-dire suivant $P_i = P_{\max} \cdot i^{-\alpha_S}$. Pour simplifier, nous fixons un certain nombre de méta-paramètres : le nombre de villes est fixé à $N_S = 30$, la population maximale à $P_{\max} = 100000$ et la croissance maximale du réseau à $g_{\max} = 0.005$. Le temps final est fixé à $t_f = 30$, ce qui correspond à des distances divisées par 5 environ⁷, afin de respecter un critère empirique : cela correspond à un passage du Paris-Lyon en une dizaine d'heures au début du 19ème siècle à deux heures aujourd'hui, mis en évidence par exemple par [thevenin2013mapping]. Nous négligeons aussi les effets de réseau au second ordre en fixant $w_N = 0$.

Nous explorons une grille de l'espace des paramètres $(\alpha_S, \phi_0, \gamma_s, w_G, d_G, \gamma_G)$. Nous utilisons les indicateurs introduits en 6.1 pour quantifier le comportement du modèle dans l'espace des paramètres. En Fig. 41, nous montrons l'évolution d'indicateurs dans le temps ainsi que des mesures agrégées, pour une grande partie de l'espace des paramètres couvert. Nous montrons les résultats pour $\alpha_S = 1$, valeur la plus proche d'un système de villes réel (par rapport à 0.5 et 1.5, voir la revue systématique des estimations de la loi rang-taille faite par [10.1371/journal.pone.01839]).

Trajectoires

L'évolution de la centralité de proximité moyenne dans le temps est visualisée en Fig. 41 (haut) pour $w_G = 0.001$, et à (γ_G, ϕ_0) variables. Le comportement n'est pas sensible à d_G (voir graphique complet en A.11). Cette évolution témoigne d'une transition en fonction du niveau de hiérarchie : lorsque celui-ci décroît, on observe l'émergence de trajectoires où la centralité moyenne croît dans le temps, ce qui

⁷ En effet, on peut calculer que le facteur multiplicatif minimal pour la distance est de $(1 - g_{\max})^{t_f}$, ce qui donne pour ces valeurs $(1 - 0.05)^{30} \simeq 0.214$, c'est à dire une division par 5 du temps de trajet.

correspond à des situations où l'ensemble des villes bénéficie d'accroissements d'accessibilité.

En terme d'entropie des populations, dont nous traçons la trajectoire temporelle en Fig. 41 (bas), l'ensemble des paramètres donne une entropie décroissante, c'est à dire des comportements de convergence des trajectoires des villes dans le temps⁸.

Lorsqu'on s'intéresse à la complexité des trajectoires d'accessibilité, on note pour des valeurs de $\phi_0 > 1.5$ un maximum de la complexité en fonction de la distance d'interaction d_G , stable lorsque w_G et γ_G varient. Cette échelle intermédiaire peut être interprétée comme produisant des sous-systèmes régionaux, assez grands pour développer chacun un certain niveau de complexité, et assez isolés pour ne pas uniformiser les trajectoires sur l'ensemble de l'espace. Nous reconstruisons ainsi une non-stationnarité spatiale, typiquement observée en 4.1, et rejoignons le concept de niche écologique⁹ localisée dans l'espace : les sous-systèmes émergents, relativement indépendants, sont de bons candidats pour être porteurs de processus de co-évolution.

C (AB) : développer

Enfin, le comportement des corrélations de rang pour l'accessibilité révèle que la distance d'interaction augmente systématiquement le nombre d'inversions de hiérarchie, ce qui correspond en un sens à une augmentation de la complexité globale du système. Le paramètre de hiérarchie diminue quant à lui cette corrélation, ce qui veut dire qu'une évolution plus hiérarchique affectera un plus grand nombre de villes dans l'aspect qualitatif de leur trajectoires.

Corrélations

Tournons nous à présent vers les motifs de corrélation produits par le modèle, illustrés en Fig. 43.

En fonction de la distance, les profils de ρ_d pour les trois couples de variables montrent que des valeurs moyennes et grandes de la distance d'interaction ($d_G > 50$) induisent des populations totalement décorrélées aux centralités et accessibilités. Pour des petits d_G , un profil décroissant puis nul confirme l'existence d'effets locaux forts, où des villes très proches s'influenceront fortement. Le comportement de la correlation entre accessibilité et centralité est plus difficile à interpréter, et peut être dû aux phénomènes d'auto-correlation¹⁰. Son

⁸ En effet, l'entropie pour la variable de population exprime la dispersion de la distribution des populations, et donc une décroissance de celle-ci exprime une tendance à la concentration dans le temps.

⁹ Comme nous l'avons déjà présenté en 5.1, une niche écologique au sens de [holland2012signals] correspond à l'écosystème relativement indépendant au sein de laquelle il y a co-évolution entre les espèces.

¹⁰ qui ne sont pas calculables, car il s'agirait de décomposer $\rho \left[\sum_{i \neq j} \frac{1}{d_{ij}}; \sum_{i \neq j} P_j \exp(-d_{ij}/d_G) \right]$. Il est possible par exemple d'approximer $\rho[X+Y; Z]$ sous la condition que $\varepsilon = \sigma_Y/\sigma_X \ll 1$ au premier ordre par

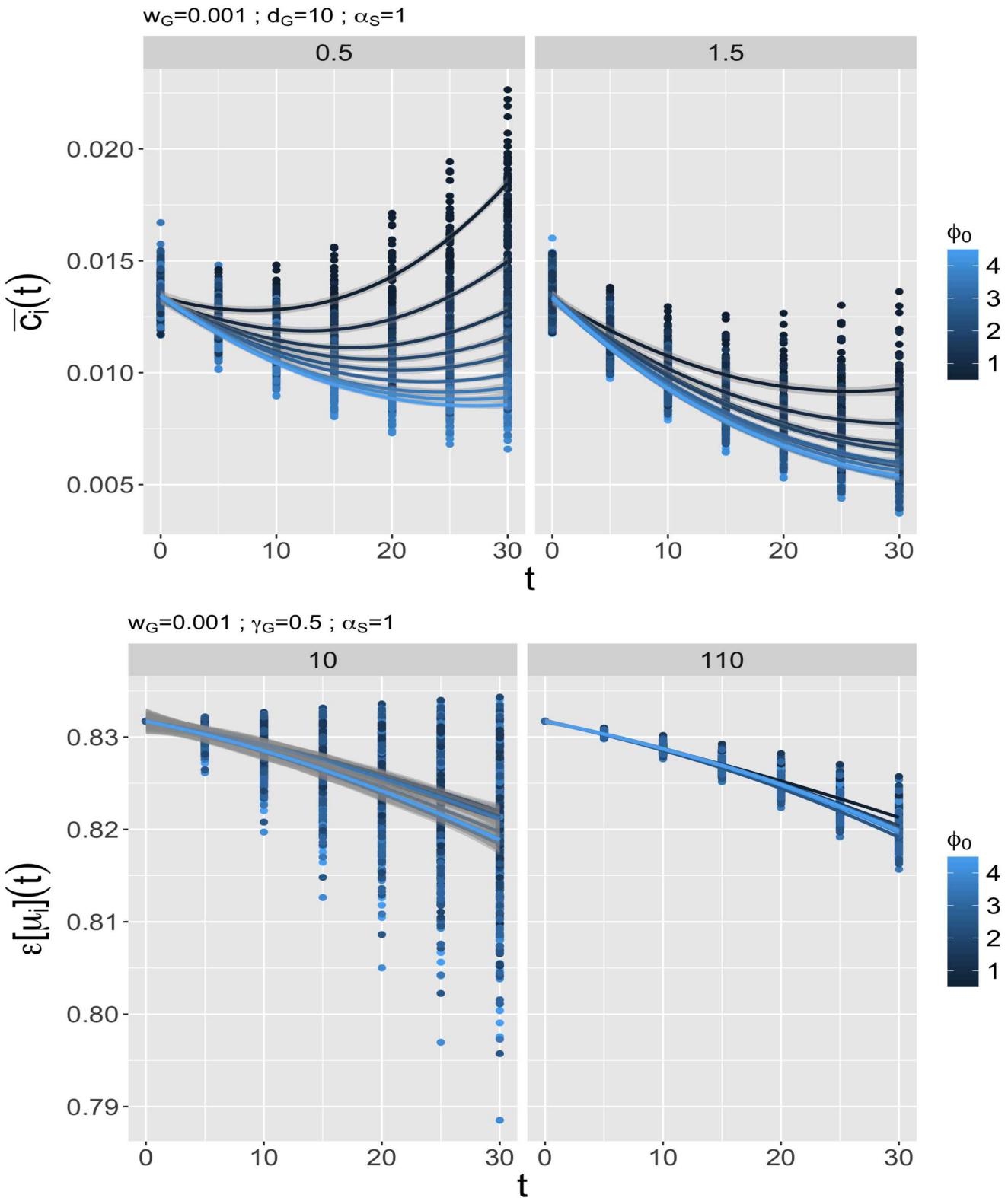


FIGURE 41 : Comportement temporel du modèle de co-evolution avec réseau abstrait sur un système de villes synthétique. (Haut) Moyenne des centralités de proximité, en fonction du temps, pour d_G (colonnes), γ_G (lignes) et φ₀(couleur) variables, à w_G = 0.001 fixé; (Bas) Entropie de populations, en fonction du temps, pour d_G (colonnes), γ_G (lignes) et φ₀(couleur) variables, à w_G = 0.001 fixé Se référer au texte pour l'interprétation.

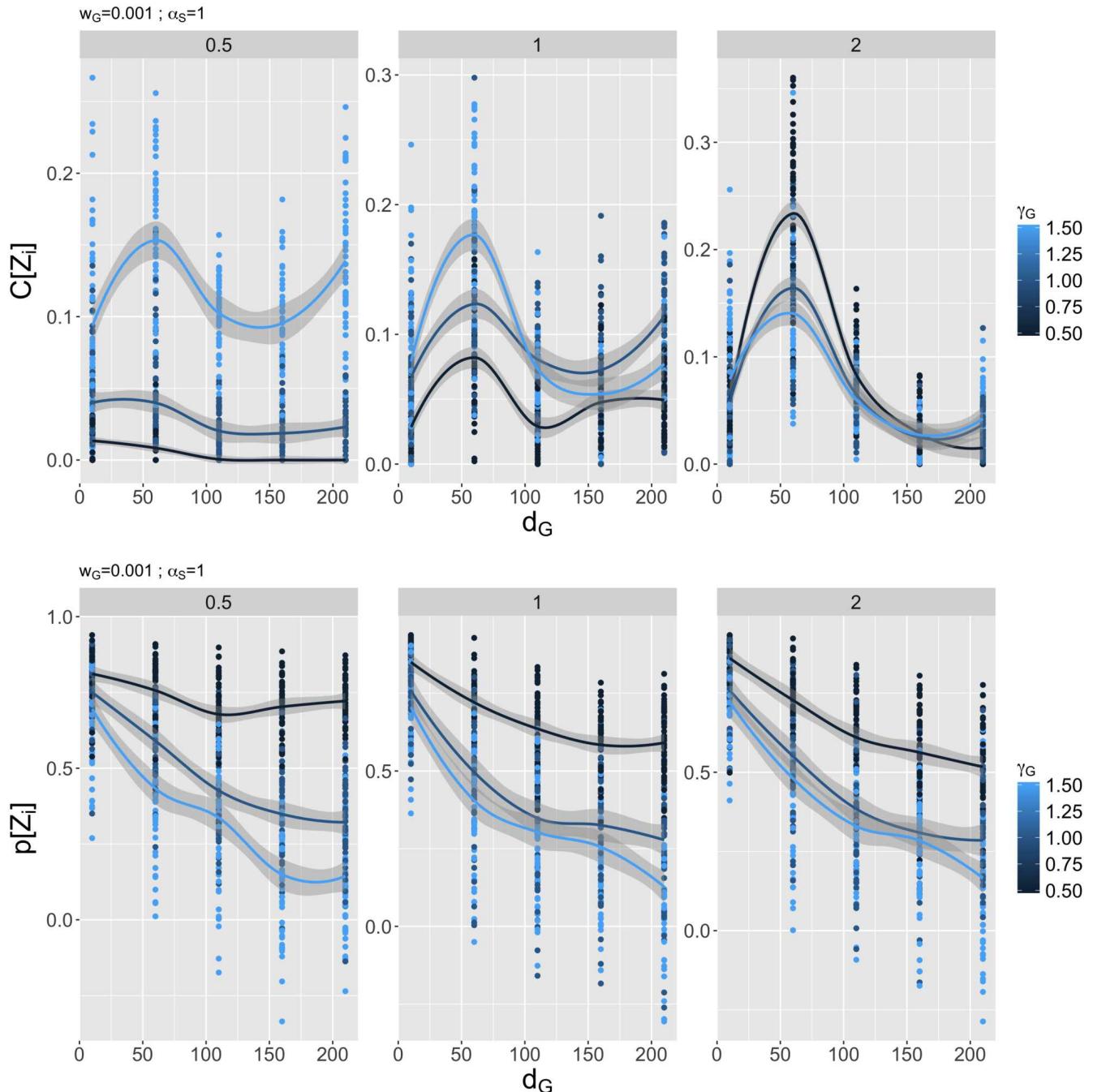


FIGURE 42 : Comportement agrégé du modèle de co-evolution (Haut) Complexité des accessibilités, en fonction de d_G , pour ϕ_0 (colonnes), w_G (lignes) et γ_G (couleur) variables; (Bas) Corrélation de rang des accessibilités, pour les mêmes paramètres.

niveau ne dépend pas de la distance mais de d_G , et est décroissant pour finir à une corrélation négative.

Enfin, concernant les corrélations retardées, on observe toujours une corrélation synchronisée, et a priori l'existence de τ_+ ou τ_- dans certains configurations, mais difficile à valider visuellement. Nous ajoutons donc ici un critère basé sur un test statistique : pour $\tau = \text{argmax}_{\rho_\tau}$ ou $\tau = \text{argmin}_{\rho_\tau}$ selon leur existence, un test de Kolmogorov-Smirnov est utilisé pour comparer les distributions de ρ_τ et de ρ_0 . S'il donne une valeur inférieur à 0.01, les distributions sont jugées similaires et il n'existe pas de relation entre les variables correspondantes.

On a systématiquement une déviation positive de la corrélation entre population et accessibilité pour les retards positifs, en croissance jusqu'au retard maximal, ce qui pourrait être un marqueur du renforcement des dynamiques de population par la centralité, fait stylisé exhibé pour le système de ville Français par [bretagnolle:tel-00459720]. La correlation entre population et accessibilité est constante, probablement par l'auto-corrélation, et n'entre pas en jeu dans la définition des régimes. Pour des valeurs intermédiaires de d_G et les fortes valeurs de γ_G , on observe également une très légère déviation pour les retards négatifs : pour ces régimes, on a causalité circulaire et le modèle capture une co-évolution dans ce sens. L'accessibilité quant à elle cause fortement la centralité pour $d_G = 10$, puis la tendance s'inverse pour les grands d_G . Pour $d_G = 10$, cela va dans le sens du lien entre population et centralité, et il n'y a dans ce cas pas co-évolution mais adaptation des populations au réseau. Pour les régimes intermédiaire, on a circularité directement entre population et centralité, tandis que pour $d_G > 110$ il y a "circularité indirecte", puisque accessibilité cause centralité qui cause population (qui entre en jeu dans la centralité). Ainsi, le modèle capture au moins trois régimes de co-évolution distincts, en fonction de la distance d'interaction et du niveau de hiérarchie¹¹.

C (FL) : il faut harmoniser les termes ici : relation univoque est plus générique ?

C (AB) : trop abstrait

Synthèse

Les faits stylisés marquants qui ressortent de l'exploration du modèle synthétique sont les suivants :

1. On révèle l'existence d'une échelle spatiale intermédiaire permettant l'évolution de niches relativement indépendantes, correspondant à un niveau de complexité des trajectoires maximal.
2. Les corrélations retardées mettent en évidence au moins trois régimes différents d'interaction, que l'on interprète comme un

¹⁰ $\rho[X+Y; Z] \simeq (\rho[X; Z] + \epsilon \rho[Y; Z]) \cdot \left(1 - \frac{1}{2} \rho[X; Y] \epsilon - \frac{\epsilon^2}{2}\right)$, mais cette hypothèse est trop restrictive pour être valable sur l'ensemble de la somme.

¹¹ Une analyse plus systématique par apprentissage non-supervisé comme en 4.2 et en 7.2 est laissée pour de futures développements.

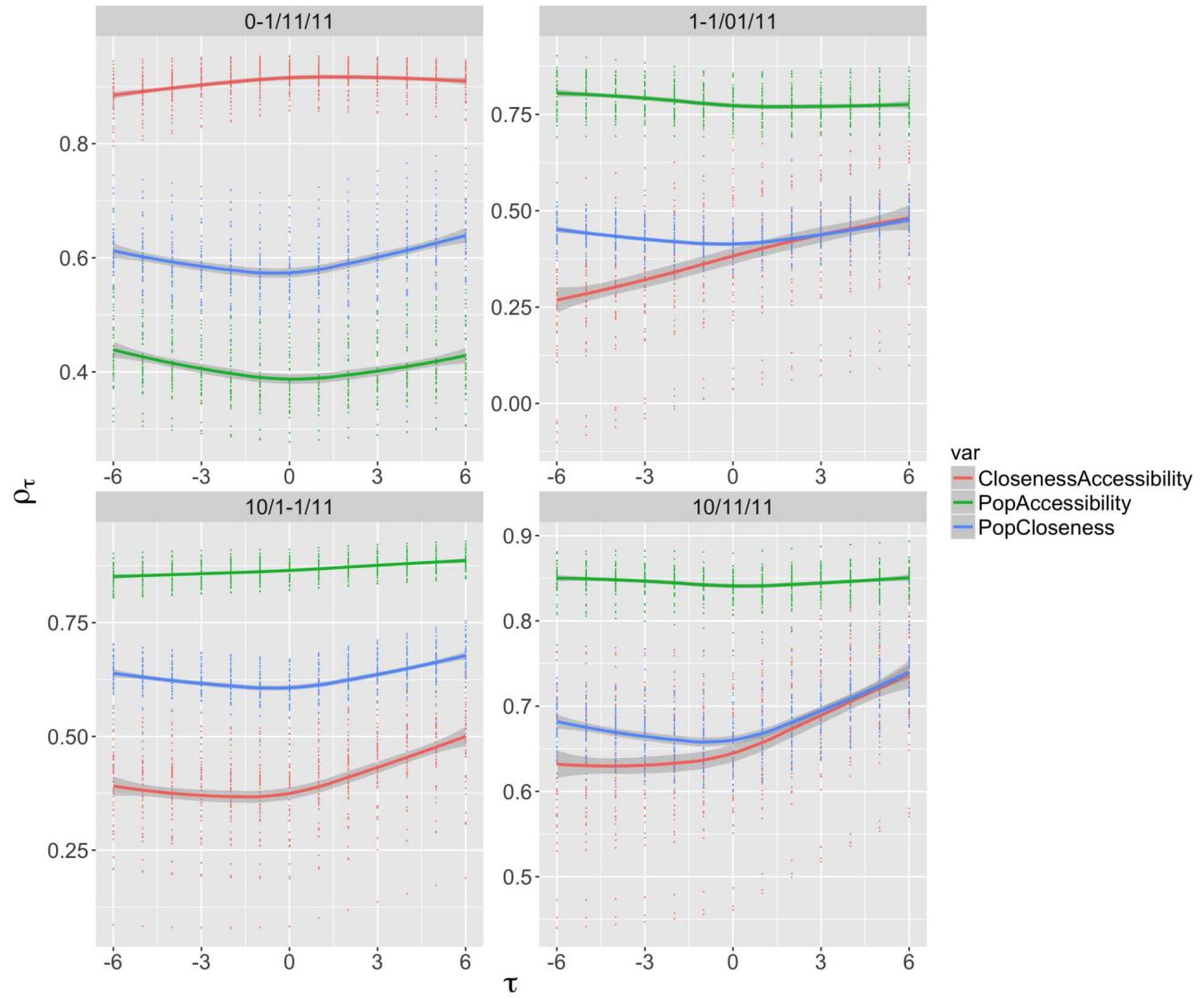


FIGURE 43 : Corrélation retardées.

C (FL) : c'est crucial, peut être plutôt dans une partie théorique "à quoi peut-on s'attendre"?

régime d'adaptation, un régime de co-évolution direct et un régime de co-évolution indirecte.

reproduction du first mover advantage? [levinson2011does]

6.2.3 Applications au Système de Villes Français

Le modèle est ensuite appliqué au système de villes français sur des données dynamiques sur le temps long : la base Pumain-INED pour les populations, couvrant de 1831 à 1999 [[pumain1986fichier](#)], avec le réseau ferré dynamique de 1840 à 2000 [[thevenin2013mapping](#)]. Cette durée temporelle découle de la logique des effets de structure sur le temps long, comme développé en 1.1. Cette application vise d'une part à tester la capacité du modèle à reproduire une dynamique de co-évolution réelle, et d'autre part à extraire une information thématique sur les processus via les valeurs calibrées des paramètres.

Données de Réseau

Nous travaillons sur les données de réseau ferré construites par [[thevenin2013mapping](#)]. Le réseau ferré français est particulièrement intéressant en conjonction avec les données de population déjà présentées, puisque la période couverte est relativement similaire, et que comme le rappelle [[thevenin2013mapping](#)], ce moyen de transport a à toute période concrétisé l'implication d'acteurs publics et privés importants, tout en correspondant à différents processus selon les époques, d'une gestion plutôt décentralisée à une centralisation très forte plus récemment, et différentes concrétisations technologiques avec par exemple l'émergence récente de la grande vitesse [[zembri1997fondements](#)]. Pour chaque date de la base de donnée de population, nous extrayons le graphe abstrait simplifié où toutes les gares et intersections de degré supérieur à deux sont reliés par les liens abstrait avec attributs de vitesse et distance traduisant la valeur réelle, à une granularité de 1km. Cela permet également de construire les matrices de distance-temps entre les villes considérées dans le modèle.

C (AB) : mieux mettre en valeur

C (FL) : développe cette partie pour mieux la mettre en valeur

Faits stylisés

Avant de calibrer le modèle, observons les motifs de corrélation présents dans les données, en appliquant la méthode des corrélations retardées. Cette étude empirique devrait permettre d'une part de vérifier des faits stylisés connus, d'autre part d'établir une connaissance préliminaire du comportement empirique du système. Nous calculons comme précisé ci-dessus la centralité de proximité via le réseau, donnée par $T_i = \sum_j \exp -d_{ij}/d_0$, et étudions la corrélation retardée entre sa dérivée ΔT_i et celle de la population ΔP_i , donnée

par $\hat{\rho}_\tau = \hat{\rho} [\Delta P_i(t), \Delta T_i(t - \tau)]$ estimée sur une fenêtre glissante comprenant T_w dates successives. Nous montrons en Fig. 44 les résultats obtenus.

Ces résultats sont importants pour au moins deux raisons. Dans un premier temps, le comportement du nombre de corrélations significatives en fonction de T_w et de d_0 permet la recherche d'échelles de stationnarité dans le système. On observe d'une part une échelle spatiale spécifique donnant un maximum pour l'ensemble des fenêtres temporelles, à $d_0 = 100\text{km}$, ce qui suggère l'existence de sous-systèmes régionaux cohérents, dont l'existence est stable dans le temps : en effet, cette valeur correspond à la distance d'interaction. Celle-ci coïncide remarquablement avec l'échelle intermédiaire isolée dans le modèle synthétique. D'autre part, les grandes portées spatiales induisent une échelle temporelle optimale, pour $T_w = 4$ ce qui correspond à une vingtaine d'année : nous l'identifions comme l'échelle de stationnarité temporelle du système dans son ensemble et étudions les corrélations retardées pour cette valeur.

Dans un second temps, le comportement des corrélations retardées apporte une mauvaise nouvelle pour la littérature existante et pour les potentialités d'application de notre modèle. A l'échelle spatiale intermédiaire, les valeurs de ρ_+ , ρ_- n'exhibent aucune régularité. Sur l'ensemble du système, on a jusqu'en 1946 quasiment aucun effet significatif, puis aucune causalité entre 1946 et 1975 (maximum à $\tau = 0$, minimum non significatif), puis un décalage de 5 an de l'accessibilité causant la population après 1968 (l'effet restant tout de même douteux). Nous ne reproduisons pas l'effet de corrélation entre centralité dans le réseau et place dans la hiérarchie urbaine défendu par [bretagnolle2003vitesses]¹², ce qui amène à relativiser l'existence de la "co-évolution structurelle" sur le temps long décrite par BRETAGNOLLE dans [espacege02014effets]. Nous rejoignons les résultats récents de [mimeur:hal-01616746] qui montrent le non-significativité statistique de la corrélation entre taux de croissance et évolution de la couverture du réseau ainsi que l'évolution de l'accessibilité, à délai nul. Nos résultats sont moins précis pour les classes de villes étudiées (ils différencient grandes villes et petites, et travaillent sur un panel plus grand), mais plus généraux car pour un délai et une portée de l'accessibilité variables, et donc complémentaires.

C (AB) : trop abstrait

C (AB) : expliciter systématiquement le lien

C (FL) : développer et hiérarchiser ce que tu retires, c'est clairement un endroit important de ta thèse : - se raccrocher aux hypothèses sur la coévolution à cette échelle; se raccrocher aux hypothèses sur la façon de modéliser ce phénomène

¹² Tout comme [lemoy2017scaling] n'arrivent pas à reproduire, pour les profils de densité en fonction de la distance au centre des métropoles européennes, la transition permettant à [guerois2008built] de définir le péri-urbain. Ces travaux plus ou moins anciens ne sont pas reproductibles, ne fournissant ni code, ni données et ne donnant qu'une description très succincte des méthodes, et il est ainsi impossible de connaître l'origine de la divergence qualitative obtenue. Une bonne reproductibilité ainsi que la construction de comparaisons systématiques (*benchmarks*) de modèles, analyses empiriques, récentes mais aussi en validation d'études passées, nous semble une bonne solution à ce genre de problèmes.

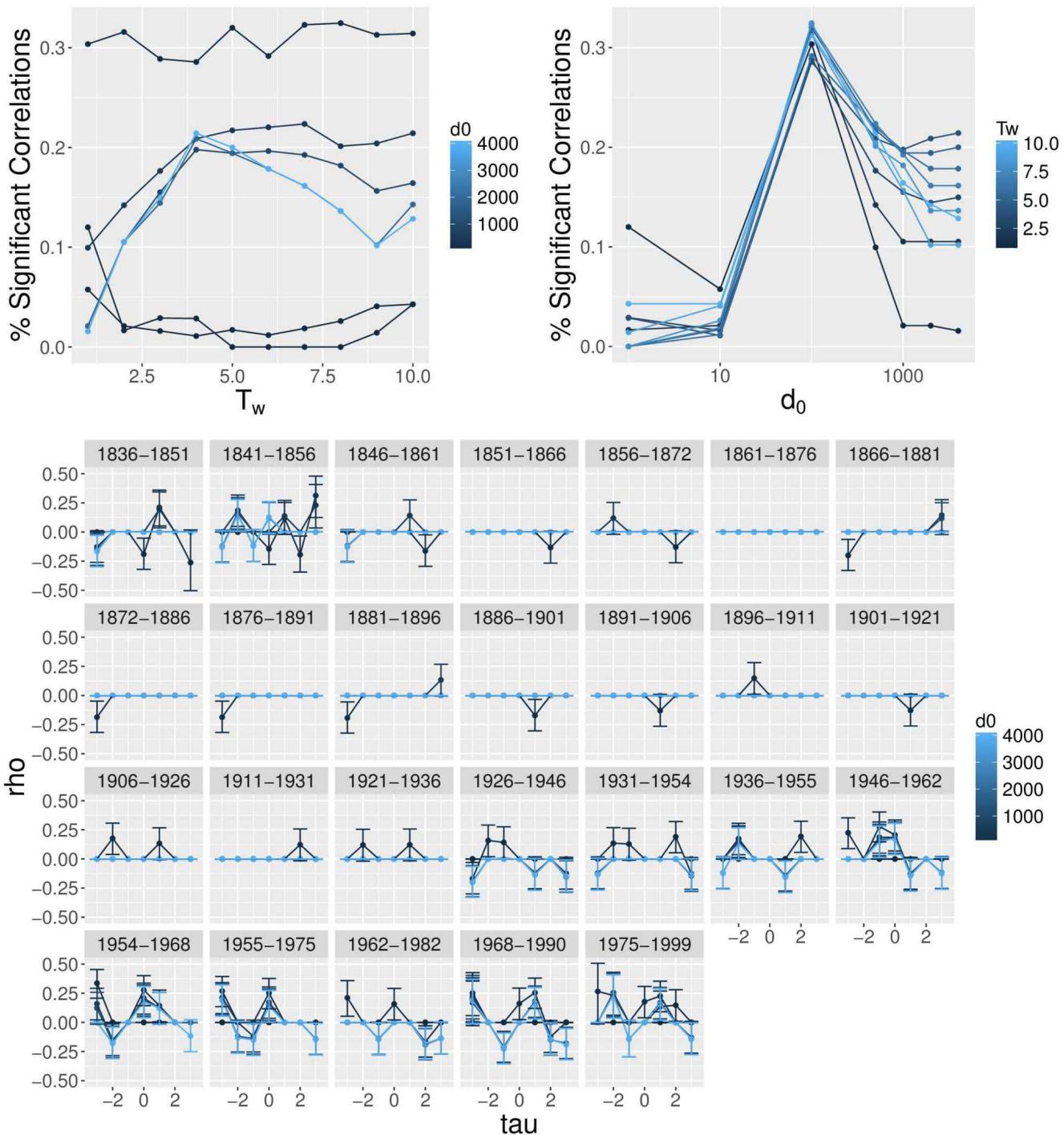


FIGURE 44 : Corrélations retardées empiriques pour le système de villes français. Les corrélations sont calculées sur une fenêtre de taille $5 \cdot T_w$, entre les taux de croissance des populations et ceux de centralité de proximité avec un paramètre de décroissance d_0 (voir texte). (Haut Gauche) Nombre de corrélation significatives (prises telles que $p < 0.1$ à 95%) en fonction de T_w pour d_0 variable; (Haut droite) Nombre de corrélations significatives en fonction de d_0 pour T_w variable; (Bas) Pour la fenêtre “optimale” $T_w = 4$, valeur de ρ_τ en fonction de τ , pour l’ensemble des périodes successives.

Calibration du modèle abstrait

Les résultats attendus de la calibration sur données réelles concernent à la fois la reproduction plus ou moins précise des dynamiques réelles de croissance de population, c'est à dire dans quelle mesure la prise en compte d'un réseau dynamique peut augmenter le pouvoir explicatif pour les trajectoires, et aussi quel est le niveau de réalisme de l'évolution de la distance par le réseau. Nous travaillons toujours avec le modèle abstrait.

EVALUATION DU MODÈLE On peut ajouter aux indicateurs utilisés précédemment un indicateur de calibration pour la distance. L'aspect particulier de l'ajustement pour les populations, qui résidait dans la présence d'une loi de puissance pour les tailles de villes rendant négligeables les performances sur les villes moyennes et les petites villes dans le cas d'une erreur cumulée, et suggérait l'ajout de l'indicateur de l'erreur sur les logarithmes , n'est pas présent pour les distances qui suivent une distribution concentrée sur un ordre de grandeur unique. Nous utilisons ainsi le logarithme de l'erreur carré sur les distances, donnée par

C (FL) : detail technique
a ce stade ?

C (FL) : cela ne donne
pas plus d'éclairage sur
tes choix

$$\varepsilon_D = \log \left[\sum_t \sum_{i,j} (d_{ij}(t) - \tilde{d}_{ij}(t))^2 \right]$$

où $d_{ij}(t)$ sont les distances observées et $\tilde{d}_{ij}(t)$ les distances simulées. Il s'agit simplement d'une erreur carré cumulée, comme utilisée pour la comparaison de matrices origine-destination dans un cas similaire de simulation d'un réseau de transport dans [jacobs2016transport].

RÉSULTATS Nous procédons à une calibration non-stationnaire, sur les objectifs ($\varepsilon_P, \varepsilon_D$), c'est à dire l'erreur carrée sur les population et celle sur les distances. L'estimation est faite par fenêtre mobile sur les périodes déjà utilisées en 4.3. Pour limiter la dimension à explorer, nous fixons $w_N = 0$ pour n'étudier les interactions qu'au premier ordre, sachant que les paramètres de réseau abstrait ($g_{max}, \gamma_S, \varphi_0$) sont pris en compte dans la calibration. La calibration est effectuée par algorithme génétique de façon similaire La Fig. 45 montre les fronts de Pareto obtenus, et la Fig. 46 l'évolution dans le temps des valeurs des paramètres pour les solution optimales.

On note la faible consistence des fronts de Pareto de manière générale, en comparaison avec ceux obtenus pour la population seule précédemment, ce qui suggère la difficulté d'optimiser conjointement trajectoire des populations et trajectoire des distances. Certaines périodes, comme 1851-1872 puis toutes celles après 1946, présentent un point optimal simultané pour les deux objectifs, ce qui pourrait aussi témoigner d'une mauvaise convergence de l'algorithme. Dans tous

les cas, la très faible significativité des corrélations empiriques observées précédemment pouvait laisser présager de ces mauvais ajustements.

Les valeurs des paramètres optimaux semblent toutefois contenir un certain signal. L'évolution de w_G et γ_G sont consistantes avec celles observées pour le modèle statique. Pour d_G , on observe des oscillations, une certaine stabilité, puis un pic pour la période 1962-1982. Il pourrait s'agir d'un "effet TGV", en cohérence avec le pic secondaire pour ϕ_0 observé au même point, puisque la construction des LGV a raccourci les distances entre les villes au plus haut de la hiérarchie (une augmentation du seuil ϕ_0 correspond à une augmentation de la sélectivité pour une diminution potentielle des distances). Le g_{\max} calibré est très représentatif de l'histoire du réseau ferré : un très fort accroissement dans les premières années, puis un accroissement stable plus tard (l'impact du TGV étant là noyé dans l'ensemble du réseau, le seul signe étant une augmentation de la déviation sur l'ensemble du front). Pour la calibration avec distance seule, la stabilisation à $g_{\max} = 0$ est un témoin de la rudimentarité du modèle. On a pu ainsi dans une certaine mesure indirectement quantifier les processus d'interaction par le réseau et ceux d'adaptation du réseau au flux, dans le cas d'un système réel.

C (FL) : discussion intéressante

Modèle avec réseau physique

Nous esquissons à présent les contours d'une spécification du modèle avec réseau physique, qui correspondrait en un sens à un modèle hybride combinant plusieurs échelles comme nous l'avons déjà argumenté. L'idée d'une telle spécification serait d'une part d'étudier l'écart de trajectoire par rapport au réseau abstrait, c'est à dire quantifier l'importance des économies d'échelles (liées aux tronçons communs) et de la congestion, ainsi que les possibles compromis à effectuer liés à la spatialisation du réseau, et d'autre part d'étudier dans quelle mesure il est possible de reproduire des réseaux réalistes en comparaison à des modèles autonomes de croissance de réseau par exemple. Ces questions sont traitées à une autre échelle et pour d'autre spécifications ontologiques au chapitre 7.

similaire à [li2014modeling]

Le réseau physique que nous implémentons cherche à satisfaire un critère de gain de temps local. Plus précisément, on suppose un auto-renforcement à la manière de [tero2010rules]. Une spécification analogue à celle utilisée précédemment suppose une croissance pour chaque lien, donnée par :

$$d(t+1) = d(t) \cdot \left(1 + g_{\max} \cdot \left[\frac{\phi}{\max \phi} \right]^{\gamma_s} \right)$$

C (AB) : expliciter l'autorenforcement ici

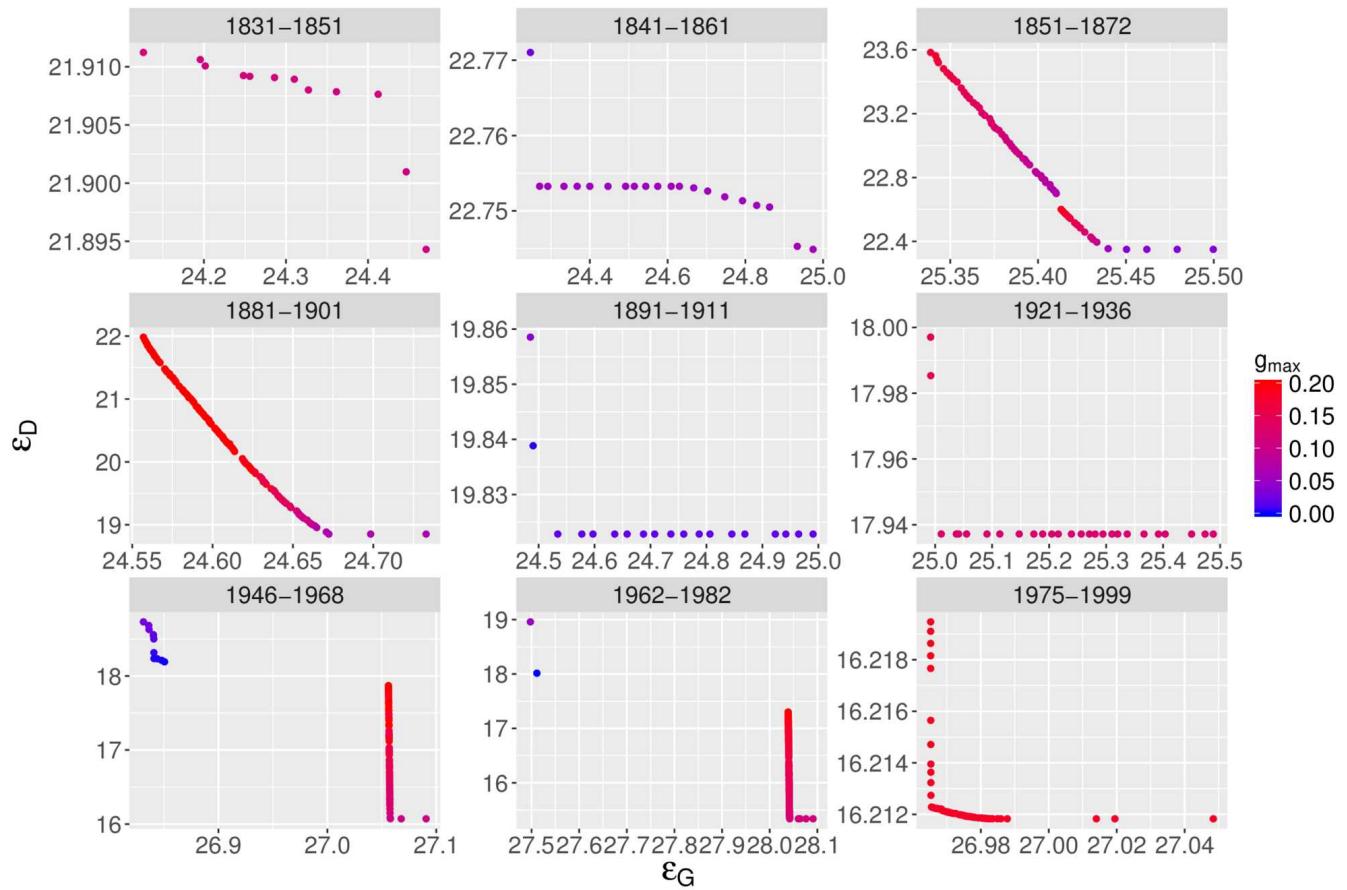


FIGURE 45 : Fronts de Pareto pour la calibration bi-objectif population et distance. Les fronts sont donnés pour chaque période de calibration, et colorés en fonction de g_{\max} .

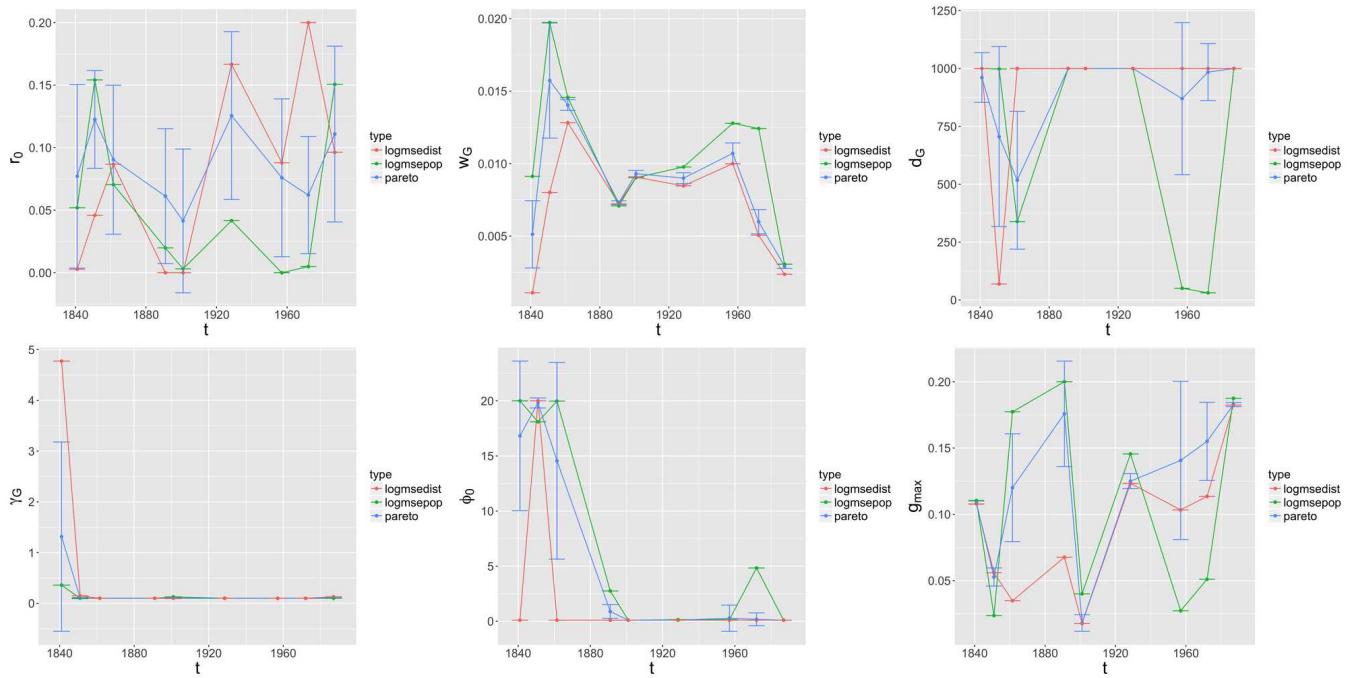


FIGURE 46 : Evolution temporelle des paramètres optimaux. Dans l'ordre de gauche à droite et de haut en bas, valeurs des paramètres ($r_0, w_G, d_G, \gamma_G, \phi_0, g_{\max}$), respectivement pour l'ensemble du front de Pareto (bleu), pour le point optimal au sens de la distance (rouge) et pour le point optimal au sens de la population (vert).

la spécification par seuil ne permettant pas une bonne convergence dans le temps.

C (FL) : répartition des pop pas différentes ?

Nous générerons un réseau initial aléatoire, en perturbant la position des sommets d'une grille dont une proportion fixée de liens a été supprimée (40%) et en y reliant les villes au plus court. Les liens ont tous même impédance, puis celle-ci évolue selon l'équation ci-dessus. Un exemple de configuration obtenue par cette spécification est donné en Fig. 47. Les bonnes propriétés de convergence (stabilisation visuelle de la structure du réseau lors d'expériences restreintes) suggèrent les potentialités offertes par cette spécification, dont l'exploration systématique est hors de notre portée ici.

Perspectives

TRAJECTOIRES PARTICULIÈRES L'étude de trajectoires particulières au sein du système de villes peut permettre de répondre à des questions thématiques spécifiques : par exemple, l'influence des villes moyennes sur la trajectoire globale du système, ou les déterminants d'une plus ou moins bonne "réussite" pour ce type de profil. Dans le cas de l'application à un système réel, la cartographie des déviations

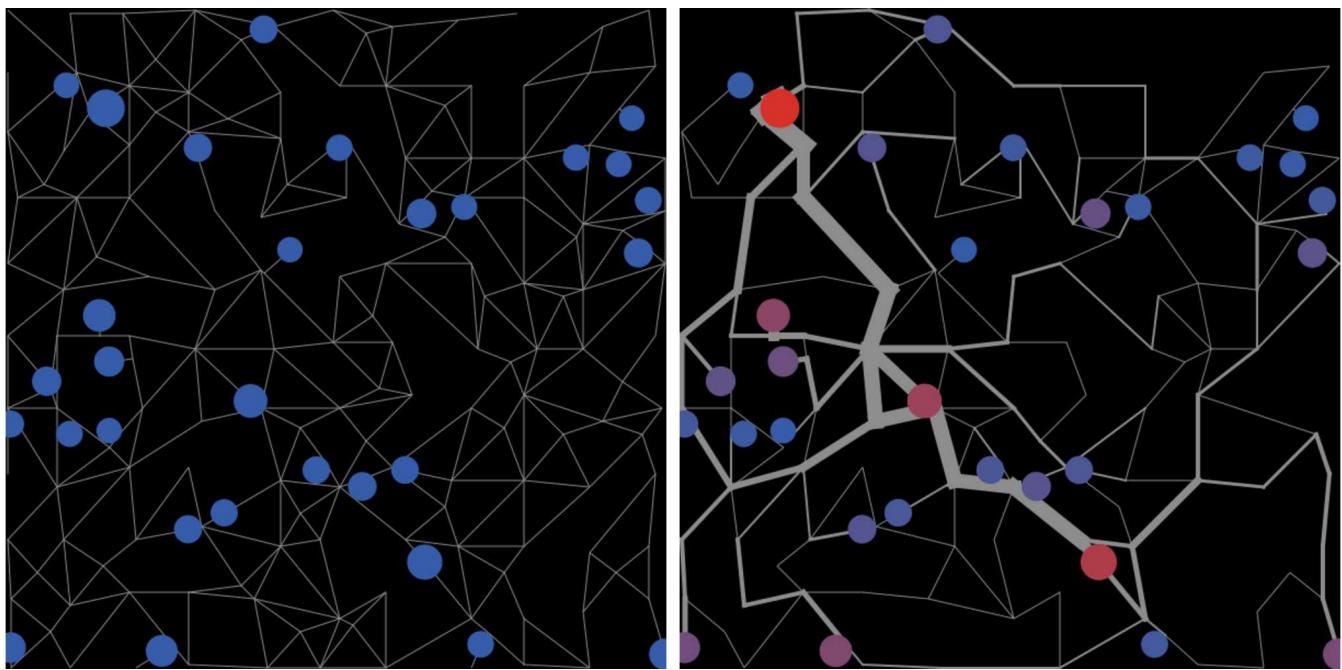


FIGURE 47 : **Example de configuration obtenue avec réseau auto-renforçant.** (*Gauche*) Configuration initiale aléatoire, capacités égales; (*Droite*) Configuration finale obtenue après 100 itérations.

au modèle dans le temps peut suggérer des particularités régionales.

C (FL) : justement quels font stylisés à reproduire?

COMPARAISON DE SYSTÈMES URBAINS On s'attend finalement également à pouvoir par l'intermédiaire de ce modèle comparer des systèmes urbains dans des contextes géographiques et politiques différents, ainsi qu'à différentes échelles. Cela devrait permettre de révéler les implications des actions de planification sur les interactions entre réseaux et territoires. Par exemple, le réseau ferre français a émergé par l'intermédiaire de multiples opérateurs, au contraire du réseau ferré à grande vitesse Chinois, pour lequel un développement précis pourrait être envisagé.

* *

*

CONCLUSION DU CHAPITRE

Cette entrée macroscopique dans les processus de co-évolution visait à les comprendre (i) au sein d'un système de villes, c'est à dire de manière agrégée et à un niveau abstrait; et (ii) sur une échelle temporelle longue, de l'ordre de la centaine d'années. Les processus considérés sont : croissance des villes entraînée par les interactions qui dépendent du réseau, effet de flux au second ordre sur ces croissances (que nous n'avons pas exploré ici), effet de rétroaction des flux sur les distances dans le réseau de manière seuillée (ce dernier étant raffiné avec un effet de la topologie du réseau dans le cas de SimpopNet). Nous démontrons dans un premier temps, par exploration systématique du modèle SimpopNet, que celui-ci est très sensible à la configuration spatiale, suggérant que les conclusions potentielles sur des processus devront toujours être contextualisées. Nous montrons également que celui-ci ne produit pas à proprement parler de co-évolution au sens de circularités causales entre réseau et villes, mais plutôt d'une adaptation des villes au réseau. Notre modèle exploré par la suite permet quant à lui, au prix d'une abstraction du réseau, de révéler de manière synthétique d'une part une échelle intermédiaire de complexité maximale suggérant l'émergence de sous-systèmes régionaux, permis par des valeurs intermédiaires de la distance d'interaction et des valeurs fortes du seuil de rétroaction pour le réseau; d'autre part l'existence d'au moins trois régimes de causalité, dont deux comprennent des causalités circulaires. L'étude des données réelles pour le système de ville français confirme bien l'existence de l'échelle régionale, ainsi que d'une échelle temporelle de stationnarité courte d'une vingtaine d'années, mais très peu de liens significatifs à celle-ci, en contradiction avec la littérature existante. La calibration du modèle sur données réelles reproduit bien les motifs connus de croissance du réseau ferré, et révèlent un "effet TGV" plus récemment, en restant modeste sur la portée des conclusions vu la faible qualité de la calibration. Nous suggérons un développement avec réseau physique, qui permet de faire le lien avec les ontologies que nous allons explorer par la suite en chapitre 7 : la co-évolution à l'échelle mesoscopique, en appuyant sur le rôle de la forme et de la fonction, et donc des mécanismes précis de développement du réseau.



* * *

*

7

CO-ÉVOLUTION À L'ECHELLE MESOSCOPIQUE

Les processus sous-jacents à la co-evolution ne sont pas exactement similaires lorsqu'on passe de l'échelle macroscopique à l'échelle mesoscopique, comme le suggèrent nos différentes analyses empiriques : par exemple, les régimes de causalités obtenus à petite échelle pour l'Afrique du Sud sont plus clairs que ceux pour les transactions immobilières et le Grand Paris. A l'échelle métropolitaine, les processus de relocalisation sont essentiels pour expliquer l'évolution de la forme urbaine, et ceux-ci peuvent partiellement être attribués aux différentiels d'accessibilité, sachant que l'évolution des réseaux répond quant à elle à des logiques complexes conditionnées par les distributions territoriales. Centralité, densité, accessibilité, autant de propriétés étant potentiellement impliquées dans les processus co-évolutifs, et propres au concept de forme urbaine. Nous faisons le choix d'appuyer le rôle de la forme urbaine à l'échelle mesoscopique, et utilisons la morphogenèse urbaine comme paradigme de modélisation de la co-évolution : le couplage fort de la forme urbaine avec le réseau par la co-évolution permet d'amener les fonctions urbaines plus explicitement.

Ce chapitre fait logiquement suite au Chapitre 5, et étend les modèles qui y ont été introduits. Différentes heuristiques de génération de réseau sont comparées dans une première section 7.1, toujours dans un paradigme de couplage simple, afin d'établir les topologies produites par différentes règles. Cette étape permet d'introduire un modèle de co-évolution par morphogenèse en 7.2, qui est calibré sur les objectifs couplé de morphologie urbaine et de topologie de réseau. Enfin, nous introduisons en 7.3 un modèle permettant l'exploration de processus complexes pour la croissance du réseau, notamment des processus endogènes de gouvernance impliquant des agents décideurs à l'échelle métropolitaine.

* * *

*

Les résultats des deux premières sections de ce chapitre ont été présentés à CCS 2017 comme [raimbault:halshs-01590624], et paraîtront prochainement de façon synthétique comme chapitre d'ouvrage [] ; la structure du

modèle et des résultats préliminaires pour la troisième section ont été présentés à ECTQG 2015 comme [le2015modeling].

7.1 MODÈLES DE CROISSANCE DE RÉSEAU

Nous proposons dans un premier temps d'étudier en détails les processus de croissance de réseau pour l'échelle mesoscopique. L'idée est de comprendre les propriétés intrinsèques des différentes heuristiques de croissance de réseau. Cet exercice est d'une part intéressant en lui-même puisqu'il n'existe pas à notre connaissance de comparaison systématique de modèles de morphogenèse des réseaux spatiaux : si [xie2009modeling] propose par exemple une revue du point de vue de l'économie des réseaux, celle-ci ne prend pas en compte certaines disciplines d'une part (voir Chapitre 2), et ne compare pas les performances des modèles par des implémentations dédiées comparables.

7.1.1 Comparer les heuristiques de croissance de réseau

Pour la croissance du réseau en tant que telle, de nombreuses heuristiques existent pour générer un réseaux sous certaines contraintes. Comme déjà développé précédemment notamment en 2.1, des modèles économiques de croissance de réseau au heuristiques d'optimisation locale, aux mécanismes géographiques ou à la croissance de réseau biologique, chacun a ses avantages et particularités propres. Nous avons déjà testé en 5.3 une heuristique basée sur la rupture de potentiel d'interaction. Pour pouvoir comparer "toutes choses égales par ailleurs" les différentes heuristiques de génération de réseau, il est nécessaire de les explorer à densité fixée, même si le sens thématique des résultats ne peut avoir de valeur ni sur le temps long, ni pour la coévolution.

L'importance d'heuristiques pouvant capturer une structure topologique permettant un certain compromis entre performance, congestion et coût, est montrée par des analyses empiriques comme [2012arXiv1202.1747W] pour les réseaux de métro, qui montre que les motifs d'évolution des corrélations entre degrés témoignent d'une évolution des réseaux vers une telle topologie.

Nous précisons par la suite le cœur du modèle de croissance de réseau ainsi qu'un certain nombre d'heuristiques aux origines variées, comparées dans des conditions similaires par leur intégration à une base commune.

Base du modèle de croissance de réseau

Un processus commun aux différentes heuristiques constitue le cœur du modèle de croissance de réseau, et fait le pont entre la distribution de densité de population et le réseau. Concrètement, il s'agit d'attribuer des nouveaux centres en fonction de cette densité, et nous

faisons le choix de spécifier ce processus de manière exogène à la croissance de réseau elle-même¹.

Reprendons le contexte utilisé en 5.3, c'est à dire une grille de cellules caractérisées par leur population P_i , sur laquelle un réseau composé de noeuds et de liens se développe. La distribution de la population sera ici fixe dans le temps $P_i(t) = P_i(0)$, et le réseau évolue séquentiellement à partir d'un réseau initial.

Une étape de croissance de réseau est réalisée à intervalles de temps t_N (paramètre permettant d'ajuster les vitesses respectives d'évolution pour la population et pour le réseau). Elle correspond aux étapes suivantes, dont les deux premières raffinent la logique de [raimbault2014hybrid] (qui stipule que des centres de peuplement doivent être connectés au réseau existant de manière basique) :

1. Un nombre fixe n_N de nouveaux noeuds est ajouté. Séquentiellement, la probabilité de recevoir un nouveau noeud est donnée par

$$p_i = \frac{P_i}{P_{\max}} \cdot \frac{\delta_M - \delta_i}{\delta_M}$$

c'est à dire qu'un noeud élémentaire correspond à la conjonction des événements : (i) densité élevée de population de la cellule P_i par rapport à la population maximale par cellule P_{\max} , (ii) densité de noeuds δ_i dans un rayon r_n faible par rapport à une densité maximale de noeuds δ_M . La population des noeuds est re-attribuée à chaque étape par triangulation comme en 5.3.

2. Les nouveaux noeuds sont alors connectés par un nouveau lien, suivant le plus court chemin vers le réseau (raccord perpendiculaire ou avec le sommet le plus proche).
3. Des nouveaux liens sont ajoutés, jusqu'à atteindre un nombre maximal de liens ajoutés l_m , suivant une heuristique pouvant varier parmi : aucune (pas d'ajouts de liens), aléatoire, rupture de potentiel déterministe (voir 5.3, rupture de potentiel aléatoire [schmitt2014modelisation], coût-bénéfices [louf2013emergence], génération de réseau biologique (heuristique basée sur [tero2010rules]).

Nous fixons pour simplifier $r_n = 5$, $\delta_M = 10$ et $n_N = 20$, et les paramètres t_N et l_m seront variables.

¹ Cette étape intermédiaire se rapproche dans notre cas d'un esprit de modélisation procédurale, puisque la règle implantée cherche à reproduire une forme sans besoin des processus réels. Cela pose la question de l'équifinalité et de l'existence potentielle de modèles équivalents pour ce sous-modèle ou pour le modèle complet capturant un processus réel correspondant à celle-ci. L'utilisation de multi-modélisation également sur cette étape pourrait être une solution mais les cadres permettant de s'extraire d'un nombre arbitraire de niveaux de stationnarité ou même permettant une autonomie du modèle sur ces choix n'existent pas encore.

Heuristiques de référence

Nous considérons deux heuristiques de référence pour mieux situer celles explorées par la suite : celle composée uniquement de la base décrite précédemment, qui produit des réseaux arborescents ; et la génération de réseau aléatoire, qui consiste à créer un nombre fixe l_m de nouveaux liens entre des sommets choisis aléatoirement, puis à planariser le réseau final².

Heuristique euclidienne

Cette heuristique, dont la rationnelle repose sur des idées de rupture de potentiel gravitaire, correspond à la méthode développée en 5.3. Il s'agit d'une méthode proche de celle introduite par [schmitt2014modelisation], sans l'aspect stochastique et pouvant passer à côté de phénomènes de dépendance au chemin, mais plus raffinée dans les mécanismes de potentiels gravitaires.

Rupture de potentiel aléatoire

La rupture de potentielle aléatoire est celle utilisée par SimpopNet [schmitt2014modelisation], qui reprend le modèle introduit par [blumenfeld2010network]. A chaque étape, deux villes sont tirées aléatoirement, la première selon une probabilité proportionnelle à $P_i^{Y_R}$ et la deuxième selon $V_{i_0j}^{Y_R}$ sachant que i_0 est la première ville tirée et V_{ij} sont les potentiels gravitaires euclidiens. Si $d_N(i_0, j_0)/d(i_0, j_0) > \theta_R$, c'est à dire si le détour relatif par le réseau est supérieur à un paramètre de seuil, un lien est créé entre les deux villes³. Cette création de liens est effectuée l_m fois à chaque pas de temps. Le réseau final est planarisé.

Heuristique biologique

[raimbault2015labex] explore des applications des modèles de croissance de réseau biologique (notamment *slime mould*), notamment leur capacité à produire de manière émergente des solutions optimales au sens de Pareto pour des indicateurs contradictoires, comme le coût et la robustesse. Le modèle considéré est issu de [tero2010rules].

L'intérêt d'une telle heuristique est confirmé dans certains cas par la réalité des optimisations multi-objectif : [padeiro:tel-00438092] (p. 72) illustre le prolongement du métro Parisien à Bobigny dans les années 1970, et la prise en compte des indicateurs de coût, de population desservie, de traffic attendu en heure de pointe, et de temps de trajet moyen.

² L'algorithme de planarisation consiste en la création de noeuds aux intersections éventuelles de nouveaux liens ("aplatissement" du réseau).

³ Pour rester comparable aux autres heuristiques qui n'incluent pas de vitesse des liens, les nouveaux liens sont de vitesse 1 et non v_0 comme dans l'implémentation de 6.1.

Le modèle de *slime mould* fonctionne de la façon suivante. Etant donné un réseau initial dont les liens ont des capacités uniformes, un fluide est distribué dans le réseau d'une source à un puit, établissant un flux dans chaque lien. Un équilibre des pressions du fluide au noeuds du réseau peut être établi, qui correspond à l'état stationnaire pour les flux⁴. Etant donné un équilibre des pressions, les capacités des liens évoluent en fonction du flux traversant. Une itération des équilibres et de l'évolution des tubes permet alors une convergence vers une distribution hiérarchique stable des capacités. Le détail de la procédure est décrit en Annexe A.12, suivant les détails mathématiques développés par [tero2007mathematical].

Notre logique est d'utiliser ce mécanisme pour à un instant donné déterminer un certain nombre de liens réalisés, en fonction d'une nouvelle configuration. Les avantages de l'heuristique que nous allons détailler sont notamment que (i) elle peut être utilisée de manière itérative pour traduire une évolution topologique séquentielle du réseau, en comparaison de la plupart des modèles d'investissement qui font évoluer uniquement les capacités dans le temps ; et (ii) elle traduit des processus d'auto-organisation du réseau, et produit par ailleurs des réseaux optimaux au sens de Pareto pour le cout et la robustesse.

L'application du modèle de slime-mould à la génération de réseau s'effectue de la façon suivante, en s'insérant dans le cadre global décrit précédemment :

1. A partir du réseau existant auquel on ajoute un réseau en grille (diamètres deux fois moindre pour prendre en compte la prépondérance du réseau existant) avec connexion diagonales, et dans lequel on supprime de manière aléatoire 20% des liens pour simuler les perturbations liées à la topologie, on constitue le support initial dans lequel les flux du slime-mould seront simulés
2. On procède par itération de générations successives, qui consistent pour k croissant ($k \in \{1, 2, 4\}$ en pratique) en les étapes suivantes :
 - Etant donné la distribution de la population, on itère $k \cdot n_b$ fois le modèle de slime mould pour obtenir le réseau emergent par convergence des capacités.
 - Les liens de capacité inférieure à un paramètre de seuil θ_b sont supprimés.
 - La plus grande composante connexe est conservée.

⁴ Plus précisément, le problème est équivalent à un système d'équations linéaire électrostatiques qu'il suffit de résoudre.

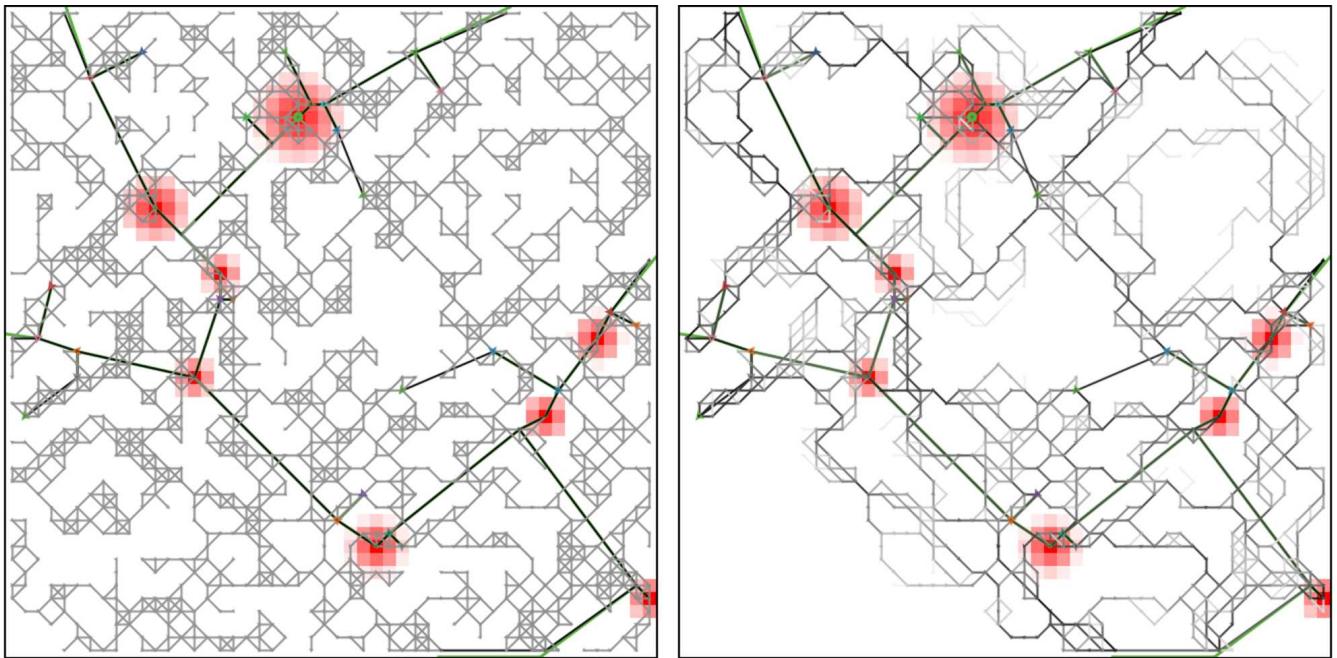


FIGURE 48 : Heuristique biologique pour la génération de réseau. Cet exemple de visualisation illustre les étapes intermédiaires pour l'ajout de lien. (*Gauche*) le réseau semi-aléatoire initial dans lequel le slime mould est lancé; (*Droite*) Même réseau après 80 itérations du slime mould, l'épaisseur des liens donnant la capacité.

3. Le réseau final est simplifié⁵ et planarisé.

Nous illustrons en Fig. 48 deux étapes de ce processus de génération, montrant la structure de base sur lequel le modèle d'autorenforcement est lancé, et la convergence des capacités des liens après un certain nombre d'étapes.

Evaluation coûts-bénéfices

La notion de cout n'est pas présente de manière explicite dans l'ensemble des heuristiques de croissance présentées jusqu'ici - elle l'est de manière implicite dans les potentiels de gravité par le paramètre d'atténuation de la distance, ainsi que dans le slime-mould puisque celui-ci génère des réseaux compromis entre robustesse et cout. Nous ajoutons donc une heuristique simple qui est centrée sur le cout des tronçons de réseau lors de leur extension. Il s'agit de celle étudiée par [louf2013emergence], qui se base sur des arguments d'économie des transports. Suivant une logique d'analyse coûts-bénéfices par les acteurs du développement du réseau, les liens sont réalisés séquentiellement pour les couples de villes non-connectées ayant un coût

⁵ L'algorithme de simplification consiste en un remplacement des séquences de liens dont les sommets hors extrémités ont tous degré 2 par un lien unique.

TABLE 15 : Résumé des paramètres de croissance de réseau pour l'ensemble des heuristiques. Nous donnons également les processus correspondant, les bornes typiques de variation et leur valeur par défaut.

Heuristique	Paramètre	Nom	Processus	Domaine	Défaut
Base	l_m	liens ajoutés	croissance	[0; 100]	10
	d_G	distance gravitaire	potentiel]0; 5000]	500
	d_0	forme gravitaire	potentiel]0; 10]	2
	k_h	poids gravitaire	potentiel	[0; 1]	0.5
	γ_G	hiérarchie gravitaire	potentiel	[0.1; 4]	1.5
Rupture aléatoire	γ_R	hiérarchie aléatoire	hiérarchie	[0.1; 4]	1.5
	θ_R	seuil aléatoire	rupture	[1; 5]	2
Coût-Bénéfices	λ	seuil aléatoire	compromis	[0; 0.1]	0.05
Biologique	n_b	itérations	convergence	[40; 100]	50
	θ_b	seuil biologique	seuil	[0.1; 1.0]	0.5

minimal, avec un coût de la forme $d_{ij} - \lambda/V_{ij}$, où le paramètre λ est le compromis entre coût de construction et gain de potentiel connecté.

Paramètres

Nous résumons les paramètres que nous ferons varier par la suite en Table 15. Un “paramètre” supplémentaire, ou plutôt un métaparamètre, est le choix de l’heuristique pour l’ajout des liens.

7.1.2 Résultats

Initialisation du modèle

Le modèle est initialisé sur configuration synthétiques ou semi-synthétiques, avec une grille de taille $N = 50$:

1. La densité de population est initialisée soit par mélange d’exponentielles, dont les centres (noeuds du réseau) suivent la configuration d’un système de ville synthétique comme fait en 6.1 ; soit à partir d’une configuration réelle extraite du raster de densité pour la France. Nous utiliserons la deuxième option dans les explorations systématiques ici.
2. Dans le second cas, un nombre fixe de noeud du réseau sont générés et localisés de manière préférentielle selon la densité (voir 5.3)⁶. Nous n’initialisons pas sur réseau réel puisqu’il s’agira de la cible de calibration, mais imposons un squelette initial synthétique pouvant être interprété comme un réseau archaïque.

⁶ Pour éviter les effets de bord d’un réseau n’ayant aucune connexion avec l’extérieur, nous ajoutons un nombre fixe n_e de noeuds (que nous prenons $n_e = 6$) à des points aléatoires sur le bord du monde.

3. Un réseau initial est créé par connection des noeuds comme détaillé en 5.3.

Réseaux générés

Une illustration visuelle des différentes topologies générées est donnée en Fig 49. Cela permet de comparer les particularités de chacune des heuristiques. Par exemple, les liens formés par la rupture aléatoire en comparaison à la rupture déterministe témoignent de la dépendance au chemin et produisent un réseau moins redondant, tandis que la rupture déterministe renforce le lien le plus fort entre les deux grandes villes proches. L'heuristique basée sur le coût donne des réseaux denses de manière très localisées, mais évite les liens trop longs. Enfin, l'heuristique biologique produit un maillage dense dans la sous-région où les interactions sont les plus fortes.

Plan d'expérience

Détaillons un plan d'expérience pour explorer l'espace des réseaux générés par les différentes heuristiques. La génération de réseau est faite à densité de population constante, sur configurations réelles classifiées morphologiquement en 4.1. Nous considérons 50 grilles réelles de densité, correspondant à des zones en France, classées dans 5 classes morphologiques. La description de celles-ci est donnée en Annexe A.12, et montre qu'elles couvrent un ensemble de morphologies allant d'établissements très localisés et dispersés à des structures polycentriques, et des configurations intermédiaires.

Etant donné les plages de paramètres données précédemment pour chacune des heuristiques, nous comparons l'espace faisable pour une exploration basique en criblage LHS de l'espace des paramètres, pour l'ensemble des grilles de densité, avec 5 répétitions par point de paramètre⁷.

Topologies obtenues

Les réseaux sont caractérisés ici par les indicateurs suivants : centralité de chemin moyenne \bar{bw} et centralité de proximité moyenne \bar{cl} , diamètre r , longueur moyenne de chemin \bar{l} , vitesse relative v_0 . Pour visualiser les espaces faisables et les comparer aux réseaux réels par la suite, nous réduisons l'espace dans un hyperplan principal, à partir des points obtenus dans les simulations. La composition des deux premières composantes est la suivante : $PC1 = -0.51\bar{bw} - 0.45\bar{l} + 0.57v_0 - 0.43r + 0.05\bar{cl}$ et $PC2 = -0.45\bar{bw} + 0.17\bar{l} + 0.33v_0 + 0.8r + 0.1\bar{cl}$. La première va caractériser des réseaux où les chemins sont courts, tandis que la deuxième exprime des réseaux à distance

⁷ Correspondant à environ 240,000 répétitions du modèle. Le jeu de données issu des simulations est disponible à <http://dx.doi.org/10.7910/DVN/0BQ4CS>.

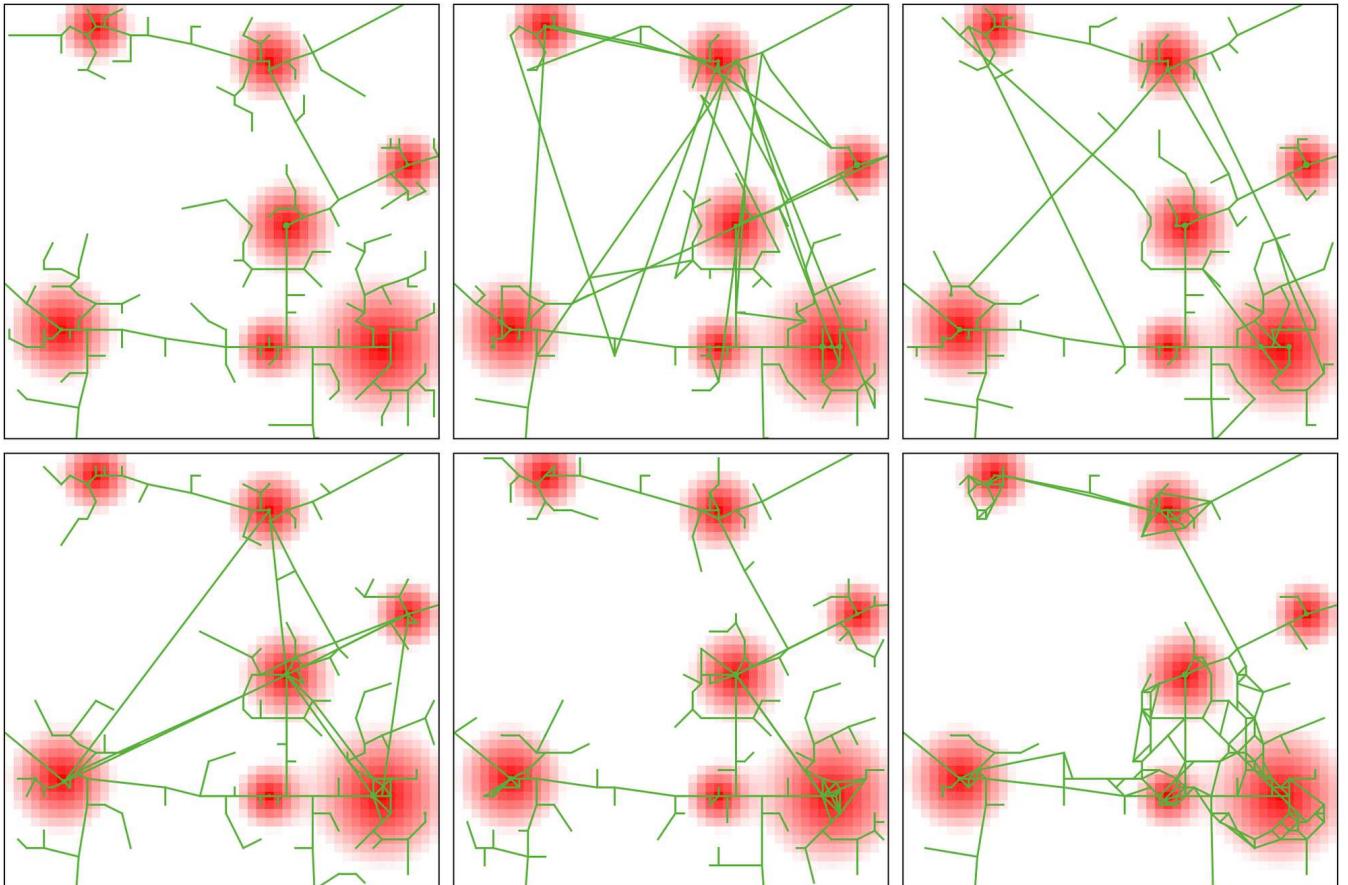


FIGURE 49 : Exemples de réseaux obtenus par les différentes heuristiques. Les réseaux sont obtenus pour une même configuration de densité composée de 7 centres, et du même réseau initial les reliant. Nous prenons $l_m = 10$ et fixons la taille finale à 200 noeuds. Les paramètres gravitaires sont $d_G = 2000$, $d_0 = 3$, $\gamma_G = 0.3$, $k_h = 0.6$. Dans l'ordre de gauche à droite et de haut en bas : réseau par connexion seule ; réseau aléatoire ; rupture de potentiel aléatoire avec $\gamma_R = 2$ et $\theta_R = 1.6$; rupture de potentiel déterministe ; coût bénéfice avec $\lambda = 0.009$; biologique avec $n_b = 50$ et $\theta_b = 0.6$.

moyenne plus grande, donc plus étalés, mais plus efficents en termes de v_0 .

Le nuage de points de l'espace topologique faisable, obtenu avec le plan d'expérience décrit ci-dessus, est donné en Fig. 50. La couverture est permise par la complémentarité des différents nuages pour chaque heuristique. Par exemple, l'heuristique aléatoire est à l'opposée complète de la référence, en termes de la première composante : le réseau arborescent de référence induit logiquement un plus grand nombre de détours, et donc des chemins plus longs. La rupture aléatoire permet de couvrir une grande plage sur PC1 et occupe une place privilégiée pour les faibles valeurs de PC2.

Pour mieux comprendre la complémentarité des approches, on peut quantifier l'intersection des nuages de points de la Fig. 50 par une méthode simple : en divisant le plan en une grille (qu'on prend de taille 20x20), les proportions p_{ij} de points de chaque heuristique j pour chaque cellule i peuvent être agrégées en un indexe de concentration $h = \sum_i p_i^2$ dont la distribution décrit les équilibres dans les régions de l'espace. On obtient pour les cellules un premier quartile à 0.54, une médiane à 0.76 et un troisième quartile à 1. Pour comparaison, dans le cas de deux types de points seulement, une répartition 65-35% donne un indice de 0.55 et une répartition un indice de 0.75, ce qui veut dire qu'au moins la moitié des cellules ont plus de trois quarts de points dans une unique catégorie. Cela confirme la conclusion de forte complémentarité des heuristiques.

Comparaison aux réseaux réels

Nous utilisons les mesures sur réseaux routiers réels calculées en 4.1 pour calculer une distance des configurations générées aux configurations observées, en considérant les réseaux réels correspondant aux configurations de densité utilisées pour l'initialisation. Nous prenons pour un point de paramètre le minimum de distance euclidienne sur les vecteurs d'indicateurs pour l'ensemble des points réels⁸. Cette comparaison est possible car les indicateurs sont normalisés.

Les résultats de comparaisons aux points réels sont donnés en Fig. 51. Nous donnons une représentation en nuage de points et les histogrammes de distribution des distances, sur l'ensemble des grilles et par classe morphologique. On constate qu'une dizaine de configurations réelles (1/5ème) se retrouvent à grande distance du nuage de points simulés, mais que les autres tombent à distance faible ou à l'intérieur du nuage de point. Encore une fois, les différentes heuristiques sont complémentaires pour approcher un plus grand nombre de points. Concernant les distances, l'aléatoire est le plus mauvais

⁸ C'est à dire si $d(1,2) = \sqrt{(\bar{bw}_1 - \bar{bw}_2)^2 + (\bar{cl}_1 - \bar{cl}_2)^2 + (\bar{l}_1 - \bar{l}_2)^2}$, on considère $d_{\min} = \min_j d(S, R_j)$ si S est le point simulé et R_j l'ensemble des points réels. Nous conservons ici uniquement les indicateurs \bar{bw} , \bar{cl} et \bar{l} , pour des raisons de normalisation.

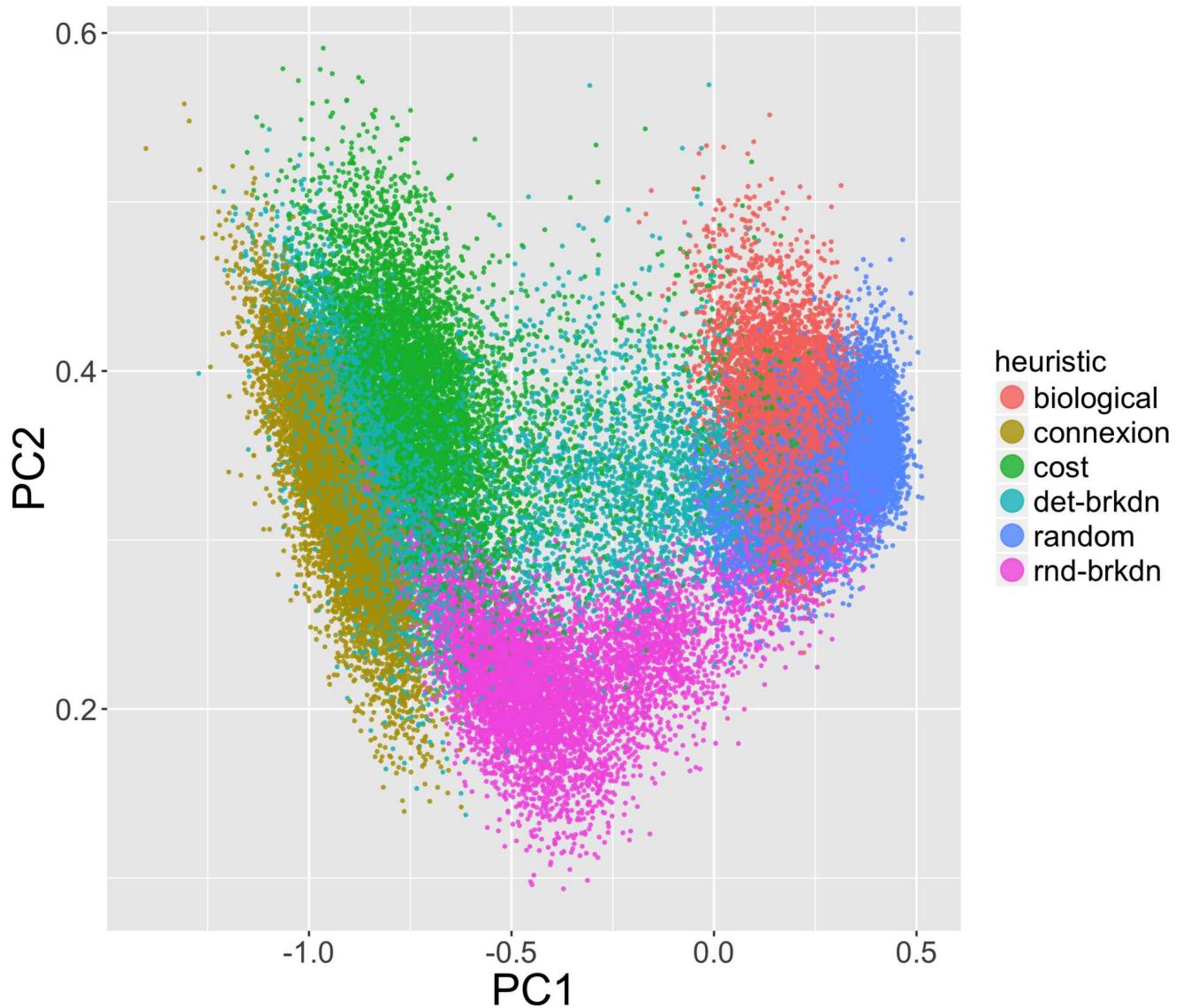


FIGURE 50 : Espace topologique faisable pour les différentes heuristiques de génération. Les nuages de points couvrent des régions complémentaires de l'espace topologique. La même figure conditionnée à la classe morphologique de densité est donnée en Appendice A.12.

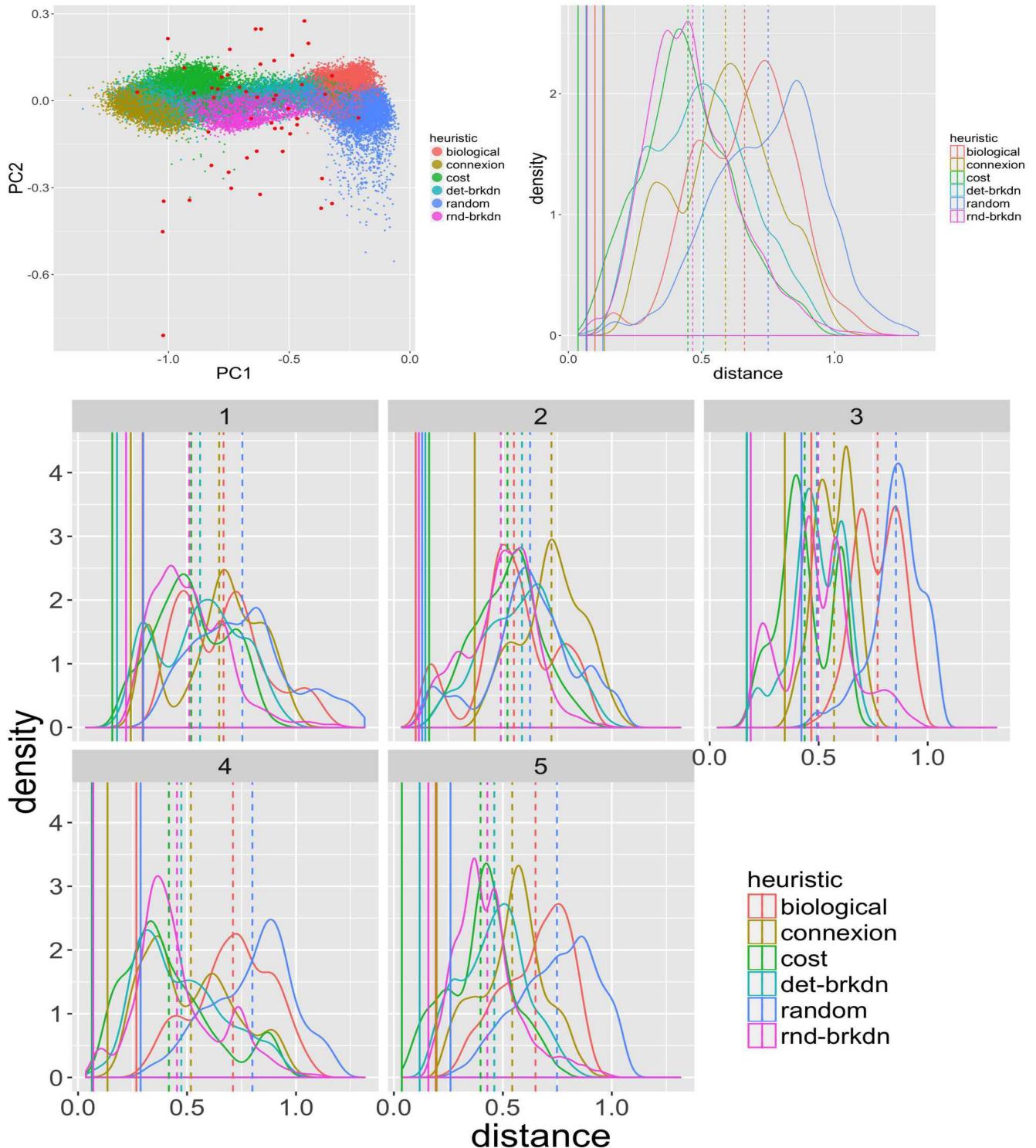


FIGURE 51 : Comparaison aux réseaux réels. (Haut Gauche) Nuage de point des configurations simulées (couleur en légende) et des configurations réelles (en rouge), dans un plan principal tel que $PC1 = 0.12\bar{bw} - 0.09\bar{cl} + 0.98\bar{l}$ et $PC2 = -0.20\bar{bw} - 0.97\bar{cl} - 0.06\bar{l}$. (Haut Droite) Distribution des distances d_{min} pour l'ensemble des points simulés, par heuristique (couleur). Les lignes verticales pointillées donnent la moyenne et les lignes solides le minimum pour chaque distribution. (Bas) Mêmes histogrammes, conditionnés par classe morphologique pour la distribution de densité.

en termes de mode et de moyenne, suivi par le biologique, la référence (connection), la rupture déterministe puis la rupture aléatoire et le coût qui sont à peu près équivalentes. Chacune réalisent des distances minimales très faibles.

En conditionnant par les classes morphologiques, nous voyons que les classes 3, 4 et 5 donnent le plus de difficultés à l'ensemble des heuristiques en termes de minimum - hors il s'agit des configurations avec établissements très localisés ou population diffuse (voir A.12) : il est donc plus facile de reproduire les configurations réelles de réseau dans le cas de structures polycentriques. Dans tous les cas, l'heuristique biologique est peu performante, mais il n'est pas directement possible de savoir si cela est du à sa sous-exploration et aux paramètres fixés ou à sa dynamique intrinsèque.

7.1.3 *Discussion*

Si le modèle slime-mould a été montré comme traduisant de manière simplifiée une génération de réseaux robustes, son utilisation pour la planification a été mise en question, notamment pour sa non prise en compte de facteurs extérieurs et de l'environnement urbain [adamatzky2010road]. Nos résultats semblent confirmer ces analyses, puisque cette heuristique est la moins performante au sens de la distance aux réseaux réels.

Nous avons donc exploré et comparé différentes heuristiques de génération de réseau, à densité fixée. Nous en retirons les enseignements suivants :

- Les différents modèles produisent des réseaux complémentaires dans un espace d'indicateurs.
- De même, ils sont complémentaires pour s'approcher des configurations des réseaux réels, tout en présentant des performances différentes. Des configurations de densité très localisées ou diffuses correspondent à des réseaux plus difficiles à reproduire, en comparaison aux structures polycentriques.

Disposant de ces modèles de croissance de réseau, nous allons pouvoir les utiliser en couplage avec un modèle de densité, afin de développer un modèle de coévolution à l'échelle mesoscopique, qui fera l'objet de la section suivante.

* * *

*

7.2 CO-ÉVOLUTION À L'ÉCHELLE MESOSCOPIQUE

Les établissements urbains et les réseaux de transport ont été montrés comme co-évolutif, dans les différentes approches thématiques, empiriques, et de modélisation des systèmes territoriaux développées jusqu'ici. Comme on l'a vu, les approches modélisant ces interactions dynamiques entre réseaux et territoires sont peu développées. Nous proposons dans cette section de réaliser une première entrée à une échelle intermédiaire, en s'intéressant aux propriétés morphologiques et fonctionnelles des systèmes territoriaux de manière stylisée. Nous introduisons un modèle dynamique et stochastique de morphogenèse urbaine qui couple l'évolution de la densité de population dans les cellules d'une grille avec l'évolution d'un réseau routier.

7.2.1 *Description du Modèle*

Structure générale

Les principes généraux du modèle sont les suivants. Avec un taux de croissance global fixé, une nouvelle population s'agrège préférentiellement à un potentiel local, dont la dépendance à diverses variables explicatives est contrôlé par des paramètres. Celles-ci sont la densité locale, la distance au réseau, les mesures de centralité dans le réseau et l'accessibilité généralisée. [[doi:10.1080/13658816.2014.893347](https://doi.org/10.1080/13658816.2014.893347)] montre dans le cas de Stockholm la très forte corrélation entre les différents types de centralité et le type d'usage du sol, ce qui confirme l'importance de considérer les centralités comme variables explicatives pour le modèle à cette échelle. Nous généralisons ainsi le modèle de morphogenèse étudié dans 5.2, avec des mécanismes d'agrégation similaires à ceux utilisés par [raimbault2014hybrid]. Une diffusion continue de la population complète l'agrégation pour traduire les processus de répulsion généralement dus à la congestion. A cause des différentes échelles de temps impliquées dans l'évolution de l'environnement urbain et des réseaux, le réseau croît à pas de temps fixes, suivant le sous-modèle développé en 7.1 : une première règle fixe assure la connectivité des patches nouvellement peuplés au réseau existant. Les différentes heuristiques de génération de réseau sont ensuite incluses dans le modèle. Nous nous attendons à une complémentarité de celles-ci, puisque par exemple le modèle gravitaire sera plus typique d'une évolution de réseau planifiée, tandis que le modèle biologique traduit des processus auto-organisés de croissance de réseau. La Fig. 52 résume la structure générale du modèle de morphogenèse.

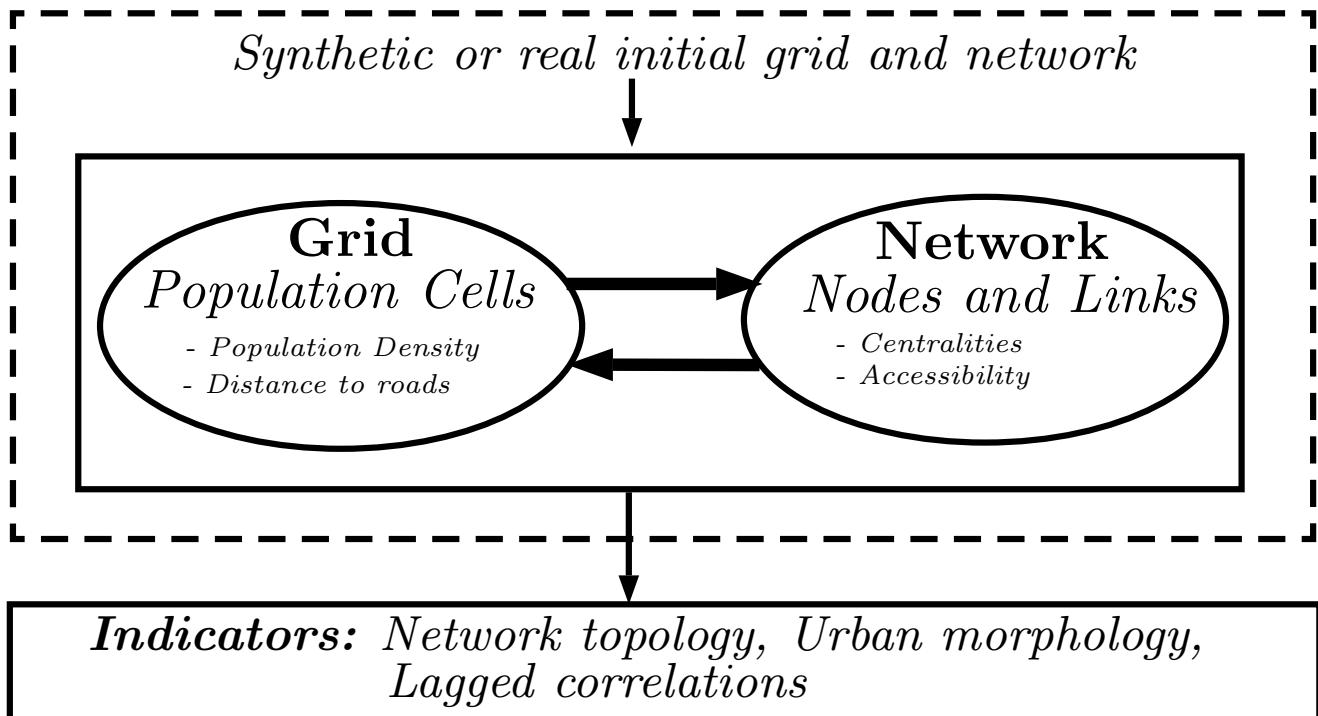


FIGURE 52 : Structure du modèle de co-évolution à l'échelle mesoscopique.

Formalisation

Le modèle est basé sur une grille carrée de population de côté N , dont les cellules sont définies par les populations (P_i). Un réseau routier s'y superpose de la même manière qu'en 7.1. Nous supposons une distribution de population à l'instant initial ainsi qu'un réseau.

L'évolution des densités se base sur une fonction d'utilité, influencée par des caractéristiques locales de la forme et de la fonction urbaine, que l'on appelle *variable explicative*. Soit $x_k(i)$ une variable explicative locale pour la cellule i , qui sera parmi les variables suivantes :

- population P_i
- distance aux routes
- centralité de chemin
- centralité de proximité
- accessibilité.

Pour les trois dernières, celles-ci sont définies comme précédemment pour les noeuds du réseau, puis associées aux cellules en prenant la valeur du noeud le plus proche, pondérée par une fonction décrois-

sante en fonction de la distance à celui-ci⁹. Nous considérons alors les variables explicatives normalisées définies par $\tilde{x}_k(i) = x_k(i) - \min_j x_k(j) / (\max_j x_k(j) - \min_j x_k(j))$.

L'utilité d'une cellule est alors donnée par une agrégation linéaire¹⁰

$$U_i = \sum_k w_k \cdot \tilde{x}_k(i)$$

où les \tilde{x}_k sont les variables explicatives locales normalisées, et w_k des paramètres de poids, qui permettent de pondérer les différentes influences.

Un pas de temps d'évolution du modèle comporte alors les étapes suivantes :

1. Evolution de la population selon des règles similaires au modèle de morphogenèse développé en 5.2. Etant donné un taux de croissance exogène N_G , les individus sont ajoutés de manière indépendante suivant une agrégation faite selon la probabilité $U_i^\alpha / \sum_k U_k^\alpha$, suivie d'une diffusion aux voisins de force β , effectuée n_d fois.
2. Croissance du réseau selon les règles décrites en 7.1, sachant que celle-ci a lieu si le pas de temps est un multiple d'un paramètre t_N , qui permet d'intégrer un différentiel d'échelles temporelles entre la croissance de la population et celle du réseau.

Les paramètres du modèle que nous ferons varier sont donc :

- les paramètres d'agrégation-diffusion α, β, N_g, n_d , résumés en Table 12,
- les paramètres de poids des variables explicatives w_k , au nombre de 4, compris dans $[0; 1]$,
- et les paramètres de croissance de réseau des différentes heuristiques, résumés en Table 15.

Les indicateurs de sortie du modèle sont les indicateurs de morphologie urbaine, les indicateurs topologique du réseau, et les corrélations retardées entre les différentes variables explicatives.

⁹ C'est à dire de la forme $x_k = x_k^{(n)}(\operatorname{argmin}_j d(i, j)) \cdot \exp(-\min_j d(i, j)/d_0)$, avec $x_k^{(n)}$ variable correspondante pour les noeuds, l'indice j étant pris sur l'ensemble des noeuds, et le paramètre de décroissance d_0 étant dans notre cas fixé à $d_0 = 1$ pour garder la caractéristique que les variables de réseau sont essentiellement significatives proche de celui-ci.

¹⁰ Une alternative étant par exemple une fonction de Cobb-Douglas, qui revient à une agrégation linéaire sur les logarithmes des variables.

7.2.2 Résultats

Implémentation

Le modèle est implémenté en NetLogo, vu l'hétérogénéité des aspects à prendre en compte, et ce langage se montrant particulièrement efficace pour coupler une grille de cellules à un réseau. Les indicateurs de morphologie urbaine sont calculés grâce à une extension NetLogo spécifiquement développée (voir E).

Plan d'expérience

Nous proposons de nous concentrer sur la capacité du modèle à capturer les relations entre réseaux et territoires, et en particulier la coévolution. Pour cela, nous chercherons si (i) le modèle est capable de reproduire, en plus des indicateurs de forme, les matrices de corrélation statiques calculées en 4.1; et (ii) le modèle produit une variété de relations dynamiques au sens des régimes de causalité développés en 4.2.

Le modèle est initialisé sur configurations entièrement synthétiques, avec une taille de grille 50. Les configurations sont générées par mélange d'exponentielle : $N_c = 8$ centres sont localisés de manière aléatoire, et une population leur est attribuée selon une loi d'échelle $P_i = P_0 \cdot (i + 1)^{-\alpha_s}$ avec $\alpha_s = 0.8$ et $P_0 = 200$. La population de chaque centre est distribuée à l'ensemble des cellules avec un noyau exponentiel de forme $d(r) = P_{max} \exp(-r/r_0)$ où le paramètre r_0 est déterminé pour fixer la population à P_i , avec $P_{max} = 20$ (densité au centre)¹¹. Le squelette de réseau initial est généré comme détaillé en 7.1.

Nous explorons un échantillonnage LHS de l'espace des paramètres, avec 10 répétitions pour environ 7000 points de paramètres, correspondant à un total autour de 70000 répétitions du modèle¹², effectuées sur grille de calcul par l'intermédiaire d'OpenMole.

Calibration statique et dynamique

Le modèle est calibré au premier ordre, sur les indicateurs de forme urbaine et de mesure de réseau, ainsi qu'au second ordre sur les corrélations entre ceux-ci. Les données réelles utilisées sont toujours celles introduites en 4.1, qui ont le rappelle sont basées sur les données de population raster Eurostat et le réseau routier issu d'OpenStreetMap. Nous utilisons ici l'ensemble des points de l'Europe.

Nous introduisons un processus *ad hoc* de calibration pour pouvoir tenir compte des deux premiers moments, que nous détaillons

¹¹ On a en effet $P_i = \iint d(r) = \int_{\theta=0}^{2\pi} \int_{r=0}^{\infty} d(r) r dr d\theta = 2\pi P_{max} \int_r r \cdot \exp(-r/r_0) = 2\pi P_{max} r_0^2$, et donc $r_0 = \sqrt{\frac{P_i}{2\pi P_{max}}}$.

¹² Pour lesquelles les résultats de simulation sont disponibles également à <http://dx.doi.org/10.7910/DVN/0BQ4CS>.

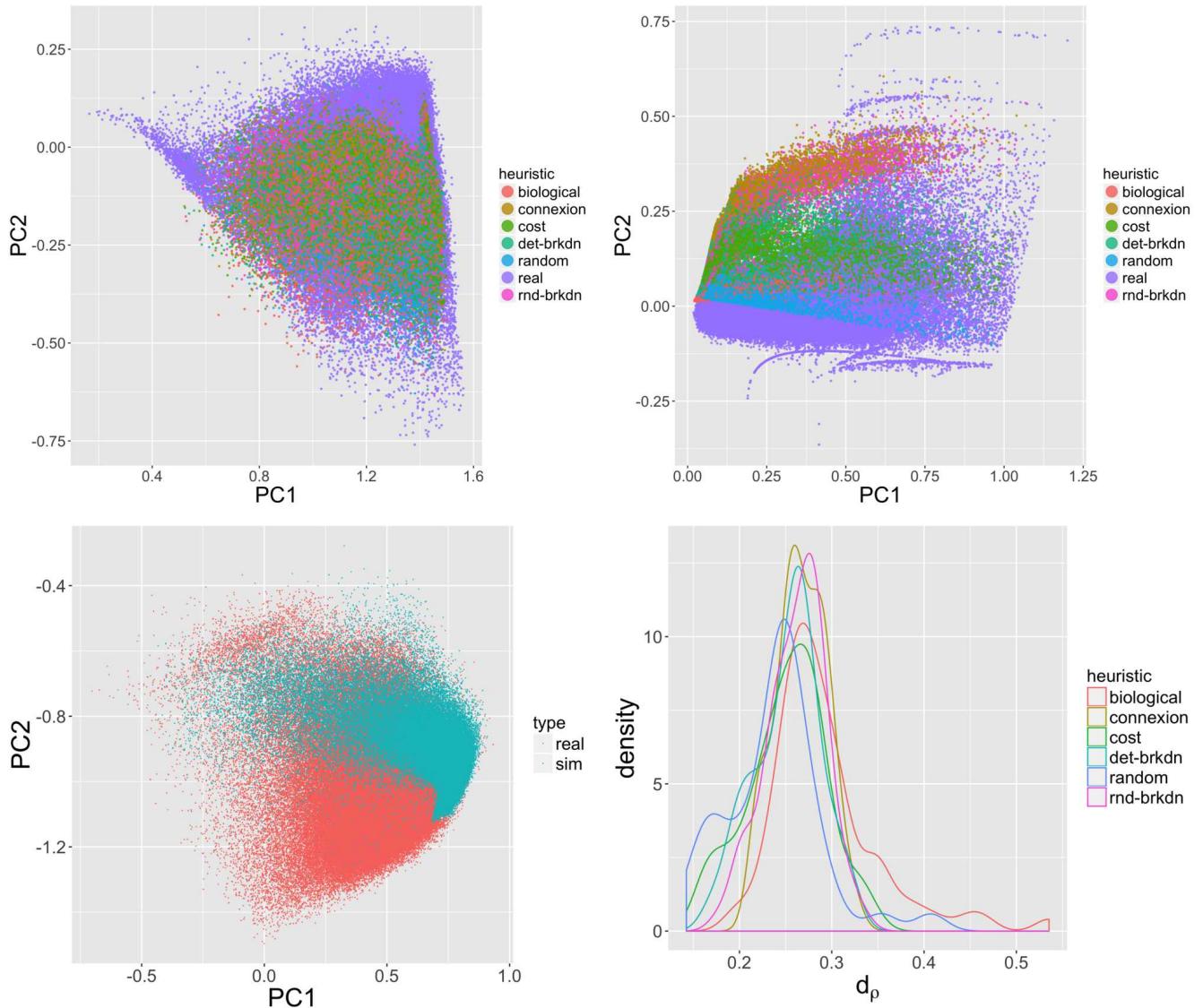


FIGURE 53 : Calibration du modèle de morphogenèse au premier et au second ordre. (Haut Gauche) Nuages de points simulés et observés dans un plan principal pour les indicateurs de forme urbaine. (Haut Droite) Nuages de points simulés et observés dans un plan principal pour les indicateurs de réseau. (Bas Gauche) Nuages de points simulés et observés dans un plan principal pour l'ensemble des indicateurs. (Bas Droite) Distributions des distances sur les corrélations d_ρ , pour les différentes heuristiques.

ci-dessous. Des procédures plus élaborées sont utilisées par exemple en économie, comme [watson1993measures] qui utilise le bruit de la différence entre deux variables pour obtenir la même structure de covariance pour les deux modèles correspondants, ou en finance, comme [frey2001copulas] qui définit une notion d'équivalence entre modèles à variables latentes qui incorpore l'égalité de la structure d'interdépendance entre variables. Nous évitons ici d'ajouter des modèles supplémentaires, et considérons simplement une distance sur les matrices de corrélation. La procédure est la suivante :

- Les points simulés sont ceux issus de l'échantillonnage, avec les valeurs moyennes sur les répétitions.
- Afin de pouvoir estimer des matrices de corrélation entre indicateurs pour les données simulées, nous faisons l'hypothèse que les second moments sont continus en les paramètres du modèle, et découpons pour chaque heuristique l'espace des paramètres en zones pour grouper les points de paramètres¹³, ce qui permet d'estimer pour chaque groupe les indicateurs et la matrice des corrélations.
- Pour chaque estimation ainsi menée, qu'on note \bar{S} (indicateurs) et $\rho[S]$ (corrélations), on peut alors calculer la distance aux points réels sur les indicateurs $d_I(R_j) = d(\bar{S}, R_j)$ et sur les matrices de corrélation $d_\rho(R_j) = d(\rho[S], \rho[R_j])$ où les R_j sont les points réels avec leurs corrélations correspondantes¹⁴, et d une distance euclidienne normalisée par le nombre de composantes.
- Nous considérons alors la distance agrégée définie comme $d_A^2(R_j) = d_I^2(R_j) + d_\rho^2(R_j)$. En effet, comme développé empiriquement et analytiquement en Annexe A.13, la forme des fronts de Pareto pour les deux distances considérées suggère la pertinence de cette agrégation. Le point réel le plus proche du point simulé est alors celui au sens de cette distance.

La Fig. 53 résume les résultats de la calibration. Les indicateurs morphologiques sont plus aisément approchés que ceux de réseau, pour lesquels une partie des nuages simulés ne se superpose pas avec les points observés. Nous retrouvons une certaine complémentarité dans les heuristiques de réseau. En considérant l'ensemble des indicateurs, peu de points simulés tombent loin des points observés, mais une proportion significative de ceux-ci est hors d'atteinte de la simulation. Ainsi, la capture simultanée de la morphologie et de la topologie se fait au prix d'une moins grande précision.

¹³ Chaque paramètre étant découpé en $15/k$ segments égaux avec k nombre de paramètres : nous avons constaté empiriquement que cela permettait d'avoir toujours un nombre minimal de mesures dans chaque zone.

¹⁴ Estimées on le rappelle en 4.1, par fenêtre centrée sur le point, qu'on prend ici pour $\delta = 4$.

Nous obtenons toutefois une bonne reproduction des matrices de corrélation, comme présenté en Fig. 53 (histogramme de d_p , bas droite). La moins bonne heuristique pour les corrélations est la biologique en termes de maximum, tandis que l'aléatoire produit d'assez bons résultats : cela pourrait par exemple être dû à la reproduction des corrélations quasi nulles, accompagnant un effet de structure dû à l'ajout initial des noeuds qui impose déjà une certaine corrélation. Au contraire, l'heuristique biologique introduit des processus supplémentaires qui peuvent éventuellement bénéficier au réseau en termes d'indépendance (ou selon le point de vue opposé être préjudiciable en termes de corrélations). En tout cas, cette application démontre que notre modèle est capable à la fois de s'approcher de configurations réelles pour les indicateurs et pour leurs corrélations.

Régimes de causalité

Nous étudions d'autre part les corrélations retardées dynamiques entre les variations des différentes variables explicatives des cellules (population, distance au réseau, centralité de proximité, centralité de chemin, accessibilité). Nous appliquons la méthode des régimes de causalité introduite en 4.2. La Fig. 54 résume les résultats obtenus par l'application de cette méthode sur les résultats de simulation du modèle de co-évolution. Le nombre de classes induisant une transition est plus faible que pour le modèle RDB, traduisant un plus faible degré de liberté, et nous fixons dans ce cas $k = 4$. Les profils des centroïdes permettent de comprendre la capacité du modèle à capturer plus ou moins une co-évolution.

Les régimes obtenus apparaissent moins divers que ceux obtenus en 4.2 ou pour la co-évolution macroscopique en 6.2. Certaines variables ont naturellement une forte corrélation simultanée, fortuite par leur définitions, comme la centralité de proximité et l'accessibilité, ou la distance à la route et la centralité de proximité. Dans l'ensemble des régimes, la population détermine significativement l'accessibilité. Le régime 1 correspond à une détermination entière du réseau par la population. Le second est partiellement circulaire, de par l'effet des routes sur la population. Le régime 3 est intéressant, la centralité de chemin causant négativement l'accessibilité : cela veut dire que dans cette configuration, l'évolution couplée du réseau et de la population vont dans le sens d'une diminution de la congestion. De plus, comme la population cause la centralité de proximité, il y a également circularité et donc co-évolution dans ce cas. En le localisant dans le diagramme de phase, ce régime est assez dispersé et rare, au contraire par exemple du régime 1 qui occupe une grande partie de l'espace pour une importance faible de la route ($w_{road} \leq 0.3$). Cela confirme que la co-évolution produite par le modèle est ponctuelle et non une caractéristique toujours vérifiée, mais qu'il est toutefois capable d'en générer dans des régimes particuliers.

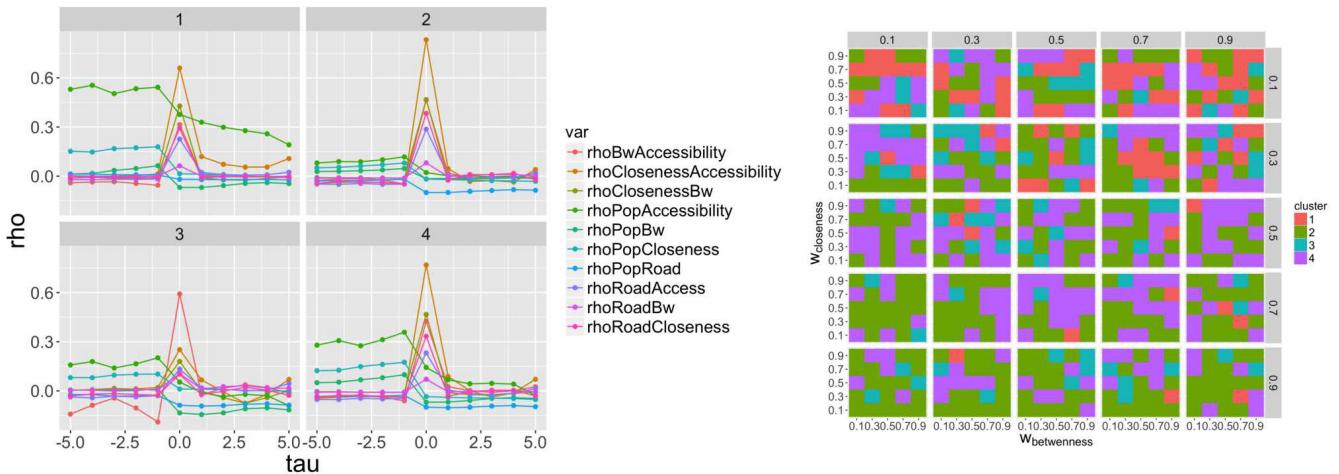


FIGURE 54 : Régimes de causalité pour le modèle de co-évolution. (Gauche) Trajectoire des centres des classes en termes de $\rho[\tau]$ entre les différentes variables explicatives. (Droite) Diagramme de phase des régimes dans l'espace des paramètres w_k , représenté ici comme la variation des diagrammes pour (w_{bw}, w_{cl}) , selon les variations de w_{road} (en ligne) et de w_{pop} (en colonne).

7.2.3 Discussion

Nous avons ainsi proposé un modèle de co-évolution à l'échelle mesoscopique, se basant sur un paradigme de multi-modélisation pour l'évolution du réseau. Le modèle est capable de reproduire un certain nombre de situations observées au premier et au second ordre, capturant une représentation statique des interactions entre réseaux et territoires. Il dégage également différents régimes dynamiques de causalité, en étant toutefois moins riche que le modèle simple étudié plus tôt : ainsi, une structure plus élaborée en terme de processus se paie en flexibilité d'interaction entre ceux-ci. Cela suggère une tension entre "performance statique" et "performance dynamique" des modèles.

Une question ouverte est dans quelle mesure un modèle de réseau pur avec attachement préférentiel des noeuds reproduirait des résultats proches des nôtres. Le couplage complexe entre agrégation et diffusion (montré en 5.2) ne pourrait pas être inclut aisément, et le modèle ne pourrait dans tous les cas répondre à des problématiques concernant le couplage des dynamiques. Cette réflexion rejoint la question des représentations territoriales que nous aborderons en ouverture.

Nous proposons pour le moment une dernière incursion dans la co-évolution à l'échelle mesoscopique, en développant un nouveau modèle qui complexifie considérablement l'influence du territoire sur le réseau, en prenant en compte des processus de gouvernance.

* *

*

7.3 MODÉLISATION DE LA GOUVERNANCE DU SYSTÈME DE TRANSPORT

Cette section se propose de donner des pistes vers une modélisation plus complexe de la coévolution, toujours à l'échelle macroscopique. Nous avons vu en 1.1 que les processus de gouvernance relevaient d'un niveau qui couple intrinsèquement les réseaux et les territoires. Nous avons par ailleurs étudié le cas particulier d'une Méga-région urbaine (MCR) en 1.2, et vu dans quelle mesure ce contexte était propice à une complexité des interactions. Nous introduisons donc ici un modèle de coévolution à l'échelle d'une MCR, qui vise en particulier à endogénérer certains processus de gouvernance du réseau de transport. Ce modèle étend en particulier celui introduit par [le2010approche] puis développé par [lenechet:halshs-00674059].

7.3.1 Contexte

Mega-régions urbaines et gouvernance

Nous rappelons qu'une Méga-région urbaine est un réseau de villes fortement connecté en termes de flux économiques et de population, au sein duquel les relocalisations se font à moindre coût, supposant une certaine cohérence géographique [hall2006polycentric]. Il s'agit du dernier "régime urbain" qui a émergé au sein des systèmes de villes, et il pourrait s'agir d'une trajectoire plus plausible que des villes monocentriques toujours plus grandes pour les agglomérats urbains considérables. [neuman2009futures] souligne que la soutenabilité future de ces MCR sera intimement liée à leur capacité à apprendre de nouveaux schémas de gouvernance, au sens d'une adaptabilité et flexibilité accrue des processus de gouvernance. [innes2010strategies] suggère par ailleurs que des stratégies impliquant auto-organisation par le dialogue entre acteurs sont un moyen de répondre efficacement à la complexité de la gouvernance d'une MCR. Nous proposons par la suite de répondre partiellement à cette question du lien entre structure de gouvernance et évolution de la MCR, par l'intermédiaire du modèle que nous développons.

Modélisation de la coévolution par des processus de gouvernance

Le rôle des processus de gouvernance dans les modèles couplant l'évolution des réseaux de transport à l'évolution de l'usage du sol a déjà été considéré de différents points de vue dans les approches de modélisation. [Xie2011] introduit un modèle économique théorique d'investissement dans les infrastructures. Deux niveaux de gouvernance, local et centralisé, sont considérés dans le modèle. Pour la provision d'une nouvelle infrastructure qui doit être partagée entre deux zones contiguës (l'espace étant à une dimension), un jeu entre

des agents de gouvernance détermine à la fois le niveau de décision et l'attribution de proportion du stock à chaque zone. Les gouvernements veulent soit maximiser l'utilité agrégée (gouvernement Pigovien), ou bien inclure des stratégies politiques explicites pour satisfaire un électeur médian. L'exploration numérique du modèle montre que ces processus sont équivalents à des compromis entre coûts et bénéfices, et que le niveau de gouvernance effectif dépend de l'état du réseau.

[**xie2011governance**] propose une version plus simple de ce modèle du point de vue de la gouvernance mais couplé à un modèle de transport plus réaliste : il couple sur un réseau synthétique croissant un modèle de traffic avec un modèle de prix et un modèle d'investissement, et montre que sous l'hypothèse d'une centralisation, un équilibre entre la demande et la performance du réseau peut être atteint, mais que les investissements ne sont pas efficaces sur le long terme, avec une perte plus importante pour les investissements décentralisés.

[**li2016integrated**] couple un modèle d'investissement de réseau avec un modèle de traffic et de localisation, et montre que les solutions stationnaires obtenues sont plus performantes qu'une approche en Recherche Opérationnelle pour la conception du réseau en termes d'accessibilité totale.

Concernant la croissance du réseau seule, [**jacobs2016transport**] propose un modèle de simulation dans lequel les alternatives entre investissements plausibles (par des investisseurs différents) sont évaluées avec un modèle de choix discrets dont la fonction d'utilité prend en compte les retours sur investissement mais également des variables à optimiser comme l'accessibilité. Il est appliqué à la croissance du réseau ferré néerlandais au 19ème siècle, et démontré capable de reproduire assez fidèlement le réseau historique.

Certains de ces modèles, en particulier [**Xie2011**], se fondent sur la Théorie des Jeux pour modéliser le comportement des acteurs. Celle-ci a déjà été largement appliquée à des questions de modélisation en sciences sociales ou politiques pour des questions impliquant des agents cognitifs en interaction avec des intérêts individuels [**ordeshook1986game**]. Ce cadre a par ailleurs été utilisé pour étudier les investissements en termes de transports, comme par exemple par [**Roumboutsos2008209**] qui utilise la notion d'équilibre de Nash pour comprendre les choix des opérateurs publics ou privés quant à l'intégration de leur système dans le système de mobilité plus global. Nous utiliserons des paradigmes de Théorie des Jeux pour intégrer la gouvernance de manière simple dans notre modèle.

Le but de cette section est donc de se placer dans la lignée de ces différents modèles, et de proposer un modèle de co-évolution au sein duquel la croissance du réseau est intégrée de manière endogène, par la modélisation des processus de gouvernance impliqués.

7.3.2 *Le Modèle Lutecia*

Nous décrivons à présent le modèle LUTECIA¹⁵, dans sa structure générale, puis dans la spécification que nous développerons par la suite.

Structure globale du modèle

Le modèle couple de manière complexe un module pour l'évolution de l'usage du sol à un module de croissance de réseau de transport. Les sous-modèles (ou modules), détaillés par la suite, incluent en particulier un modèle de gouvernance pour régir l'évolution du réseau. L'inclusion d'un modèle endogène de provision d'infrastructures basé sur les augmentation itératives de l'accessibilité, au sein d'un modèle Luti, consiste en la contribution principale du modèle LUTECIA.

L'accessibilité, que nous prendrons ici comme un potentiel d'accès des actifs aux emplois, est au cœur du modèle. En effet, les agents micro-économiques se relocalisent afin de maximiser leur accessibilité, tandis que les décisions de nouvelles infrastructures de transport sont prises par des agents de gouvernance selon un critère de maximisation de l'augmentation d'accessibilité dans leur zone.

Dans sa structure la plus générale, le modèle LUTECIA est composé de cinq sous-modèles, parmi lesquels nous n'en développerons que trois ici, vu notre cas d'application au Delta de la Rivière des Perles. Les sous-modèles sont les suivants :

- LU correspond à l'usage du sol : il opère la relocalisation des actifs et des emplois étant donné les conditions courantes d'accessibilité.
- T correspond à Transport : il évalue les conditions de transport (flux, congestion) dans la région urbaine.
- EC correspond à l'évaluation de la coopération : il évalue le ou les agents qui procéderont à la construction d'une nouvelle infrastructure.
- I correspond à provision d'infrastructure : il détermine la localisation de la nouvelle infrastructure de transport en fonction d'un critère de maximisation d'accessibilité.

¹⁵ L'appellation est un acronyme lié à sa structure détaillée par la suite. Nommer les modèles est une opération délicate puisqu'elle induit une certaine réification voire personification, dans tous les cas relève d'un certain fétichisme. Celle-ci peut potentiellement perturber la place du modèle au sein du processus de production de connaissance et faire du modèle une fin en soi. Nous sommes convaincus qu'une dénomination endogène via les usages du modèles par la communauté est plus approprié. Nous faisons ici une exception vu l'histoire particulière de sa genèse.

- A correspond aux agglomérations d'économie : il évalue la productivité des firmes, selon l'accessibilité aux emplois.

Nous étudierons par la suite le couplage entre les sous-modèles LU-EC-I : nous supposons au premier ordre pas d'effet significatif de la congestion, et donc pas de rôle de la modélisation du transport ; et par ailleurs prenons des hypothèses simples sur le plan économique et négligeons les agglomérations d'économie.

Des échelles de temps imbriquées sont incluses dans le modèle : un échelle courte, correspondant à la mobilité quotidienne qui produit les flux dans le réseau de transport et aux productivités des entreprises (modules T et A) ; une échelle intermédiaire pour les dynamiques de localisation des actifs et emplois (module LU) ; et une longue échelle de temps pour l'évolution du réseau (modules EC et I). Les niveau de stochasticité sont pris en conséquence : les échelles les plus petites ont des dynamiques déterministes tandis que la plus longue présente un comportement aléatoire.

Description détaillée du modèle

DESCRIPTION DE L'ENVIRONNEMENT La Méga-région Urbaine est modélisée avec un zonage spatial à deux niveaux. L'environnement du modèle est composé par une grille, dont les cellules sont les unités élémentaires pour quantifier l'usage du sol. Nous supposons que chaque cellule k est caractérisée au temps t par le nombre d'actif y résidant $A_k(t)$ et son nombre d'emplois $E_k(t)$. À un niveau supérieur, la MCR est décomposée en unités administratives qui correspondent au niveau de gouvernance des villes, auxquelles sont attribués M agents abstraits appelés *maires* : M_k désigne ainsi la zone administrative à laquelle chaque cellule appartient. Nous supposons de plus l'existence d'un agent de gouvernance global qui correspond à une autorité typiquement régionale, au niveau de la MCR.

De manière complémentaire à cette configuration d'usage du sol et de gouvernance, nous introduisons un réseau de transport $G = (V, E)$ localisé dans l'espace par les coordonnées de ses noeuds (x_v, y_v) , et caractérisé par une vitesse v_G relative aux déplacements dans l'espace euclidien. Sous l'hypothèse que le réseau peut être rejoint à tout endroit sur les liens, il induit de manière univoque une distance-temps géographique, que nous décrivons par la matrice des plus courts temps entre chaque cellule $D = (d_{k,k'}(t))$. L'accessibilité des actifs aux emplois est alors définie pour chaque cellule comme une accessibilité de Hansen, avec un paramètre de décroissance de la distance λ qui capture la distance typique domicile-travail, par

$$X_k^{(A)} = A_k \cdot \sum_{k'} E_{k'} \exp(-\lambda \cdot d_{k,k'})$$

L'accessibilité des emplois aux actifs est définie de manière similaire. La dynamique est considérée de façon discrete : $t \in \{t_0 = 0, \dots, t_f\}$, avec les pas de temps correspondant à une échelle à laquelle l'usage du sol évolue en moyenne, i.e. de 5 à 10 ans.

EVOLUTION DE L'USAGE DU SOL Pour le module d'usage du sol, le modèle s'inspire du modèle de Lowry [[lowry1964model](#)]. Les relocalisations d'une proportion fixe d'actifs et d'emplois sont supposées à l'équilibre à l'échelle d'un pas de temps. En comparaison, l'évolution de l'infrastructure de transport est largement plus lente [[wegerer2004land](#)]¹⁶. Les actifs et les emplois se relocalisent selon des utilités qui prennent en compte à la fois l'accessibilité et la forme urbaine. En effet, l'un des moteurs de l'étalement urbain peut être interprété comme une répulsion des résidents par la densité. Pour agréger les deux effets de façon simple, nous prenons une fonction de Cobb-Douglas pour l'utilité

$$U_k^{(A)} = X_k^{(A)^{\gamma_A}} \cdot F_k^{(A)^{1-\gamma_A}} \quad (11)$$

ce qui est équivalent à une agrégation linéaire du logarithme des variables explicatives. Les emplois suivent une expression analogue avec un paramètre de poids spécifique γ_E . L'utilité est influencée ici uniquement par l'accessibilité et par un indicateur de forme urbaine locale nommé *facteur de forme*. Nous le définissons dans le cas des actifs par $F_k^{(A)} = \frac{1}{A_k \cdot E_k}$, ce qui signifie que la population est repoussée par la densité. La combinaison de l'effet positif de l'accessibilité à celui négatif de la densité produit une tension entre des objectifs contradictoires, permettant un certain niveau de complexité déjà dans le sous-modèle d'usage du sol seul. Le facteur de forme pour les emplois est pris comme $F_k^{(E)} = 1$ pour simplifier et suivant la logique que les emplois peuvent s'agréger bien plus que les logements.

Les relocalisation sont ensuite faites de manière déterministe suivant un modèle de choix discret, qui donne les valeurs des actifs à l'étape suivante comme

$$A_i(t+1) = \alpha \cdot \sum_i A_i(t) \cdot \frac{\exp(\beta U_i(A))}{\sum_i \exp(\beta U_i(A))} \quad (12)$$

où β est le paramètre de choix discrets qui peut être interprété comme un "niveau d'aléatoire"¹⁷. α est la fraction fixe d'actif se relocalisant. Les emplois suivent une expression similaire.

¹⁶ Nous ne considérons pas ici les valeurs foncières, les loyers ou les coûts de transport, qui sont au cœur des modèles d'Economie Urbaine comme le modèle d'Alonso ou de Fujita par exemple (voir [[lemoy2017exploring](#)] pour une approche récente basée-agents à ceux-ci).

¹⁷ Quand $\beta \rightarrow 0$, toutes les cellules de destination ont une probabilité égale depuis l'ensemble des cellules d'origine, tandis que $\beta \rightarrow \infty$ donne un comportement totalement déterministe vers la cellule avec meilleure utilité.

Evolution du réseau : processus de gouvernance

HYPOTHÈSES Le sous-modèle pour la gouvernance suit les hypothèses suivantes :

- Trois niveaux de gouvernance sont inclus, qui sont un acteur central (la région, ou le gouvernement régional), les acteurs locaux (municipalités) qui agissent seuls, et les acteurs locaux qui coopèrent, ce qui constitue un niveau abstrait intermédiaire.
- Sous l'hypothèse qu'une nouvelle infrastructure doit être construite, la planification peut être soit par le haut (région) soit par le bas (acteurs locaux). Nous supposons que les processus derrière la détermination du niveau de décision sont bien trop complexes (puisque il incluent généralement des processus politiques) pour être pris en compte par le modèle. Cette étape est donc déterminée de manière exogène selon une loi uniforme à paramètre fixe¹⁸.
- Si la décision est prise au niveau local, des négociations entre les acteurs ont lieu. Les concernant, nous supposons :
 - L'initiateur de la nouvelle infrastructure peut être n'importe quel acteur local, mais les villes riches ont plus de chance de construire.
 - Les négociations pour des possibles collaborations n'ont lieu qu'entre acteurs voisins, ce qui est en cohérence avec des segments d'infrastructure de longueur moyenne.
 - Pour cette raison, et d'autant plus que les jeux à n joueurs présentent des comportements chaotiques quand n augmente [2016arXiv161208111S], nous ne considérons des négociations qu'entre deux acteurs uniquement. De plus, la probabilité de coopération endogène peut alors être directement interprétée.
- Pour rester simple, le stock total d'infrastructure construit à un pas de temps de gouvernance est constant, et les temps de décision sont également fixés¹⁹

EVOLUTION DU RÉSEAU Les étapes pour le développement du réseau de transport sont les suivantes :

1. A chaque pas de temps, 2 nouveaux segments de route de longueur l_r sont construits. Le choix entre le niveau local et global est déterminé par un tirage uniforme avec une probabilité

¹⁸ Une piste alternative pour endogéniser ce processus est proposée dans les développements.

¹⁹ Voir également la discussion pour de possibles relaxations de ces hypothèses.

ξ . Dans le cas d'une construction locale, les routes sont attribuées successivement aux maires avec des probabilités ξ_i qui sont proportionnelles au nombre d'emplois de chacun, ce qui signifie que les zones plus riches auront plus de routes.

2. Les zones devant construire chacun une route entrent en négociations. Les stratégies possibles pour les acteurs (zones en négociation, $i = 0, 1$, les stratégies étant notées S_i) sont de ne pas collaborer (NC) et de collaborer (C). Les stratégies sont choisies simultanément (jeu non-coopératif), de manière aléatoire selon des probabilités déterminées comme détaillé ci-dessous. Pour les combinaisons (C, NC) et (NC, C), les routes sont construites séparément et l'agent qui voulait collaborer perd un certain montant investi dans le processus de collaboration. Pour (NC, NC) les deux agissent séparément et pour (C, C) un développement commun est mené.
3. Selon le niveau de gouvernance et les stratégies choisies, l'infrastructure optimale correspondante est construite.

EVALUATION DE LA COOPÉRATION Détaillons à présent la manière dont les probabilités de coopération sont détaillées. Nous notons $Z_i^*(S_0, S_1)$ l'infrastructure optimale en termes de gain d'accessibilité pour la zone i avec $(S_1, S_2) \in \{(NC, C), (C, NC), (NC, NC)\}$ qui sont déterminées de manière heuristique pour chaque zone séparément (voir détails d'implémentation), et Z_C^* l'infrastructure optimale commune calculée sur l'union des deux zones avec une infrastructure composée de deux segments élémentaires. Cette dernière correspond au cas où les deux stratégies sont C. Les accessibilités marginales pour la zone i et l'infrastructure Z sont définies comme $\Delta X_i(Z) = X_i^Z - X_i$. Nous introduisons des coûts de construction, notés I pour un segment de route, supposés uniformes dans l'espace. Nous introduisons de plus un coût de collaboration J qui correspond à un coût partagé pour construire une infrastructure plus grande.

La détermination des probabilités donnant la composition des stratégies mixtes se base sur la matrice de gain, qui donne les gains d'utilité pour chaque joueur et chaque combinaison de décisions. La matrice de gain du jeu est la suivante, avec κ une constante de normalisation ("prix de l'accessibilité"), et les joueurs étant notés $i \in \{0; 1\}$ (tel que $1 - i$ désigne le joueur opposé à i)

$o i$	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	$\begin{cases} U_0 = \kappa \cdot \Delta X_0(Z_0^*) - I - \frac{J}{2} \\ U_1 = \kappa \cdot \Delta X_1(Z_1^*) - I \end{cases}$	$U_i = \kappa \cdot \Delta X_i(Z_i^*) - I$

Pour simplifier, nous supposerons les paramètres de coût redimensionnés à une accessibilité ce qui revient à avoir $\kappa = 1$. Nous verrons par ailleurs que seul des différentiels d'utilité étant déterminants, le coût de construction I ne joue finalement pas de rôle. Cette matrice de gain est utilisée dans deux jeux traduisant des processus complémentaires :

- Le jeu de coordination dans lequel les joueurs ont une stratégie mixte, et pour lequel nous considérons l'équilibre de Nash²⁰ pour les probabilités correspondantes. Ce jeu implique une compétition entre les joueurs.
- Une heuristique selon laquelle les joueurs prennent leur décision suivant un modèle de choix discrets. Celle-ci implique uniquement une maximisation du gain d'utilité et une compétition indirecte seulement.

Notons $p_i = \mathbb{P}[S_i = C]$ la probabilité de chaque joueur de collaborer.

EQUILIBRE DE NASH L'équilibre de Nash à stratégie mixte pour ce jeu non-coopératif peut être obtenu en toute généralité. Nous détaillons le calcul en Annexe A.14. En écrivant $U_i(S_i, S_{1-i})$ la matrice de gain complète, on a l'expression des probabilités

$$p_{1-i} = -\frac{U_i(C, NC) - U_i(NC, NC)}{(U_i(C, C) - U_i(NC, C)) - (U_i(C, NC) - U_i(NC, NC))}$$

Ce qui donne avec les expressions des utilités données précédemment,

$$p_i = \frac{J}{\Delta X_{1-i} Z_C^* - \Delta X_{1-i} Z_{1-i}^*}$$

Cette expression peut être interprétée de la façon suivante : dans ce jeu compétitif, la chance qu'un joueur coopère décroît quand le gain

²⁰ Un équilibre de Nash est un point de stratégies dans un jeu discret non-collaboratif pour lequel aucun joueur ne peut améliorer son gain en changeant sa stratégie [ordeshook1986game].

de l'autre joueur augmente, et d'une certaine manière contre-intuitif, s'accroît quand le coût de collaboration augmente.

Cela impose également des conditions de faisabilité pour J et les gains d'accessibilité pour conserver une probabilité : celles-ci sont $J \leq \Delta X_{1-i}(Z_C^*) - \Delta X_{1-i}(Z_{1-i}^*)$ (équivalent à une condition coût-bénéfice) et $\Delta X_{1-i}(Z_C^*) \leq \Delta X_{1-i}(Z_{1-i}^*)$.

DÉCISIONS PAR CHOIX DISCRETS Avec les mêmes fonctions d'utilité, un modèle d'utilité aléatoire pour un choix discret permet également d'obtenir des expressions des probabilités. On a pour le joueur i le différentiel d'utilité entre le choix C et le choix NC donné par

$$U_i(C) - U_i(NC) = p_{1-i} (\Delta X_i Z_C^* - \Delta X_i Z_i^*) - J$$

Sous l'hypothèse classique d'un modèle d'utilité aléatoire distribuée en loi de Gumbel [[ben1985discrete](#)], on a $P[S_i = C] = \frac{1}{1 + \exp[-\beta_{DC}(U_i(C) - U_i(NC))]}$, où β_{DC} est le paramètre de choix discrets (que nous fixerons grand $\beta_{DC} = 400$, en supposant un certain déterminisme à ce niveau, puisqu'on a ensuite un deuxième niveau aléatoire).

On substitue l'expression de p_{1-i} dans l'expression de p_i , ce qui conduit p_i à vérifier l'équation suivante

$$p_i = \frac{1}{1 + \exp \left(-\beta_{DC} \cdot \left(\frac{\Delta X_i Z_C^* - \Delta X_i Z_i^*}{1 + \exp(-\beta_{DC}(p_i \cdot (\Delta X_{1-i}(Z_C^*) - \Delta X_{1-i}(Z_{1-i}^*)) - J))} - J \right) \right)}$$

Nous démontrons (voir Annexe [A.14](#)) qu'il existe toujours une solution $p_i \in [0, 1]$, et nous la résolvons numériquement pour déterminer la probabilité de collaboration dans le modèle.

Implémentation du modèle

L'ensemble des paramètres du modèle est rappelé en Table [16](#). Nous ne donnons ici que les paramètres qui n'ont pas été fixés explicitement précédemment, et il s'agit des paramètres privilégiés sur lesquels l'exploration et l'application du modèle sera faite.

Le modèle est implémenté en Netlogo, pour des raisons d'ergonomie vu son niveau de complexité, ainsi que les possibilités d'exploration interactives. Une attention particulière a été portée aux points suivants :

- Les calculs des matrices de distance sont nécessaires pour chaque segment d'infrastructure potentiel, ce qui rend le module de gouvernance très couteux sur le plan computationnel. Nous utilisons donc un calcul des plus courts chemins basé sur la programmation dynamique, inspiré de [[tretyakov2011fast](#)], mettant à jour directement la matrice des distances plutôt que de recalculer les plus courts chemins à chaque fois.

TABLE 16 : Résumé des paramètres du modèle LUTECIA. Nous donnons également les processus correspondant, les bornes typiques de variation et leur valeur par défaut.

Sous-modèle	Paramètre	Nom	Processus	Domaine	Défaut
Usage du sol	λ	Portée de l'accessibilité	Accessibilité	$]0; 1]$	0.001
	γ_A	Exposant de Cobb-Douglas actifs		$[0; 1]$	0.85
	γ_E	Exposant de Cobb-Douglas emplois		$[0; 1]$	0.85
	β	Exposant choix discrets	Relocalisation	$[0; +\infty]$	1
Transports	α	Taux de relocalisation		$[0; 1]$	0.05
	v_G	Vitesse du réseau	Hiérarchie	$[1; +\infty[$	5
	J	Coût de collaboration	Planification	$[0; 0.005]$	0.001
Gouvernance	l_r	Longueur de l'infrastructure		$]0; \sqrt{2} \cdot K[$	2

- Le réseau est pour cette raison représenté de manière duale, sous forme vecteur et raster. Le passage de l'un à l'autre et leur cohérence est assuré.
- Pour la détermination de l'infrastructure optimale, l'ordre de grandeur du nombre total d'infrastructures à explorer est un $O(l_r \cdot N)$, si N est le nombre de patches et en supposant que l'ensemble des infrastructures potentielles ont leur extrémités au centre d'une cellule²¹. Cela augmente considérablement le coût computationnel opérationnel, et nous utilisons une heuristique explorant un nombre fixé N_I d'infrastructures choisies aléatoirement.

Plus de détails d'implémentation sont donnés en Annexe A.14.

Validation du modèle

Différentes expériences nous permettent de valider le modèle dans une certaine mesure. Nous adoptons une stratégie modulaire, c'est à dire par tests relativement indépendants des sous-modèles pour commencer.

Nous travaillons sur des systèmes synthétiques. Les configurations de populations et d'emplois suivent des mélanges d'exponentielles. Nous donnons en Annexe A.14 les détails des paramètres d'initialisation.

USAGE DU SOL Les dynamiques d'usage du sol ont toujours un état stationnaire lorsque le réseau n'évolue pas. Nous démontrons l'existence de l'équilibre en A.14. Par ailleurs, les expériences numériques

²¹ Pour chaque cellule, on aura une infrastructure pour chaque autre cellule à un rayon l_r , ce qui asymptotiquement revient au périmètre du cercle $2\pi l_r$. Par ailleurs, comme précisé en A.14, on suppose une heuristique d'accrochage aux infrastructures existantes pour garder un réseau cohérent.

montrent que le modèle converge assez rapidement. Les expériences ciblant l'usage du sol uniquement et qui sont détaillées en A.14 fournissent les résultats suivants :

- Une grande diversité de trajectoires morphologiques dans le temps, c'est à dire l'évolution des indicateurs morphologiques pour la distribution de la population et des emplois, est obtenue en jouant sur les paramètres $\gamma_A, \gamma_E, \lambda, \beta$, ainsi que sur la structure d'un réseau statique.
- De même, ces trajectoires ne convergent pas vers les mêmes formes et on a donc une diversité des formes finales obtenues.
- Il est possible de minimiser à $\alpha = 1$ fixé la quantité totale de relocalisation. Nous jouerons toutefois sur ce paramètre pour contrôler la vitesse d'étalement urbaine, et prendrons typiquement des valeurs autour de 0.1.

GOUVERNANCE Afin de comprendre l'influence des paramètres de gouvernance sur les formes produites par le modèle, nous menons une expérience simple dans le cas d'un système bi-centrique, sans réseau initial. Les paramètres du modèle d'usage du sol sont fixés à des valeurs standard $\gamma_A = \gamma_E = 0.8, \beta = 2; \lambda = 0.001, \alpha = 0.16$ et la longueur des tronçons est fixée à $l_r = 2$. Nous considérons uniquement le jeu à choix discrets. La situation de référence est donnée par un niveau de décision uniquement régional, correspondant à $\xi = 1$. Nous la comparons à deux situations dans lesquelles le niveau de décision est uniquement local ($\xi = 0$) mais pour lequel nous forçons les probabilités de collaboration à des valeurs extrêmes par l'intermédiaire du coût de coopération, pris respectivement comme $J = 0$ et $J = 0.005$.

La configuration initiale ainsi que trois exemples de formes de réseau obtenues pour chacune des configurations sont montrés en Fig. 55. Les formes de réseau sont visuellement²² différentes et témoignent de caractéristiques de structure particulière. Dans le cas de la décision régionale, un arc structurant relie les deux centre, à partir duquel se branchent des ramifications d'abord perpendiculaires puis parallèles. La structure obtenue dans le cas d'un local collaboratif est également arborescente mais comporte moins de branches, les prolongement se faisant majoritairement à la suite des branches existantes. Enfin, comme on pouvait s'y attendre, le réseau non-collaboratif paraît moins optimal en termes de couverture que les deux premiers, et présente des redondances. Concernant la structure urbaine, on obtient que les niveaux locaux conservent plus la structure bicentrique en comparaison au niveau régional (voir la position des centres finaux par rapport à la position initiale) : via le réseau, la prise de décision

²² Cette expérience préliminaire n'implique pas d'exploration intensive, et il n'est donc pas possible de traduire ces conclusions de manière robuste en termes de statistiques des indicateurs.

au niveau régional a plus de potentiel pour créer de nouvelles centralités.

7.3.3 Application au Delta de la Rivière des Perles

Il a été suggéré par [liao2017ouverture] qu'une forme de gouvernance multi-niveau a récemment émergé en Chine, dans le contexte des activités économiques. Nous tentons par notre modèle de tester la pertinence de ce paradigme au regard de la structure urbaine de la MCR.

Initialisation du modèle

Nous travaillons sur une configuration raster simplifiée (cellules de 5km) pour la population du Delta de la Rivière des Perles, ainsi que sur le réseau d'autoroute stylisé. Nous considérons le réseau routier uniquement puisque, selon [hou2011transport], il s'agit du moteur principal des changements dans les motifs d'accessibilité en comparaison au réseau ferré dont le développement accéléré est récent. Les réseaux sont stylisés à partir du plan donné par [hou2011transport] qui reproduit les documents officiels de la province du Guangdong en 2010. Nous considérons ainsi le réseau autoroutier en 2010 et celui planifié à cette époque. La Fig. 56 illustre la configuration stylisée pour le Delta de la Rivière des Perles.

Procédure de calibration

Lors de l'application d'un modèle si complexe à une situation semi-réelle, il faut rester vigilant. Il est important de choisir les processus adéquats ainsi que le niveau de granularité à reproduire. En particulier, notre modèle produit des motifs d'usage du sol relativement précis, mais utilise leur approximation comme base de la croissance du réseau, dont l'évolution qualitative permet d'informer sur les processus de gouvernance. Nous proposons pour cela de "calibrer" sur la forme d'une infrastructure donnée, au sens de déterminer des configurations de paramètres pour lesquelles en probabilité les morceaux successifs d'infrastructure sont les plus proches d'une infrastructure visée.

Pour calibrer sur les réseaux produits par la simulation, il s'agit de comparer à un réseau de référence. C'est un problème difficile, puisque différentes mesures de proximité avec différentes signification peuvent être utilisées. Les mesures géométriques s'intéressent à la proximité spatiale des réseaux. Pour un réseau $(E, V) = ((e_j), (v_i))$, une distance basée sur les noeuds est donnée par $\sum_{i \neq i'} d^2(v_i, v_{i'})$. Une mesure plus précise qui n'est pas biaisée par d'éventuels noeuds intermédiaires est donnée par l'aire cumulée entre chaque paire de liens $d_A = \sum_{j \neq j'} A(e_j, e_{j'})$ (il ne s'agit pas d'une distance à propre-

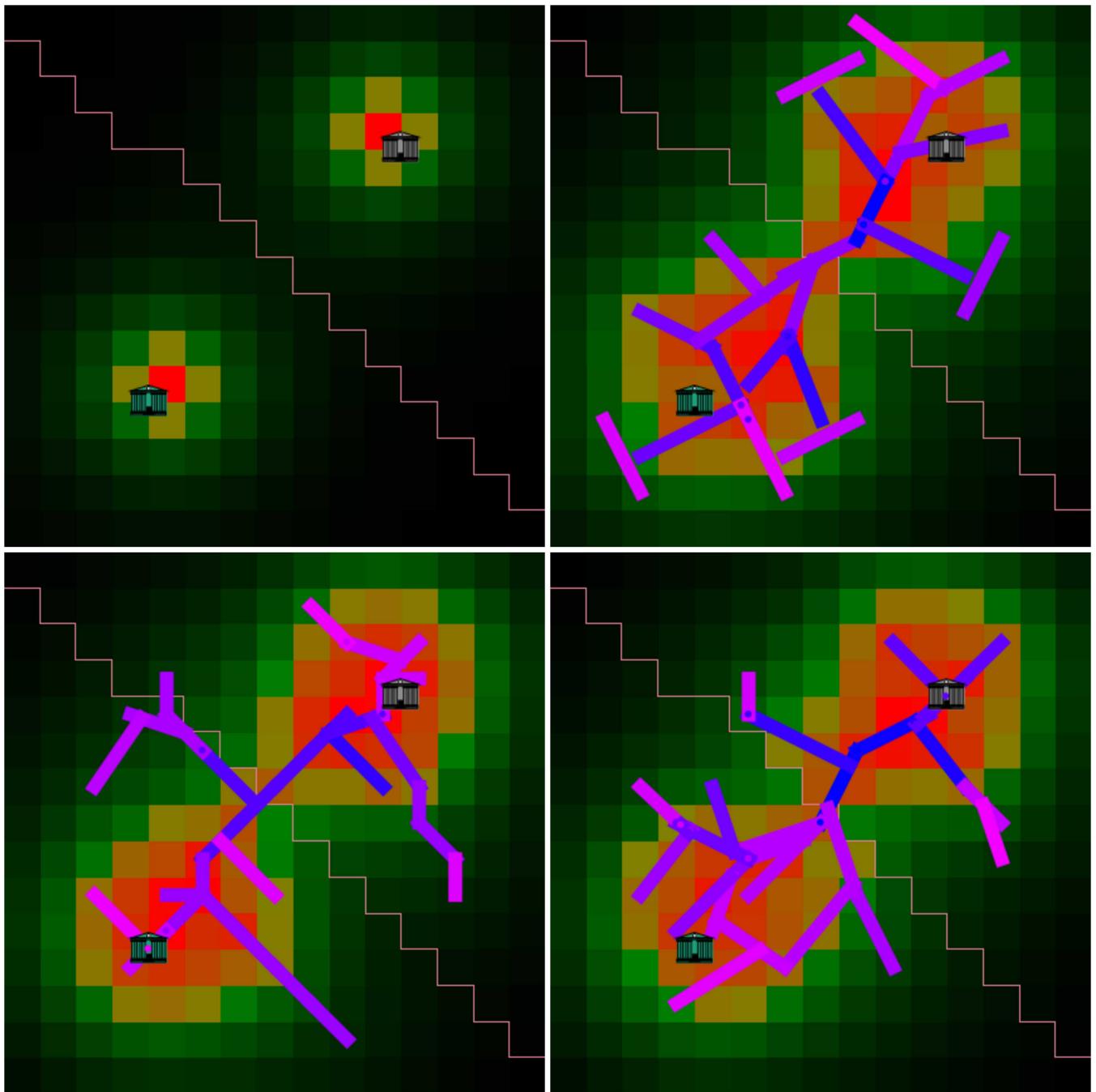


FIGURE 55 : Formes de réseau obtenues pour différents niveaux de gouvernance. Le modèle est initialisé sur une configuration synthétique symétrique à deux centres (Haut gauche). Les paramètres pour l'évolution de l'usage du sol sont $\gamma_A = \gamma_E = 0.8; \beta = 2; \lambda = 0.001; \alpha = 0.16$, et pour l'évolution du réseau $l_r = 2$ et un jeu à choix discrets. L'évolution est stoppée à stock constant $S = 50$ et l'exploration heuristique faite pour $N_e = 200$. (Haut droite) Niveau de décision régional ($\xi = 1$) ; (Bas gauche) Niveau local ($\xi = 0$) et bas niveau de collaboration, obtenu avec un fort coût de coopération $J = 0.005$; (Bas droite) Niveau local et haut niveau de collaboration, avec $J = 0$.

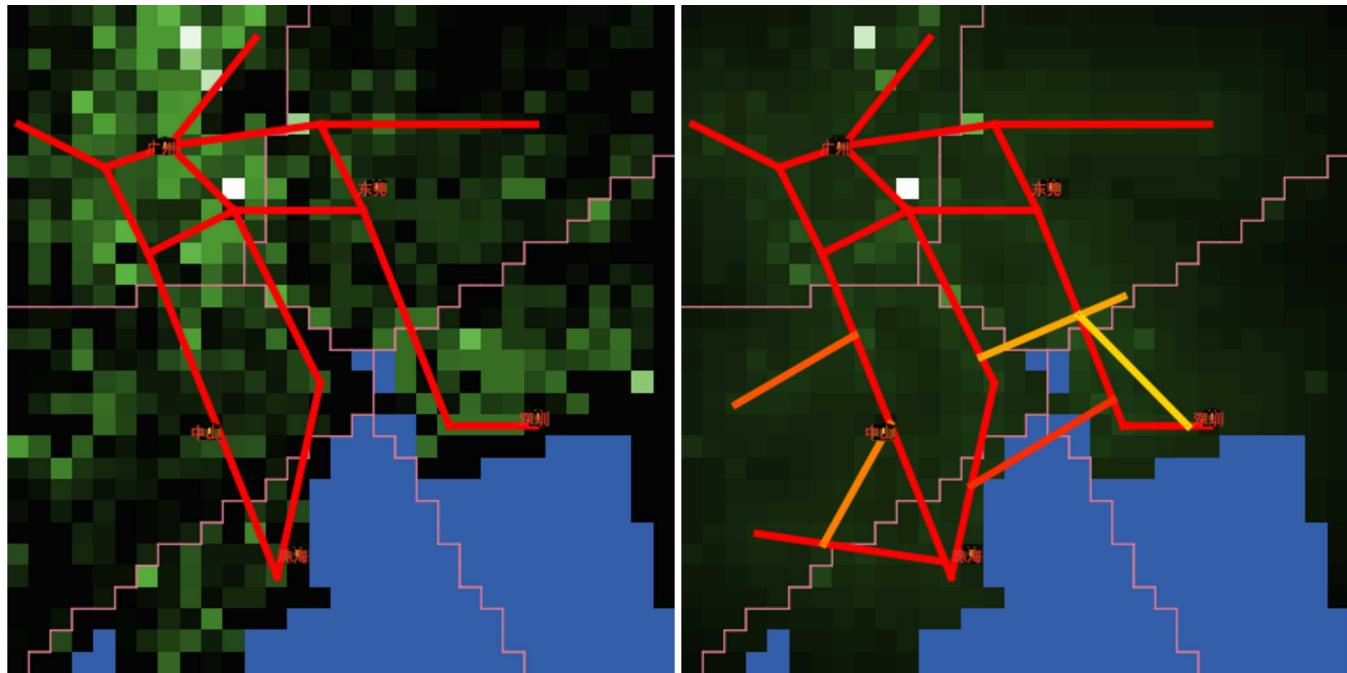


FIGURE 56 : Exemple d'application au Delta de la Rivière des Perles. (Gauche) Initialisation avec le raster de population 2010, agrégé à la résolution 5km, et le réseau autoroutier simplifié ; (Droite) Etat après 6 pas de temps ($\alpha = 1$)

ment parler), où $A(e, e')$ est l'aire du polygone fermé constitué en reliant les sommets des liens. Nous considérerons cette dernière pour la calibration.

Calibration

Les expériences que nous menons sont à usage du sol fixé, le niveau de détail requis pour des données plus anciennes et plus récentes, voir des projections, pour la population et les emplois n'étant pas permis par les données à notre disposition.

Nous faisons varier les paramètres de gouvernance, incluant le type de jeu, avec $l_r = 2$ fixé, et explorons un échantillonnage LHS de 4000 points dans l'espace de ces paramètres, avec 10 répétitions du modèle pour chaque point. Les deux expériences menées correspondent à des configurations cible différentes :

- pas de réseau initial et réseau de 2010 comme cible, dans l'esprit d'extrapoler la configuration de gouvernance la plus probable ayant mené à la configuration actuelle ;
- réseau 2010 initial, et réseau planifié comme cible : extrapolation de la configuration de gouvernance de la planification.

Nous obtenons des résultats qualitativement similaires pour les deux expériences, suggérant qu'il n'y a pas eu de transition de type

de gouvernance entre réseau passé et réseau futur. Les résultats sont illustrés en Fig. 57. On obtient, à l'étude du graphe de d_A en fonction de ξ , que le niveau régional est le plus fidèle pour reproduire la forme du réseau. Par contre, les jeux de choix discrets et de compétition se comporte différemment, et le jeu compétitif est le plus proche de la réalité quand ξ diminue : les relations entre acteurs locaux seraient a priori de nature plus compétitive qu'égoïste. Quand on étudie la variation de la distance en fonction du niveau de collaboration observé, on obtient une forme intéressante en cloche inversée, c'est à dire que les situations les plus probables sont soit celles où il n'y a que de la collaboration, soit celles où il n'y en a pas du tout, mais pas de situations intermédiaires. Enfin, la comparaison des distributions statistiques des distances entre les configurations cibles et les types de jeux montre que la différence entre les jeux n'est considérable que pour le réseau réel mais pas le réseau planifié (conclusion difficile à interpréter).

Nous tirons donc de cette expérience les conclusions suivantes, à prendre bien sûr avec prudence :

- Une compétition entre les acteurs est plus probable qu'un comportement égoïste dans le cas de décisions locales.
- Les compromis de collaboration forment des réseaux moins probables que les situations avec pleine collaboration ou avec aucun collaboration.

Ces conclusions peuvent être mises en perspective avec la compétition accrue au sein du Delta révélée par [xu2005city]. Ainsi, cette application du modèle permet d'inférer indirectement des processus de gouvernance.

Discussion

Bien que le modèle ait encore à être exploré plus en profondeur et pour l'ensemble de ses modules, certains développements possibles peuvent retenir notre attention.

NIVEAU DE DÉCISION ENDOGÈNE Une extension pertinente serait l'étude de l'émergence de zones administratives par agrégation, c'est à dire l'émergence de nouveaux niveaux de gouvernance dans une région métropolitaine polycentrique. L'exemple de la Métropole du Grand Paris en est une bonne illustration, puisqu'elle se situe entre les collectivités locales et la région ainsi que l'état [gilli2009paris]. Une extension du modèle avec des règles de fusion des entités est une direction potentielle pour étudier cette question.

COMPÉTITION POUR UNE RESSOURCE EXTERNE L'influence des territoires extérieurs ou des externalités sur l'évolution d'une MCR

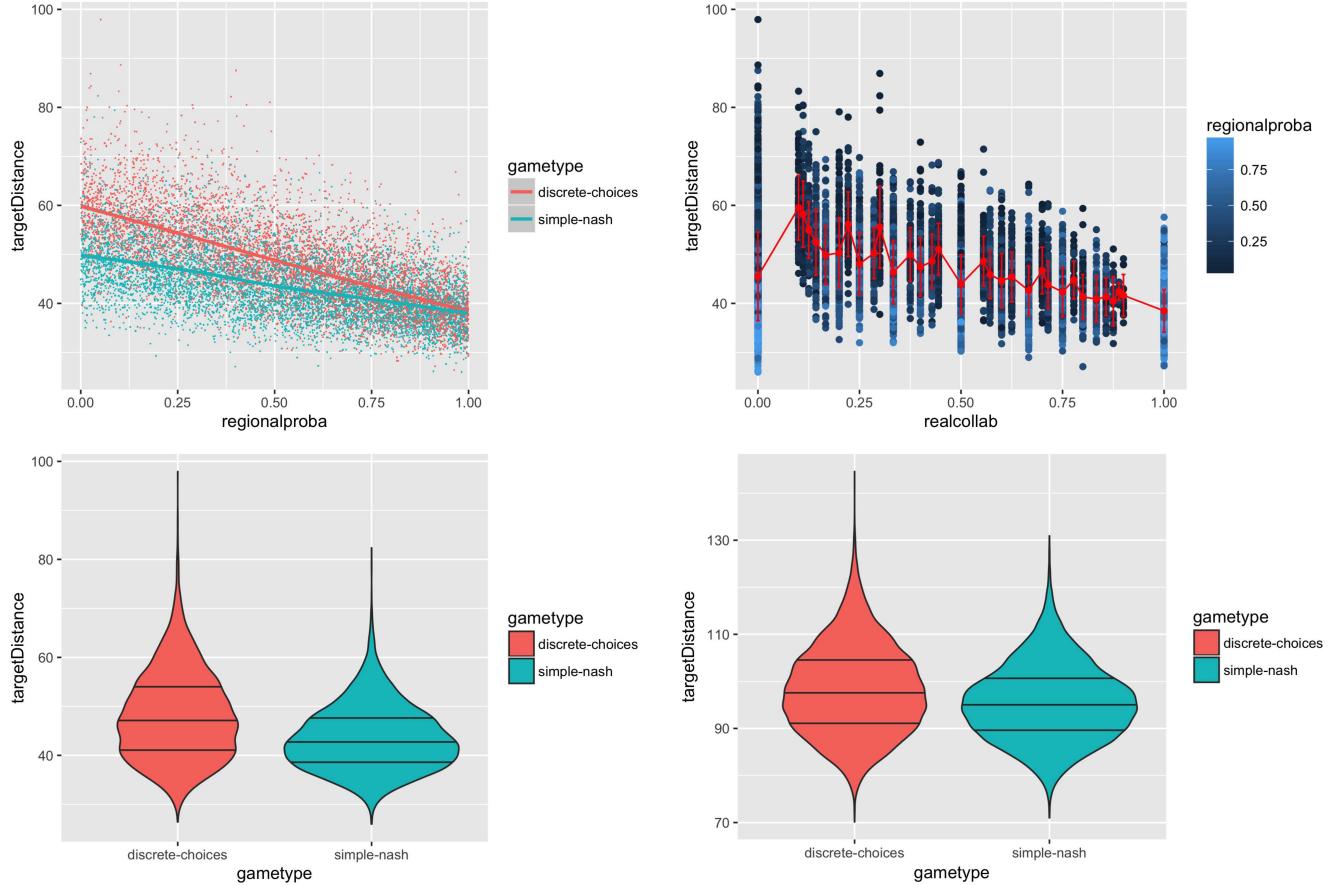


FIGURE 57 : **Calibration du modèle à usage du sol fixé.** On prend $\alpha = 0$ pour ne faire évoluer que le réseau, et échantillonnons l'espace des paramètres de gouvernance. (*Haut Gauche*) Distance d_A au réseau cible (`targetDistance`), dans le cas du réseau réel, en fonction de la probabilité de décision régionale ξ (`regionalproba`), pour les deux types de jeu (couleur). (*Haut Droite*) Distance d_A en fonction de la probabilité de collaboration observée (`realcollab`); la courbe rouge donne les moyennes avec les barres d'erreur. (*Bas Gauche*) Distribution statistique de la distance en fonction du type de jeu, dans le cas du réseau réel; (*Bas Droite*) dans le cas du réseau planifié.

est une question ouverte. Dans le cas d'une ressource commune, localisée dans l'emprise de la MCR, des dynamiques de compétition ou de collaboration peuvent s'instituer entre acteurs pour son exploitation. Ce modèle est une solution pour étudier cette situation de manière stylisée, et réaliser ainsi une expérience contrôlée sur les dynamiques de co-évolution, qui permettrait de répondre à des questions plus générales quant au rôle de l'isolation territoriale dans les processus de co-évolution.

Nous avons ainsi posé les premières briques de modèles visant à une intégration plus complexe des processus de co-évolution, en introduisant le modèle Lutetia qui a ensuite été validé de manière préliminaire et dont les potentialités ont été démontrée par l'application au cas d'étude du Delta de la Rivière des Perles.

* * *

*

CONCLUSION DU CHAPITRE

Cette deuxième entrée sur les modèles de co-évolution, à l'échelle mesoscopique, a été l'occasion d'explorer le couplage entre forme urbaine et fonctions au travers du couplage entre territoire et réseau. En comparaison aux modèles macroscopiques, les processus pris en compte ici sont beaucoup plus variés et complémentaires.

Un premier modèle de morphogenèse inclut différentes heuristiques pour la croissance du réseau, qui sont nécessaires et complémentaires pour capturer toute l'étendue possibles des configurations de réseau générées. Nous montrons que le modèle est capable de se rapprocher de situation observées, pour la forme urbaine, le réseau, ainsi que pour les corrélations statiques entre ces indicateurs, tout en nécessitant un compromis entre ces différents objectifs. En termes de régimes de causalité, et donc de capture de dynamiques co-évolutives, le modèle est capable d'en capturer dans certaines situations précises, mais on tire de cette expérience une leçon fondamentale pour les modèles de co-évolution : une fidélité des processus ou des configurations statiques doit se faire au prix de la flexibilité des régimes dynamiques produits. Cela peut être un effet structurel des modèles, ou plus intéressant, une restriction des régimes existants dans les situations réelles. Cela ouvre ainsi des avenues pour la recherche future.

Nous avons ensuite fait le pari d'introduire un modèle plus complexe, incluant une ontologie pour les processus de gouvernance pour l'évolution du réseau de transport. Nous menons des premières expériences de validation du modèle sur données synthétiques, et proposons une application au cas du Delta de la Rivière des Perles, renouvelant le regard que nous en avons apporté en 1.2. Nous montrons par exemple qu'il est possible d'extrapoler des paramètres liés au niveau de collaboration entre acteurs. Cette section permet ainsi d'introduire une nouvelle façon de considérer la co-évolution, prenant en compte l'intégralité du cadre conceptuel développé en 1, et ouvre également de nombreuses perspectives de recherche.

★ ★

*

CONCLUSION DE LA PARTIE III

Towards operational Models : what is possible ; what is desirable ; etc.

Vers des Modèles Opérationnels de Coévolution

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.

Quatrième partie

OUVERTURE

Un bâtiment n'est jamais utilisé de la façon pour laquelle il a été conçu : l'intégration de cette réalité fera la différence entre un bon et un excellent architecte. L'utilisation fonctionnelle effective donne sens à la forme, tout en dépendant de celle-ci. Il en est de même pour une construction de connaissances. Nous prenons à présent du recul et ouvrons des perspectives à la fois empirique et théoriques.

INTRODUCTION DE LA PARTIE IV

8

ECHELLES ET ONTOLOGIES

La richesse des interactions entre réseaux et territoires, développée dans le Chapitre 1, est que celle-ci se produisent à différentes échelles, entre ces échelles, et par des intermédiaires très variés, au sens des agents ou structures impliquées mais aussi de leur caractéristiques, ceux-ci allant de la congestion des réseaux aux dynamiques sur le temps long en passant par les re-localisations des activités par exemple. Le cas de Zhuhai développé en 1.2 illustre la complexité d'une trajectoire locale et régionale, d'une bifurcation politique induisant l'instauration de la Zone Economique Spéciale par XI JINPING conditionnée à une bifurcation historique bien plus ancienne liée à la colonisation européenne qui a conduit au statut actuel de Macao, à une bifurcation socio-historico-géographique en terme d'accessibilité régionale et une nouvelle position centrale de la ville dans la Mega-city Region du Delta de la Rivière des Perles. Nous avons dans le Chapitre 4 étudié empiriquement les manifestations morphologiques des interactions à l'échelle mesoscopique, mais également mis en évidence des effets de structure à cette même échelle sur un temps long dans le cas de l'Afrique du Sud. Quelle échelle minimale est-il pertinent de considérer, autrement dit l'étude de l'échelle microscopique peut-elle nous apporter de l'information? Et peut-on clarifier certaines ontologies, ou au moins un certain degré de précision ou de complexité requis dans celles-ci? Ce chapitre cherche à répondre à ces interrogations par le biais d'études empiriques. Ainsi, nous tentons de préciser itérativement la structure des modèles futurs, mais aussi leur non-structure.

Dans une première section 8.2, nous explorons empiriquement un jeu de données à l'échelle microscopique sur le trafic routier en Ile-de-France, en ayant notamment à l'esprit la notion d'équilibre des flux de trafic qui est une hypothèse particulièrement répandue dans la modélisation du trafic. Nous démontrons que cet équilibre n'a aucun fondement empirique, ce qui amène à questionner son application à des situations réelles, et que les trajectoires microscopiques du système sont chaotiques. Cela nous permettra d'une part de confirmer nos choix épistémologiques de modèle loin de l'équilibre typique d'une appréhension de la complexité, d'autre part de confirmer que cette échelle n'est pas pertinente. Nous continuons sur le traffic routier dans une deuxième section 8.3, en nous concentrer sur la composante du prix de transport via le proxy du prix de vente du carburant, et ces liens potentiels avec les caractéristiques socio-économiques des territoires, dans le cas des Etats-Unis avec une granularité spatiale

au comté et temporelle à la journée. Nous obtenons le résultat assez inattendu des deux échelles endogènes proprement définies, correspondant aux échelles mesoscopique et macroscopique, mais aussi la mise en évidence de la superposition de processus de gouvernance à des processus locaux.

* * *

*

La première section de ce chapitre est inédite. La suite de ce chapitre est entièrement adapté d'articles : la section 8.2 a été publiée en anglais comme [raimbault2017investigating]; la section 8.3 également en anglais en collaboration avec A. BERGEAUD comme [raimbault2017cost].

8.1 REPRÉSENTATION TERRITORIALES

“Quel compromis sur les objets et échelles sont réalisés lors de la création d'un modèle de simulation?” Cette question pourtant fondamentale à toute construction d'un modèle, reste aujourd'hui largement ouverte, et est empruntée ici à la description d'une session spéciale sur la simulation au congrès du CIST 2018¹.

8.1.1 *Representations de systèmes territoriaux*

8.1.2 *Dimension intrinsèque d'un système territorial*

8.1.3 *Etendre les ontologies*

★ ★

★

¹ Voir l'appel à communication à <https://cist2018.sciencesconf.org/resource/page/id/22>. Le CIST a pour but “de formaliser et organiser le champ interdisciplinaire des sciences du territoire” (voir <http://www.gis-cist.fr/cist/missions/>). Groupement de divers laboratoires et instituts dans des domaines très divers, il cherche à développer la connaissance des territoires dans un souci d'interdisciplinarité crucial.

8.2 EQUILIBRE UTILISATEUR STATIQUE

L'Equilibre Utilisateur Statique est un cadre puissant pour l'étude théorique du trafic. Malgré l'hypothèse de stationnarité des flots qui intuitivement limite son application aux systèmes de trafic réels, de nombreux modèles opérationnels qui l'implémentent sont toujours utilisés sans validation empirique de l'existence de l'équilibre. Nous étudions celle-ci sur un jeu de données de trafic couvrant trois mois sur la région parisienne. L'implémentation d'une application d'exploration interactive de données spatio-temporelles permet de formuler l'hypothèse d'une forte hétérogénéité spatiale et temporelle, guidant les études quantitatives. L'hypothèse de flots localement stationnaires est invalidée en première approximation par les résultats empiriques, comme le montrent une forte variabilité spatio-temporelle des plus courts chemins et des mesures topologiques du réseau comme la centralité de chemin. De plus, le comportement de l'index d'autocorrelation spatiale pour les motifs de congestion à différentes portées spatiales suggère une évolution chaotique à l'échelle locale, en particulier lors des heures de pointe. Nous discutons finalement les implications de ces résultats empiriques et proposons des possibles développements futurs basés sur l'estimation de la stabilité dynamique au sens de Lyapounov des flots de trafic.

8.2.1 Contexte

La modélisation du trafic a été largement étudiée depuis les travaux séminaux de Wardrop ([[wardrop1952road](#)]) : les enjeux économiques et techniques justifient entre autre le besoin d'une compréhension fine des mécanismes régissant les flots de trafic à différentes échelles. Différentes approches aux objectifs différents coexistent aujourd'hui, parmi lesquels on trouve par exemple les modèles dynamiques de micro-simulation, généralement opposés aux techniques de basant sur l'équilibre. Tandis que la validité des modèles microscopiques a été étudiée de façon conséquente et leur application souvent questionnée, la littérature est relativement pauvre en études empiriques assurant l'hypothèse d'équilibre stationnaire du cadre de l'Equilibre Utilisateur Statique (EUS). De nombreux développements plus réalistes on été documentés dans la littérature, tels l'Equilibre Utilisateur Dynamique Stochastique (EUDS) (voir pour une description par exemple [[han2003dynamic](#)]). A un niveau intermédiaire entre les cadres statiques et stochastiques se trouve l'Equilibre Utilisateur Stochastique Restreint, pour lequel les choix d'itinéraire des utilisateurs sont contraints à un ensemble d'alternatives réalistes ([[rasmussen2015stochastic](#)]). D'autres extensions prenant en compte le comportement de l'utilisateur via des modèles de choix ont été proposé plus récemment, comme [[zhang2013dynamic](#)] qui inclut à la fois l'influence de la ta-

rification routière et de la congestion sur le choix avec un modèle Probit. La relaxation d'autres hypothèses restrictives comme la maximisation pure de l'utilité par l'utilisateur ont aussi été introduites, tels l'Equilibre Utilisateur Borné décrit par [mahmassani1987boundedly]. Dans ce cadre, l'utilisateur est satisfait si son utilité tombe dans un intervalle et l'équilibre est achevé lorsque chaque utilisateur est satisfait. Les dynamiques résultantes sont plus complexes comme révélé par l'existence d'équilibres multiples, et permet de rendre compte de faits stylisés spécifiques comme des évolutions irréversibles du réseau comme développé par [guo2011bounded]. D'autres modèles d'attribution de trafic inspirés d'autres domaines ont également été plus récemment proposés : dans [puzis2013augmented], une définition étendue de la centralité de chemin qui combine linéairement la centralité des flots non-constraints avec une centralité pondérée par le temps de parcours permet d'obtenir une forte corrélation avec les flots de trafic effectifs, fournissant ainsi un modèle d'attribution de trafic. Cela fournit également des applications pratiques comme l'optimisation de la distribution spatiale des capteurs de trafic.

Malgré ces nombreux développements, de nombreuses études et applications concrètes se reposent toujours sur l'Equilibre Utilisateur Statique. La région parisienne utilise par exemple un modèle statique (MODUS) pour gérer et planifier le trafic. [leurent2014user] introduit un modèle statique de flots qui inclut les recherches locales et le choix du parking : il est légitime de s'interroger, en particulier à de si faibles échelles, si la stationnarité de la distribution des flots est une réalité. Une example d'exploration empirique des hypothèses classiques est donné par [zhu2010people], pour lequel les choix d'itinéraires révélés sont étudiés. Les conclusions questionnent le "premier principe de Wardrop" qui implique que les utilisateurs choisissent parmi un ensemble d'alternatives parfaitement connu. Dans le même esprit, nous étudions l'existence possible de l'équilibre en pratique. Plus précisément, l'EUS suppose une distribution stationnaire des flots sur l'ensemble du réseau. Cette hypothèse reste valable dans le cas d'une stationnarité locale, tant que l'échelle temporelle d'évolution des paramètres est considérablement plus grande que les échelles typiques de voyage. Le second cas qui est plus plausible et de plus compatible avec les cadres théoriques dynamiques est testé ici.

C (FL) : ?

La suite de ce travail s'organise ainsi : la procédure de collection de données ainsi que le jeu de données sont décrits ; nous présentons ensuite une application interactive pour l'exploration du jeu de données, dans le but de fournir une intuitions sur les motifs présents ; puis nous donnons divers résultats d'analyses quantitatives allant dans le sens d'indices convergents pour une non-stationnarité des flots de trafic ; nous discutons finalement les implications de ces résultats et des développements possibles.

C (FL) : smt (?) : tu ne dis pas ce que tu veux faire ni pourquoi.

8.2.2 Résultats

Collecte des données

CONSTRUCTION DU JEU DE DONNÉES Nous proposons de travailler sur l'étude de cas de la région métropolitaine de Paris. Un jeu de données ouvert a été construit, comprenant les liens autoroutiers dans la région, par collecte des données publiques en temps réel des temps de parcours (disponible sur www.sytadin.fr). Comme rappelé par [bouteiller2013open], la disponibilité de jeux de données ouverts pour les transports est loin d'être la règle, et nous contribuons ainsi à une ouverture par la construction de notre jeu de données. La procédure de collecte de données consiste en les points suivants, exécutés toutes les deux minutes par un script python :

- récupération de la page web brute donnant les informations de trafic
- parsing du code html afin de récupérer les identifiants des liens de trafic et les temps de parcours correspondants
- insertion des liens dans une base sqlite avec le temps courant.

Le script automatisé de collection des données continue d'enrichir la base au fur et à mesure du temps, permettant des développements futurs de ce travail sur un jeu de données plus large, et une réutilisation potentielle pour des travaux scientifiques ou opérationnels. La dernière version du jeu de données au format sqlite est disponible en ligne sous une Licence *Creative Commons*².

DESCRIPTION DES DONNÉES Une granularité de deux minutes a été obtenue pour une période de trois mois (de février 2016 à avril 2016 inclus). La granularité spatiale est en moyenne de 10km, les temps de trajet étant fournis pour les liens majeurs. Le jeu de données contient 101 liens. La variable brute utilisée est le temps de trajet effectif, à partir duquel il est possible de construire la vitesse de trajet et la vitesse relative de trajet, définie comme le rapport entre temps de trajet optimal (temps de trajet sans congestion, pris comme le temps minimal sur l'ensemble des pas de temps) et le temps de trajet effectif. La congestion est calculée par inversion d'un fonction BPR³ simple avec exposant 1, i.e. en prenant $c_i = 1 - \frac{t_{i,\min}}{t_i}$ avec t_i temps de trajet effectif dans le lien i et $t_{i,\min}$ temps de trajet minimal.

Méthodes and Résultats

VISUALISATION DES MOTIFS SPATIO-TEMPORELS DE CONGESTION

Notre approche étant entièrement empirique, une bonne connaissance

² à l'adresse http://37.187.242.99/files/public/sytadin_latest.sqlite3

³

des motifs existants pour les variables de traffic, en particulier de leur variations spatio-temporelles, est crucial pour guider toute analyse quantitative. En s'inspirant de la littérature étudiant la validation empirique de modèles, plus précisément les techniques de *Modélisation orientée-motifs* introduites par [grimm2005pattern], nous nous intéressons au motifs macroscopiques à des échelles temporelles et spatiales données : d'une manière équivalente aux faits stylisés qui sont dans cette approches extraits d'un système avant de tenter de le modéliser, nous devons explorer les données de manière interactive dans le temps et l'espace afin d'identifier des motifs pertinents et les échelles associées. Une application web interactive a ainsi été implémentée pour explorer les données, à l'aide des packages R `shiny` et `leaflet`⁴. Cela permet une visualisation dynamique des motifs de congestion sur l'ensemble du réseau ou dans une zone particulière grâce au zoom. L'application est accessible en ligne à l'adresse <http://shiny.parisgeo.cnrs.fr/transportation>. La Figure 58 présente une capture d'écran de l'interface. La conclusion majeure de l'exploration interactive des données est qu'une grande hétérogénéité spatiale et temporelle est la règle. Le motif temporel le plus récurrent, la périodicité journalière des heures de pointe, est perturbée pour une proportion non négligeable de jours. En première approximation, les heures creuses peuvent être approchées par une distribution localement stationnaire des flots, tandis que les heures de pointe sont trop courtes pour pouvoir impliquer la validation de l'hypothèse d'équilibre. Concernant l'espace, aucun motif spatial particulier n'émerge clairement. Cela signifie que dans le cas d'une validité de l'équilibre utilisateur statique, les méta-paramètres régissant son établissement doivent varier à des échelles temporelles plus courtes qu'un jour. Nous postulons au contraire que le système de traffic est loin de l'équilibre, en particulier pendant les heures de pointe pendant lesquelles des transitions de phase critiques à l'origine des embouteillages émergent.

C (FL) : lesquelles et pourquoi ?

VARIABILITÉ SPATIO-TEMPORELLE DES TRAJETS A la suite de l'exploration interactive des données, nous proposons de quantifier la variabilité spatiale des motifs de congestion pour valider ou invalider l'intuition que si l'équilibre existe par rapport au temps, il est fortement dépendant de l'espace et localisé. La variabilité spatio-temporelle des plus courts chemins de trajet est une première façon d'étudier la stationnarité des flots d'un point de vue de théorie des jeux. En effet, l'Equilibre Utilisateur Statique est la distribution stationnaire des flots sous laquelle aucun utilisateur ne peut augmenter son temps de trajet en changeant son itinéraire. Une forte variabi-

⁴ le code source de l'application et des analyses est disponible sur le dépôt ouvert du projet à
<https://github.com/JusteRaimbault/TransportationEquilibrium>

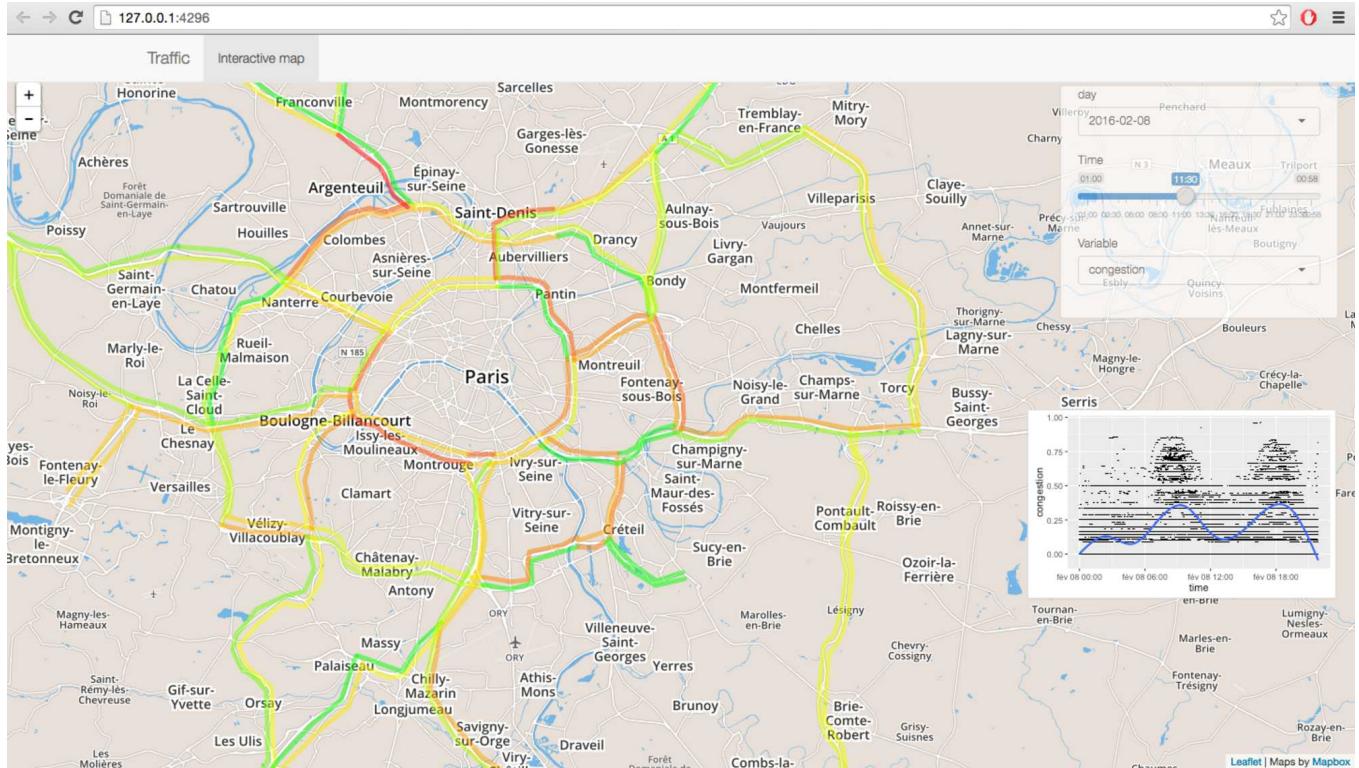


FIGURE 58 : Capture de l'application web. Nous avons développé celle-ci pour permettre l'exploration spatio-temporelle des données de trafic pour la région Parisienne. Il est possible de choisir date et heure (précision de 15min sur un mois, réduite par rapport au jeu de données initial pour des raisons de performance). Le graphe en insert résume les motifs de congestion pour la journée courante, en donnant en fonction du temps l'ensemble des valeurs (points noirs) et leur lissage (courbe bleue).

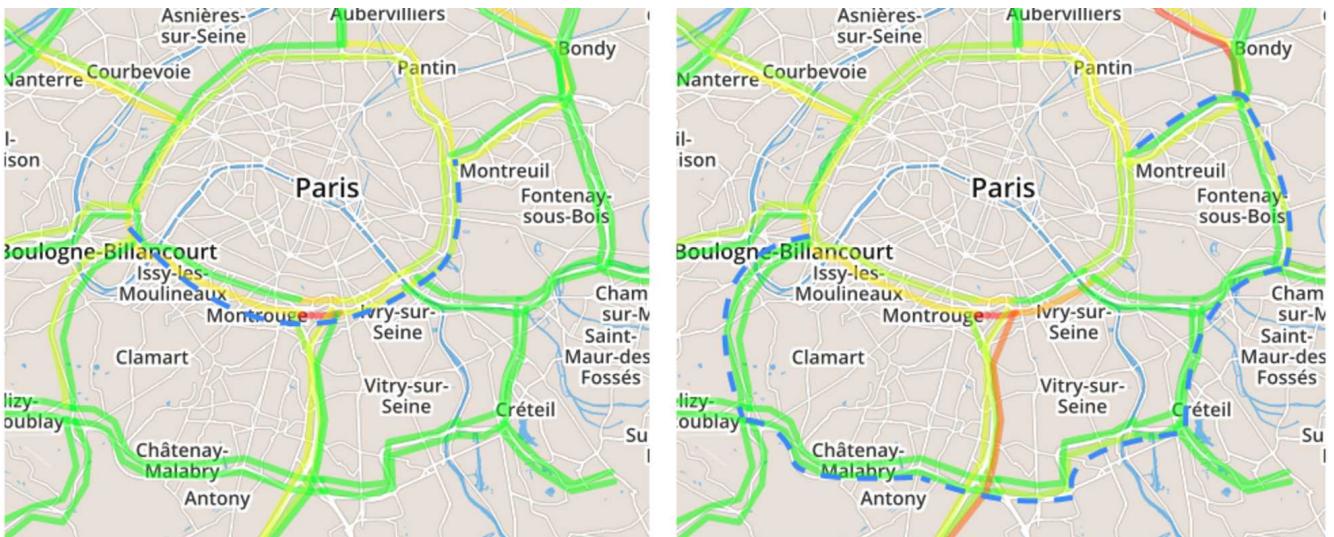


FIGURE 59 : Variabilité spatiale d'un plus court chemin en temps de trajet (trajet du plus court chemin en pointillé bleu). Dans un intervalle de seulement 10 minutes, entre le 11/02/2016 00 :06 (à gauche) et le 11/02/2016 00 :16 (à droite), le plus court chemin entre Porte d'Auteuil à l'ouest et Porte de Bagnolet à l'est, augmente en distance effective de $\simeq 37\text{km}$ (avec une augmentation du temps de trajet de seulement 6 minutes), à cause d'une forte perturbation sur le périphérique parisien.

lité spatiale des plus courts chemins sur de courtes échelles spatiales révèle ainsi une non-stationnarité, puisque un même utilisateur prendra un chemin complètement différent après un court laps de temps et ne contribuera plus au même flot que précédemment. Une telle variabilité est en effet observée sur un nombre non-négligeable de chemins pour chaque jour du jeu de données. La figure ?? montre un exemple de variation spatiale extrême d'un trajet pour une paire Origine-Destination particulière.

L'exploration systématique de la variabilité du temps de trajet sur l'ensemble du jeu de données, et des distances de trajet associées, confirme, comme présenté en figure , que la variation absolue du temps de trajet présente fréquemment une forte variation de son maximum sur l'ensemble des paires O-D, jusqu'à 25 minutes avec une moyenne temporelle locale autour de 10 minutes. La variabilité spatiale correspondante entraîne des détours allant jusqu'à 35km.

STABILITÉ DES MESURES DE RÉSEAU La variabilité des trajectoires potentielles observée dans la section précédente peu être confirmée par l'étude de la variabilité des propriétés du réseau. En particulier, les mesures topologiques de réseau capturent les motifs globaux dans un réseau de transport. Les mesures de centralité et de connectivité des noeuds sont des indicateurs classiques pour la description des réseaux de transport comme rappelé par [bavoux2005geographie]. La littérature en transports a développé des mesures de réseau élaborées et opérationnelles, comme des mesures de robustesse pour identifier

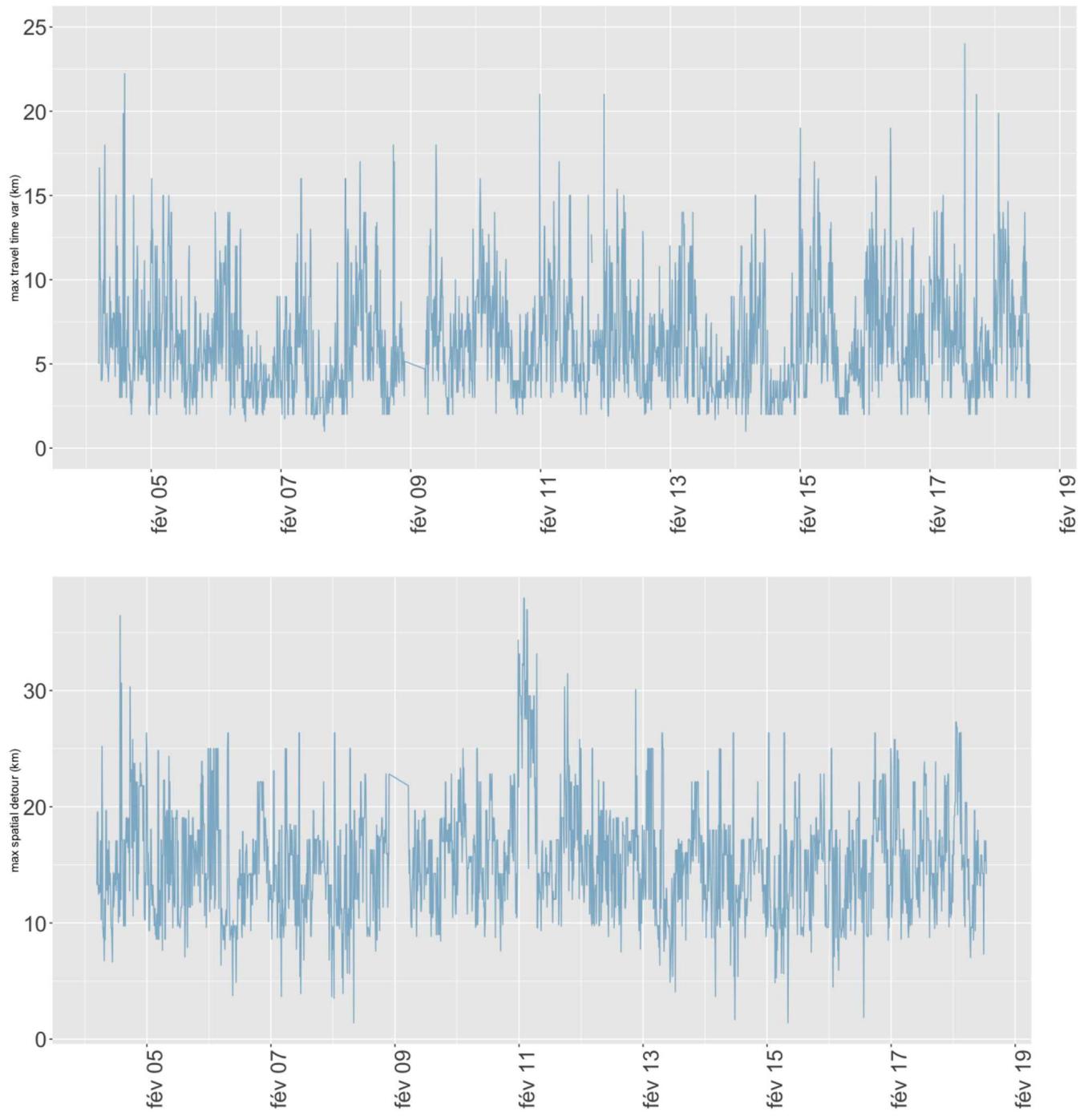


FIGURE 60 : Variabilité maximale du temps de trajet (en haut) en minutes et de la distance de trajet correspondante (en bas) pour un échantillon de deux semaines. Le graphe représente le maximum sur l'ensemble des paires Origine-Destination de la variabilité absolue entre deux pas de temps consécutifs. Les heures de pointe induisent une forte variabilité du temps de trajet, allant jusqu'à 25 minutes et une variabilité de distance jusqu'à 35km.

les liens critiques et mesurer la résilience globale du réseau aux perturbations (un exemple parmi d'autres est l'indice de *Robustesse du Réseau Effective* introduit dans [sullivan2010identifying]).

Plus précisément, nous étudions la centralité de chemin du réseau de transport, défini pour un noeud comme le nombre de plus courts chemins passant par celui-ci, i.e. par l'équation

$$b_i = \frac{1}{N(N-1)} \cdot \sum_{o \neq d \in V} \mathbb{1}_{i \in p(o \rightarrow d)} \quad (13)$$

où V est l'ensemble des sommets du réseau de taille N , et $p(o \rightarrow d)$ est l'ensemble des noeuds sur le plus court chemin entre les sommets o et d (le plus court chemin étant calculé avec le temps de trajet effectif). Cette mesure de centralité est plus adaptée que d'autre dans notre cas, comme la centralité de proximité qui n'inclut pas la congestion potentielle comme la centralité de chemin.

Nous montrons en Fig. 61 la variation relative absolue du maximum de la centralité de chemin, pour la même fenêtre temporelle que les indicateurs empiriques précédents. Plus précisément, elle est définie par

$$\Delta b(t) = \frac{|\max_i(b_i(t + \Delta t)) - \max_i(b_i(t))|}{\max_i(b_i(t))} \quad (14)$$

où Δt est le pas de temps du jeu de données (la plus petite fenêtre temporelle sur laquelle une variabilité peut être capturée). Cette variation relative absolue a une signification directe : une variation de 20% (qui est atteinte un nombre significatif de fois comme montré en Figure 61) implique dans le cas d'une variation négative, qu'au moins cette proportion de trajectoires potentielles ont changé et que la potentielle congestion locale a décrue de la même proportion. Dans le cas d'une variation positive, un seul noeud a capturé au moins 20% des trajets. Sous l'hypothèse (qu'on ne tente pas de vérifier ici et qu'on peut également supposer non vérifiée comme montré par [zhu2010people]), mais que l'on utilise comme un outil pour donner une intuition sur la signification concrète de la variabilité de la centralité) que les utilisateurs choisissent rationnellement le plus court chemin, et supposant que la majorité des trajets est réalisées, une telle variation de la centralité implique une variation similaire dans les flots effectifs, conduisant à la conclusion qu'ils ne peuvent être stationnaires ni dans le temps (au moins sur une échelle plus grande que Δt) ni dans l'espace.

HÉTÉROGÉNÉITÉ SPATIALE DE L'ÉQUILIBRE Afin d'obtenir un point de vue différent sur la variabilité spatiale des motifs de congestion, nous proposons d'utiliser un indice d'auto-corrélation spatiale, l'indice de Moran (défini par exemple dans [tsai2005quantifying]).

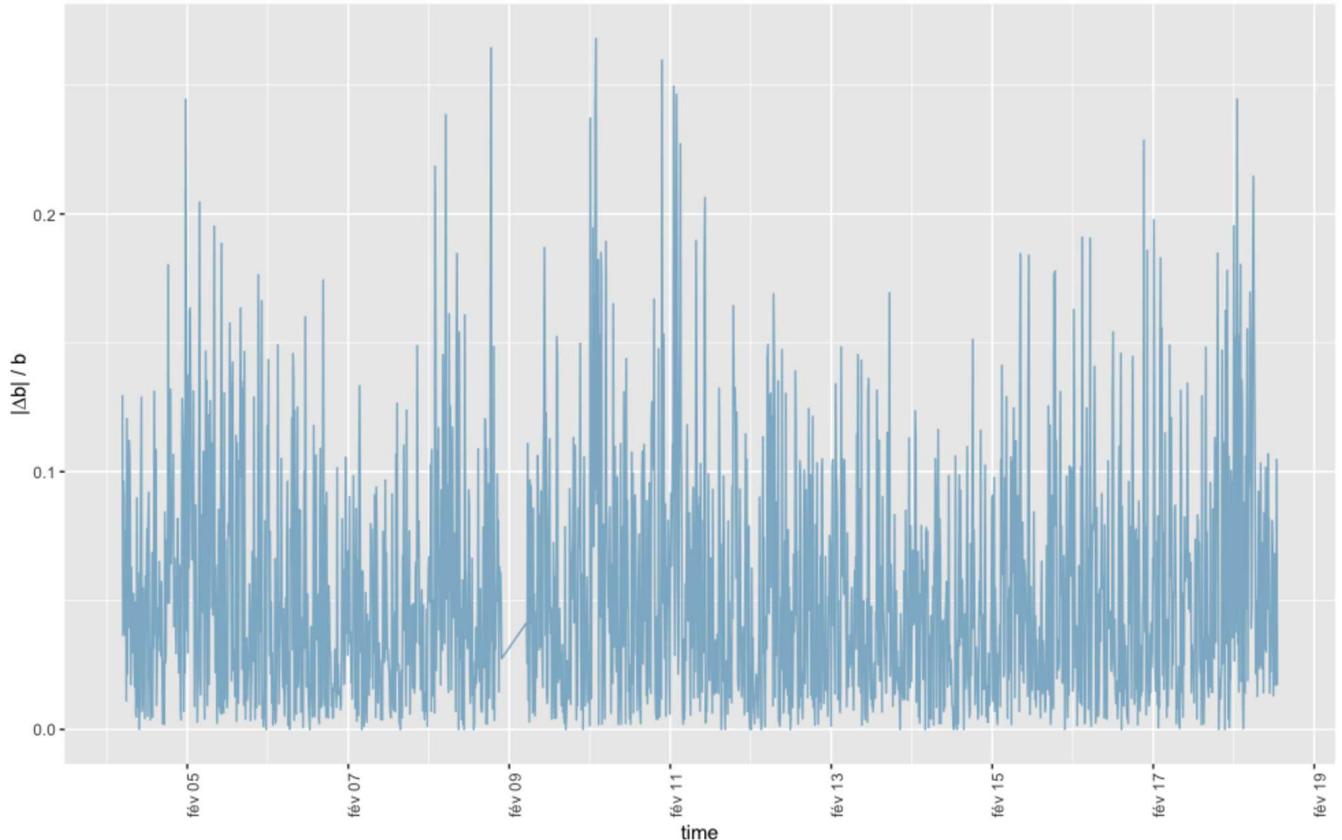


FIGURE 61 : Stabilité temporelle du maximum de la centralité de chemin. Le graphe montre dans le temps la dérivée normalisée du maximum de la centralité de chemin, qui capture ses variations relatives à chaque pas de temps. La valeur maximale de 25% correspond à de très fortes perturbations du réseau sur les liens correspondants, puisque cela implique qu'au moins cette proportion d'utilisateurs prenant le lien dans des conditions précédentes doivent prendre un trajet complètement différent.

Utilisé plus généralement en analyse spatiale, avec diverses applications allant de l'étude de la forme urbaine à la quantification de la ségrégation, il peut être appliqué à toute variable spatiale. Il permet d'établir des relations de voisinage et révèle la consistance spatiale locale d'un équilibre s'il est appliqué à une variable de traffic localisée. A un point donnée de l'espace, l'auto-corrélation locale pour la variable c est calculée par

$$\rho_i = \frac{1}{K} \cdot \sum_{i \neq j} w_{ij} \cdot (c_i - \bar{c})(c_j - \bar{c}) \quad (15)$$

où K est une constante de normalisation égale à la somme des poids spatiaux fois la variance de la variable et \bar{c} est la moyenne de la variable. Dans notre cas, nous choisissons des poids spatiaux de la forme $w_{ij} = \exp\left(\frac{-d_{ij}}{d_0}\right)$ avec d_0 distance typique de décroissance. L'auto-corrélation est calculée sur la congestion des liens, localisée au centre du lien. Elle capture ainsi les corrélations spatiales dans un rayon du même ordre que la distance de décroissance autour du point i . La moyenne sur l'ensemble des points fournit l'indice d'auto-corrélation spatiale I . Une stationnarité des flots devrait impliquer une stabilité temporelle de l'index.

La figure 62 présente l'évolution temporelle de l'auto-corrélation spatiale pour la congestion. Comme attendu, on observe une forte décroissance de l'auto-corrélation avec la distance de décroissance, à la fois sur l'amplitude et les moyennes temporelles. La forte variabilité temporelle implique de courtes échelles temporelles pour des fenêtres potentielles de stationnarité. Pour une distance de décroissance de 1km, en comparant l'auto-corrélation à la congestion (ajustée à l'échelle du graphe pour lisibilité), on observe que les fortes corrélations coïncident avec les heures creuses, tandis que les heures de pointe correspondent à une décroissance des corrélations. Notre interprétation, combinée avec la variabilité observée des motifs spatiaux, est que les heures de pointe correspondent à un comportement chaotique du système, puisque les bouchons peuvent émerger dans n'importe quel lien du réseau : la corrélation disparaît alors puisque l'espace des phases atteignables pour un système dynamique chaotique est rempli uniformément par les trajectoires, de façon équivalente à des vitesses relatives qui apparaîtraient comme aléatoires et indépendantes.

8.2.3 Discussion

Implications théoriques et pratiques des conclusions empiriques

Nous formulons l'interprétation que les implications théoriques de ces résultats empiriques n'impliquent pas nécessairement un rejet

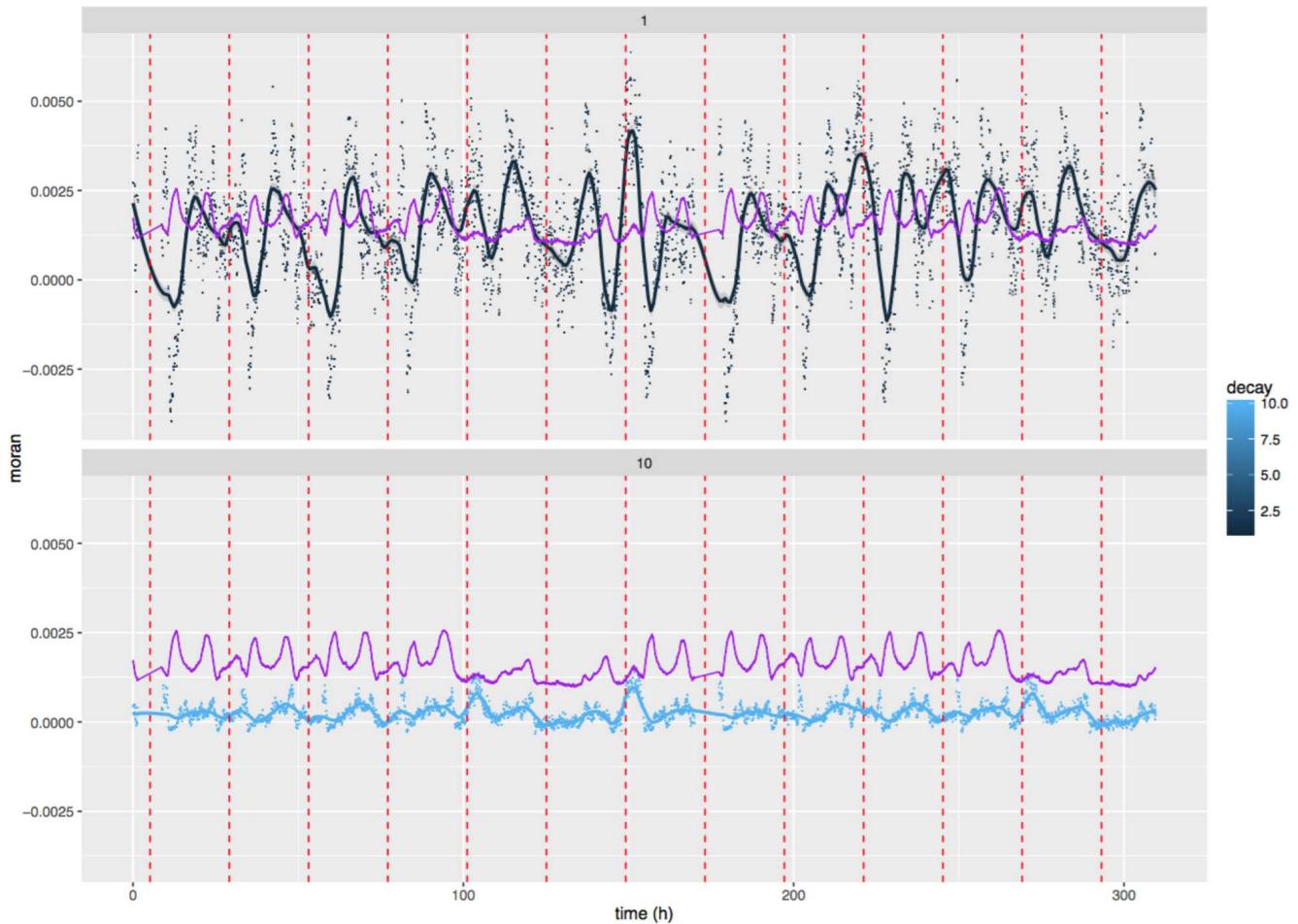


FIGURE 62 : Auto-corrélations spatiales pour les vitesses relatives sur deux semaines. Le graphe montre les valeurs de l'auto-corrélation dans le temps, pour des valeurs variables (1,10km) de la distance de décroissance. les valeurs intermédiaires de la distance de décroissance donnent une déformation relativement continue entre ces deux extrêmes. Les points sont lissés sur une fenêtre temporelle de 2h pour faciliter la lecture. Les lignes pointillées verticales correspondent à minuit de chaque jour. La courbe violette donne la vitesse relative, ajustée à l'échelle pour établir la correspondance entre les heures de pointe et les variations de l'auto-corrélation.

total du cadre de l'Equilibre Utilisateur Statique, mais révèlent plutôt un besoin de plus fortes connexions entre la littérature théorique et les études empiriques. Si chaque nouveau cadre théorique introduit est généralement testé sur un cas ou plus, il n'existe pas de comparaisons systématiques de chacun sur des jeux de données de grande taille et variés, et pour des objectifs d'application différents (prédition du traffic, reproduction de faits stylisés, etc.), à l'image des revues systématiques qui sont la règle en évaluation thérapeutique par exemple [bastian201oseventy]. Cela implique cependant des pratiques de partage des données et des modèles plus larges que celles existant couramment. La connaissance précise des potentialités d'application d'un cadre donné peut induire des développements inattendus comme l'intégration dans des modèles plus larges.

L'exemple des études des interaction entre Transport et Usage du Sol (modèles *LUTI*) est une bonne illustration d'un cas où le EUS peut toujours être utilisé avec des motivations plus larges que la modélisation du traffic. [kryvobokov2013comparison] décrit deux modèles *LUTI*, dont l'un inclut deux équilibres pour les modèles de transport à quatre temps et pour l'évolution de l'usage du sol (localisation des ménages et emplois), l'autre étant dynamique. La conclusion est que chaque modèle a ses avantages au regard de l'objectif poursuivi, et que le modèle statique peut être utilisé pour comparer des politiques sur le temps long, tandis que le modèle dynamique fournit de l'information plus précise à de plus petites échelles temporelles. Dans le premier cas, un module de transport plus compliqué aurait été plus difficile à inclure, ce qui est un avantage du EUS dans ce cas.

Concernant les applications pratiques, il semble naturel que les modèles statiques ne devraient pas être utilisés pour la prédition et la gestion du traffic sur de petites échelles temporelles (semaine ou jour) et que des efforts doivent être faits pour implémenter des modèles plus réalistes. Cependant, l'utilisation des modèles par la communautés des ingénieurs et des planificateurs n'est pas directement reliée aux enjeux académiques et à l'état de l'art dans le domaine. Dans le cas particulier de la France et des modèles de mobilité, [commenges2013invention] a montré que les ingénieurs allaient jusqu'au point de construire des problèmes inexistant et d'implémenter les modèles correspondants qu'ils avaient importé d'un contexte géographique totalement différent (la planification aux Etats-Unis). L'utilisation d'un cadre ou d'un type de modèle a des raisons historiques qui peuvent être difficiles à surmonter.

C (FL) : je ne pense pas que tu as prouvé cela : tu ne dis pas comment sont utilisées ces modèles

Vers des interprétations de la non-stationnarité

Une hypothèse qu'on peut formuler concernant l'origine de la non-stationnarité des flots dans le réseau, au regard de l'exploration des données et des analyses quantitatives, est que le réseau est au moins la moitié du temps fortement congestionné et dans un état critique.

Les heures creuses sont les plus grandes fenêtres temporelles potentielles de stationnarité spatiale et temporelle, mais couvre moins de la moitié du temps. Comme déjà interprété dans le comportement de l'indicateur d'auto-corrélation, un comportement chaotique pourrait être à l'origine d'une telle variabilité lors des heures congestionnées. A la manière d'un fluide supercritique qui condense sous une perturbation externe infinitésimale, l'état d'un lien peut qualitativement changer par un petit incident, produisant une perturbation du réseau qui se propage et peut même s'amplifier. L'effet direct des événements du traffic (incidents signalés ou accidents) ne peut pas être étudié sans source de données extérieure, et un enrichissement de la base de données dans cette direction pourrait être intéressante. Cela permettrait d'établir la proportion de perturbations qui paraissent avoir un effet direct et quantifier un niveau de caractère critique de la congestion du réseau dans le temps, ou d'étudier plus précisément des phénomènes localisés comme les conséquences d'un incident de traffic sur la voie opposée.

Développements

Le travail futur pourra être planifié dans la direction d'une étude de la stabilité temporelle sur des zones du réseau, i.e. l'étude quantitative précise de la non-stationnarité des heures de pointes découverte ci-dessus. Pour cela nous proposons de calculer numériquement la stabilité de Liapounov du système dynamique régissant les flots de trafic, par l'intermédiaire d'algorithmes numériques comme ceux décrits par [goldhirsch1987stability]. La valeur des exposants de Liapounov fournit l'échelle de temps sur laquelle le système instable s'éloigne de l'équilibre. Leur comparaison avec la durée des heures de pointe et le temps de trajet moyen, sur différentes zones spatiales et différentes échelles, devrait fournir plus d'information sur une possible validité de l'hypothèse de stationnarité locale. Cette technique a déjà été introduite à une autre échelle dans les études de transport, comme e.g. [tordeux2016jam] qui étudie la stabilité des modèles de régulation de vitesse à l'échelle microscopique pour éviter l'émergence de congestion.

D'autres directions de recherche peuvent consister en le test des autres hypothèses du EUS (comme le choix rationnel du plus court chemin, qui serait cependant difficile à tester à un tel niveau d'agrégation, impliquant l'utilisation de modèles de simulation calibrés et cross-validés sur le jeu de données pour comparer différentes hypothèses, sans toutefois nécessairement une validation ou invalidation directe de l'hypothèse), ou le calcul empirique des paramètres dans les cadres d'Equilibre Utilisateur Stochastique ou Dynamique.

Conclusion

Nous avons décrit une étude empirique ayant pour but une étude simple, mais selon notre point de vue nécessaire, de l'existence de l'équilibre utilisateur statique, plus précisément de sa stationnarité dans le temps et l'espace pour un réseau routier métropolitain principal. Un jeu de données de congestion du trafic est construite par collection de données, pour le réseau du Grand Paris sur 3 mois avec une granularité temporelle de 2 minutes. L'exploration interactive du jeu de données via une application web permettant la visualisation spatio-temporelle aide à guider les analyses quantitatives. La variabilité spatio-temporelle des plus courts chemins et de la topologie du réseau, en particulier la centralité de chemin, révèle que l'hypothèse de stationnarité ne tient généralement pas, ce qui est confirmé par l'étude de l'auto-corrélation spatiale de la congestion du réseau. Nous suggérons que nos résultats soulignent un besoin général de plus grandes connexions entre les études théoriques et empiriques, puisque cette étude permet de chasser les incompréhensions théoriques sur l'Equilibre Utilisateur Statique, et guider le choix d'application potentielles.

★ ★

★

8.3 TRANSPORT ROUTIER ET DÉTERMINANTS DES COÛTS

C (FL) : quel rapport avec la problématique ?

C (JR) :
[orfeuil2012grand] p307,
fait le lien entre prix
essence et crise immo-
bilier : marqueur des
interactions entre trans-
port et urbanisme

La géographie des prix du carburant a de nombreuses applications variées, de son impact significatif sur l'accessibilité à son rôle comme indicateur d'équité territoriale et de politique de transports. Dans cette section, nous étudions les variations spatio-temporelles des prix du carburant aux Etats-Unis à une résolution très fine, par l'utilisation d'un nouveau jeu de données, donnant les prix journaliers sur deux mois pour une proportion significative des stations essence. Les données ont été collectées par l'intermédiaire d'une technologie de crawling à grande échelle élaborée spécifiquement, que l'on décrira. Nous étudions l'influence de variables socio-économiques, en utilisant des méthodes complémentaires : la Régression Géographique Pondérée pour tenir compte de la non-stationnarité spatiale, et une modélisation économétrique linéaire pour conditionner à l'Etat et tester des caractéristiques au niveau du Comté. La première fournit une portée spatiale optimale qui correspond globalement à l'échelle de stationnarité, et une influence significative des variables comme le revenu moyen ou le salaire par travail, avec un comportement spatial dont la non simplicité confirme l'importance des particularités géographiques. D'autre part, la modélisation multi-niveaux révèle un très fort effet Etat, alors que les caractéristiques spécifiques au Comté gardent un impact significatif. A travers la combinaison de ces méthodes, nous démontrons la superposition d'un processus de gouvernance avec un processus spatial socio-économique local. Nous discutons une application potentielle importante qui est l'élaboration de politiques de régulation automobiles localement paramétrées.

8.3.1 Contexte

Quels sont les déterminants des prix du carburant ? Par l'utilisation d'une nouvelle base de données des prix des carburant au niveau de la station, collectée pendant deux mois, nous explorons leur variabilité dans le temps et l'espace. Une variation du coût du carburant peut avoir de nombreuses causes, du prix brut du pétrole au politiques fiscales locales et au caractéristiques géographiques, chacun ayant des effets hétérogènes dans l'espace et le temps. Bien que l'évolution du prix moyen du carburant dans le temps soit un indicateur suivi avec attention et analysé par de nombreuses institutions financières, sa variabilité dans l'espace reste relativement non-explorée dans la littérature. Cependant, de telles différences peuvent refléter des variations dans des indicateurs socio-économiques plus indirects comme des inégalités territoriales, des singularités géographiques ou des préférences des consommateurs.

Il n'existe à notre connaissance pas de cartographie systématique dans le temps et l'espace des prix de vente à l'échelle d'un pays. La raison principale est probablement que la disponibilité des données a pu être un obstacle important. Il est aussi probable que la nature de la question joue un rôle, puisque celle-ci se trouve à l'interface de plusieurs disciplines. Alors que les économistes étudient l'élasticité des prix et leur mesure dans différents marchés, la géographie des transports, par des méthodes comme les prix des transports intégrés aux modèles spatiaux, met une emphase plus grande sur la distribution spatiale que sur des mécanismes précis de marché. Toutefois, des exemples de travaux relativement liés peuvent être trouvés. Par exemple, [rietveld2001spatial] étudie l'impact de différences de prix transfrontalières et leur implications pour une taxation spatiale graduelle aux Pays-Bas. A l'échelle du pays, [rietveld2005fuel] fournit des modèles statistiques pour expliquer les variabilités des prix entre les pays Européens. [macharis2010decision] modélise l'impact d'une variation spatiale des prix sur les motifs d'intermodalité, ce qui implique que l'hétérogénéité spatiale des prix du carburant a un impact fort sur le comportement des utilisateurs. Avec une approche similaire par la géographie des transports, [gregg2009temporal] étudie la distribution spatiale des émissions à l'échelle des Etats américains. La géographie des prix du carburant a également d'importantes répercussions sur les coûts effectifs, comme le montre [combes2005transport] en déterminant les coûts réels de transport pour les différentes aires urbaines françaises. De façon plus proche de notre travail, et en utilisant des données similaires en Accès Ouvert pour la France, [gautier2015dynamics] étudie les dynamiques de transmission des prix bruts du pétrole aux prix de vente. Toutefois, ils n'introduisent pas de modèle spatial explicite de diffusion des prix et n'étudient pas de dynamiques spatio-temporelles.

Dans cette section nous adoptons une approche différente en procédant à une analyse spatiale exploratoire des prix du carburant aux Etats-Unis. Nous montrons que la majorité des variations s'observent entre les Comtés et non dans le temps, malgré les évolutions du baril brut pendant la période considérée. Nous employons pour cela une analyse spatiale de la distribution des prix. Les résultats majeurs obtenus sont les suivants : d'une part nous montrons l'existence de motifs spatiaux significatifs dans des grandes régions US, d'autre part nous montrons que même si la majorité des variations observées par les politiques des Etats, et en particulier le niveau de taxation, certaines caractéristiques à l'échelle du Comté restent significatives.

Données

Notre jeu de données contient l'information journalière des prix des carburants à l'échelle de la station essence pour l'ensemble du territoire US métropolitain. Ces informations sont construites à partir des

prix reportés par les utilisateurs et couvre pratiquement l'ensemble des stations essence aux Etats-Unis. Nous commençons par décrire la collection des données et donnons des statistiques de ce jeu de données nouveau.

Collection de données hétérogènes à grande échelle

La disponibilité de nouveaux types de données a conduit à des évolutions significatives dans de nombreuses disciplines (e.g. l'analyse des réseaux sociaux en ligne ([tan2013social])) à la géographie (e.g. les nouvelles approches de la mobilité urbaine ou les perspectives de ville plus "intelligentes" ([batty2013big])) en incluant l'économie pour laquelle la disponibilité de données exhaustives à l'échelle individuelle ou de l'entreprise est vu comme une révolution dans le champ. La plupart des études impliquant ces nouvelles données sont à l'interface des disciplines concernées, ce qui est à la fois un avantage mais aussi une source de complications. Par exemple les malentendus entre physique et sciences urbaines décrites par [dupuy2015sciences] sont en particulier causées par des attitudes différentes au regard des données non conventionnelles ou des interprétations et ontologies différentes pour celles-ci. La collection et l'utilisation des nouvelles données est donc devenu un enjeu essentiel en sciences sociales. La construction des tels jeux de données est cependant loin d'être évidente de par la nature incomplète et bruitée de la donnée. Des outils techniques spécifiques doivent être implémentés mais sont souvent conçus pour surmonter un problème donné et sont difficiles à généraliser. Nous développons un tel outil qui remplit les contraintes suivantes typiques de la collection de données à grande échelle : (i) un niveau raisonnable de flexibilité et de généralité ; (ii) une performance optimisée par la collection parallélisée ; (iii) l'anonymat des jobs de collection pour éviter le plus possible tout biais dans le comportement de la source de données. L'architecture, à un assez haut niveau, a la structure suivante :

- Un ensemble indépendant des tâches fait tourner en continu des proxies socks pour envoyer les requêtes via tor.
- Un manager suit les tâches de collection en cours, répartit la collection entre les sous-tâches et en lance des nouvelles lorsque cela s'avère nécessaire.
- Les sous-tâches peuvent être toute application prenant comme argument les adresses de destination, elles procèdent à la collecte, au parsing et au stockage des données collectées.

L'application est ouverte et ses modules sont réutilisables : le code source est disponible sur le dépôt du projet.⁵ Nous avons construit

⁵ à <https://github.com/JusteRaimbault/EnergyPrice>

TABLE 17 : Statistiques descriptives des prix des carburants (\$ par gallon)

Moyenne	Dev. Std.	p10	p25	p50	p75	p90
2.28	0.27	2.02	2.09	2.21	2.39	2.65

notre jeu de données en utilisant l'outil en continu pendant deux mois pour collecter des données crowdsourcées disponibles de diverses sources en ligne.

Jeu de données

Le jeu de données contient autour de $41 \cdot 10^6$ observations uniques des prix de vente au niveau de la station, s'étendant sur une période du 10 janvier 2017 au 19 mars 2017, correspondant à 118,573 station service uniques. Pour chacune, nous disposons d'une localisation géographique précise (résolution à la ville). En moyenne nous avons 377 informations de prix par station. Les prix correspondent à un mode d'achat unique (par carte de crédit, les autres modes comme l'argent liquide représentant moins de 10% sur des jeux tests, ils ont été abandonnés dans le jeu de données final) et quatre types de carburant possibles : Diesel (18% des observations), Regular (34%), Midgrade (24%) et Premium (24%). La meilleure couverture des stations est pour le carburant Regular avec en moyenne 4,629 données de prix par Conté. Nous choisissons pour cette raison de concentrer l'étude sur ce type de carburant, en gardant à l'esprit que des développements futurs avec le jeu de données pourraient inclure des analyses comparatives des types de carburant. Notre jeu de données final contient ainsi 14,192,352 observations provenant de 117,155 stations service, suivies pendant 68 jours. Nous agrégeons de plus les données par jour, en prenant la moyenne du prix observé par gallon, pour obtenir un panel de 5,204,398 observations station-jour.⁶ La table ?? donne des statistiques descriptives basiques sur les données de prix, montrant que la distribution des prix est fortement concentrée avec une faible skewness (le ratio du 99th au 1st quantiles est 1.6). Enfin, dans l'analyse spatiale, nous utiliserons également des données socio-économiques au niveau du Conté, disponible par le US Census Bureau. Nous utiliserons les plus récentes disponibles (ce qui dans la plupart des cas implique d'utiliser le Census de 2010).

⁶ Le panel n'est pas équilibré puisque les prix ne sont pas reportés chaque jour pour chaque station. Une station moyenne possède l'information de prix pour 44 jours (sur 68).

8.3.2 Résultats

Motifs spatio-temporels des prix

Avant de se consacrer à une étude plus systématique de la variation des prix des carburants, nous proposons une première introduction exploratoire pour donner une idée de sa structure spatio-temporelle. Cette exercice est une étape cruciale pour guider les analyses suivantes, mais aussi pour comprendre leurs implications dans le contexte géographique. Afin d'explorer les données, nous construisons une application web basique permettant de cartographier les données dans l'espace et le temps. Elle est disponible à . Nous montrons également de carte au niveau du Conté à la figure ?? pour le prix moyen sur l'ensemble de la période. On voit clairement apparaître des motifs régionaux, avec les régions du centre sud et du sud est ayant les prix les plus bas et la côte Pacifique et le nord est les prix les plus hauts. Bien évidemment , une carte agrégée sur l'ensemble de la période n'apporte guère d'information sur les variations temporelles des données. Comme nous allons le montrer plus en détails par la suite, la majorité des variations des prix des carburants a lieu dans l'espace. une décomposition de la variance des prix donne seulement 11% de la variance totale expliquée par les variations intra-station. De la même manière, le coefficient de corrélation de rang de Spearman entre le prix des stations pour le carburant regular entre le premier jour du jeu de données et le dernier jour est de 0.867, et l'hypothèse nulle que ces deux informations sont indépendantes est fortement rejetée.

Puisque la majorité de la variation des prix est inter-station, nous nous intéressons maintenant principalement aux corrélations spatiales. Nous conduisons l'analyse à l'échelle du Conté pour diverse raisons. D'une part une décomposition des prix des carburants inter et intra-Conté montre que plus de 85% de la variance est inter-Conté, d'autre part car la localisation des stations n'est pas assez fiable pour permettre une granularité plus fine, et enfin car la majorité des variables socio-économiques est à ce niveau. Nous étudions donc l'autocorrelation spatiale des prix à l'échelle du Conté. L'autocorrelation spatiale peut être vue comme une indicateur d'hétérogénéité spatiale que nous mesurons par l'index de Moran ([tsai2005quantifying]), avec des poids spatiaux de la forme $\exp(-d_{ij}/d_0)$ avec d_{ij} étant la distance entre les entités spatiales i et j, et d_0 un paramètre de décroissance donnant la portée spatiale des interactions que l'estimation prend en compte. Nous montrons en Fig. ?? ses variation pour chaque jour ainsi que comme fonction du paramètre de decay. Les fluctuations dans le temps de l'index de Moran journalier pour les valeurs basses et moyennes du paramètre de decay, confirme les spécificité géographiques au sens de régimes de corrélation changeant localement. Celles-ci sont logiquement atténuées pour les longues portées,

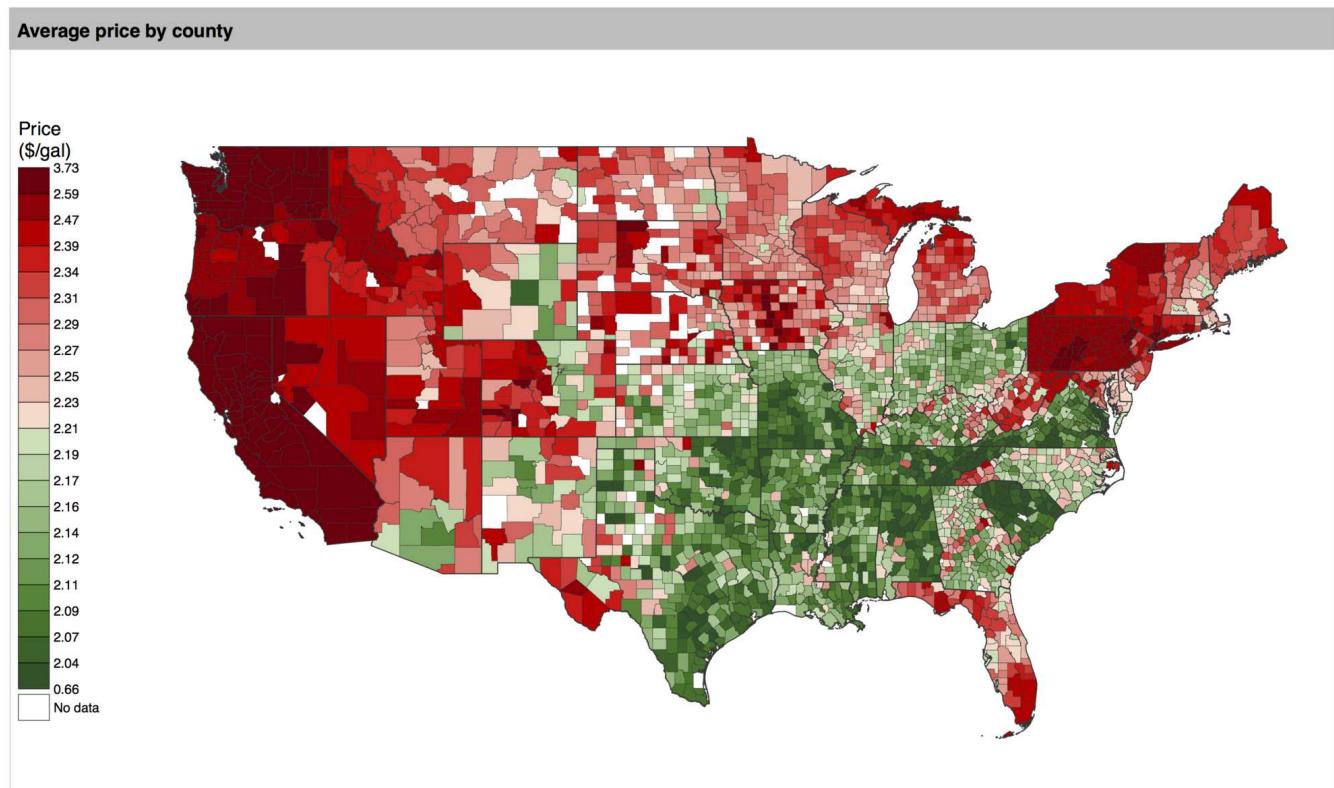


FIGURE 63 : Carte du prix moyen par comté. Le prix est donné pour du carburant régulier, et la moyenne temporelle est prise sur l'ensemble de la période.

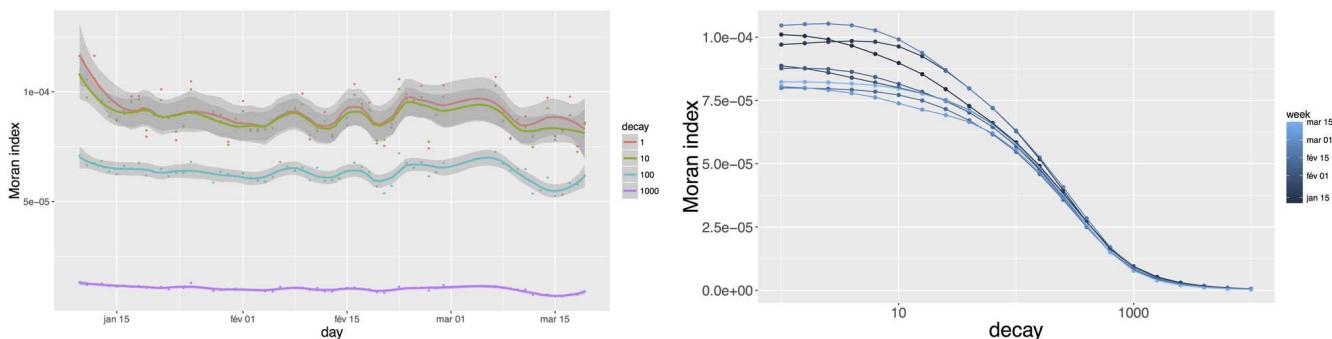


FIGURE 64 : Comportement de l'index d'autocorrelation spatiale de Moran. (Gauche) Evolution dans le temps de l'index de Moran, calculé sur des fenêtres journalières, pour différentes valeurs du paramètre de décroissance. (Droite) Index de Moran en fonction du paramètre de décroissance, calculé sur des fenêtres hebdomadaires.

puisque les corrélations des prix diminuent avec la distance. Le comportement de l'autocorrelation spatiale en fonction du paramètre de decay est particulièrement intéressant : nous observons une premier changement de régime autour de 10km (d'un régime constant à un régime linéaire par morceau), et une seconde transition importante autour de 1000km, les deux constants sur des fenêtres temporelles à la semaine. Nous postulons que celles-ci correspondent au échelles spatiales typiques des phénomènes observés : le régime bas serait les spécificités locales et l'intermédiaire le processus au niveau de l'Etat. Ce comportement confirme que les pris sont non-stationnaires dans l'espace, et que pour cette raison des techniques statistiques appropriées doivent être utilisées pour étudier les variables jouant un rôle à différents niveaux. Les deux parties suivantes suivent cette idée et étudient des variables explicatives potentielles des prix locaux du carburant, utilisant deux techniques différentes qui correspondent à deux paradigmes complémentaires : la régression géographique pondérée qui met l'emphase sur les effets de voisinage, et des régressions multi-niveaux prenant en compte les limites administratives.

Régression Géographique Pondérée

La question de la non-stationnarité des processus géographiques a toujours été une source d'analyses agrégées biaisées ou de mauvaises interprétations lorsque des conclusions générales sont appliquées à des cas locaux. Pour le prendre en compte dans les modèles statistiques, de nombreuses techniques ont été proposées, parmi lesquelles la simple mais très élégante Régression Géographique Pondérée (GWR), qui estime des régressions non-stationnaires en pondérant les observations dans l'espace de manière similaire aux techniques d'estimation de densité par noyaux. Elle a été introduite dans un article séminal par [brunsdon1996geographically] et a été utili-

sée et développée en conséquence depuis. L'avantage considérable de cette technique est qu'une portée spatiale optimale au sens de la performance du modèle peut être déduite pour dériver un modèle qui traduit des effets des variables variant dans l'espace, révélant ainsi des effets locaux qui peuvent se produire à différentes échelles spatiales ou à travers les frontières. Nous procédons à un multi-modeling pour trouver le meilleur modèle et le noyau ainsi que la portée spatiale associés. Plus précisément, nous suivons les étapes suivantes : (i) tous les modèles linéaire potentiels à partir des cinq variables candidates sont générés (revenu, population, salaire par emploi, emploi par tête, emplois) ; (ii) pour chaque modèle et chaque forme de noyau candidate (exponentiel, gaussien, bisquare, escalier), nous déterminons la portée optimale au sens à la fois de la cross-validation et du critère d'Information d'Akaike corrigé (AICc) qui quantifie l'information contenue dans le modèle ; (iii) nous ajustons les modèles avec cette portée. Nous choisissons le modèle avec le meilleur AICc, en l'occurrence $\text{price} = \beta \cdot (\text{income}, \text{wage}, \text{percapjobs})$ pour une portée de 22 voisins et un noyau Gaussien,⁷ avec un AICc de 2,900. La différence médiane d'AICc avec l'ensemble des autres modèles est 122. Le coefficient de détermination global est 0.27, ce qui est relativement bon en comparaison du meilleur R-squared de 0.29 (obtenu pour le modèle avec l'ensemble des variables, qui surfe clairement avec un AICc de 3010) ; de plus la dimension effective est inférieure à 5 puisque 90% de la variance est expliquée par les trois premières composantes principales pour les variables normalisées.

Les coefficients et le R-squared local pour le meilleur modèle sont montrés en Fig. ???. La distribution spatiale des résidus (qui n'est pas montrée ici), semble globalement distribuée aléatoirement, ce qui confirme d'une certaine façon la cohérence de l'approche. En effet, si une structure géographique distinguable était trouvée dans les résidus, cela signifierait que le modèle géographique ou les variables considérées ont échoués à traduire la structure spatiale. Nous pouvons à présent proposer une interprétation des structures spatiales obtenues. Tout d'abord, la distribution spatiale de la performance du modèle révèle des régions où ces indicateurs socio-économiques simples expliquent relativement bien les prix, et celles-ci sont localisées sur la côte ouest, la frontière sud, la région nord-est des lacs à la côte est, et une bande de Chicago au sud du Texas. Les coefficients correspondants ont des comportements différents selon les zones, suggérant différents régimes.⁸ Par exemple, l'influence du revenu dans chaque région semble s'inverser quand la distance à la côte augmente (du nord au sud-est dans l'ouest, du sud au nord au Texas, de l'est à l'ouest & l'est), ce qui pourrait témoigner de différentes

C (FL) : r2

⁷ on note que la forme du noyau n'a pas plus d'influence tant que des fonctions décroissant graduellement sont utilisées.

⁸ Nous commentons leur comportement dans les zones où le modèle a une performance minimale, que nous fixons arbitrairement à un R-squared local de 0.5.

spécialisations économiques. Au contraire, le changement de régime pour les salaires montre une rupture notable entre l'ouest (sauf autour de Seattle) et le centre et l'est, qui ne correspond pas directement à des politiques d'Etat locales puisque le Texas est coupé en deux par exemple. De la même façon, les emplois par capita montrent une opposition entre est et ouest, qui pourrait être due par exemple à des différences culturelles. Ces résultats sont toutefois difficiles à interpréter directement, et doivent être compris comme la confirmation que les particularités géographiques importent, puisque les régions diffèrent dans le régime du rôle de chacune des variables socio-économiques simples. Une connaissance plus précise pourrait être obtenue par des études géographiques ciblées incluant des études de terrain qualitatives et des analyses quantitatives, qui sont au delà de la portée de cette étude exploratoire et laissée à une éventuelle recherche future.

Enfin, nous extrayons l'échelle spatiale des processus étudiés, c'est à dire en calculant la distribution de la distance aux plus proches voisins avec la portée optimale. On obtient approximativement une distribution log-normale, de médiane 77km et d'interquartile 30km. Nous interprétons cette échelle comme l'échelle de stationnarité spatiale du processus de prix en relation avec les agents économiques, qui peut également être comprise comme la portée des marchés cohérents de compétition entre les stations service.

Régressions multi-niveaux

Comme notre base initiale permet de regarder au niveau des variables $x_{i,s,c,t}$, le prix du carburant au jour t , dans la station i , dans l'Etat s et dans le Comté c , nous commençons par estimer des régressions à effets fixes en grande dimension, suivant le modèle :

$$x_{i,s,c,t} = \beta_s + \varepsilon_{i,s,c,t} \quad (16)$$

$$x_{i,s,c,t} = \beta_c + \varepsilon_{i,s,c,t} \quad (17)$$

$$x_{i,s,c,t} = \beta_i + \varepsilon_{i,s,c,t} \quad (18)$$

(19)

Où $\varepsilon_{i,s,c,t}$ contient une erreur idiosyncrasique⁹ et un effet fixe jour. Cette première analyse confirme que la majorité de la variance peut être expliquée par un effet fixe Etat et que d'intégrer des niveaux plus fins a un effet négligeable sur la performance du modèle mesurée par le R-squared.

Nous nous tournons à présent vers une analyse différente, visant à capturer les variables explicatives qui rendent compte des variations spatiales du carburant. Nous considérons le modèle linéaire suivant :

$$\log(x_i) = \beta_0 + X_i \beta_1 + \beta_{s(i)} + \varepsilon_i, \quad (20)$$

⁹ C'est à dire étant propre à chaque individu.

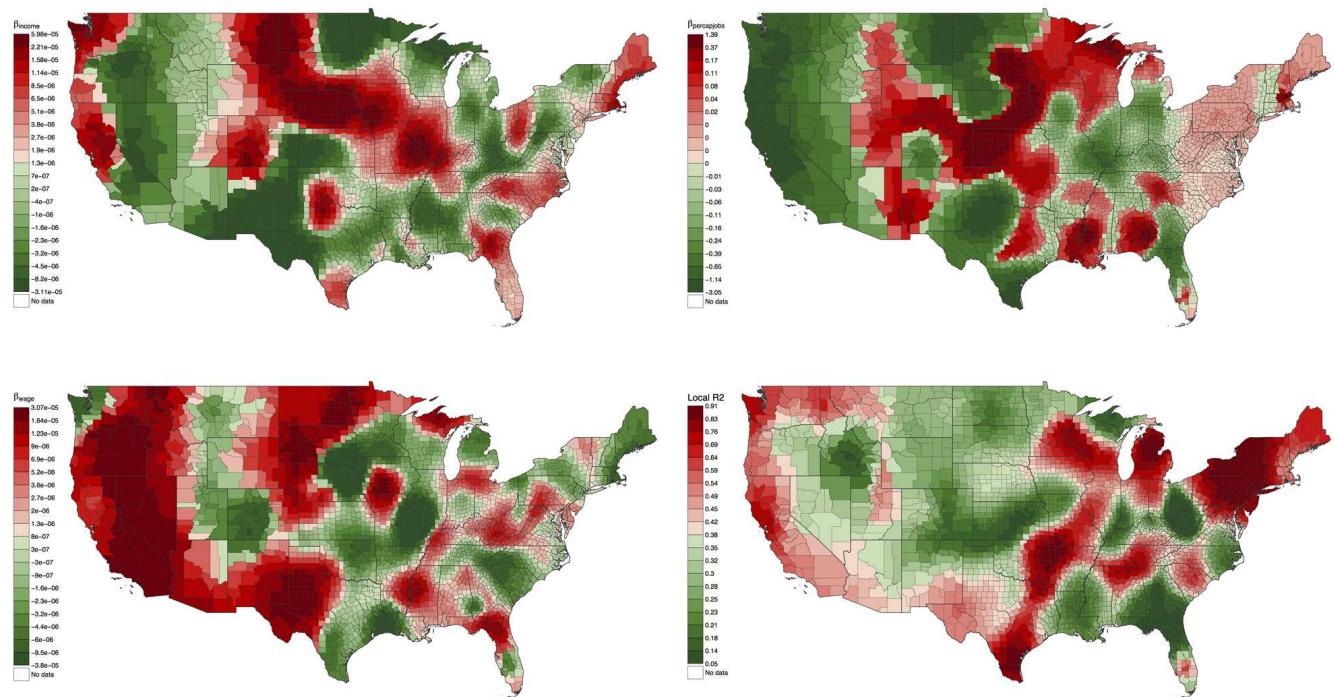


FIGURE 65 : **Résultats des analyses GWR.** Pour le meilleur modèle au sens de l'AICc, les cartes donnent la distribution spatiale des coefficients estimés, dans l'ordre de gauche à droite et de haut en bas, β_{income} , $\beta_{percapjobs}$, β_{wage} , et finalement les valeurs du R² local.

où x_i dénote le prix moyen mesuré du carburant dans le Conté i agrégé sur l'ensemble des jours, X_i est un ensemble de variables spécifiques au Conté et $s(i)$ est l'état dans lequel se trouve le Conté de telle façon que $\beta_{s(i)}$ capture toute la variation spécifique aux Etats. Enfin ε_i est un terme d'erreur satisfaisant $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ si $s(i) \neq s(j)$. ce regroupement de l'erreur standard au niveau de l'état est motivé par les résultats de la partie précédente, montrant que l'autocorrélation spatiale des prix du carburant au niveau de l'état est toujours potentiellement forte. Cette spécification vise à capturer les effets de variables socio-économiques variées au niveau du Conté après que l'effet fixe Etat aie été retiré. Les résultats sont présentés en Table ???. La première colonne montre que la regression du logarithme des prix sur un effet fixe Etat est déjà suffisant pour expliquer 74% de la variance. Cela est majoritairement du aux taxes sur les carburants qui sont fixées au niveau de l'Etat aux Etats-Unis. En fait, une régression du log-prix sur le niveau de taxe donne un R-squared de 0.33%. Les variables explicatives restantes montrent que les Contés urbains denses ont des prix plus élevés, mais que le prix décroît avec la population. Ce résultat paraît raisonnable, les zones désertiques ayant en moyenne des prix plus hauts. Les prix augmentent avec le revenu total, décroissent avec le niveau de pauvreté et décroisse avec le niveau de vote pour un candidat républicain. Ce dernier point suggère un lien circulaire : les Contés qui utilisent beaucoup la voiture auront tendance à voter pour un politicien qui promouvra des politiques favorable à son usage. L'ajout de ces variables explicatives augmente légèrement le R-squared, ce qui suggère que même après avoir enlevé l'effet fixe Etat, la prix du carburant peut être expliqué par des caractéristiques socio-économiques locales.

8.3.3 Discussion

SUR LA COMPLÉMENTARITÉ DES MÉTHODES ÉCONOMÉTRIQUES ET DES MÉTHODES D'ANALYSE SPATIALE Un aspect important de cette contribution est méthodologique. Nous montrons que pour explorer un nouveau panel de données, les géographes et les économistes prennent des approches différentes, menant à des conclusions génériques similaires par des chemins différents. Des études ont déjà combiné les GWR et les régressions multi-niveau ([chen2012using]), ou les ont comparées en terme de performance de modèle ou de robustesse ([lee2009determinants]). Nous prenons ici un point de vue multi-disciplinaire et combinons des approches répondant à des questions différentes, GWR ayant pour but de trouver des variables explicatives précises et de mesurer le rôle de l'auto-corrélation spatiale, tandis que les modèles économétriques expliquent plus précisément les effets des différents facteurs à plusieurs niveaux (Etat, Conté) mais prennent ces caractéristiques géographiques comme exogènes. Nous

TABLE 18 : Régressions au niveau du comté

	(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 : Cette table donne les résultats d'une régression des Moindres Carrés Ordinaire pour le modèle présenté en équation (20). La densité est mesurée comme le nombre d'habitants au mile-carré et le revenu total est donné en dollars. La pauvreté est mesurée comme le nombre de personnes sous le seuil de pauvreté par habitants. On étudie aussi l'influence du pourcentage de personnes noires et de la part de personnes ayant voté pour Donald Trump aux élections de 2016. La régression inclut un effet fixe Etat. Les erreurs standard robustes, agrégées au niveau de l'état, sont données entre parenthèses. ***, ** and * indiquent respectivement les niveaux de significativité 0.01, 0.05 and 0.1.

postulons que les deux sont nécessaires pour comprendre toutes les dimensions du phénomène étudié.

PROPOSITION DE POLITIQUES DE RÉGULATION LOCALISÉES Une autre application de ce type d'analyse est d'aider à une meilleure conception de politiques de régulation de la voiture. Les problèmes environnementaux et de santé requièrent de nos jours un usage raisonnable de celle-ci, dans les villes avec le problème de la pollution atmosphérique, mais aussi globalement pour réduire les émissions de CO₂. [fullerton2002can] montre qu'une taxation des carburants et des voitures peut être équivalente à une taxation des émissions. [brand2013accelerating] souligne le rôle des incitations pour une transition vers des transports décarbonés. Cependant, de telles mesures ne peuvent pas être uniformes d'un Etat à l'autre ou même entre les Contés, pour des raisons évidentes d'équité territoriale : des zones avec des caractéristiques socio-économiques différentes ou avec différentes aménités doivent contribuer selon leur possibilité et préférences. La connaissance des dynamiques locales des prix et leur déterminants, ce en quoi notre étude est une étape préliminaire, peut être une voie vers des régulations localisées prenant en compte la configuration socio-économique et inclure un critère d'équité.

Conclusion

Nous avons décrit une première étude exploratoire des prix des carburants aux US dans le temps et l'espace, utilisant une nouvelle base de données au niveau de la station s'étendant sur deux mois. Notre premier résultat est de montrer la grande hétérogénéité spatiale des processus de prix, par une exploration interactive des données et des analyses d'auto-corrélation. Nous procédons à deux études complémentaires des déterminants potentiels : GWR révèle des structures spatiales et des particularités géographiques, and fournit une échelle caractéristique des processus autour de 75km ; les régressions multi-niveaux montrent que même si la majorité des variations sont expliquées par les caractéristiques des Etats, et majoritairement par le niveau de taxation fixé par l'Etat, il existe toujours des spécificités socio-économiques au niveau du Conté qui peuvent expliquer la variation spatiale des prix du carburant.

* * *

*

CONCLUSION DU CHAPITRE

Cette collection d'études empiriques nous permet à la fois d'illustrer par des cas concrets nos considérations générales sur les réseaux et territoires, mais aussi de clarifier les échelles et ontologies qu'il nous est pertinent d'utiliser. Comme développé par 8.2, l'échelle microscopique dans le temps et l'espace, pour les objets du traffic routier ici, présente des dynamiques chaotiques, rendant peu réaliste l'intégration de cette échelle dans des modèles qui rendraient comptes d'interactions à de plus grandes échelles. Si cet aspect est pris en compte, c'est généralement sous la forme de congestion, qui est agrégée à une échelle supérieure et pour laquelle soit les conséquences des propriétés chaotiques ont été lissées (ce qui peut être un problème pour les modèles d'équilibre), soit elles sont calibrées empiriquement et l'échelle inférieure n'a donc pas d'ontologie dans le modèle. Nous prendrons ce parti dans nos modèles impliquant un transport routier. Ensuite dans 8.3, toujours concernant le réseau de transport routier, mais selon le point de vue d'un ancrage nodal dans les territoires par les stations essence, en relation avec diverses caractéristiques socio-économiques de ces territoires, nous démontrons d'une part l'existence d'échelles endogènes, correspondant à l'échelle mesoscopique et l'échelle macroscopique, et d'autre part la complexité des processus d'interaction mis en jeu de par leur non-stationnarité déjà démontrée en 4.1 mais aussi par la superposition d'effets territoriaux locaux à des effets liés à la gouvernance. La dernière section ?? permet de conforter ces conclusions de par l'existence d'effet causaux significatif à une échelle mesoscopique dans le temps et dans l'espace. Nous ferons ainsi les choix de modélisation de séparer les échelles, les modèles macroscopiques (comme celui déjà introduit en 4.3) visant à capturer la non-stationnarité en regardant la dynamique à un niveau supérieur en étudiant des variables simples, les modèles mesoscopiques visant à traduire les processus de morphogenèse locaux. Ceux-ci seront introduits dans le chapitre suivant. L'existence d'effets causaux nous confortent dans la recherche de régimes de causalité dans les modèles de coévolution, comme introduits en 4.2, ce qui sera fait en chapitre 7. Enfin, les processus de gouvernance feront l'objet d'une attention particulière dans la modélisation proposée en ??.

* * *

*

9

CADRE THÉORIQUE

La théorie est un élément essentiel de toute construction scientifique, en particulier en Sciences Humaines pour lesquelles la définition des objets et questions de recherche sont plus ouverts mais aussi plus déterminants des directions de recherche alors prises. L'esprit de notre travail n'est pas de produire une théorie unifiée, mais des pistes pour des *Théories Intégrées*, c'est à dire s'appuyant sur une intégration horizontale et verticale au sens de la feuille de route [2009arXiv0907.2221B] , mais aussi permettant une intégration des domaines de connaissance et une réflexivité, au sens qui seront précisés en section 9.3. Nous développons dans ce chapitre un cadre théorique à plusieurs niveaux. Il émerge naturellement de l'interaction des différentes composantes de la connaissance développées jusqu'ici. Dans sa partie thématique, il s'agit donc d'une clarification et unification d'hypothèse ainsi que de conclusions éparses.

C (FL) : affirmation gratuite

C (FL) : dont il convient de rappeler les grands principes

Nous proposons d'abord de construire une *Théorie Géographique* , en quelque sorte un cadre théorique même si nous postulons qu'une Théorie propre a une plus grande portée de par son intégration forte avec les autres domaines de connaissance, qui fixera les objets étudiés et leur nature réelle (leur ontologie), ainsi que leur interrelations. Celle-ci permettra de produire des hypothèses précises qu'on cherchera à confirmer ou infirmer par la suite. Rester à un niveau thématique apparaît cependant ne pas être suffisant pour obtenir des lignes directrices générales sur le type de méthodologies et d'approches à utiliser. Plus précisément, même si certaines théories impliquent un usage plus naturel de certains outils¹, au niveau plus subtil de la mise en contexte au sens de l'approche prise pour implémenter la théorie (comme modèles ou analyses empiriques), la liberté de choix d'objets et d'approches en sciences sociales peut conduire à l'utilisation de techniques inappropriées ou des questionnements inadaptés (voir la section 3.1 pour l'exemple de l'usage inconsidéré des données massives et du calcul). Nous développons pour cela dans une seconde section (??) un cadre théorique à un niveau plus abstrait, visant à formaliser les entreprises de modélisation dans une certaine structure algébrique afin de capturer des articulations fondamentales entre diverses approches. Enfin, nous élaborons dans une dernière section

C (FL) : éviter ce genre de phrase

C (FL) : cela ressemble plus à des pensées qu'à un cadre d'analyse

¹ pour donner un exemple basique, une théorie mettant l'emphase sur la complexité des relations entre agents dans un système conduira généralement à utiliser de la modélisation basée agent et des outils de simulation, tandis qu'une théorie basée sur un équilibre macroscopique favorisera l'usage de dérivations mathématiques exactes.

(9.3) un cadre de connaissances appliqué visant à expliciter des processus de production de connaissance sur les systèmes complexes. Celui-ci est illustré par une analyse fine de la genèse de la Théorie Evolutive des Villes, puis est ensuite appliqué de manière réflexive à l'ensemble de notre travail.

C (FL) : ne pas dire cela ; l'éviter

C (FL) : digression inutile - A x : (IR) pas d'accord car aspect essentiel de mon positionnement sur la production de connaissance, c'est l'illustration récursive en quelque sorte

C (FL) : B

Ce chapitre sera éventuellement le plus délicat à la lecture, d'une part car il est fortement dépendant de la majorité des points thématiques traités précédemment et devrait être lu progressivement selon les concepts introduits (on touche encore aux limitations de la présentation linéaire), et d'autre part car les constructions théoriques introduites sont à un niveau d'abstraction progressif : en quelque sorte, chaque théorie est un cadre méta pour la précédente. On touche alors la question de la réflexivité, et dans quelle mesure celles-ci peuvent s'appliquer à elles-mêmes, en gardant à l'esprit que la séparation entre les niveaux n'est pas directement évidente : par exemple le cadre formel pour les systèmes socio-techniques pourrait être appliqué comme une formalisation du cadre de connaissances. Dans tous les cas, il faut comprendre la démarche à la fois comme une synthèse et comme une ouverture.

* * *

*

La première section de ce chapitre reprend un court passage de [raimbault2017knowledge] ; la deuxième est entièrement inédite. La troisième a été proposée par [raimbault:halshs-01505084] puis développée et appliquée dans [raimbault2017knowledge], et son application réflexive a été présentée par [raimbault2017co].

9.1 CONTRIBUTIONS ET PERSPECTIVES

9.1.1 Définition de la co-évolution

ÉCHELLES SPATIALES ET NON-STATIONNARITÉ

RÉGIMES DE CO-ÉVOLUTION pourquoi serait-ce un bon proxy ?

perspective vers réseaux causaux [seth2005causal] [castellacci2013dynamics]

L'application de notre approche doit être menée précautionneusement concernant le choix des échelles, processus et objets d'étude. Typiquement, elle ne sera pas du tout adaptée à la quantification de processus spatio-temporels dont l'échelle temporelle de diffusion est de l'ordre de celle de la fenêtre d'estimation : l'hypothèse de stationnarité est basique. On peut proposer de procéder à des estimations par fenêtres glissantes, mais il faudrait ensuite élaborer une technique de correspondance spatiale pour traquer la propagation des phénomènes. Un exemple d'application concrète à l'impact thématique fort serait une caractérisation d'une composante fondamentale de la Théorie Evolutive des Villes, la diffusion hiérarchique de l'innovation entre les villes [pumain2010theorie], en analysant les potentielles dynamiques spatio-temporelles des classifications de brevets comme celle introduite par [10.1371/journal.pone.0176310]. Il faut noter toutefois qu'il s'agit de questions méthodologiques relativement ouvertes, dont une des manifestations est le lien potentiel entre le caractère non-ergodique des systèmes urbains [pumain2012urban] et une caractérisation ondulatoire de ces processus. [co2002evolution] diffusion de l'innovation

Une autre direction de développement et d'applications potentiels se révèle en se tournant vers l'échelle plus locale, et d'explorer une hybridation avec les techniques de Regression Géographique Pondérée [brunsdon1998geographically]. La détermination par validation croisée ou Critère d'Akaike d'une portée spatiale optimale pour la performance de ce type de modèles pourrait être adaptée dans notre cas pour déterminer une échelle locale optimale sur laquelle les corrélations retardées sont les plus significatives, ce qui permettrait de s'extraire du problème de la non-stationnarité prioritairement par l'aspect spatial.

APPLICABILITÉ EMPIRIQUE Nos différents cas d'étude empiriques témoignent de la difficulté voire de l'impossibilité de mettre en place les méthodes testées sur des données synthétiques ou uniquement théorique. Prenons l'illustration de la méthode des régimes de causalité : sur les données d'Ile-de-France en 1.2, sur une échelle temporelle courte et une portée spatiale restreinte, son application suggère l'existence de différents régimes. Sur les données sud-africaines

C (AB) : Phrase diderot - NB : a la fin de ta these je te suggere de revenir a cette phrase car au fond ta "definition" de coevolution ne fait pas vraiment echo a cette phrase : il faudra le dire, comme pistes de recherche. -
 A : si quand même : voir def coevol multiscalaire. apres en effet on ne fait pas de modeles multiscalaire et donc pas la composante lente de l'evol - mais quand meme modeles temps long non stationnaires !

en 4.2, on n'est pas capable de classifier les relations entre différentes variables, notamment à cause de l'autocorrélation de l'accessibilité, en on dévoie la méthode à l'étude unique 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 ??, qu'on peut qualifier de pire au regard de l'esprit initial de la méthode, 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, la méthode est bien loin de fonctionner comme attendu, 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. En résumé, la co-évolution est si 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 difficile à caractériser.

Dans cette perspective, et au regard des résultats plus concluants obtenus par l'intermédiaire de la modélisation, une direction future de recherche, bien au delà de la portée de notre travail de par l'envergure de l'entreprise, consiste en l'élaboration de méthodes hybrides qui viseraient à compléter les données manquantes par l'intermédiaires de modèles développés. Plus précisément,

9.1.2 Systèmes de villes et échelle macroscopique

IMPLICATIONS THÉORIQUES Nos résultats soutiennent l'hypothèse que les réseaux de transports physique sont nécessaire pour expliquer la morphogenèse des systèmes territoriaux, au sens où certains aspects sont entièrement contenus dans les réseaux et ne peuvent pas être approchés par des proxy abstraits. Nous avons montré en effet sur un cas relativement simple que l'intégration des réseaux physiques dans certains modèles améliore effectivement leur pouvoir explicatif même lorsqu'on contrôle pour l'overfitting. Cela peut être compris comme une direction pour étendre la Théorie Evolutive des Villes de PUMAIN [pumain1997pour], qui considère les réseaux comme médiateurs des interactions dans les systèmes de villes mais ne met pas d'accent précis sur leur aspect physique et les possibles motifs spatiaux en résultant comme des bifurcations ou des différenciations induites par le réseau. Le développement d'une sous-théorie

C (FL) : trop fort

C (FL) : cela tu ne le montres pas

se concentrant sur ces aspects est une direction intéressante suggérée par ces résultats empiriques et de modélisation. Nous explorerons cette piste en section 9.2.

SPÉCIFICITÉ DU SYSTÈME URBAIN Le modèle n'a pas encore été testé sur d'autres systèmes urbains et d'autres étendues temporelles, et les développements futurs devront étudier quelles conclusions obtenues ici sont spécifiques au système de villes français sur ces périodes, et lesquelles sont plus générales et pourraient être plus générales dans les systèmes de villes. L'application du modèle à d'autres systèmes de villes rappelle également la difficulté de définir les systèmes urbains. Dans notre cas, une forte biais doit être induit par le fait de considérer la France seule, comme Lille doit être fortement influencée par Bruxelles par exemple. L'étendue et l'échelle de tels modèles est toujours un sujet délicat. Nous reposons ici sur la cohérence administrative et celle de la base de données, mais la sensibilité à la définition du système et à son étendue doivent encore être testés.

C (FL) : c'est loin d'être le seul exemple (on parle de métropoles transfrontalières)

RÉSEAU MULTI-COUCHES La considération d'un seul mode de transport pour le système réel est bien sûr réductrice, et une direction immédiate de développement est d'une part le test du modèle avec des matrices de distance réelles pour d'autres types de réseaux, comme le réseau autoroutier qui a connu un essor considérable en France entre 1950 et 1999. Cette application nécessite la mise en place d'une base dynamique pour la croissance du réseau couvrant 1950 à 2015, les bases classiques (IGN ou OpenStreetMap n'intégrant pas la date d'ouverture des tronçons). Une extension naturelle du modèle consisterait alors en la mise en place d'un réseau multi-couches, approche typique pour représenter des systèmes de transport multi-modaux [gallotti2014anatomy]. Chaque couche du réseau de transport devrait avoir une dynamique co-évolutive avec les populations, avec possiblement l'existence d'une dynamique inter-couches.

RÉSEAU PHYSIQUE Ces extensions sont aussi l'objet de [mimeur:tel-01451164], qui produit 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 coévolution. 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:tel-01451164] relève de la modélisation procédurale et non d'indicateurs de structure. C'est probablement pour la même raison que [schmitt2014modelisation] 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.

9.1.3 Territoires et échelle mesoscopiques

INTÉGRATION DANS UN MODÈLE DE CROISSANCE MULTI-SCALAIRE La question du caractère générique du modèle est également ouverte, c'est à dire s'il fonctionnerait de la même manière pour reproduire des formes urbaines sur des systèmes très différents comme les Etats-Unis ou la Chine. Un premier développement intéressant serait de le tester sur ces systèmes et à des échelles légèrement différentes (cellules de taille 1km par exemple). Enfin, nous pensons qu'un gain de connaissance important concernant la non-stationnarité des systèmes urbains serait rendu possible par son intégration dans un modèle de croissance multi-échelles. Les motifs de croissance urbaine ont été prouvés empiriquement exhibant un comportement multi-échelle [zhang2013identifying]. Ici à l'échelle mesoscopique, la population totale et le taux de croissance sont fixés par les conditions exogènes de processus se produisant à l'échelle macroscopique. C'est particulièrement le but des modèles spatiaux de croissance comme le modèle Favaro-Pumain [favaro2011gibrat] de déterminer de tels paramètres par les relations entre villes comme agents. On pourrait conditionner le développement morphologique de chaque zone aux valeurs des paramètres déterminés au niveau supérieur. Dans ce contexte, il faudrait être prudent sur le rôle de la retroaction bottom-up : la forme urbaine émergente devrait-elle influencer le comportement macroscopique à son tour ? De tels modèles complexes multi-scalaires sont prometteurs mais doivent être considérés avec précaution.

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 consequences, 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.

9.2 UNE THÉORIE GÉOGRAPHIQUE

RAFFESTIN souligne dans sa préface de [offner1996reseaux] qu'une théorie géographique articulant espaces, réseaux et territoires n'a jamais été formulée de manière cohérente, chaque approche ayant une vision réduite à certaines composantes seulement et ne visant pas à construire une théorie globalement cohérente. Une piste que nous proposons d'introduire ici est la conjonction des approches de la Théorie Evolutive et de la Morphogenèse, pour à la fois produire une théorie multi-scalaire et intégrant pleinement réseaux et territoires.

9.2.1 *Fondations*

Territoires Humains en Réseau

Notre premier pilier a déjà été construit précédemment lors de l'exploration thématique en Chapitre 1. Nous nous basons sur la notion de *Territoire Humain* élaborée par RAFFESTIN comme la base de la définition d'un système territorial. Elle permet de capturer les systèmes complexes géographiques humains dans l'ensemble de leur caractéristiques concrètes et abstraites, ainsi que dans leur représentations. Par exemple, un territoire métropolitain peut être appréhendé simplement par l'étendue fonctionnelle des flux pendulaires journaliers, ou par l'espace perçu ou vécu des différentes populations, le choix dépendant de la question précise à laquelle on cherche à répondre. Le territoire de RAFFESTIN devrait correspondre à un système cohérent de *synergetic inter-representation networks*, qui est à la fois une théorie et un modèle pour la cognition spatiale des individus et des sociétés, construite par *Portugali* et *Haken* (voir [portugali2011sirn] pour une présentation synthétique). Elle postule que les représentations sont le produit du couplage fort entre les individus des cognitions et de leurs comportements individuels et collectifs. Cette approche au territoire est bien sûr un choix délibéré et que d'autres entrées, possiblement compatibles, peuvent bien sûr être prises [murphy2012entente]. Le ciment de ce pilier est renforcé par la théorie territoriale des réseaux de DUPUY, fournissant la notion de territoire humain en réseau, comme un territoire humain dans lequel un ensemble de réseaux transactionnels potentiels ont été réalisés, ce qui s'accorde par ailleurs avec les visions du territoire comme un lieu des réseaux [champollion:halshs-00999026]. Nous n'utiliserons pas les implications du développement de la notion de *lieu*, celles-ci étant trop éparses (voir définition de [hypergeo]), et à cause de la redondance avec le territoire dans la vision de lien complexe entre représentation et réalité physique. Nous ferons pour ce premier pilier l'hypothèse fondamentale, déjà introduite en chapitre 1, que les réseaux réels sont des éléments nécessaires des systèmes territoriaux.

Théorie Evolutive des Villes

Le second pilier de notre construction théorique est la théorie évolutive des villes de PUMAIN, en relation étroite avec l'approche complexe que nous prenons de manière générale. Celle-ci a déjà été présenté en détails ainsi que ses implications explorées en Chapitre 4. Ici, cette théorie nous permet d'interpréter les systèmes territoriaux comme systèmes complexes adaptatifs avec les implications listées ci-dessus.

Morphogenèse Urbaine

La notion de morphogenèse a été déjà explorée en profondeur et selon un point de vue interdisciplinaire en 5.1. Nous rappelons ici certains grands axes et dans quelles mesure ceux-ci contribuent à la construction de notre théorie. La morphogenèse a été particulièrement soulignée par TURING dans [turing1952chemical] lorsqu'il proposait d'isoler des règles chimiques élémentaires qui pourraient mener à l'émergence de l'embryon et à sa forme. La morphogenèse d'un système consiste en des règles d'évolution auto-cohérentes qui produisent l'émergence de ses états successifs, i.e. la définition précise de l'auto-organisation, avec la propriété supplémentaire qu'une architecture émergente existe, au sens de relations causales circulaires entre la forme et la fonction. Les progrès vers la compréhension de la morphogenèse de l'embryon (en particulier l'isolation de processus particuliers induisant la différentiation de cellules à partir d'une unique) sont relativement récents grâce à l'application des approches complexes en biologie intégrative [delile2016chapitre]. Dans le cas des systèmes urbains, l'idée de morphogenèse urbaine, i.e. de mécanismes auto-cohérents qui produiraient la forme urbaine, est plutôt utilisé dans les champs de l'architecture et de l'urbanisme [hachi2013master] (comme e.g. la grammaire générative du "Pattern Language" d'ALEXANDER), en relation avec des théories de la forme urbaine [moudon1997urban]. Cette idée peut être poussée jusqu'à de très petites échelles comme celle du bâtiment [whitehand1999urban] mais nous l'utiliserons plus à une échelle mesoscopique, en termes de changements d'usage du sol à une échelle intermédiaire des systèmes territoriaux, avec des ontologies similaires à la littérature de modélisation de la morphogenèse urbaine (par exemple [bonin2012modele] décrit un modèle de morphogenèse urbaine avec différentiation qualitative, tandis que [makse1998modeling] donne un modèle de croissance urbaine basé sur une distribution monocentrique de la population perturbée par des bruits corrélés). La notion de morphogenèse sera importante dans notre théorie en lien avec la modularité et l'échelle. La modularité d'un système complexe consiste en sa décomposition en sous-modules relativement indépendants, et la décomposition modulaire d'un système peut être vue comme un moyen de supprimer les correlations

non intrinsèques [2015arXiv150904386K] (pour donner une image, penser à une diagonalisation par blocs d'un système dynamique du premier ordre). Dans le cadre de la conception et du contrôle de systèmes cyber-sociaux à grande échelle, des problèmes similaires surgissent naturellement et des techniques spécifiques sont nécessaires pour le passage à l'échelle des techniques simple de contrôle [2017arXiv170105880W]. L'isolation d'un sous-système fournit une échelle caractéristique correspondante. Isoler des processus de morphogenèse possibles implique une extraction contrôlée (conditions au bord contrôlées par exemple) du système considéré, ce qui correspond à un niveau de modularité et donc à une échelle. Quand des processus auto-cohérents ne sont pas suffisants pour expliquer l'évolution d'un système (dans des variations raisonnables des conditions initiales), un changement d'échelle est nécessaire, causé par une transition de phase implicite dans la modularité. L'exemple de la croissance métropolitaine en est une très bonne illustration : la complexité des interactions au sein de la région métropolitaine sera croissante avec sa taille et la diversité des fonctions urbaines, ce qui conduit à un changement de l'échelle nécessaire pour comprendre les processus. L'émergence d'un aéroport international pourra dans certains cas influencer fortement le développement local, ce qui correspondra à une intégration significative dans un système plus vaste. Les échelles caractéristiques et la nature des processus pour lesquels ces changements ont lieu peuvent être des questions précisément approchées par l'angle de la modélisation. Il est important de noter qu'un sous-système territorial pour lequel la morphogenèse prend sens et dont les frontières sont bien définies peut être vu comme un *système auto-poiétique* au sens étendu de BOURGINE dans [bourgine2004autopoiesis], i.e. comme un réseau de processus qui s'auto-reproduisent² en régulant leur conditions aux bords, ce qui souligne la notion de frontière sur laquelle nous allons finalement nous attarder.

Co-évolution

Notre dernier pilier consiste en une clarification de la notion de *co-evolution*, sur laquelle HOLLAND apporte un éclairage pertinent à travers son approche des systèmes complexes adaptatifs (CAS) par une théorie des CAS comme agents dont la propriété fondamentale est de traiter des signaux grâce à leur frontières [holland2012signals]. Dans cette théorie, les systèmes complexes adaptatifs forment des agrégats à différents niveaux hiérarchiques, qui correspondent à différents niveaux d'auto-organisation, et les frontières sont intriquées horizontalement et verticalement de manière complexe. Cette approche introduit la notion de *niche* comme un sous-système relativement indépendant au sein duquel les ressources circulent (de la même façon

² qui ne sont toutefois pas cognitifs, ne rendant pas ces systèmes morphogénétiques vivants au sens de auto-poiétique et cognitif

que des communautés dans un réseau) : de nombreuses illustrations telles les niches écologiques ou économiques peuvent être données. Les agents au sein d'une niche sont dits en *coévolution*. Empiriquement, les résultats obtenus témoignant d'une co-évolution à l'échelle mesoscopique comme en 4.2, confirment l'existence de niches pour certains aspects des systèmes territoriaux. La co-évolution implique ainsi de fortes interdépendances (impliquant des processus causaux circulaires) et une certaine indépendance au regard de l'extérieur de la niche. La notion est naturellement flexible puisqu'elle dépendra des ontologies, de la résolution, des seuils, etc. que l'on considère pour définir le système. Nous postulons vu les indices d'existence obtenus dans les résultats empiriques, mais aussi les modèles reproduisant les processus de manière crédible sous une hypothèse d'isolation raisonnable, que ce concept peut se transmettre à la théorie évolutive urbaine et correspond à la notion de co-évolution décrite par PUMAIN : des agents co-évolutifs dans un système de villes consistent en une niche et ses flots, signaux et limites et sont donc des entités co-évolutives au sens de HOLLAND. Cette notion sera importante pour nous dans la définition des sous-systèmes territoriaux et de leur couplage. Nous gardons à l'esprit les potentialités et limitation du parallèle entre systèmes biologiques et systèmes sociaux décrits en 3.3.

9.2.2 Une théorie des systèmes territoriaux co-évolutifs en réseau

Nous synthétisons les différents piliers en une théorie géographique autonome des systèmes territoriaux pour lesquels les réseaux jouent un rôle central pour la co-évolution des composantes du système. Pour les définitions des termes et les références, se référer à la section précédente. La formulation ici est voulue minimaliste.

Définition 1 - Système Territorial. *Un système territorial est un ensemble de territoires humains en réseau, c'est à dire des territoires humains au sein desquels et entre lesquels des réseaux réels existent.*

Le territoire est bien un élément du système territorial, qui de manière plus générale connecte différents territoires par les réseaux. A cette étape la complexité et le caractère évolutif et dynamique des systèmes territoriaux sont impliqués par les partis pris mais pas une partie explicite de la théorie. We supposerons pour simplifier une définition discrète des dimensions temporelles, spatiales et ontologiques, sous des hypothèses de modularité et de stationnarité locale. Cet aspect, à la fois pour le discret et la stationnarité, correspond à une simplification ontologique de la supposition d'une "échelle minimale" à laquelle les sous-systèmes fournissent une décomposition modulaire simple du système global. Elle reflète nos conclusions empiriques obtenues en Chapitre 8 et les modèles développés par la

C (FL) : a mettre bien plus haut : c'est assez fondamental

suite. On suppose également ergodicité locale, pour obtenir grâce à la démonstration proposée en 4.1 la propriétés de non-ergodicité globale typique des systèmes urbains.

Proposition 1 - Echelle discrètes. *Supposant une décomposition modulaire discrète d'un système territorial, l'existence d'un ensemble discret (τ_i, x_i) d'échelles temporelles et fonctionnelles pour le système territorial est équivalent à la stationnarité temporelle locale d'une spécification par système dynamique stochastique du système.*

Preuve (Tentative). Nous partons de l'hypothèse que tout système territorial peut être représenté par un ensemble de variables aléatoires, ce qui revient à avoir des objets et états bien définis et utiliser le Théorème de Transfert sur les événements des états successifs. Si $X = (X_j)$ est la décomposition modulaire, on a nécessairement quasi-indépendance des composantes au sens que $\text{Cov}[dX_j, dX_{j'}] \simeq 0$ à tout moment. Les transitions de stationnarité globales induise des transitions dans chaque module, qui sont conservées si elles correspondent effectivement à un transition dans le sous-système. On obtient ainsi les échelles temporelles comme temps caractéristiques des sous-dynamiques. Les échelles fonctionnelles sont les étendues correspondantes dans l'espace d'état. ■

C (FL) : à discuter

Cette proposition postule une représentation des dynamiques du système dans le temps. On peut noter que même en l'absence de représentation modulaire, le système dans son ensemble vérifiera la propriété. Cette définition des échelles permet d'introduire explicitement des boucles de rétroaction, puisqu'on peut par exemple conditionner l'évolution d'une échelle à celle d'une autre qui la contient, et ainsi l'émergence et la complexité, rendant la théorie compatible avec la théorie évolutive urbaine.

Hypothèse 1 - Imbrication des échelles et des sous-systèmes. *Des réseaux complexes de retroaction existent à la fois entre et à l'intérieur des échelles [bedau2002downward]. De plus, un emboîtement horizontal et vertical des limites ne sera généralement pas hiérarchique.*

Au sein de ces imbrications de sous-systèmes nous pouvons isoler des composantes en co-évolution en utilisant la morphogenèse. La proposition suivante est une conséquence de l'équivalence entre l'indépendance d'une niche et sa morphogenèse. La morphogenèse fournit la décomposition modulaire (sous hypothèse de stationnarité locale) nécessaire pour l'existence de l'échelle, donnant des sous-systèmes minimaux indépendants de manière verticale (échelle) et horizontale (espace).

Proposition 2 - Co-évolution des composantes. *Les processus morphogénétiques d'un système territorial sont une formulation équivalente de l'existence de sous-systèmes co-évolutifs.*

Nous formulons finalement la dernière hypothèse clé qui met les réseaux réels au centre des dynamiques co-évolutives, introduisant leur nécessité pour expliquer les processus dynamiques des systèmes territoriaux.

Hypothèse 2 - Nécessité des réseaux. *L'évolution des réseaux ne peut pas être expliquée simplement par la dynamique des autres composantes territoriales et réciproquement, i.e. les sous-systèmes territoriaux co-évolutifs contiennent les réseaux réels. Ceux-ci peuvent ainsi être à l'origine de changements de régime (transitions entre régimes stationnaires) ou de bifurcations plus conséquentes dans les dynamiques de l'ensemble du système territorial.*

9.2.3 Contextualisation

Sur de longues échelles temporelles, une co-évolution globale a été montrée pour le système ferroviaire français par [bretagnolle:tel-00459720]. A de plus petites échelles celle-ci est moins évidente (débat sur les effets structurants) mais nous supposons la présence d'effets co-évolutifs à toutes les échelles. Des exemples régionaux peuvent illustrer ce fait : Lyon n'a pas les mêmes relations dynamiques avec Clermont qu'avec Saint-Etienne, et la connectivité de réseau a probablement un rôle à y jouer (parmi les effets des dynamiques intrinsèques des interactions, et de la distance par exemple). A une plus petite échelle encore, nous partons du principe que les effets sont encore moins observables, mais précisément à cause du fait que la co-évolution est plus forte et les bifurcations locales se produisent avec une plus grande amplitude et une plus grande fréquence que dans les systèmes macroscopiques où les attracteurs sont plus stables et les échelles de stationnarité plus grandes. Nous pour cela que nous avons tenté d'identifier des bifurcations ou des transitions de phase dans des modèles jouets, des modèles hybrides, et des analyses empiriques, à différentes échelles, sur différents cas d'études et avec différentes ontologies.

Une difficulté dans notre construction est l'hypothèse de stationnarité locale, qui est essentielle pour formuler des modèles à l'échelle correspondante. Même si cela paraît une hypothèse raisonnable à plusieurs échelles et a déjà été observé des données empiriques [sanderson1992système], nous devrons le vérifier dans nos études empiriques. En effet, cette question est au centre des efforts de recherche courants pour appliquer les techniques d'apprentissage profond aux systèmes géographiques : BOURGINE a récemment développé un cadre pour extraire des motifs des systèmes complexes adaptatifs. En utilisant un théorème de représentation [knight1975predictive], tout processus stationnaire discret est un *Modèle de Markov Caché*. Etant donné la définition d'un état causal comme $\mathbb{P}[\text{future}|A] = \mathbb{P}[\text{future}|B]$, la partition des états du système par la relation d'équivalence correspondantes permet de produire un *Réseau Récurrent* qui est suffisant pour détermi-

ner l'état suivant du système, puisqu'il s'agit d'une fonction *déterministe* des états précédents et des états cachés [shalizi2001computational] : $(x_{t+1}, s_{t+1}) = F[(x_t, s_t)]$. L'estimation des états cachés et de la fonction récurrente capture ainsi entièrement par apprentissage profond le comportement dynamique du système, i.e. l'information complète sur ses dynamiques et les processus internes. Les questions sont ensuite si les hypothèses de stationnarité peuvent être réglées par augmentation des états du système, et si des données hétérogènes et asynchrones peuvent être utilisées pour initialiser des séries temporelles assez longues pour une estimation correcte du réseau de neurones ou de tout autre type d'estimateur. Ces questions sont reliées à l'hypothèse de stationnarité pour la première et à la non-ergodicité pour la seconde.

★ ★

★

9.3 UN CADRE DE CONNAISSANCES APPLIQUÉ

La complexité de la production de connaissance sur des systèmes complexes est bien connue, mais il n'existe toujours pas de cadre de connaissance qui rendrait à la fois compte d'une certaine structure de la production de connaissance à un niveau épistémologique et serait directement applicable à l'étude et au management des systèmes complexes. Nous posons ici les bases d'un tel cadre, en commençant par analyser en détail l'étude de cas de la construction d'une théorie géographique des systèmes territoriaux complexes, au travers de méthodes mixtes, plus précisément des analyses qualitatives d'entretiens et une analyse quantitative de réseau de citation. Nous pouvons par cela construire de manière inductive un cadre qui considère les entreprises de production de connaissance comme des perspectives, dont les composantes sont en co-évolution au sein de domaines de connaissances complémentaires. Nous discutons finalement des applications et développements potentiels.

La compréhension des processus et des conditions de production de la connaissance scientifique est une question toujours globalement ouverte, à laquelle des monuments de l'épistémologie comme la Critique de la Raison Pure de Kant, ou plus récemment l'étude par Kuhn de la "structure des révolutions scientifiques" [kuhn1970structure] ou le positionnement de Feyerabend pour une diversité des approches [feyerabend1993against], ont apporté des éléments de réponse d'un point de vue philosophique. Un matériau plus empirique a été apporté également récemment avec les analyses quantitatives de la science, dans un sens une *épistémologie quantitative* qui va bien plus loin que des indicateurs bibliométriques purs [cronin2014beyond]. Les contributions s'intéressant à la complexité, c'est à dire étudiant des systèmes complexes en un sens très large, peuvent témoigner de la production de cadre de travail très divers qui peuvent être vus comme des éléments élémentaires de réponse à la question à un autre niveau ci-dessus. Nous utiliserons par la suite le terme *Cadre de Connaissances*, pour tout cadre tel ayant une composante épistémologique s'intéressant à la nature de la connaissance et à sa production. Pour illustrer, nous pouvons mentionner de tels cadres dans différents domaines, à différents niveaux, et avec des but différents. Par exemple, [durantin2017disruptive] explore les potentialités de coupler l'ingénierie avec des paradigmes du design to encourager l'innovation disruptive. Toujours en Gestion de Connaissances, utilisant la contrainte de l'innovation comme un avantage pour appréhender la nature complexe de la connaissance, [carlile2004transferring] introduit les notions de frontières des domaines de connaissance et de processus de production. Introduisant également un framework meta, mais dans le champ de l'ingénierie des systèmes, [geminio2004framework] recommande l'utilisation de grammaires pour comparer les techniques de Modélisation Concep-

tuelle. Les cadres de meta-modélisation peuvent aussi être compris comme des cadres de connaissance. [cottineau2015modular] décrit un cadre de multi-modélisation pour le test d'hypothèses dans la simulation des systèmes complexes socio-techniques. [golden2012modeling] postule une formulation unifiée de la notion de système, ce qui inclut nécessairement différents types de connaissance sur un système correspondant à la description de ses différents composants.

Une explication possible pour une telle richesse est la nature fondamentalement réflexive de l'étude des Systèmes Complexes : à cause du choix plus grand pour la méthodologie et sur quels aspects du système mettre l'emphase, une partie significative d'une entreprise de modélisation ou de design est une exploration à un niveau meta. De plus, les études de la production de connaissance sont profondément ancrées dans la complexité, comme Hofstadter a bien souligné dans [hofstadter1980godel] en rappelant l'existence de "boucles étranges", c'est à dire de boucles de rétroaction permettant la reflexivité comme une théorie s'appliquant à elle-même, dans ce qui constitue l'intelligence et l'esprit. L'intelligence artificielle est de fait un champ crucial au regard de nos réflexions, comme ses progrès impliquent une compréhension plus fine de la nature de la connaissance. [2017arXiv170401407M] introduit un meta-cadre pour une typologie générale des approches en intelligence artificielle, ce qui correspond à un cadre de connaissance non au sens propre mais dans un cas particulier d'application.

Le niveau des cadres présentés ci-dessus peut être très général mais reste conditionné à un certain champ ou discipline, et à une certaine approche ou méthodologie. Il n'existe à notre connaissance pas de cadre réalisant un exercice difficile, qui est de capturer une certaine structure de production de la connaissance à un niveau épistémologique, mais qui est conjointement pensée dans une perspective très appliquée, avec des conséquences directes pour la conception et la gestion de systèmes complexes. La contribution de cette partie propose de poser les bases pour un cadre réalisant cela dans le cas des Systèmes Complexes. Pour y parvenir, nous partons du postulat que la tension entre ces deux objectifs contradictoires est un atout pour éviter d'une part une généralité globale impossible et d'autre part une spécificité due à un domaine qui serait trop restrictive. En se basant sur l'idée des domaines de connaissance introduite par [livet2010], son aspect central est une approche cognitive de la science qui implique des processus de co-evolution entre les domaines de connaissance et leur supports. Une première ébauche de ce cadre a été présentée par [raimbault:halshs-01505084], dans le cas particulier des systèmes complexes territoriaux comme étudiés par la géographie théorique et quantitative. Nous proposons de l'introduire ici par une démarche inductive, c'est à dire en partant d'une étude de cas concrète

qui a largement inspiré la construction du cadre, pour finir avec sa description générique.

La suite de cette section est organisée de la façon suivante : nous détaillons d'abord les études de cas, plus précisément une étude détaillée d'une théorie géographique des systèmes urbains complexes : la théorie évolutive des villes, puis un court exemple d'ingénierie qui permet d'illustrer les possibilités de transfert des concepts. Nous spécifions ensuite les définitions et formulons le cadre épistémologique. Nous discutons ensuite les questions d'applicabilité, des développements potentiels comme une version mathématique du cadre, puis une application réflexive du cadre à notre sujet d'étude.

9.3.1 *Etude de cas*

Genèse de la Théorie Evolutive Urbaine

La première étude de cas rappelle la construction de la *Théorie Evolutive Urbaine*³, une théorie géographique qui considère les systèmes territoriaux par une perspective complexe, développée depuis une vingtaine d'années environ. Nous étudions sa genèse par l'utilisation de méthodes mixtes, c'est à dire à la fois des interviews semi-dirigées avec des contributeurs principaux, et une analyse bibliométrique quantitative des publications principales. Les interviews ont été menées en suivant les standards méthodologiques classiques [legavre1996neutralite] pour assurer une interférence limitée des expériences de l'interviewer, mais sans le faire disparaître complètement afin de permettre un contexte précis favorable à la fluidité de l'interviewé. Nous utilisons ici des interviews⁴ avec Pr. D. Pumain qui a introduit et développé majoritairement la théorie, et Dr. R. Reuillon, dont la recherche sur le calcul intensif et distribué et l'exploration de modèles a été une pierre d'angle des développements les plus récents.

Pour commencer il est important de se rappeler un aperçu rapide du contenu de la théorie évolutive. Pour cela, se référer à la présentation qui en est faite en introduction du Chapitre 4, qui en donne la substantifique moelle.

³ L'ambiguïté de l'adjectif *évolutive* fait gagner la théorie en subtilité, puisqu'il s'applique aussi bien au sens premier c'est à dire aux entités urbaines étudiées, mais aussi à un sens meta à la théorie elle-même, ce qui confirme un certain niveau de réflexivité de la théorie qui est essentiel comme développé en 3.3. Pour traduire le terme en anglais, il a été choisi "Evolutionary Urban Theory" par [pumain2006evolutionary], mais "Evolutive Urban Theory" convient aussi, mais il semble dans tous les cas difficile de transférer l'ambiguïté lors de la traduction.

⁴ Toutes les deux d'une durée environ une heure. Le son et les transcripts sont disponibles sous une Licence CC à <https://github.com/JusteRaimbault/Interviews> [raimbault2017entretiens]. Les interviews sont en français et la traduction anglaise des passages cités dans l'article original est assurée par l'auteur.

La caractéristique frappante dans cette construction est l'équilibre entre les différents *types* de connaissance, desquelles une typologie sera le point de départ de notre construction. La relation entre les considérations théoriques et les cas d'étude empiriques est fondamental. En effet l'article séminal [pumain1997pour] est déjà positionné comme "un plaidoyer pour une théorie [...] moins ambitieuse, mais qui ne néglige pas les aller-retours avec l'observation". Nous pouvons maintenant nous tourner vers les entretiens pour mieux comprendre les implications de l'intrication des différents types de connaissance. D. Pumain retrace les idées germinales à son travail de maîtrise en 1968, quand "tout a commencé avec une question de données". L'intérêt pour les villes, et pour le *changement dans les villes*, a été conduit par la disponibilité d'un jeu de données raffiné sur les flux migratoires à différentes dates. Egalement rapidement, est venue "la frustration des méthodes qui manquaient", mais l'accès au centre de calcul (*outil technique*) a permis le test de méthodes et modèles nouvellement introduits, liés à l'approche de la complexité par Prigogine. Les méthodes restaient toutefois limitées pour capturer l'hétérogénéité des interactions spatiales. Un besoin progressivement spécifié et une rencontre fortuite, avec "une dame qui travaillait sur les réseaux de neurones et les modèles agents à la Sorbonne", a conduit à une bifurcation et un nouveau niveau d'interaction entre modèles, théorie et connaissance empirique : en 1997, deux articles séminaux, l'un donnant la base théorique, l'autre introduisant le premier modèle Simpop, étaient publiés simultanément. A partir de ce point, il était clair que toute entreprise de modélisation était conditionnée à une connaissance empirique de cas d'étude géographiques et à des hypothèses théoriques à tester. Les méthodes et les outils techniques ont alors pris aussi un rôle nécessaire, avec des méthodes d'exploration de modèles spécifiques développées avec le logiciel Open-Mole. R. Reuillon raconte qu'un saut qualitatif de connaissances a été rendu rapidement possible quand les méthodes d'exploration systématiques ont été introduites pour comprendre le comportement du modèle SimpopLocal. A la base, les géographes n'étaient pas sûrs si le modèle fonctionnait seulement, dans le sens où il produisait les faits stylisés attendus comme l'émergence de la hiérarchie d'un système de villes. Des trajectoires satisfaisantes ont été trouvées par l'utilisation d'algorithmes génétiques de calibration, en calcul distribué sur grille [schmitt2014half]. L'existence de multiples solution équivalentes pour les valeurs des paramètres est une barrière pour des questions concrètes de nécessité ou suffisance d'un mécanisme donné du modèle agent. Ce besoin, venant du domaine de la connaissance empirique et théorique géographique, a mené à la conception d'un algorithme spécifique : le Calibration Profile, qui est une avancée méthodologique dans l'exploration de modèles [reuillon2015]. Ce cercle vertueux a été continué avec la famille de modèles Ma-

rius [cottineau2014evolution] et l'algorithme Parameter Space Exploration [10.1371/journal.pone.0138212]. R. Reuillon évalue son impact du point de vue d'un informaticien : "Je ne suis pas sûr si les géographes étaient immédiatement conscients de la portée du résultat, c'était du lourd, les gens qui bossaient avec nous l'ont directement vu." Cette vision positive est confirmée par D. Pumain, qui souligne les bénéfices de ces nouvelles méthodes pour la connaissance Géographique, et que c'était la première fois qu'une recherche menait à des publications à la frontière de la connaissance à la fois en géographie et en informatique.

En prenant du recul, émerge une typologie de domaines dans laquelle de la connaissance a été créée mais également nécessaire pour les autres domaines dans la genèse de la Théorie Evolutive Urbaine. La récolte des données et la construction de jeux de données est un premier pré-requis pour toute connaissance supplémentaire. A partir des données on extrait des faits stylisés empiriques, desquels sont déduits des hypothèses théoriques. La Théorie peut être testée pour falsification, dans le domaine empirique mais aussi par les modèles, par exemple par des expériences ciblées dans les modèles de simulation. De nouvelles méthodes sont alors développées pour mieux les explorer. Les outils sont cruciaux à chaque étape, pour implémenter un modèle, faire de la fouille de données ou collecter et formater les données par exemple. L'analyse précédente montre comment ces domaines sont interdépendants, et sont dans un sens *co-évolutifs*.

Nous supportons cette analyse qualitative par une analyse quantitative bibliométrique modeste. L'idée est d'étudier la structure du coeur du réseau de citations des publications principales construisant la Théorie Evolutive Urbaine. Nous construisons le réseau de citations comme décrit en Fig. 66, en utilisant l'outil de collection de données fournit par [raimbault2016indirect]⁵. Partant des deux publications séminales [pumain1997pour] et [sanderson1997simpop], le réseau de citation inverse est obtenu à profondeur 2 (les références citant ces références initiales, et celles citant les citantes), en filtrant à la première étape sur les auteurs pour avoir au moins un des principaux contributeurs de la théorie (que nous prenons comme *Pumain*, *Sanders* et *Bretagnolle*, en accord avec l'entretien avec D. Pumain). Les noeuds de degré 1 sont supprimés, pour obtenir uniquement le coeur du réseau d'*ego*. On peut noter qu'il ne manque pas de lien entre les noeuds du premier niveau, puisque tous les liens citants ont été récupérés. Le réseau a une densité de 0.019, ce qui est plutôt élevé pour un réseau de citation, et la signature d'un haut niveau de dépendance entre les publications. En partant de deux noeuds distincts, nous aurions pu avoir en théorie des composantes connexes distinctes, mais comme attendu le réseau n'en a qu'une de par la nature fortement interconnectée des

⁵ L'ensemble du code et des données pour cette analyse sont disponibles à <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/QuantEpistemo>

deux aspects. Pour analyser la structure de manière plus fine, nous détectons les communautés en utilisant l'algorithme de clustering de Louvain, et évaluons la modularité dirigée de la partition comme donnée par [nicosia2009extending]. Nous montrons en Fig. 66 une visualisation du réseau. Nous obtenons 7 communautés avec une valeur de modularité de 0.39. Pour assurer que cette valeur est significative, nous procédons à des simulations de Monte Carlo et distribuons de manière aléatoire les liens de citation 100 fois, en calculant à chaque fois la modularité des communautés dans le réseau aléatoire. Nous obtenons une modularité moyenne dirigée de $\bar{m} = 0.002 \pm 0.015$, rendant la modularité du réseau réel hautement significative (plus de 200 déviations standard). Nous analysons le contenu des communautés en examinant leur publications du premier niveau. Nous trouvons que les communautés sont globalement cohérentes avec les typologies des domaines : une pour les méthodes, trois sur la modélisation spatio-temporelle des systèmes urbains qui mélange empirique et modélisation, une conceptuelle, une sur les modèles Simpop, et une dernière sur les lois d'échelle qui est complètement empirique. Les *Data Papers* ne sont pas encore une pratique courante en géographie et des articles spécifiques au domaine des données ne peuvent être trouvés dans le réseau. Un taux de citation accru entre papiers du même domaine est dans tous les cas attendu à cause du standard scientifique de toujours situer une contribution au regard des travaux similaires. La valeur significative de la modularité confirme que les domaines sont cohérents au regard d'une certaine structure endogène de la production de connaissance.

Ingénierie

Après l'aperçu sur les domaines de connaissances extraits dans l'étude de cas précédente, nous proposons de prendre un point de vue similaire sur un exemple assez différent plus en relation avec la technologie et l'ingénierie. Nous interprétons ainsi des questions d'ingénierie liées au système de transport métropolitain parisien au travers du prisme des domaines de connaissance. En prenant l'exemple de l'automatisation progressive de la ligne 1, considérée largement comme une prouesse technique, de nombreuses études intégrant modélisation et études empiriques ont été conduite en préliminaire [belmonte2008automatisation]. L'utilisation et l'adaptation de méthodes particulières comme la modélisation basée-agent est cruciale pour le développement de transports autonomes innovants [balbo2016positionnement]. Dans ce problème d'ingénierie, des solutions techniques comme les portes papillères de quai peuvent être vues comme des outils qui évoluent également, et sont nécessaires pour qu'une nouvelle approche conceptuelle (*le transport automatique*) soit implémentée [foot2005faut]. Mais ils peuvent aussi interagir avec d'autres aspects de la connaissance conceptuelle, comme le management et l'organisation au sein de l'opé-

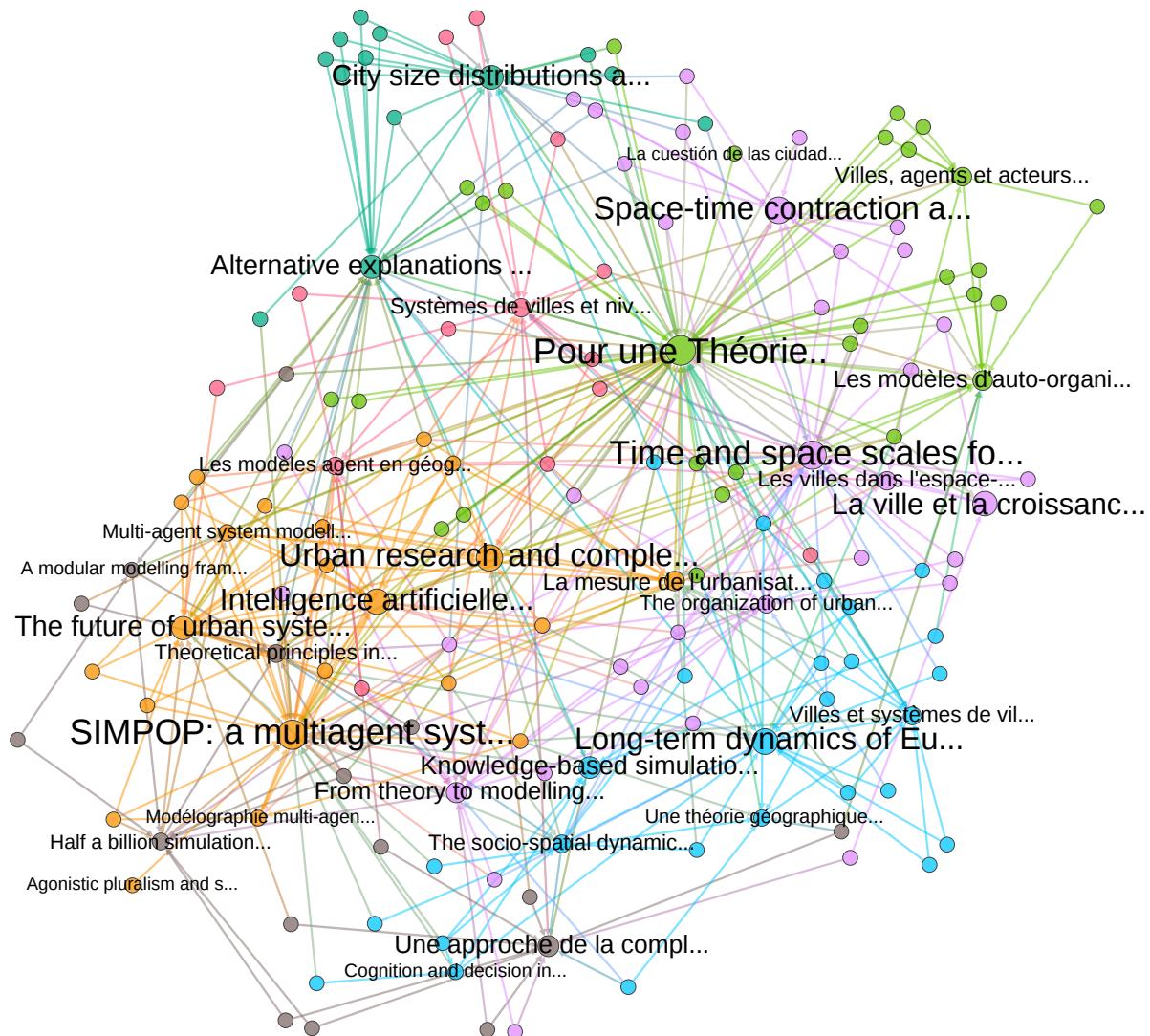


FIGURE 66 : Réseau de citations des publications principales de la Théorie Evolutive Urbaine. Le réseau est construit de la manière suivante : à partir des deux publication séminales [[pumain1997pour](#)] et [[sanderson1997simpop](#)], nous récupérons les publications les citant, filtrons sous la condition d'un des contributeurs principaux appartenant aux auteurs, récupérons encore les publications citantes et filtrons. Les noeuds sont les publications ($|V| = 155$), leur taille correspondant à la centralité de vecteur propre, et les liens sont les liens de citation dirigés ($|E| = 449$). La couleur donne les communautés obtenues par l'algorithme de clustering de Louvain (7 communautés, modularité 0.39).

rateur [[foot1994ratp](#)]. L'aspect multi-dimensionnel complexe de l'innovation pour de tels systèmes avait déjà été souligné depuis longtemps comme le montre [[hatchuel1988stations](#)]. D'autres aspects techniques, comme des problèmes d'ingénierie civile [[moreno2016etude](#)], sont aussi mise en jeu pour développer une telle nouvelle approche, et ils nécessitent au moins les domaines empiriques et de modélisation, voire plus. Cet exemple relativement court illustre comment l'in-

terprétation par domaines de connaissance peut être appliquée à l'ingénierie et au management de systèmes complexes industriels. Des détails spécifiques seraient nécessaires pour une application plus en profondeur, mais nous proposons ici une preuve de concept.

9.3.2 Cadre de Connaissances

Nous pouvons à présent formuler le cadre de manière inductive. Comme déjà évoqué, il tire l'idée de domaines de connaissance en interaction du cadre introduit par [livet2010], mais étend ces domaines et prend une nouvelle position épistémologique, se concentrant sur les dynamiques co-évolutives entre agents et connaissances.

CONTRAINTE Pour être particulièrement adapté à l'étude et au management de la complexité, nous postulons que le cadre doit répondre à certaines contraintes, en particulier pour prendre en compte et même favoriser la *nature intégrative de la connaissance*, comme illustré par l'importance de l'interdisciplinarité et de la diversité dans les cas d'étude. Le cadre doit ainsi être favorable aux points suivants :

- Intégration des disciplines, puisque les Systèmes Complexes sont par essence à la croisée de champs multiples
- Intégration des domaines de connaissance, c'est à dire qu'aucun type particulier de connaissance ne doit être privilégié dans le processus de production⁶
- Intégration des types de méthodologie, en particulier dépasser les frontières artificielles entre méthodes "quantitatives" et "qualitatives", qui sont particulièrement fortes en sciences sociales et humanités classiques.

FONDATIONS ÉPISTÉMOLOGIQUES Le positionnement épistémologique du cadre est celui développé dans la première section de 3.3. Nous rappelons l'importance de la *perspective* [giere2010scientific], composée des agents, des objets représentés, du but et du medium (le modèle). L'approche par agents est fondamentale pour la cohérence du cadre.

DOMAINES DE CONNAISSANCE Nous postulons les domaines de connaissance suivants, avec leurs définitions :

- **Empirique.** Connaissance empirique d'objets du monde réel.
- **Théorique.** Connaissance conceptuelle plus générale, impliquant des constructions cognitives.

⁶ ce qui n'est pas incompatible avec des spécifications fonctionnelles très strictes, puisque des chemins divers sont possibles pour atteindre le même état final fixé

- **Modélisation.** Le modèle est le *medium* formalisé de la Perspective Scientifique, aussi divers que la classification de VARENNE des fonctions des modèles [[varenne2010simulations](#)] (voir ci-dessous).
- **Données.** Information brute qui a été collectée.
- **Méthodes.** Structures génériques de production de connaissance.
- **Outils.** Proto-méthodes (implémentation des méthodes) et supports des autres domaines.

Nous prenons le parti de séparer Outils et Méthodes, pour insister sur le rôle de support des outils, et car le développement des deux est lié mais pas identique. De la même façon, le domaine des Données et le domaine Empirique sont distincts, car des nouveaux jeux de données n'impliquent pas systématiquement une nouvelle connaissance de faits empiriques, même si la construction des outils de captation de données souvent requiert une connaissance empirique. Le domaine de la Modélisation a un rôle central puisque nous postulons que *toute connaissance d'un système complexe nécessite un modèle*.

CO-ÉVOLUTION DES CONNAISSANCES Nous pouvons à présent formuler l'hypothèse centrale de notre cadre, qui est partiellement contenu dans le positionnement par rapport au Perspectivisme. Nous postulons que *toute construction de connaissance scientifique sur un système complexe*⁷ est une perspective au sens de GIERE. Elle est composée de contenu de connaissance dans chacun des domaines, qui *co-évolue* entre eux et avec les autres éléments de la perspective, en particulier les agents cognitifs. La notion de co-évolution est prise au sens de [[holland2012signals](#)], c'est à dire d'entités étant fortement interdépendantes au sein de niches avec des relations causales circulaires et qui ont une certaine indépendance avec l'extérieur dans leur frontières. Nous notons l'importance de l'émergence faible au sens de BEDAU [[bedau2002downward](#)] dans la construction de la perspective à partir de la co-évolution de ses composants, comme il s'agit d'un niveau supérieur autonome qui peut être compris en lui-même, comme

⁷ Nous sommes convaincus que cet aspect intriqué de la production de connaissance est nécessairement présent pour les Systèmes Complexes, en écho à la remarque sur la réflexivité en introduction de la section. Même des *modèles simples* de systèmes complexes impliquent une complexité conceptuelle qui nécessite que la complexité de la connaissance soit présente pour être traduite. Cette dernière hypothèse pourrait liée à la nature de la complexité et la relation entre la complexité computationnelle et la complexité au sens de l'émergence faible, qui est suggérée par exemple par [[2014arXiv1403.7686B](#)] qui explique l'émergence et la décohérence depuis le niveau quantique par la NP-complétude de la résolution des équations fondamentales. Ces considérations sont bien au delà de la portée de cette section (voir [3.3](#) pour une réflexion plus approfondie), et nous prenons comme une hypothèse que les systèmes complexes nécessitent de la connaissance complexe, tandis que de la connaissance simple (au sens de domaines et agents non co-évolutifs) *peut* exister pour des systèmes simples.

la connaissance scientifique peut être. Il faut aussi noter qu'une perspective n'a pas nécessairement des composants dans tous les domaines, mais devraient généralement en avoir dans la plupart.

L'aspect social de la production de connaissance n'est pas inclus dans les domaines de connaissance, mais dans les agents et leur relation. [roth2010social] montre une co-évolution des réseaux sociaux et des réseaux sémantiques avec l'exemple d'une communauté scientifique en biologie du développement et un environnement de blogs politiques, ce qui confirme dans notre cas la co-évolution entre les agents et les domaines.

APPLICATION Les types de modèles auquel notre cadre s'applique sont supposés être tous les modèles possibles en un sens très large, puisque **GIÈRE** désigne par modèle tout *medium* d'une perspective. Une vue fonctionnelle des modèles comme VARENNE introduit [varenne2010simulations] (introduisant une typologie des modèles par leur fonctions, par exemple les modèles explicatifs, les modèles de simulation, les modèles prédictifs, les modèles de compréhension, les modèles interactifs, etc.) est un moyen d'appréhender leur variété. Il est aussi possible de le faire en terme de classifications plus classiques, et l'appliquer au modèles mathématiques, statistiques, de simulation, de données, ou conceptuels par exemple. Concernant les contraintes données précédemment, comme toutes les connaissances sont en co-évolution, aucun domaine n'est privilégié en particulier. Aucune discipline non plus, puisque celles-ci auront leur différents aspects contenus dans les domaines, et finalement les méthodes qualitatives et quantitatives seront présentes et nécessaire dans la majorité. Nous montrons en Fig. 67 une projection des domaines de connaissance comme un réseau complet, pour illustrer de quoi peuvent être composées les relations entre domaines.

9.3.3 *Discussion*

Portée d'application

Nous insistons sur le fait que notre cadre ne prétend pas introduire une épistémologie générale de la connaissance scientifique, mais loin de cela est plutôt ciblé vers une réflexivité dans la compréhension des systèmes complexes. Le niveau de généralité est à niveau très différent, mais le but d'implications pratiques dans la compréhension de la complexité contribue à un certain caractère générique dans les applications. Il est de plus particulièrement adapté à l'étude des Systèmes Complexes, puisque des approches plus réductionnistes peuvent gérer des productions de connaissance plus compartimentées, tandis que l'intégration des disciplines et des échelles et donc des domaines de connaissance a été souligné comme crucial pour l'étude de la complexité.

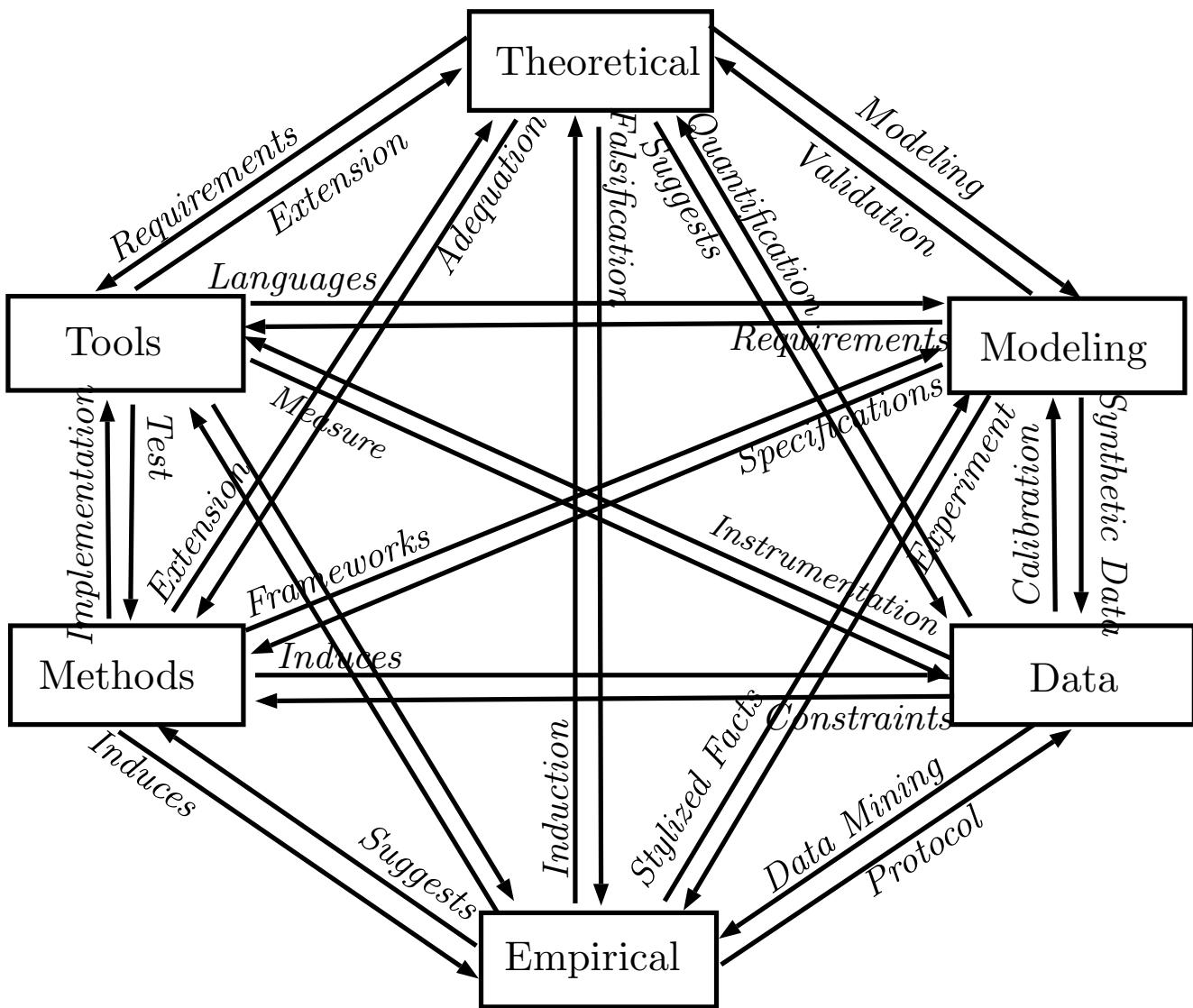


FIGURE 67 : Projection d'une perspective comme un réseau complet des domaines de connaissance. Pour illustrer les domaines et les processus d'interaction possibles entre ceux-ci, nous faisons l'exercice d'essayer de qualifier toutes les relations binaires possibles entre les domaines. Cela ne reflète en rien la structure réelle du cadre, mais est une aide pour considérer ce que les interactions peuvent être. Il faut noter que la nature des relations n'est pas toujours la même ici, certaines étant des contraintes, d'autres des transferts de connaissance, d'autres processus à l'intérieur d'autres domaines comme les données synthétiques qui est une méthodologie. Cela montre que certains domaines agissent comme catalyseurs pour les relations entre les autres, dans cette configuration de réseau, ce qui correspond en fait à une situation de co-évolution.

Vers une Formalisation

Le cadre de connaissances reste à un niveau épistémologique, et son application pourrait être formalisée de manière plus systématique. Pour cela, il faudrait reprendre partiellement le cadre développé dans la section précédente ???. Rappelons les éléments clés et comment

ceux-ci peuvent s'articuler. L'aspect principal est le couplage d'une formalisation du modèle du système avec celle de la perspective. Une perspective serait définie comme une *Dataflow Machine M* au sens de [golden2012modeling] qui donne un moyen pratique pour la présenter et pour introduire les échelles de temps et les données, à laquelle est associée une ontologie O au sens de [livet2010], i.e. un ensemble d'éléments dont chacun correspond à une entité (qui peut être un objet, un agent, un processus, etc.) du monde réel. Le motif et l'agent porteur de la perspective sont contenus dans l'ontologie s'ils font sens pour étudier le système. Décomposer l'ontologie en éléments atomiques $O = (O_j)_j$ et introduire une relation d'ordre entre les éléments des ontologies basée sur l'émergence faible ($O_j \succcurlyeq O_i$ si et seulement si O_j émerge faiblement de O_i) devrait fournir une décomposition canonique de la perspective contenant la structure du système. Le défi serait ensuite de lier cette décomposition avec la décomposition canonique de la *Dataflow Machine* postulée par [golden2012modeling], et ensuite définir les domaines de connaissance au sein de ce couplage : les données sont dans les flots des machines, le modèle est la machine, l'empirique et le théorique dans les ontologies, les méthodes dans la structure de l'arbre. Une telle entreprise avec des opérations cohérentes entre les éléments est cependant hors de notre portée pour l'instant, mais serait un développement puissant.

Nous avons étudié par des méthodes mixtes la construction d'une théorie scientifique en géographie théorique et quantitative, et à partir de cela introduit de manière inductive un cadre de connaissances visant comprendre la production de connaissances sur un système complexe comme un système complexe elle-même, plus précisément une perspective avec des composantes co-évolutives au sein de domaines de connaissances interdépendants. On peut noter que cette approche est totalement réflexive puisque plusieurs de ces composantes ont été nécessaires. Nous postulons que ce cadre peut être un outil utile pour étudier la complexité et gérer des systèmes complexes, puisqu'il explicite certains choix et directions de développements qui pourraient autrement être inconscients.

Co-construction des théories et modèles en géographie quantitative : une synthèse de nos 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 prendrons et défendrons tout au long de ce travail est celle d'une science anarchiste proposée par FEYERABEND, qui sans être prise purement 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 ne présente généralement que peu d'intérêt de par le caractère contraint de l'exercice (combien d'interminables chapitres "état de l'art" et "problématique" ou "enjeux sociaux" témoignent d'une platitude au point de vouloir arrêter la lecture d'un ouvrage par ailleurs remarquable, ce qui s'est sûrement passé dans notre cas d'ailleurs), et paraît relativement vaine vu la destinée de prendre la poussière dans une étagère obscure d'un laboratoire obscur, sans être sauvé par la mise en ligne vu la langue imposée⁸. On se rêve d'imaginer 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 scientifiques est un défi urgent à part entière, et notre ébauche de réflexivité développée en Appendice 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 [livet2010]. 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 [pumain1997pour] jusqu'aux résultats matures présentés dans [pumain2012multi]), l'étude du caractère fractal des structures urbaines (par exemple de [frankhauser1998fractal] à [frankhauser2008fractal]) ou plus récemment le projet Transmon-dyn visant à enrichir la notion de transition des systèmes de peu-

C : sur la communication pour l'extérieur
[Martinez-Conde@1082017]

⁸ Ce qui relève bien sûr par ailleurs d'une problématique bien plus complexe que la simple audience [tardy2004role] 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 que le discours du consul et son snobisme d'énarque rapportés en 1.2.

plement (ouvrage à paraître). Cette communication propose un format original en s'inscrivant dans cette lignée, par la synthèse de différents travaux empiriques et de modélisation menés conjointement avec l'élaboration d'appareils théoriques visant à mieux comprendre les relations entre territoires et réseaux de transports. L'originalité de cette contribution réside à la fois dans la synthèse de travaux très divers pourtant reliés en filigrane, et dans la proposition d'une théorie géographique spécifique s'appuyant sur cette synthèse en seconde partie.

POURQUOI UNE THÉORIE ET DES MODÈLES DE CO-ÉVOLUTION
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:tel-00459720], la littérature est très pauvre en modèles de simulation endogenéisant cette co-évolution. Une exploration algorithmique de la littérature a été menée dans [raimbault2015models], 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 [raimbault2016indirect], 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.

ETUDES EMPIRIQUES 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 [raimbault2016cautious]. 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 [raimbault2016investigating], conjointement avec l'hétérogénéité des processus pour un autre type de système [raimbault2015hybrid]. Ces faits stylisés valident pour l'instant l'utilisation de modèles de simulation complexes, pour lesquels des premiers efforts de modélisation ont ouvert la voie vers des modèles plus élaborés.

MODÉLISATION A l'échelle mesoscopique, des processus d'agrégation-diffusion ont été prouvés 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 [**raimbault2016generation**], 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 entamée [**raimbault2015labex**], 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 [**raimbault2016models**]. Cet ensemble de modèles se positionne avec un objectif de parcimonie et dans une perspective d'application en multi-modélisation. Dans une démarche basée-agent plus descriptive et donc par un modèle plus complexe, [**le2015modeling**] décrit 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. Même si ce dernier modèle est toujours en exploration, les premières études de la dynamique montre 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 premiers efforts de modélisation, bien qu'ils ne soient pas majoritairement centrés sur des modèles de co-évolution à proprement parler, supportent les premiers fondements théoriques que nous proposons par la suite.

CONSTRUCTION D'UNE THÉORIE GÉOGRAPHIQUE Nous revoyons enfin sous l'oeil de la co-evolution des domaines la théories construite en [9.2](#). Nous insistons ici sur son caractère intégratif permettant de joindre Théorie Evolutive 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 [**antelope2016interdisciplinary**]. Pour notre part, la morphogenèse 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 [[doursat2012morphogenetic](#)]). 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 morphogenèse urbaine qui suffisait à reproduire des motifs de forme stylisés [[raimbault2014hybrid](#)], nous suggèrent la possible existence de tels processus au sein des systèmes territoriaux. D'autre part, le cadre d'un 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 [[raffestin1988repères](#)], 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 Réseaux*, 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 [[dupuy1987vers](#)]. Nous spéculons 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 la présence 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 correlations, causalités (dans le cas de séries temporelles) et recherche de décompositions modulaires des systèmes.

★ ★

*

CONCLUSION DU CHAPITRE

Dans une logique de lecture linéaire, cette ouverture par l'introduction de cadres théoriques selon divers points de vue, devrait avoir synthétisé et rassuré sur les questions ouvertes a priori réglées dans leur majorité - seul la conclusion pouvant encore apporter une chute dans la narration. Il s'agit d'un malentendu, et le lecteur qui voudrait être rassuré aurait du s'arrêter au Chapitre précédent, à la fin duquel nous avions fait un tour relativement conséquent des approches proposées. Ce chapitre ouvre en fait un gouffre, et fait prendre conscience que la portée des connaissances est extrêmement embryonnaire. Pour donner une allégorie, nous serions un peu dans la situation du périphérie de Mercure et du spectre de l'atome qui étaient des détails négligeables pour la physique classique à la fin du 19ème siècle, et ont mené aux gigantesques développements au cours du 20ème que sont la physique quantique et la relativité générale. 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 9.2 pointe à nouveau la question de la non-stationnarité spatio-temporelle et la non-ergodicité dans un contexte multi-échelle, que nous postulons cruciale mais très peu comprise. On distingue aussi la difficulté d'intégration de théories existantes ce qui implique une compréhension le couplage de modèle. Ce problème est au cœur du cadre formel développé par la suite ??, 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 "industriel") avec celles de systèmes complexes naturels. La possibilité de théorie intégratives est soulevée par l'introduction du cadre de connaissance 9.3, 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 question ouvertes dans un projet de recherche cohérent sur un long terme mais incluant des premières pistes concrètes immédiates, que nous présenterons en ouverture.

* * *

*

CONCLUSION DE LA PARTIE IV

Perspectives pour la co-évolution : recensement systematique des données, benchmarks plus systématiques des modèles ; etc ; exploration des modèles ; plus d'interdisc ; communication etc : justifie ces cadres.

CONCLUSION

Nous concluons en proposant de formuler les questions ouvertes fondamentales qui se dégagent de notre travail.

OUVERTURES

PERSPECTIVES THÉMATIQUES ET GÉNÉRALES

Développement Spécifiques

Le mode de communication scientifique actuel est loin d'être optimal et 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 recherche (*Journal of Brief Ideas*). 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, et encore plus les articles qui lui sont associés.

Nous proposons dans cette section un exercice pas forcément conventionnel : proposer des idées et développements possibles, en s'efforçant de concrétiser les questions de recherche et/ou points techniques autant que possible, afin que ceux-ci ne s'apparentent pas à une bouteille à la mer.

*Epistémologie Quantitative**Modèles Multi-scalaires**Vers des Modèles Opérationnels*

VERS UN PROGRAMME DE RECHERCHE

Pour une Géographie Intégrée Alternative

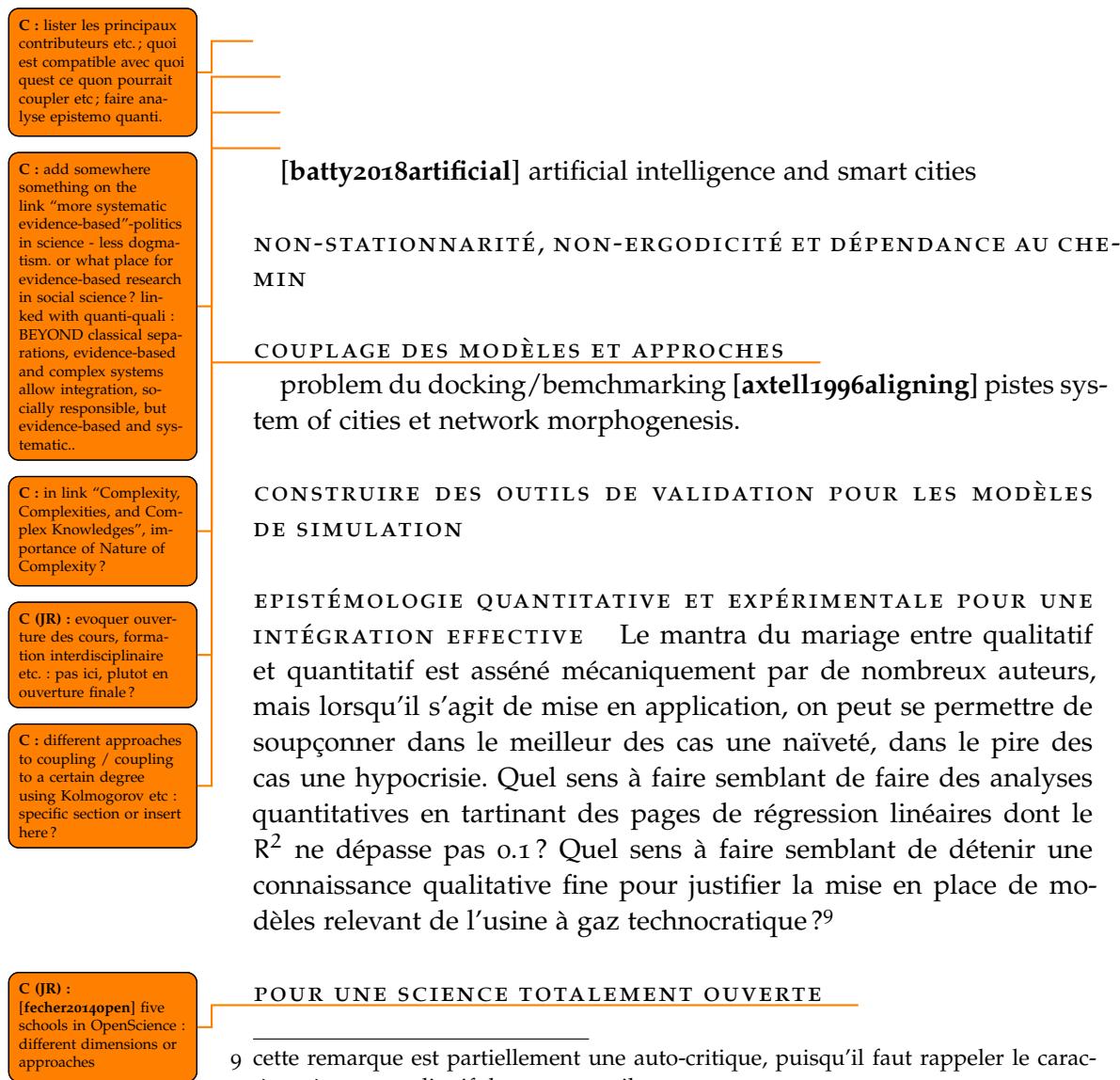
Comme déjà souligné en citant REY, les bouleversements techniques et méthodologiques qu'une discipline peut subir sont souvent accompagnés de profondes mutations épistémologiques, voire de la nature même de la discipline. Il est impossible de juger si l'état actuel des connaissances est transitoire, et s'il l'est quelle est le régime stable qui terminerait la transition s'il en existe un. La spéulation est le seul moyen de lever partiellement le voile, sachant que celle-ci sera nécessairement auto-réalisatrice : proposer des visions ou des programmes de recherche oriente les moyens et questions. L'incomplétude théorique en physique, lorsqu'il s'agit par exemple de lier relativité générale et physique quantique, c'est à dire le microscopique stochastique au macroscopique déterministe, orientent les visions du futur de la discipline qui elle-même conditionnent les actions concrètes qui dans ce domaine sont indispensables (financement du CERN ou de l'interféromètre d'ondes gravitationnelles spatial LISA). En géographie, même si les investissements techniques sont incomparables, ceux-ci existent (accès aux moyens de calcul, financement de laboratoires intégrés, etc.) et sont déterminés également par les perspectives pour la discipline. Nous proposons ici une vision et un manifeste d'une nouvelle géographie, qui est déjà en train de se faire et dont les bases sont solidement construites petit à petit. L'aventure de l'ERC Geodiversity en est l'allégorie, d'autant plus qu'elle a confirmé la plupart des directions professées par BANOS [[banos2017knowledge](#)]. L'intégration de la théorie, de l'empirique, de la modélisation, mais aussi de la technique et de la méthode, n'a jamais été aussi creusée et renforcée que dans les divers développements du projet. Sans l'accès à la grille de calcul et aux nouveaux algorithmes d'exploration permis par Open-Mole, les connaissances tirées du modèle SimpopLocal auraient été moindres, mais les développements techniques ont aussi été conduits par la demande thématique.

Nous proposons un cadre de connaissances pour les études ayant une composante quantitative, ou plus précisément se posant dans la lignée de la Géographie Théorique et Quantitative (TQG). Ce cadre tente de répondre aux contraintes suivantes : (i) transcender les frontières artificielles entre quantitatif et qualitatif ; (ii) ne pas favoriser de composante particulière parmi les moyens de production de connaissance (aussi divers que l'ensemble des méthodes qualitatives et quantitatives classiques, les méthodes de modélisation, les approches théoriques, les données, les outils), mais bien le développement conjoint de chaque composante. Nous étendons le cadre de connaissances de [[livet2010ontology](#)], qui consacre le triptyque des domaines empi-

C : sur l'evidence-based : même le sub-jectif est objectif en un sens ? question d'honnetete et d'intégrité intellectuelle - lié nature connaissance, à développer, arrêter les arnaques quel que soit le type de méthode, rigueur et reproductibilité à mettre en place.

riques, conceptuels et de la modélisation, en y ajoutant les domaines à part entière que sont les méthodes, les outils (qu'on peut voir comme des proto-méthodes) et les données. Les interactions entre chaque domaine sont détaillées, comme par exemple le passage des méthodes vers les outils qui consiste en l'implémentation, ou le passage de l'empirique aux méthodes comme prospection méthodologique. Toute démarche de production de connaissance, vue comme une *perspective* au sens de [giere2010scientific], est une combinaison complexe des six domaines, les fronts de connaissance dans chacun étant en coévolution. Nous nommons notre cadre de connaissance *Géographie Intégrée*, pour souligner à la fois l'intégration des différents domaines mais aussi des connaissances qualitatives et quantitatives, puisque les deux se fondent dans chacun des domaines.

Axes de Recherche



⁹ cette remarque est partiellement une auto-critique, puisqu'il faut rappeler le caractère très peu qualitatif de notre travail

C : brosser ici directions vers lesquelles travailler; intégrer faits dans positionnements

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 l'assène BANOS qui y dédie un de ses commandements, "le modélisateur n'est pas le gardien de la vérité prouvée", et comme rappelé en chapitre ??, 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 comme l'a introduit 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.

Les observateurs soulevant le caractère détraqué du mode actuel de publication scientifique sont nombreux. Un papier n'est pas un format compréhensible ni vraiment reproductible, et pousse au biais. Comme me le rappelait un ami qui s'est spécialisé de manière admirable dans l'acceptation de papiers extrêmement techniques par des *top-journals* économiques, écrire de façon à être accepté est "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 nécessaires à une science éthique et ouverte. De la même façon que nous soutenons qu'une présentation linéaire d'un travail de thèse est trop fortement réducteur

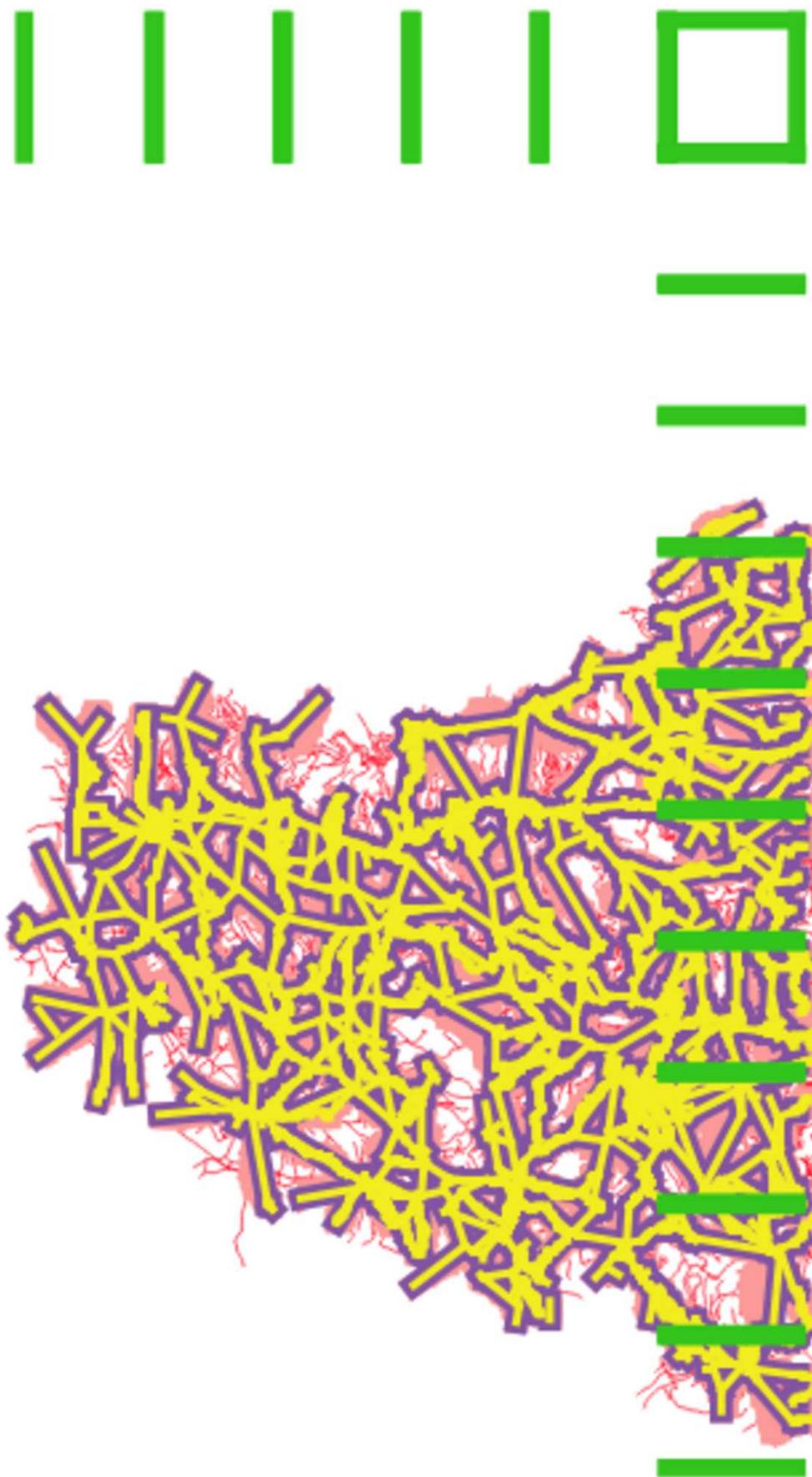
CONCLUSION

*Explorer sans relâche les systèmes géographiques...
- ARNAUD BANOS*

Le lecteur qui aura tenu jusqu'ici et qui a la mémoire solide ou bien sélective, ou encore qui aura adopté un style de lecture roman policier, se plaindra du manque d'originalité dans l'origine des citations introductives. Ce n'est pas anodin si les positions de BANOS, simples mais efficaces et profondes, ouvrent et ferment 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 Demon 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.

Notre contribution thématique est également difficile à situer et nécessitera un recul considérable pour apprécier 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 partis ou les positionnement de type "poule-et-oeuf", mais que nous avons noué une autre bien plus considérable.



Cinquième partie

APPENDICES

Les appendices sont organisées dans la logique des domaines de connaissance : après une présentation linéaire des diverses informations supplémentaires pour chaque section du texte principal, nous introduisons des développements méthodologiques (domaine des méthodes), des développements thématiques (domaine empirique), une synthèse des logiciels développés (domaine des outils), une synthèse des jeux de données construits (domaine des données). Nous concluons par une courte analyse réflexive du contenu de ce mémoire.

A

INFORMATIONS SUPPLÉMENTAIRES

Cette annexe regroupe divers matériaux supplémentaires, nécessaire à la robustesse des études mais pas à l'argumentaire général. Elle inclut par exemple des explorations plus précises de modèles et des analyses de sensibilité.

A.1 ETUDES DE CAS

A.1.1 Terrain en Chine

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. 68 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. 69 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. 70 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).



FIGURE 68: Localisation des lieux de terrain, à l'échelle du sud de la Chine. Source : OpenStreetMap.



FIGURE 69: Localisation des lieux de terrain, à l'échelle du Delta de la Rivière des Perles. Source : OpenStreetMap.

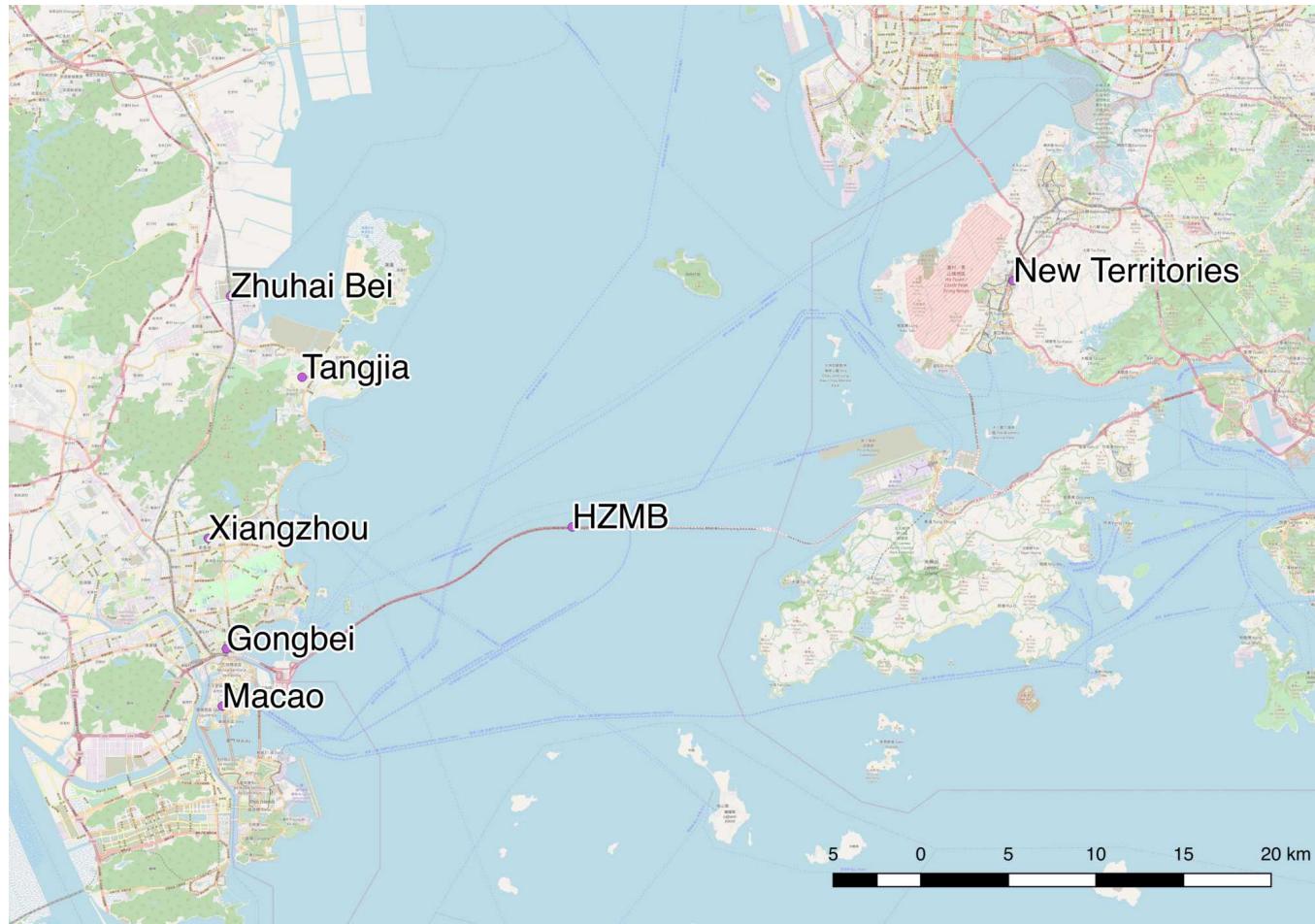


FIGURE 70: Localisation des lieux étudiés, à l'échelle de Zhuhai. Source : OpenStreetMap.

A.2 ELEMENTS DE TERRAIN

A.2.1 Carnet de Terrain

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, [goffman1989fieldwork] 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. 68 à Fig. 70. 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'accéssibilité.

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.

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.2.2 *Entretiens*

Les "entretiens" menés relèvent de l'entretien actif non-structuré [holstein2004active] lors d'une mise en situation vécue conjointement. Les difficultés linguistiques de part et d'autre ont pu rendre compliqué les dialogues et nous donnons ici une synthèse des informations acquises.

Dans cette synthèse narrative et subjective, la première personne désigne l'auteur.

12/08/2018 ? est une habitante de Guangzhou, originaire du Guanxi. Nous nous rencontrons au fond du dernier bus retournant 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 paraît-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 criti-

quons 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

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 ? 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'entrainer à Pekin. Originaire du Hebei, elle n'aime pas vivre à Beijing mais son travail l'y oblige. 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 dédicacation au travail, mais se désole d'un manque d'ouverture d'esprit et d'inventivité dans ce travail.

A.3 EPISTÉMOLOGIE QUANTITATIVE

A.3.1 Revue systématique algorithmique

DESCRIPTION DE L’ALGORITHME Soit \mathcal{A} un alphabet (un ensemble arbitraire de symboles), \mathcal{A}^* les mots correspondants (chaînes de longueur finie sur l’alphabet). Les textes de longueur finie sur celui-ci sont donc $T = \bigcup_{k \in \mathbb{N}} \mathcal{A}^{*k}$. Ce qu’on nomme une référence est pour l’algorithme un enregistrement avec des champs textuels représentant le titre, le résumé et les mots-clés. L’ensemble de références à l’itération n est ainsi noté $\mathcal{C}_n \subset T^3$: il s’agit d’un sous-ensemble de triplets de textes. Nous supposons l’existence d’un ensemble de mots-clés \mathcal{K}_n , les mots-clés initiaux étant \mathcal{K}_0 , spécifiés par l’utilisateur¹. Une itération procède de la manière suivante :

1. Un corpus intermédiaire brut \mathcal{R}_n est obtenu par une requête à un catalogue² auquel on fournit les mots-clés précédents \mathcal{K}_{n-1} .
2. Le corpus total est actualisé par $\mathcal{C}_n = \mathcal{C}_{n-1} \cup \mathcal{R}_n$.
3. Les nouveaux mot-clés \mathcal{K}_n sont extraits du corpus par Traitement du Language Naturel (NLP), étant donné un paramètre fixé N_k donnant le nombre de mot-clés extraits à cette étape.

L’algorithme s’arrête quand la taille du corpus ne varie plus (l’expérience sur les requêtes testées montre pour celles-ci que le corpus ne contient plus de nouvelles références après un certain nombre d’itérations) ou quand un nombre maximal d’itérations défini par l’utilisateur est atteint. La figure 7 synthétise le processus général.

IMPLÉMENTATION De par l’hétérogénéité des opérations requises par l’algorithme (organisation des références, requêtes au catalogue, analyse textuelle), le langage Java s’est présenté comme une alternative raisonnable. Le code source est disponible sur le dépôt ouvert du projet³. Les requêtes au catalogue, qui consistent à récupérer un ensemble de références à partir d’un ensemble de mots-clés, sont faites

¹ On pourrait également partir d’un corpus \mathcal{C}_0 , mais il s’agit plutôt de l’esprit de la méthodologie présentée dans la sous-section suivante. Nous nous en tiendrons ici pour cette exploration préliminaire en assumant le caractère arbitraire forcément biaisé de cette spécification. Le choix du corpus initial doit donc être fait en bonne connaissance des domaines existants, et fait nécessairement suite à la revue de littérature de 2.1.

² Le catalogue est une fonction fournissant des références en réponse à une requête composée d’expressions régulières de mots-clés. En pratique, nous utilisons le catalogue bibliographique en ligne Mendeley. La dépendance au catalogue devant sûrement introduire un biais que nous ne pouvons contrôler, une analyse de sensibilité ou le croisement de divers catalogues étant hors de propos pour cette analyse exploratoire.

³ à l’adresse <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/QuantEpistemo/AlgoSR>

via l'API du logiciel Mendeley [**mendeley**] qui permet un accès ouvert à une base de données conséquente. L'extraction des mots-clés est effectuée par techniques d'Analyse Textuelle (NLP) selon le processus donné dans [**chavalarias2013phylometic**], via un script Python qui utilise [**bird2006nltk**].

C : (Florent) avec quels mots clés as tu validé empiriquement la convergence de l'algorithme ?

CONVERGENCE ET ANALYSE DE SENSIBILITÉ

Une preuve formelle de convergence de l'algorithme n'est guère envisageable puisque qu'elle dépendra de la structure empirique inconue des résultats de requête et d'extraction de mots-clés. Il est donc nécessaire d'étudier le comportement de l'algorithme de manière empirique. Comme présenté en Fig. ??, l'algorithme a de bonnes propriétés de convergence mais diverse sensibilités à N_k . Nous étudions également la cohérence lexicale interne des corpus finaux et fonction du nombre de mots-clés. Comme attendu, des valeurs faibles produisent des corpus plus cohérents, mais la variabilité lorsque qu'elles augmentent reste raisonnable.

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 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.3.2 Analyse par hyperréseau

CORPUS INITIAL Le tableau ?? donne la composition du corpus par domaines.

ANALYSE DE SENSIBILITÉ L'analyse de sensibilité permettant de fixer les paramètres optimaux pour le réseau sémantique est montrée en Fig. 72.

★ ★

★

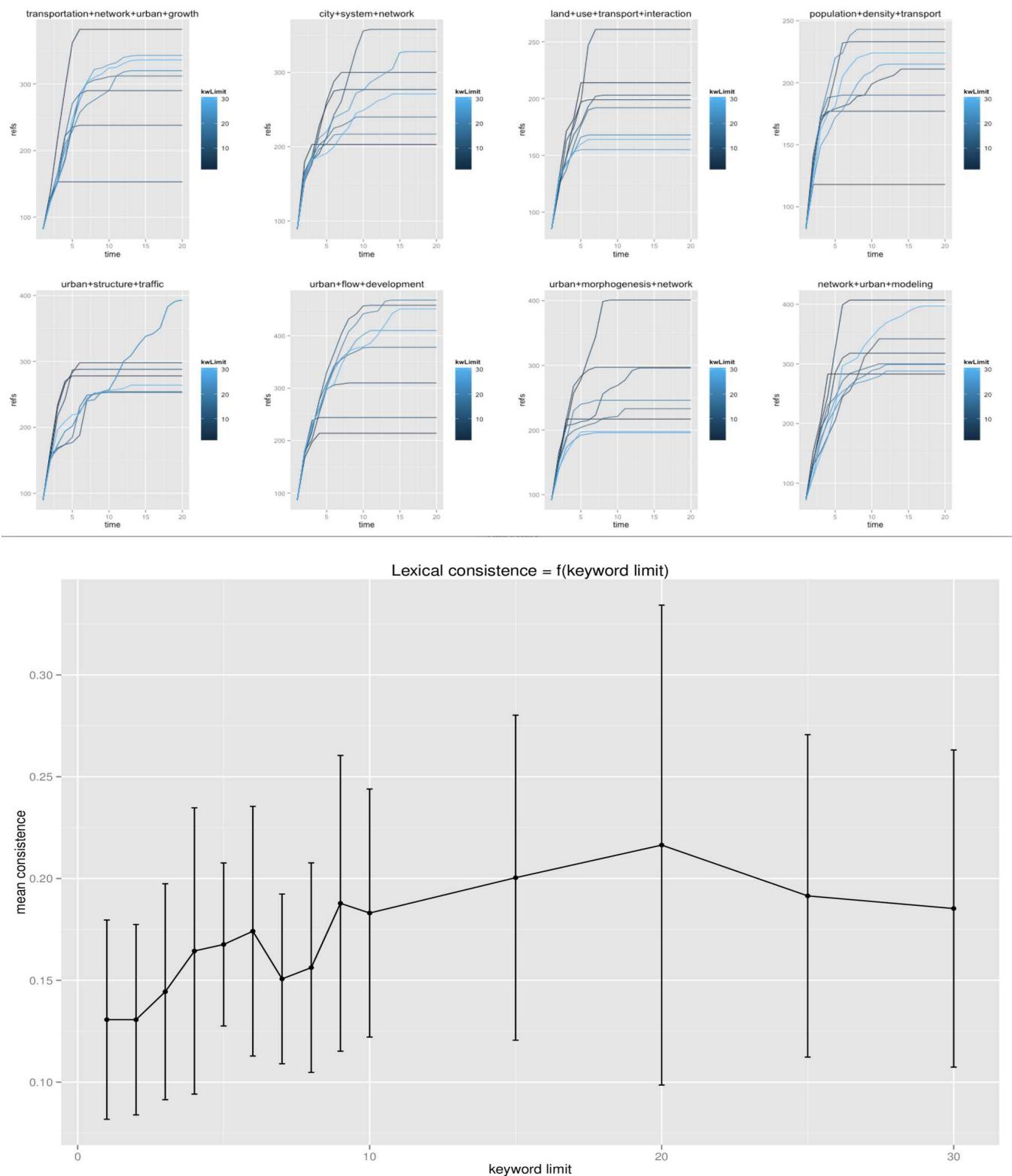


FIGURE 71: Convergence et analyse de sensibilité de l'algorithme de revue systématique. (Haut) Graphes des nombres de références en fonction de l'itération, pour différentes requêtes liées à notre thème, et pour différentes valeurs de N_k (de 2 à 30, couleur). On obtient une convergence rapide dans la majorité des cas, autour de 10 itérations étant nécessaires. Le nombre final de références semble très sensible au nombre de mots-clés selon les requêtes, ce qui confirme une forte variabilité du paysage rencontré selon les termes. (Bas) Consistance lexicale moyenne et déviation standard sur différentes requêtes, en fonction de N_k . La consistance lexicale est définie par les co-occurrences des mots-clés, comme $k = \frac{2}{N_k(N_k-1)} \cdot \sum_{i,j \in \mathcal{K}_f} |c(i) - c(j)|$, avec f temps final, $c(i)$ co-occurrence des mots dans les références. La stabilisation confirme la **consistance des corpus finaux** [3.4.2]

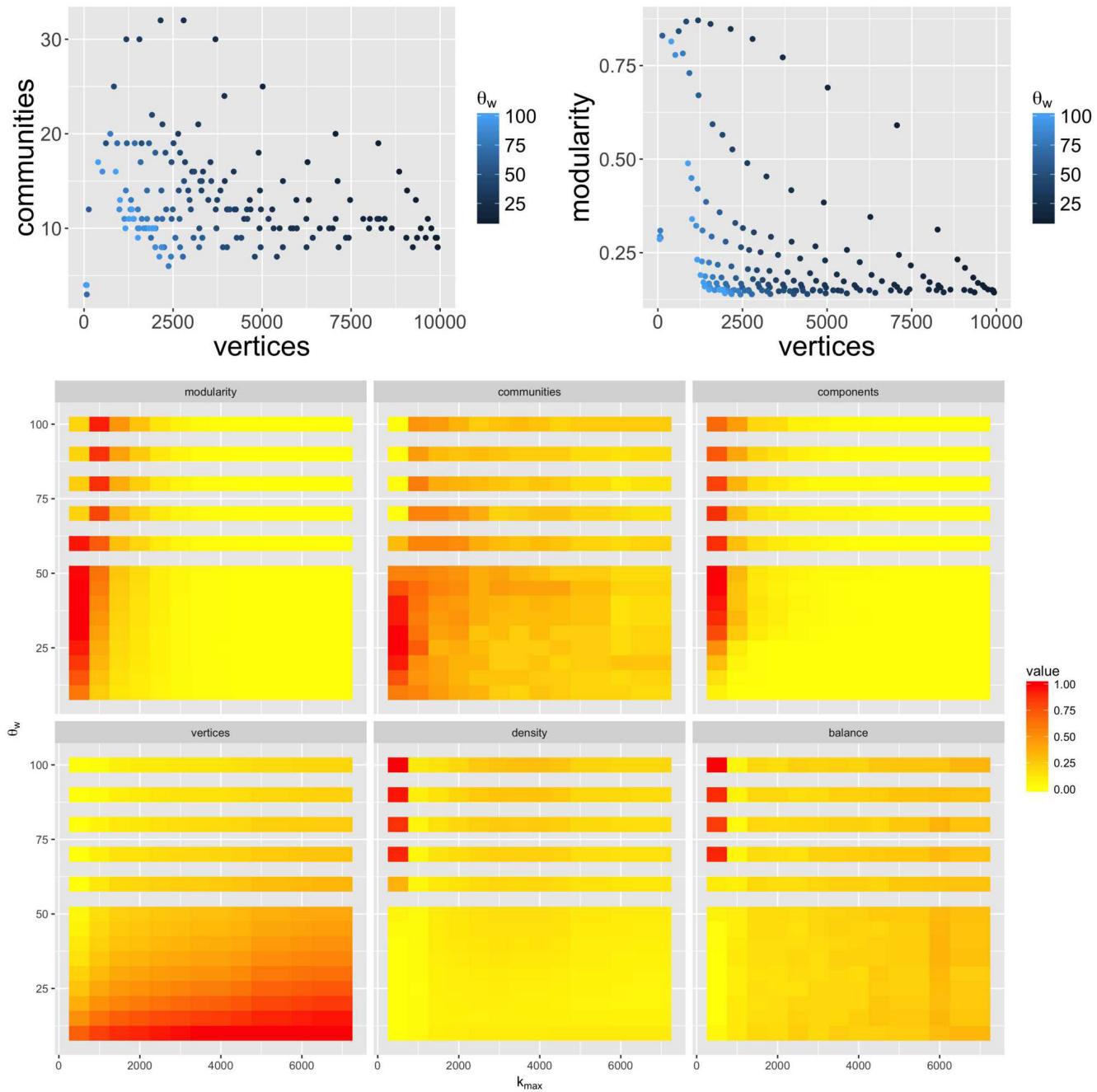


FIGURE 72: Analyse de sensibilité des propriétés modulaires du réseau sémantique en fonction des paramètres de filtrage. (Haut Gauche) Front de Pareto du nombre de communauté et du nombre de sommets (deux objectifs à maximiser), la couleur donnant la valeur de θ_w ; (Haut Droite) Front de Pareto de la modularité en fonction du nombre de sommets, pour θ_w variant; (Bas) Valeurs des objectifs possibles (modularité, nombre de communautés, nombre de composantes connexes, nombre de sommets, densité, équilibre de taille entre communautés), chaque objectif étant normalisé dans $[0; 1]$, en fonction des paramètres θ_w et k_{\max} .

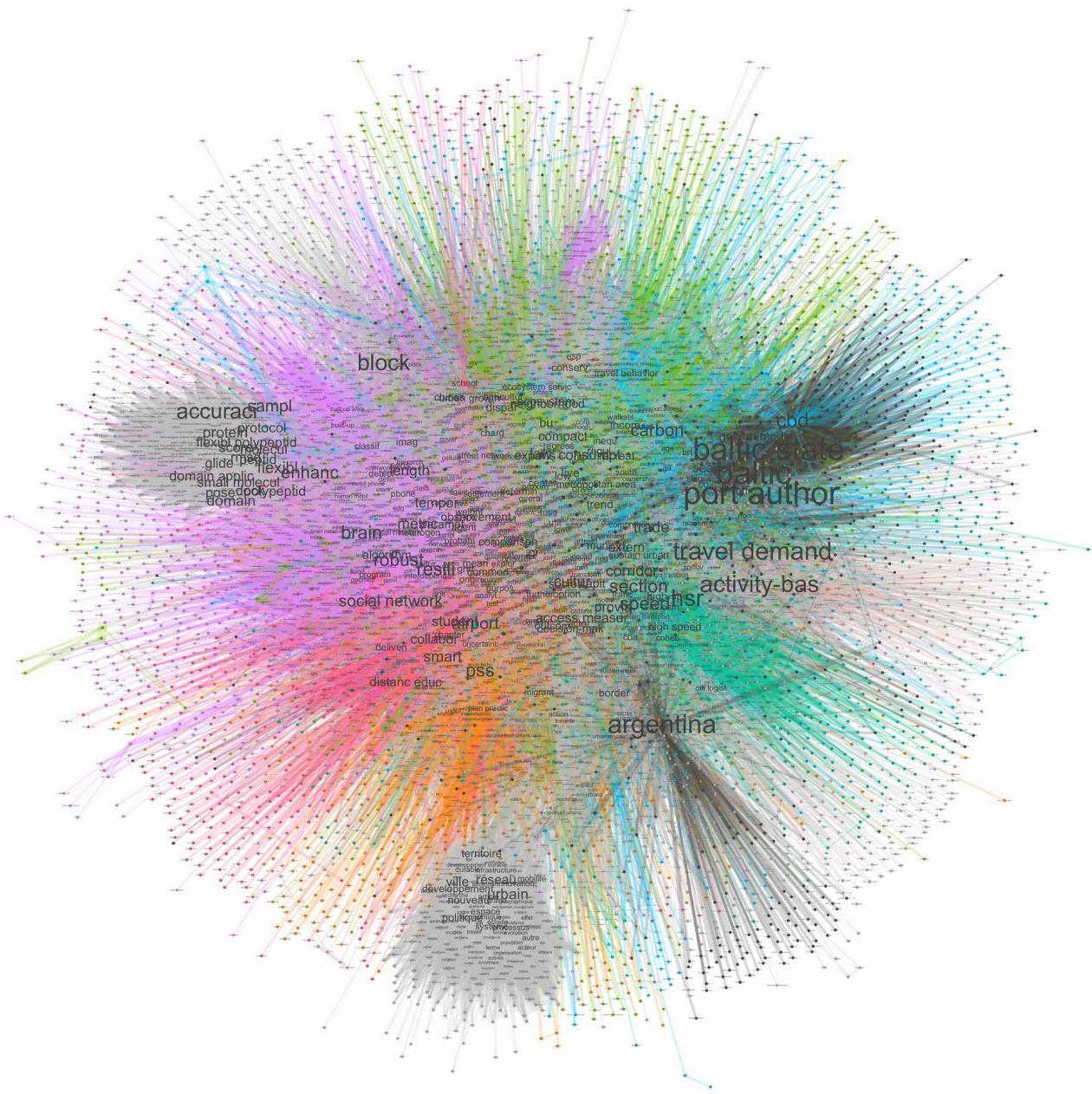


FIGURE 73: Réseau sémantique des domaines. La couleur des liens donne la communauté et la taille des mots-clés est fixée par leur degré.

A.4 MODÉOGRAPHIE

A.4.1 Méthodologie de la revue systématique

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.3).

Première revue du corpus

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.

Remarques sur la classification manuelle

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.4.2 *Meta-analyse*

Nous donnons ici les résultats numériques complets des analyses statistiques reliant caractéristiques de modèles et variables explicatives.

Modalités des variables

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).

Sélection des modèles

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. 74 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.

Ajustement des modèles

INTERDISCIPLINARITÉ L'interdisciplinarité est ajustée selon les modèles linéaires présentés en Table 19.

ECHELLE D'ESPACE L'échelle spatiale est ajustée selon le modèle linéaire dont l'ajustement est donné en Table 20.

ECHELLE DE TEMPS L'échelle de temps est ajustée selon les modèles linéaires présentés en Table 21.

ANNÉE L'année de publication est ajustée selon le modèle linéaire dont l'ajustement est donné en Table 22.

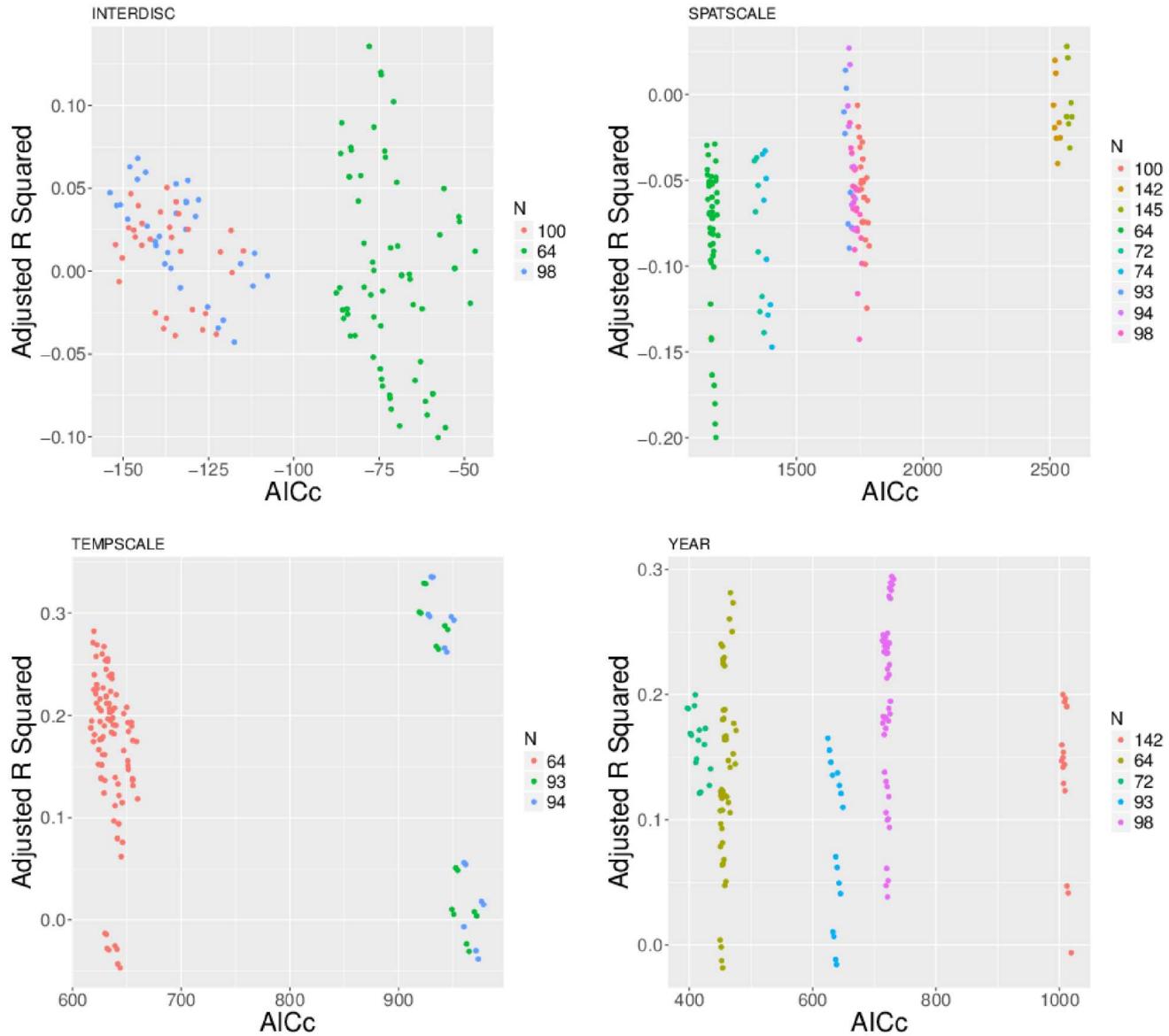


FIGURE 74: Sélection multi-objectif des modèles linéaires. Pour chaque variable à expliquer, nous représentons la position de l'ensemble des modèles linéaires dans l'espace des objectifs (critère d'Akaike corrigé AICc et R^2 ajusté). La couleur des points donne le nombre d'observations.

TABLE 19: Modèles linéaires pour l'interdisciplinarité.

	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 = 0.300	
DISCIPLINEtransportation	0.062 (-0.025, 0.149) p = 0.169	
TYPEstrong		-0.026 (-0.134, 0.081) p = 0.633
TYPEterritory		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	[22 janvier 2018 at 12:17] Thesis version 3.4.2 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)

TABLE 20: Modèle linéaire pour l'échelle spatiale.

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
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

TABLE 21: Modèles linéaires pour l'échelle temporelle.

	TEMPSCALE	
	(1)	(2)
YEAR	0.674 (-0.294, 1.643) p = 0.179	
TYPEstrong		100.271 (58.312, 142.230) p = 0.00002***
TYPERegion	-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

TABLE 22: Modèle linéaire pour l'année de publication.

	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
FMETHODdro	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
SEMCOMhedonic	-5.769 (-19.931, 8.393) p = 0.430
SEMCOMhsr	6.135 (-9.889, 22.159) p = 0.458
SEMCOMinfra planning	-4.123 (-18.910, 10.663) p = 0.588
SEMCOMnetworks	4.711 (-9.736, 19.158) p = 0.527
SEMCOMtod	[22 janvier 2018 at 12:17 – Thesis version 3.4.2] -1.653 (-15.837, 12.532) p = 0.821
Constant	2.004.945 (1.981.531, 2.028.359)

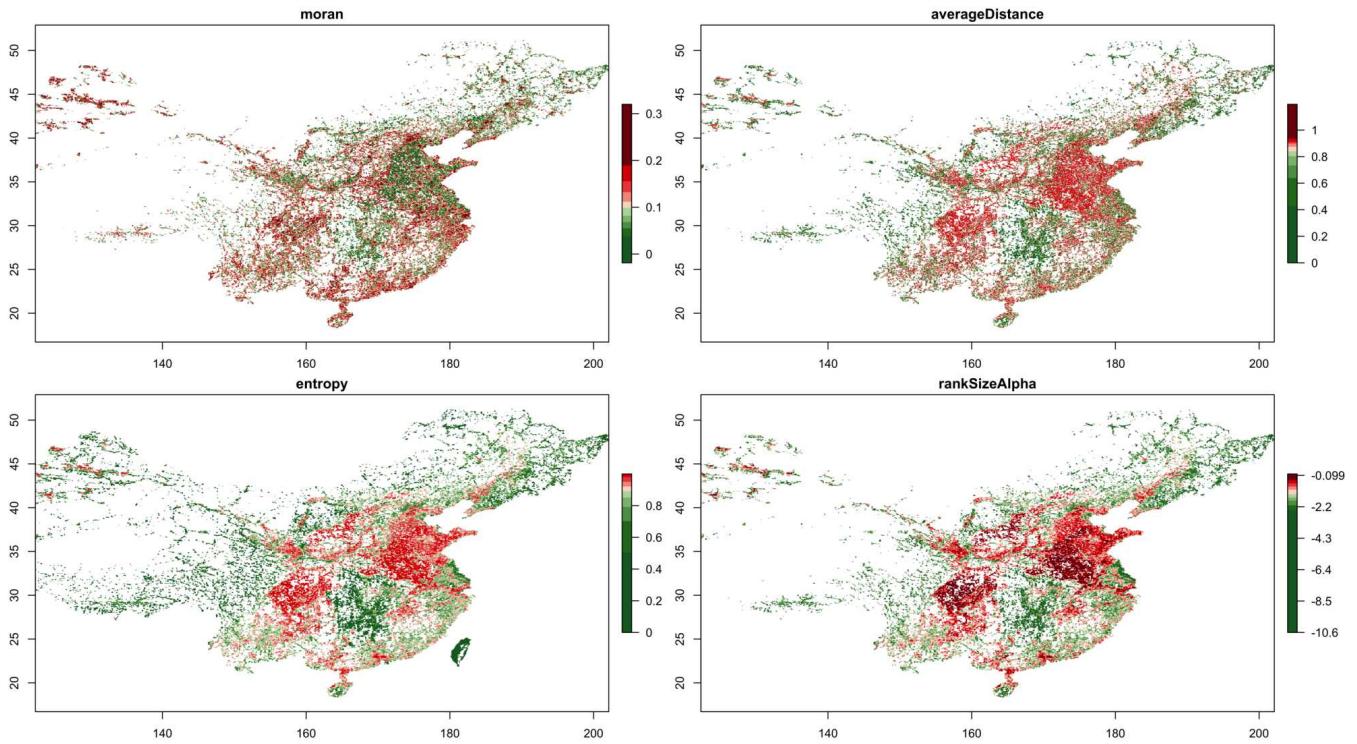


FIGURE 75: Indicateurs morphologiques pour la Chine.

A.5 CORRELATIONS STATIQUES

A.5.1 Mesures morphologiques

A.5.2 Mesures de réseau

Algorithme de Simplification du Réseau

Le réseau est d'abord agrégé à une granularité de 100m pour pouvoir être utilisé de manière cohérente avec les grilles de population. Cela permet par ailleurs d'être robuste aux imperfections locales de codage ou données très locales manquantes. Les données OSM sont importées dans une base de données pgsql en utilisant le logiciel osmosis [osmosis].

More precisely we use the following procedure :

- a background raster (which resolution r gives the snapping parameter for aggregation) is constructed from a reference raster and the extent of network. This grid gives spatial aggregation units for network nodes.
- for each feature of the road dataset, corresponding connected raster cells are stored with corresponding impedance and distance in a sparse adjacency matrix.

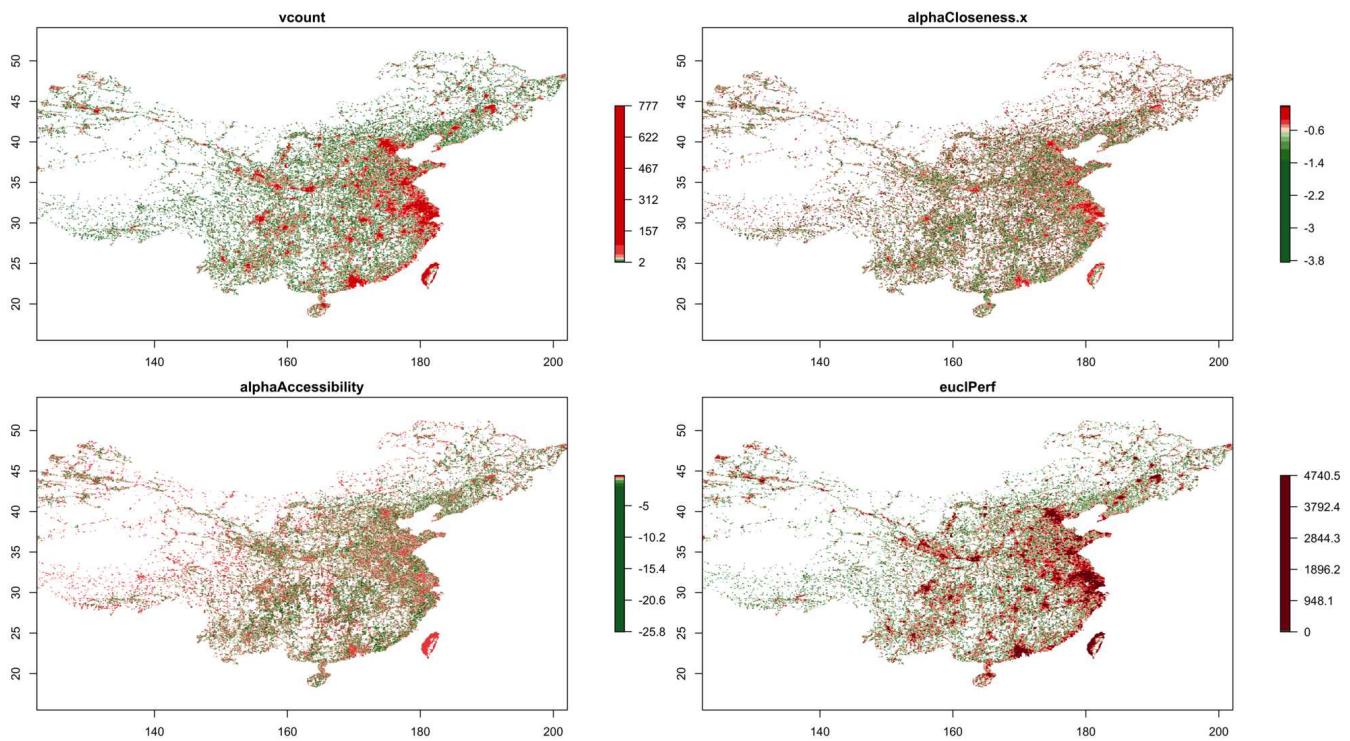


FIGURE 76: Indicateurs de réseau pour la Chine.

- Network is simplified by iterative suppression of nodes with degree two, with keeping link speed and real length to their effective value.

IMPLÉMENTATION A PostGIS database is used to store raw and simplified network, in order to perform efficient spatial requests, compared for example to initial osm data formats (osm or pbf). However the size of storage of data into this base is much higher (factor 10) so processing was parallelized between european countries. Consistence is ensured by the use of the same common density raster as simplification canvas. Final network is stored into the Postgis database for efficient indicator computation given a spatial extent.

SENSIBILITÉ AUX PARAMÈTRES DE SIMPLIFICATION Sensitivity of indicators to raster resolution and to degree simplification algorithm must still be tested to ensure the relevance of data preproces sing.

C : (Florent) y'a t'il un effet de bord dans les carrés 50x50 qui se trouvent à la frontière de 2 pays - A : pas avec nouvelle parallelisation pas par pays mais par split and merge (TODO rewrite nouvel algo)

Indicateurs de réseau

A.5.3 Sensibilité à la résolution

Nous évaluons ici la sensibilité des divers indicateurs à la taille de la grille. Nous montrons en Fig. 77 les indicateurs morphologiques et en Fig. 77 certains indicateurs de réseau, cartographiés pour la France, pour des tailles différentes de grille. Les tailles données ici, en écho à celle de 50km utilisée dans les résultats principaux, sont dans des ordres de grandeur équivalents : nous testons des fenêtres de taille 30km et 60km. Les décalages sont à chaque fois de la moitié de la fenêtre (15km et 50km respectivement). Il est possible de voir “à l’oeil” que certains indicateurs sont peu sensibles, le changement d’échelle ressemblant à un lissage du champ le plus fin : par exemple dans le cas morphologique pour l’indice de Moran, l’entropie et la hiérarchie. La distance moyenne, en fait très bruitée à l’échelle la plus faible, est nécessairement sensible à l’agrégation, ce qui est consistant avec une sensibilité attendue au lissage. Les indicateurs de réseau sont relativement robustes à la taille de la fenêtre.

Cette comparaison, d’une part est à prendre avec précaution de par la non-comparabilité directe des échelles pour les indicateurs, et d’autre part reste limitée. Nous proposons alors une méthode pour quantifier la variabilité des indicateurs à la taille de la fenêtre. Soit X_D et X_d deux champs spatiaux correspondant à deux échelles spatiales $D > d$ (qu’on prend comme des distances caractéristiques), qu’on suppose discrets en des points respectifs $(\vec{x}_i^{(D)})_{1 \leq i \leq N_D}$ et $(\vec{x}_j^{(d)})_{1 \leq j \leq N_d}$. L’idée est de comparer un lissage du champ le plus fin au champ le moins fin : si la corrélation entre ces deux valeurs est forte, il est possible de passer d’un champ à l’autre par agrégation et l’échelle de calcul n’influe ainsi pas autrement que sur la résolution finale. Soit $W_{ij} = (\exp -d_{ij}/d_0)_{ij}$ une matrice de poids spatiaux calculée par les distances euclidiennes d_{ij} entre les points $\vec{x}_i^{(D)}$ et $\vec{x}_j^{(d)}$. Alors avec $W'_{ij} = W_{ij} / \sum_j W_{ij}$, on peut calculer le lissage spatial de X_d aux points $\vec{x}_i^{(D)}$, par le produit matriciel

$$\tilde{X}_d(\vec{x}_i^{(D)}) = W' * \vec{x}_j^{(d)}$$

La corrélation est alors donnée par $\rho[\tilde{X}_d, X_D]$ évaluée sur l’ensemble des points $\vec{x}_i^{(D)}$.

La Fig. 79 donne les variations de la corrélation pour l’ensemble des couples (D, d)

A.5.4 Corrélations Spatiales

La Fig. 80 donne la distribution spatiale pour l’ensemble de l’Europe, d’un échantillon de corrélations entre indicateurs : $\rho[\alpha_{cl}, I]$, $\rho[\gamma, \alpha]$, $\rho[bw, \gamma]$.

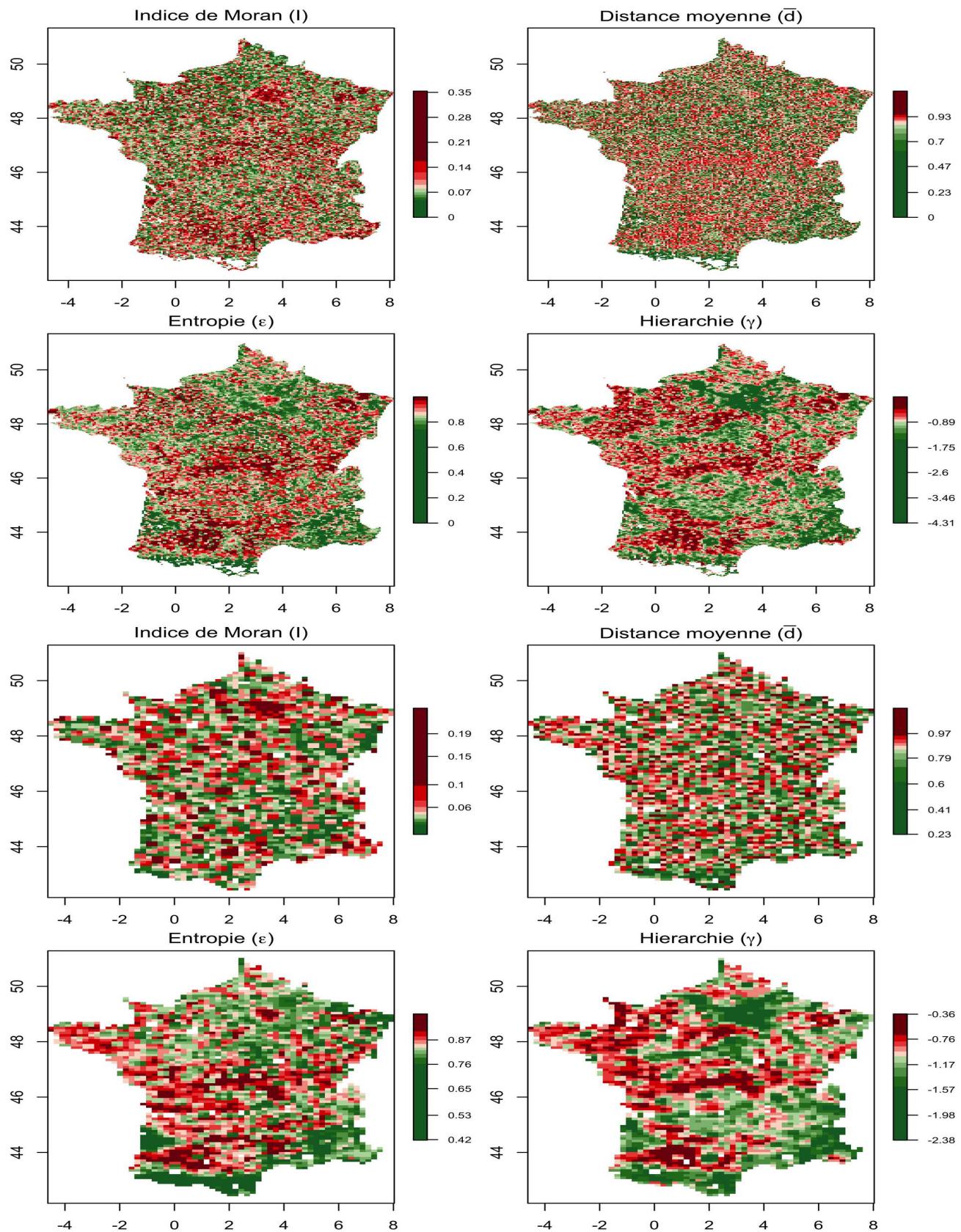


FIGURE 77: Indicateurs morphologiques pour différentes tailles de grille. Les 4 premières cartes montrent les indicateurs calculés avec une fenêtre de taille 30km, les 4 dernières avec une fenêtre de taille 100km.

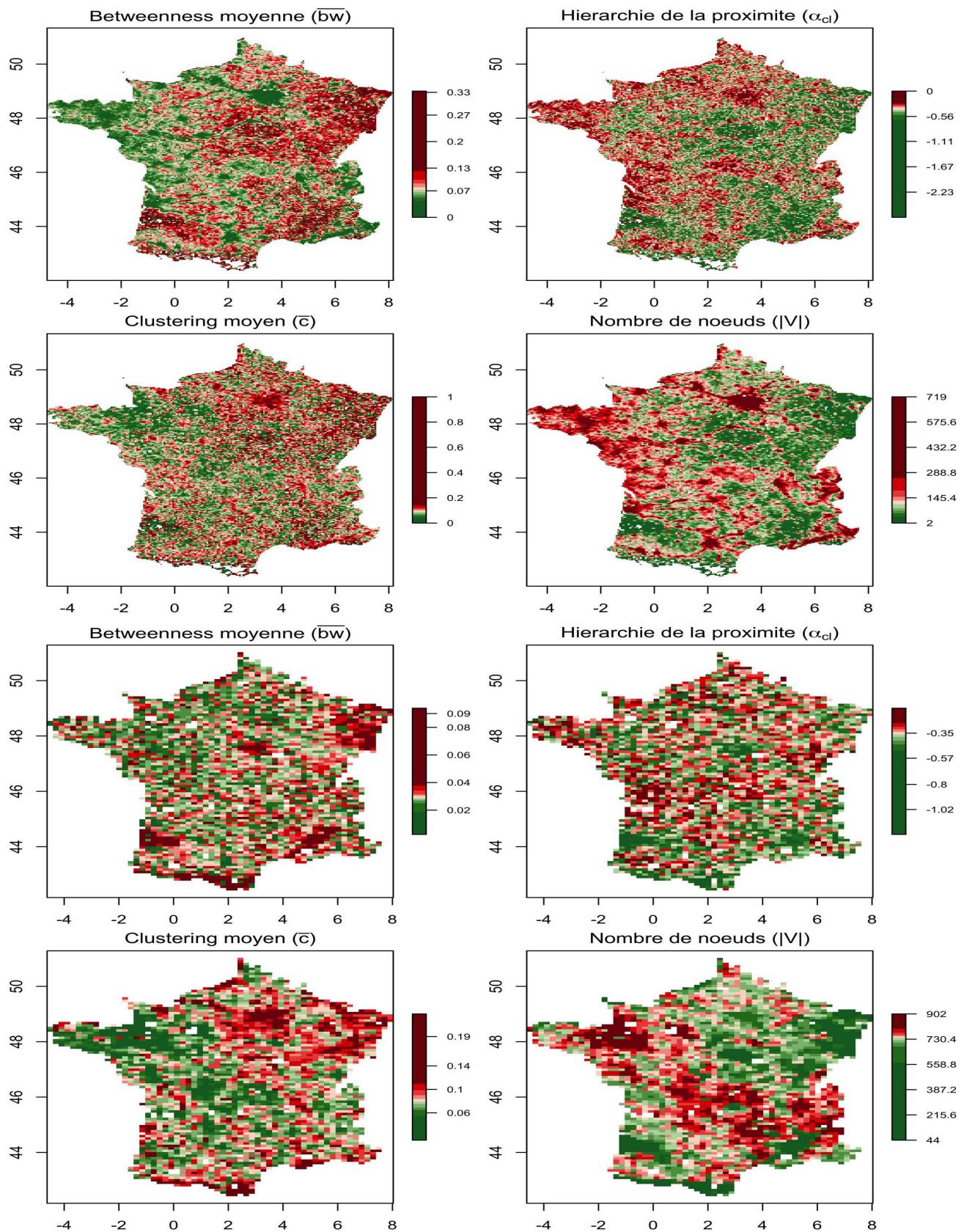


FIGURE 78: Echantillon des indicateurs de réseau pour différentes tailles de grille. Les 4 premières cartes montrent les indicateurs calculés avec une fenêtre de taille 30km, les 4 dernières avec une fenêtre de taille 100km.

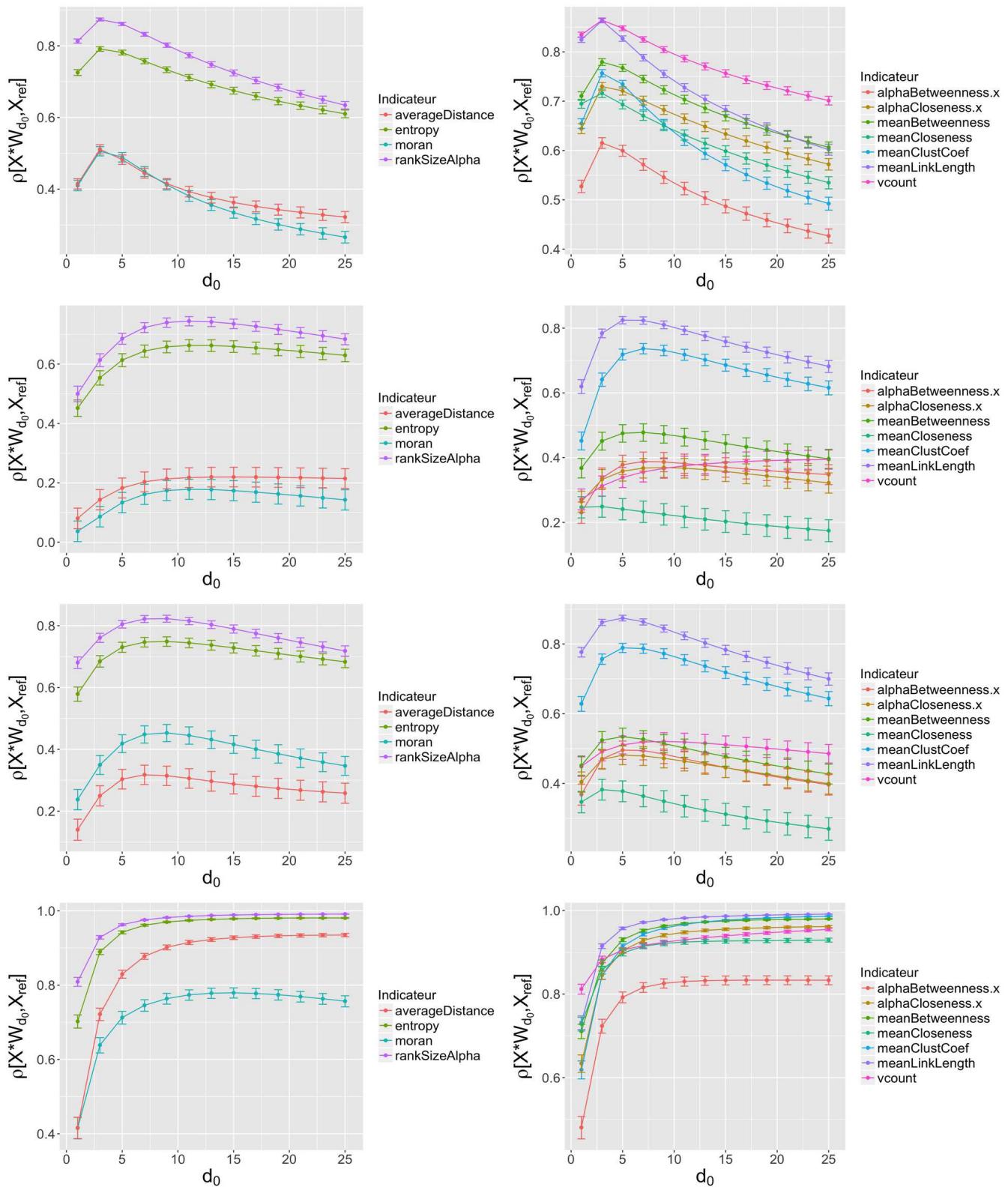


FIGURE 79: Corrélations entre indicateurs à différentes échelles.

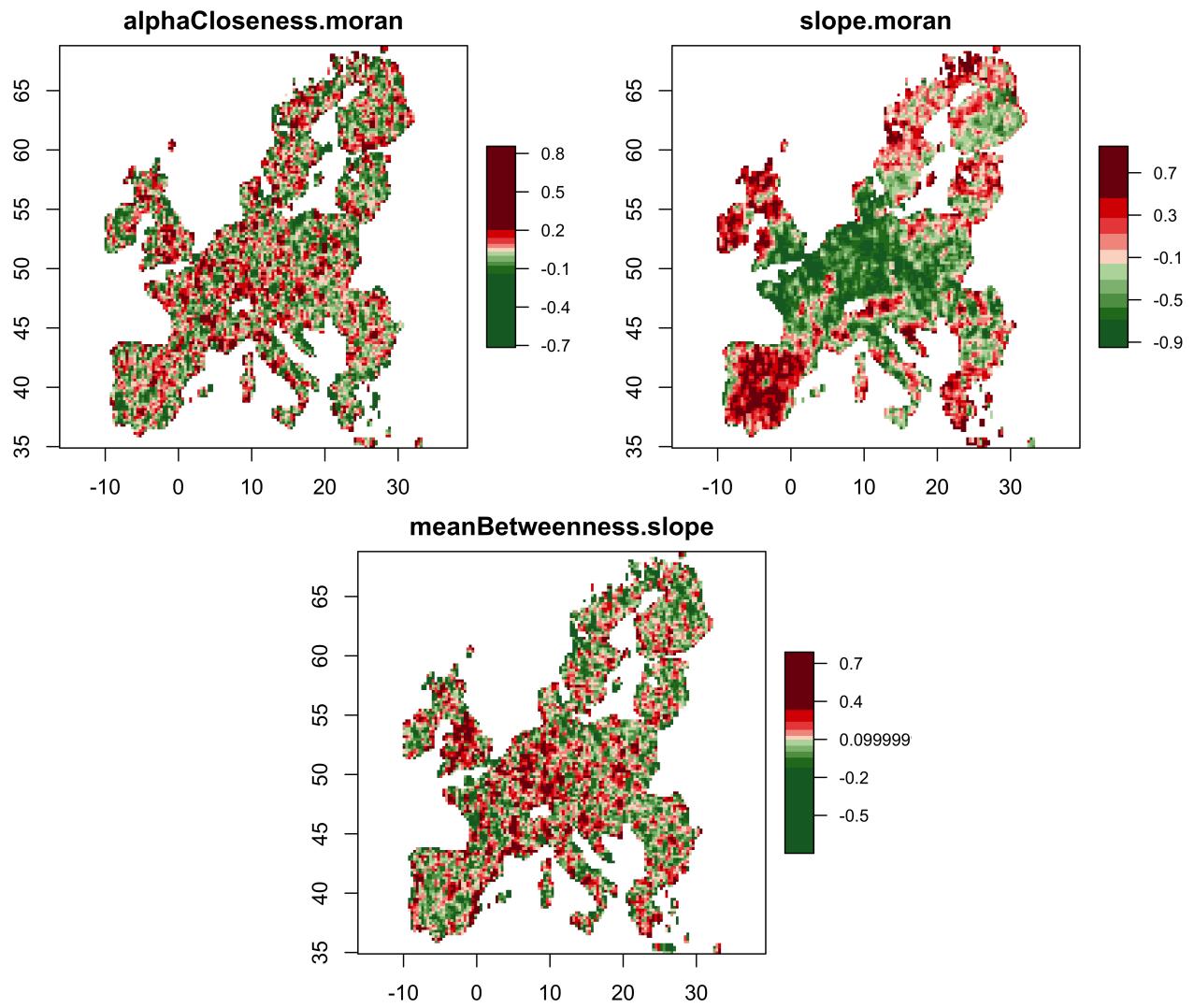


FIGURE 8O: Correlations spatiales pour l'Europe.

La Fig. 81 donne les distributions statistiques des corrélations estimées, pour différentes valeurs de δ .

A.5.5 Multi-scalarité

Estimation des corrélations pour un processus multi-scalaire

Nous proposons ici de relier le caractère multi-scalaire d'un processus stochastique spatio-temporel avec l'estimation de sa matrice de corrélation. Pour simplifier et dans le cadre où ce résultat est utilisé en texte principal, nous considérons des corrélations statiques estimées dans l'espace. Pour simplifier également, considérons des processus ayant deux échelles caractéristiques se superposant linéairement, c'est à dire s'écrivant sous la forme

$$X_i = X_i^{(0)} + \tilde{X}_i$$

avec $X_i^{(0)}$ tendance aux petites échelles ayant une distance caractéristique d'évolution d_0 , et \tilde{X}_i signal évoluant à une distance caractéristique $d \ll d_0$.

Nous pouvons alors calculer la décomposition de la corrélation entre deux processus, de manière similaire à ce qui est fait en C.5. En supposant que $\text{Cov}[X_i^{(0)}, \tilde{X}_j]$ pour tous i, j , et en notant $\varepsilon_i = \frac{\sigma[X_i^{(0)}]}{\sigma[\tilde{X}_i]}$ le rapport des écarts type entre tendance et signal, il y a

$$\begin{aligned} \rho[X_1, X_2] &= \rho[X_1^{(0)} + \tilde{X}_1, X_2^{(0)} + \tilde{X}_2] \\ &= \frac{\text{Cov}[\tilde{X}_1, \tilde{X}_2] + \text{Cov}[X_1^{(0)}, X_2^{(0)}]}{\sqrt{(\text{Var}[X_1^{(0)}] + \text{Var}[\tilde{X}_1]) (\text{Var}[X_2^{(0)}] + \text{Var}[\tilde{X}_2])}} \\ &= \frac{\varepsilon_1 \varepsilon_2 \rho[X_1^{(0)}, X_2^{(0)}] + \rho[\tilde{X}_1, \tilde{X}_2]}{\sqrt{(1 + \varepsilon_1^2)(1 + \varepsilon_2^2)}} \end{aligned}$$

En supposant $\varepsilon_i \ll 1$, on peut développer cette expression au premier ordre et obtenir

$$\rho[X_1, X_2] = \left(\varepsilon_1 \varepsilon_2 \rho[X_1^{(0)}, X_2^{(0)}] + \rho[\tilde{X}_1, \tilde{X}_2] \right) \cdot \left(1 - \frac{1}{2}(\varepsilon_1^2 + \varepsilon_2^2) \right) \quad (21)$$

L'ajout de la tendance au signal introduit ainsi une correction sur la corrélation, d'une part par la prise en compte directe de la corrélation entre tendance atténuee, et d'autre part par le terme d'interférence en facteur.

Pour appliquer ce résultat à notre problématique, supposons que $d \simeq l_0, l_0$ étant la distance minimale d'estimation des corrélations. On a par ailleurs l'échelle de stationnarité d_s qui correspond à l'échelle

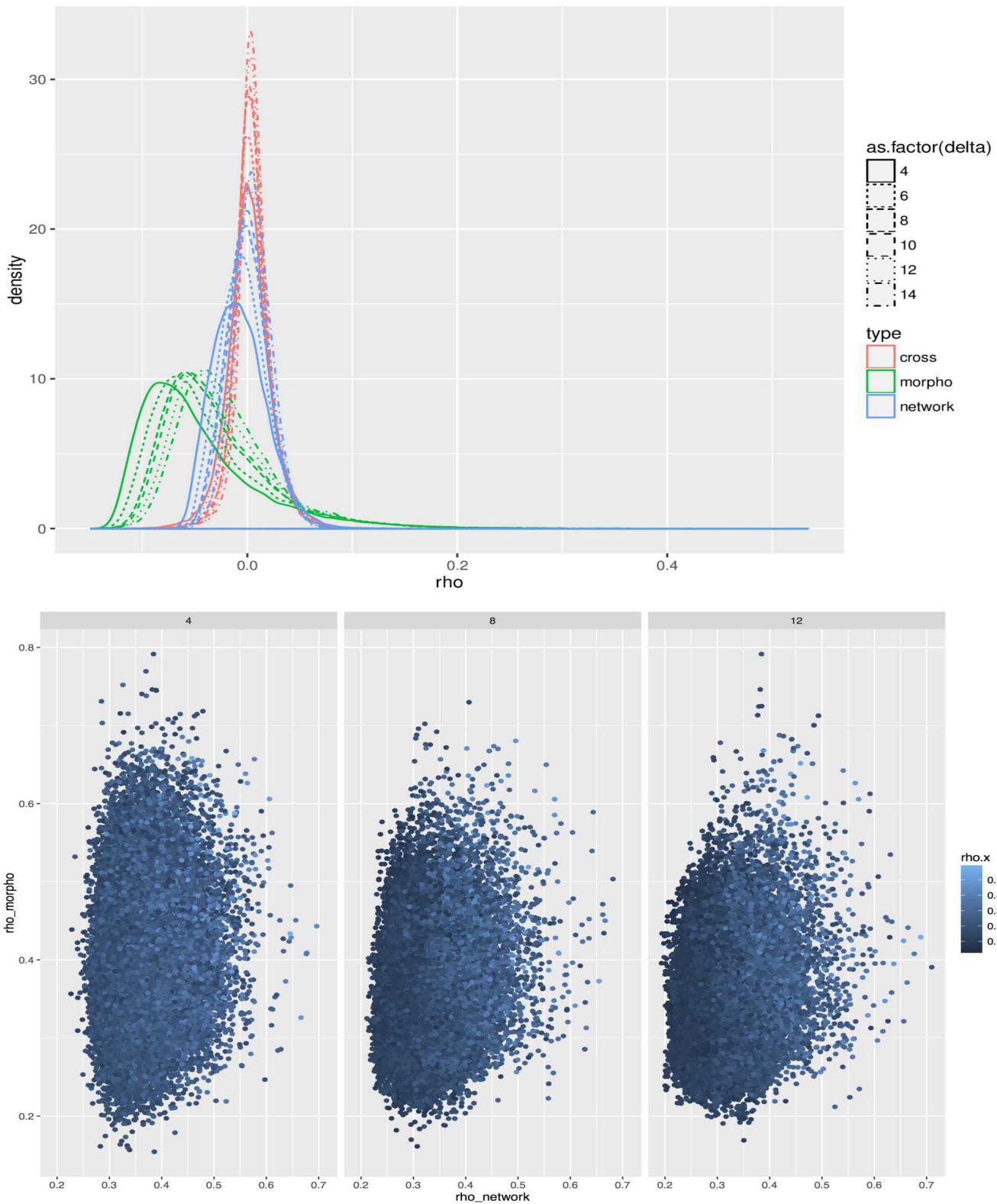


FIGURE 81: Distribution des corrélations. (Haut Gauche) Distribution statistique des corrélations, pour les différents blocs morphologiques, réseau et corrélations croisées (couleur), pour différentes valeurs de δ (type de ligne); (Bas Droite) Correlation absolues moyennes pour le réseau en fonction de la morphologie, niveau de couleur donnant la corrélation croisée, pour différentes valeur de δ .

de variation des corrélations, et selon les résultats empiriques vérifie $d_s > l_0$, significativement au moins pour certains indicateurs (par exemple hiérarchie et Moran, pour laquelle elle est de l'ordre du pays). Enfin, notons $\delta_0 = d_0/d$ l'échelle de la tendance en termes de δ . On suppose donc

$$d < d_s < d$$

Pour les valeurs de δ telles que $\delta \cdot d < d_s$, on devrait avoir $\hat{\text{Cov}}_\delta [\tilde{X}_1, \tilde{X}_2] \simeq \hat{\text{Cov}}_{\delta=1} [\tilde{X}_1, \tilde{X}_2]$ si $\hat{\text{Cov}}_\delta$ est l'estimateur sur la zone de taille δ .

Par ailleurs, on peut supposer raisonnablement que $\hat{\text{Var}}_{\delta=1} [X_i^{(0)}] \ll \hat{\text{Var}}_{\delta=d_s/d} [X_i^{(0)}]$, c'est à dire que la tendance est constante à la plus petite échelle en comparaison des variations aux échelles intermédiaires.

Sous ces hypothèses, l'estimateur de ρ devrait varier en fonction de δ selon les variations de ε_i en fonction de δ . En supposant enfin les tendances très peu corrélées (effets structurels indépendants), on conserve la correction d'interférence dans l'expression de ρ , et donc que $\rho(\delta)$ décroît pour des faibles valeurs de δ .

Nous avons ainsi démontré qu'une structure simple multi-scalaire du processus implique une variation de la corrélation estimée en fonction de δ , sous un certain nombre d'hypothèse. La réciproque n'a a priori pas de raison d'être vraie. Le lien que nous opérons ici est ainsi une illustration pour renforcer une hypothèse, qui est par ailleurs également soutenue par les résultats sur la variation de l'intervalle de confiance décrits par la suite.

Intervalle de confiance pour la corrélation

Nous dérivons ici le comportement de l'estimateur de corrélation en fonction de la taille de l'échantillon. Sous l'hypothèse de distribution normale de deux variables aléatoire X, Y , alors la transformée de Fisher de l'estimateur de Pearson $\hat{\rho}$ telle que $\hat{\rho} = \tanh(\hat{z})$ a une distribution normale. Si z est la transformée de la corrélation réelle ρ , alors un intervalle de confiance pour ρ est de taille

$$\rho_+ - \rho_- = \tanh(z + k/\sqrt{N}) - \tanh(z - k/\sqrt{N})$$

où k est une constante. Comme $\tanh z = \frac{\exp(2z)-1}{\exp(2z)+1}$, on peut développer puis réduire cette expression, pour obtenir

$$\begin{aligned} \rho_+ - \rho_- &= 2 \cdot \frac{\exp(2k/\sqrt{N}) - \exp(-2k/\sqrt{N})}{\exp(2z) - \exp(-2z) + \exp(2k/\sqrt{N}) + \exp(-2k/\sqrt{N})} \\ &= 2 \cdot \frac{\sinh(2k/\sqrt{N})}{\cosh(2z) + \cosh(2k/\sqrt{N})} \end{aligned}$$

En utilisant le fait que $\cosh u \sim_0 1 + u^2/2$ et que $\sinh u \sim_0 u$, on obtient bien que $\rho_+ - \rho_- \sim_{N \gg 0} k'/\sqrt{N}$. ■

★ ★

★

A.6 RÉGIMES DE CAUSALITÉ

A.6.1 Données Synthétiques

Séries temporelles

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

$$\mathbf{A}(\tau_0) = \begin{pmatrix} 0 & a \\ a & 0 \end{pmatrix}$$

avec $-1 < a < 1$. 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

$$\begin{cases} X(t) = a \cdot Y(t - \tau_0) + \varepsilon_X \\ Y(t) = a \cdot X(t - \tau_0) + \varepsilon_Y \end{cases}$$

En prenant la variance dans les deux équations et en faisant la différence, on obtient que nécessairement $\text{Var}[X] = \text{Var}[Y]$ car $a^2 \neq 1$. La somme donne alors $\text{Var}[X] = \text{Var}[Y] = \frac{\sigma^2}{1-a^2}$.

Nous calculons alors

$$\begin{aligned} \rho[X(t), Y(t - \tau_0)] &= \rho[aY(t - \tau_0) + \varepsilon_X, Y(t - \tau_0)] \\ &= \frac{\text{Cov}[aY(t - \tau_0) + \varepsilon_X, Y(t - \tau_0)]}{\sqrt{(a^2 \text{Var}[Y] + \sigma^2) \text{Var}[Y]}} \\ &= \frac{a \text{Var}[Y]}{|a| \text{Var}[Y] \sqrt{1 + \frac{\sigma^2}{a^2 \text{Var}[Y]}}} = \frac{a}{|a| \sqrt{1 + \frac{1-a^2}{a^2}}} \\ &= a \end{aligned}$$

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.

Morphogenèse Urbaine

La Fig. 82 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.6.2 Afrique du Sud

La Fig. 83 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.

★ ★

★

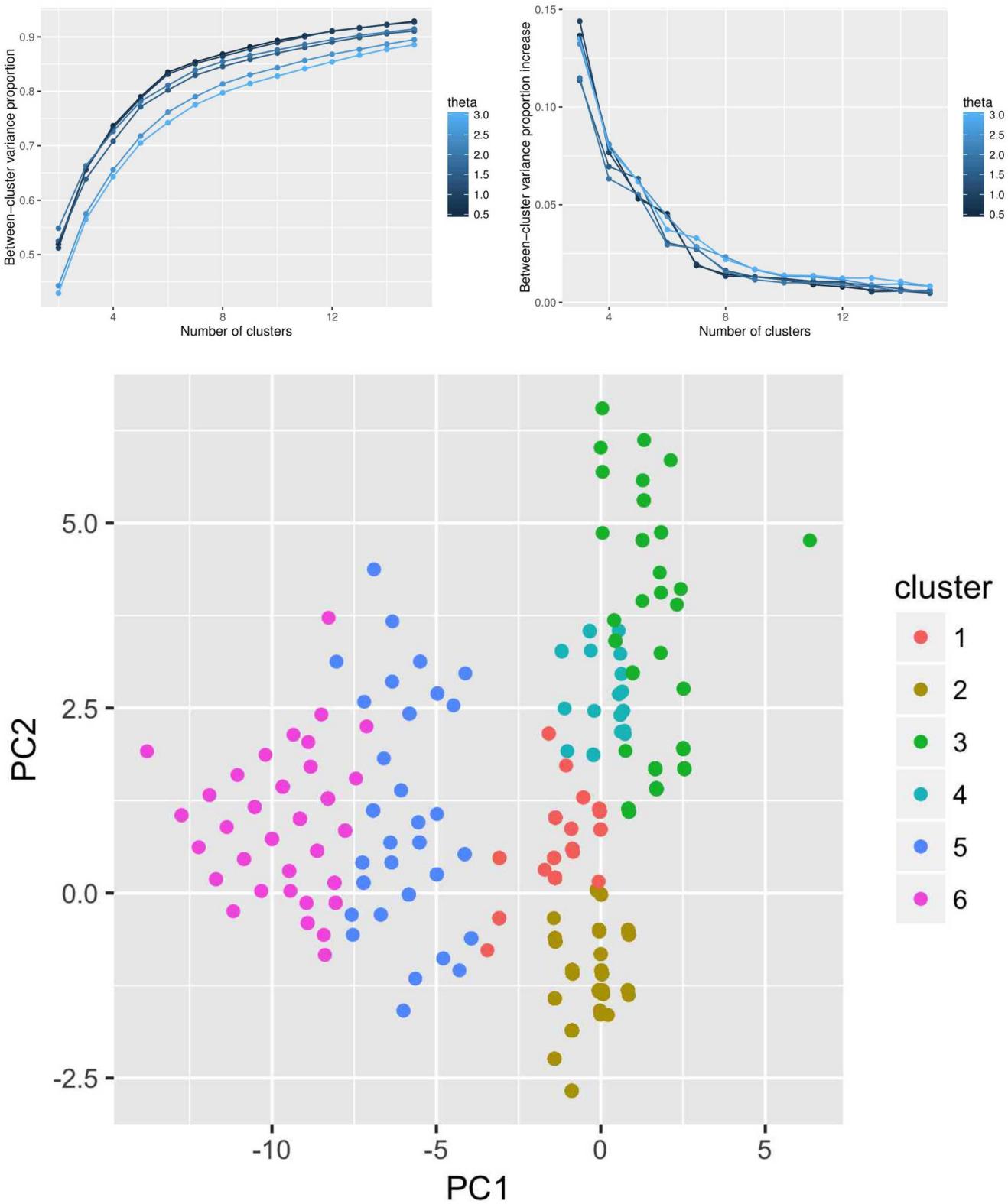


FIGURE 82: Identification de régimes d’interactions endogènes par classification non-supervisée. **(Haut Gauche)** Variance inter-cluster comme fonction du nombre de clusters. **(Haut Droite)** Dérivée de la variance inter-cluster. **(Bas)** Features dans un plan principal (81% de variance expliquée par les deux premières composantes).

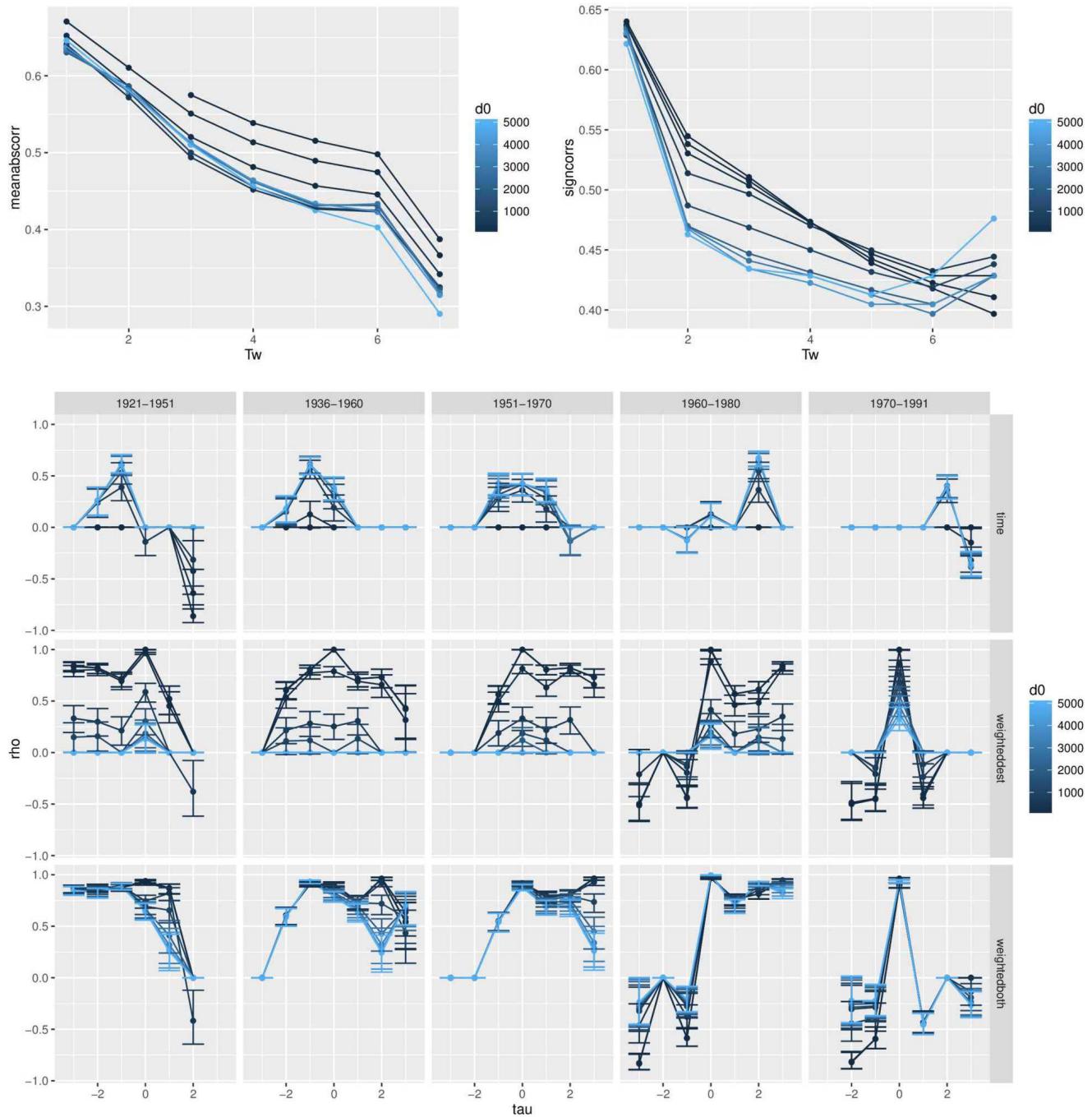


FIGURE 83: (Haut Gauche) Corrélations absolues moyennes sur l'ensemble des retards, en fonction de la taille de la fenêtre temporelle T_W (en nombre d'observations temporelles), pour différentes valeurs du paramètre de décroissance d_0 ; (Haut Droite) Proportion de corrélations significatives, en fonction de T_W pour d_0 variable; (Bas) Corrélations retardées en fonction du délai τ , pour la taille optimale $T_W = 3$, sur les différentes périodes successives (colonnes), pour les différents degrés de pondérations (première ligne $w_i = 1$, deuxième ligne $w_i = 1, w_j = P_j / \sum_k P_k$, troisième ligne $w_i = P_i / \sum_k P_k, w_j = P_j / \sum_k P_k$), et pour d_0 variable (couleur).

A.7 EFFETS DE RÉSEAU

A.8 MORPHOGENÈSE PAR AGRÉGATION-DIFFUSION

A.8.1 Figures supplémentaires pour l'exploration du modèle

Convergence

Les histogrammes pour les 81 points de paramètres pour lesquelles 100 répétitions ont été menées sont donnés en Fig. 84, pour l'index de Moran et la hiérarchie. Les autres indicateurs témoignent de propriétés de convergence similaires. L'exploration visuelle des histogrammes confirme l'analyse numérique menée dans le texte principal pour la convergence statistique.

Indicateurs

Nous donnons en Fig. 85 à Fig. 88 le comportement exhaustif des indicateurs, pour l'ensemble des paramètres variant. Ceux-ci ont été obtenus par l'exploration intensive, et les graphiques en texte principal en sont des cas particuliers. A cause de la nature complexe de la forme urbaine émergente, il n'est pas possible de prédire les valeurs de sorties sans référer à cette exploration "exhaustive" de l'espace des paramètres.

Scatterplot des indicateurs

Nous montrons finalement les nuages de points complets des indicateurs, avec les points observés, en Fig. 89. Il s'agit de l'étape préliminaire à la calibration sur les composantes principales, et nous pouvons voir ici sur quelles dimensions le modèle échoue particulièrement à s'approcher des données observées (en particulier la distance moyenne).

A.8.2 Analyse semi-analytique du modèle simplifié

Equation aux dérivées partielles

Nous proposons de dériver l'EDP dans un cadre simplifié. Pour rappeler la configuration donnée en texte principal, le système a une dimension, tel que $x \in \mathbb{R}$ avec $1/\delta x$ cellules de taille δx , et nous utilisons les valeurs attendues des populations des cellules $p(x, t) = \mathbb{E}[P(x, t)]$. Nous prenons de plus $n_d = 1$. Des valeurs plus grandes devraient impliquer des dérivées à un ordre supérieur à deux, mais les résultats qui suivent sur l'existence d'une solution stationnaire devraient être conservés.

En écrivant $\tilde{p}(x, t)$ les populations intermédiaires obtenues après l'étape d'agrégation, nous avons

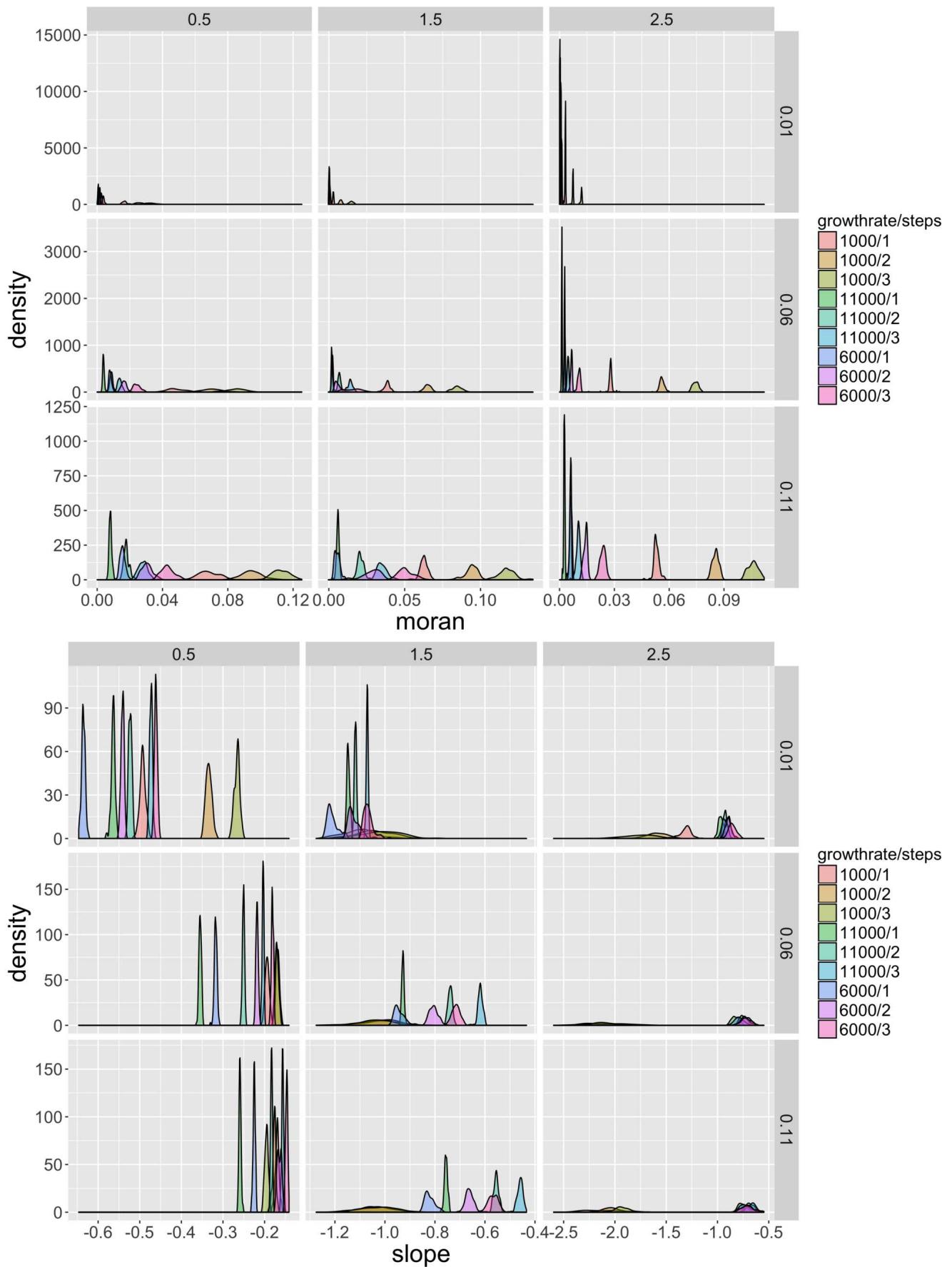


FIGURE 84: (Haut) Distributions de l'Index de Moran, pour des valeurs variables de α (colonnes), β (lignes), N_G et n_d (couleurs)

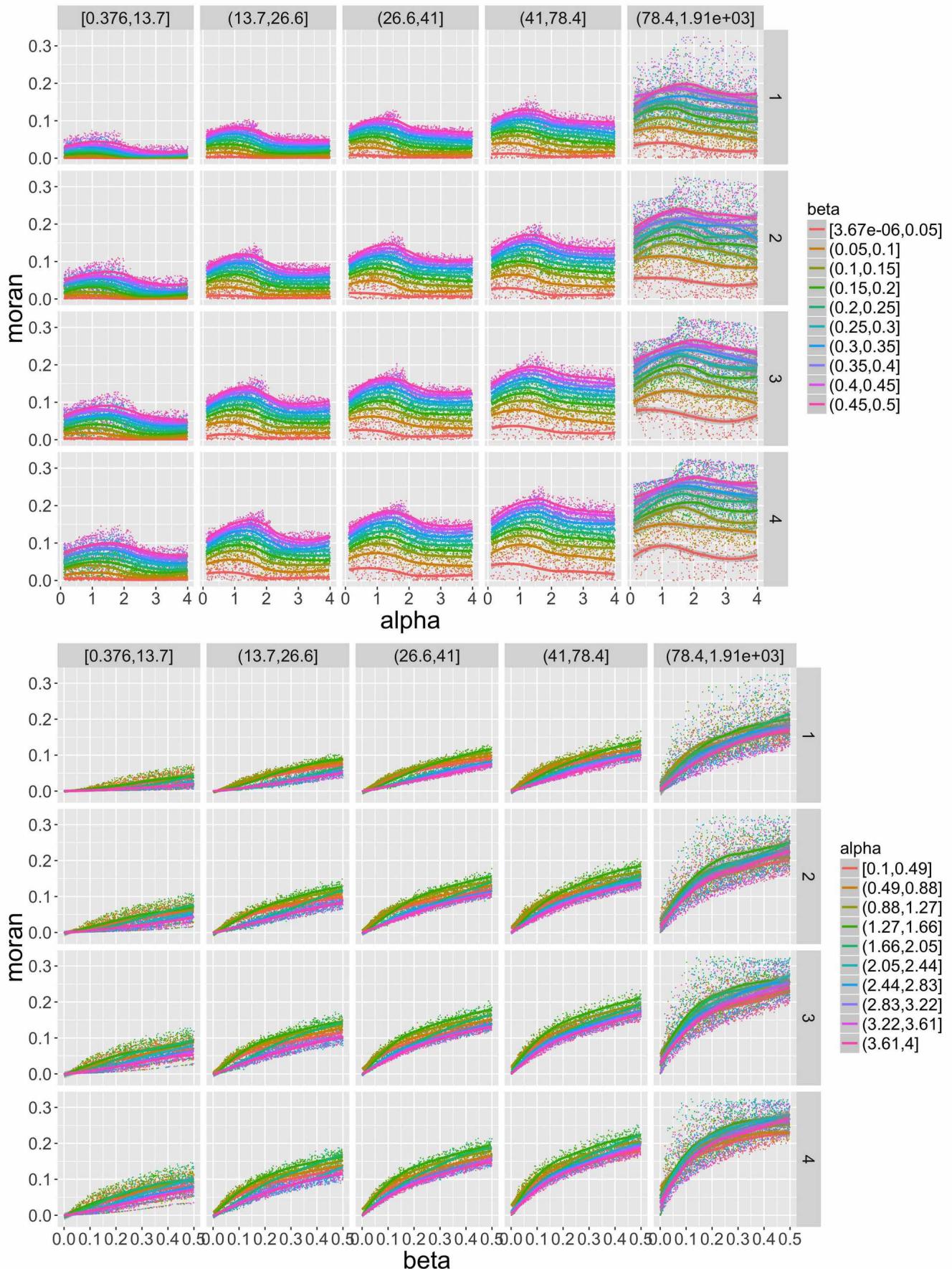


FIGURE 85: Indice de Moran en fonction de α (Haut) et β (Bas) pour β variable (resp. α) donné par la couleur, et n_d (lignes) et N_G (colonnes) variables.

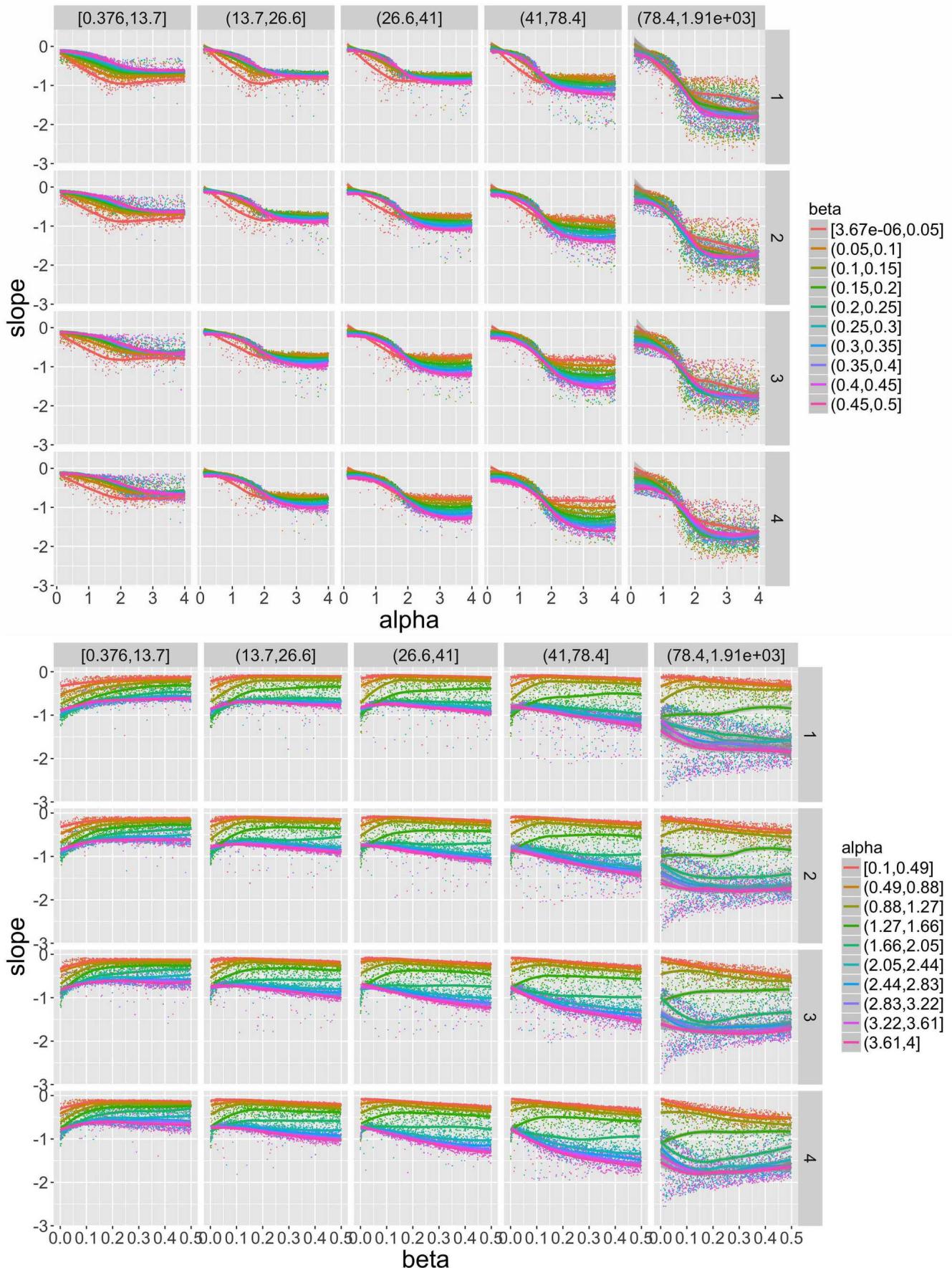


FIGURE 86: Hiérarchie en fonction de α (Haut) et β (Bas) pour β variable (resp. α) donné par la couleur, et n_d (lignes) et N_G (colonnes) variables.

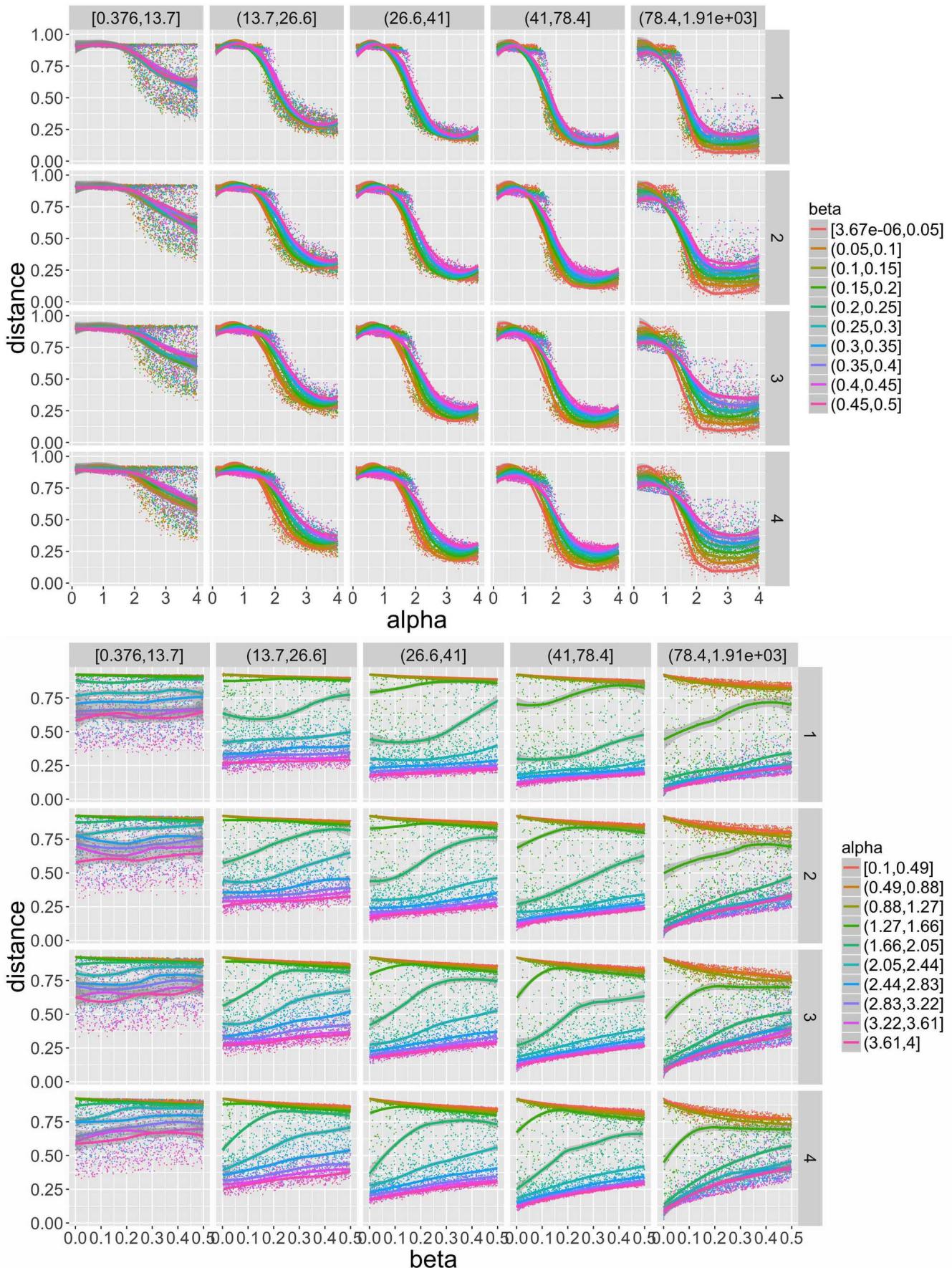


FIGURE 87: Distance moyenne en fonction de α (Haut) et β (Bas) pour β variable (resp. α) donné par la couleur, et n_d (lignes) et N_G (colonnes) variables.

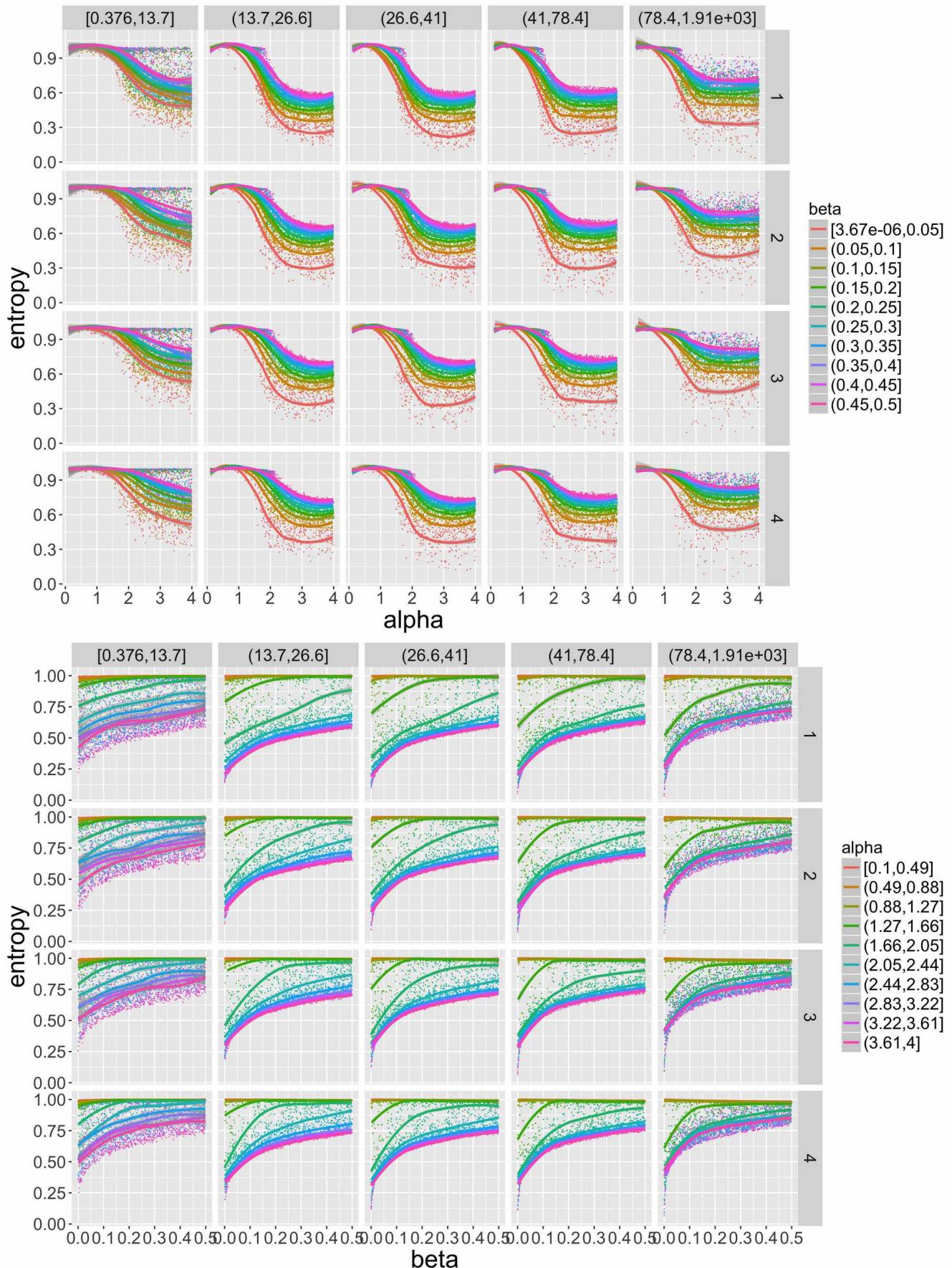


FIGURE 88: Entropie en fonction de α (Haut) et β (Bas) pour β variable (resp. α) donné par la couleur, et n_d (lignes) et N_G (colonnes) variables.

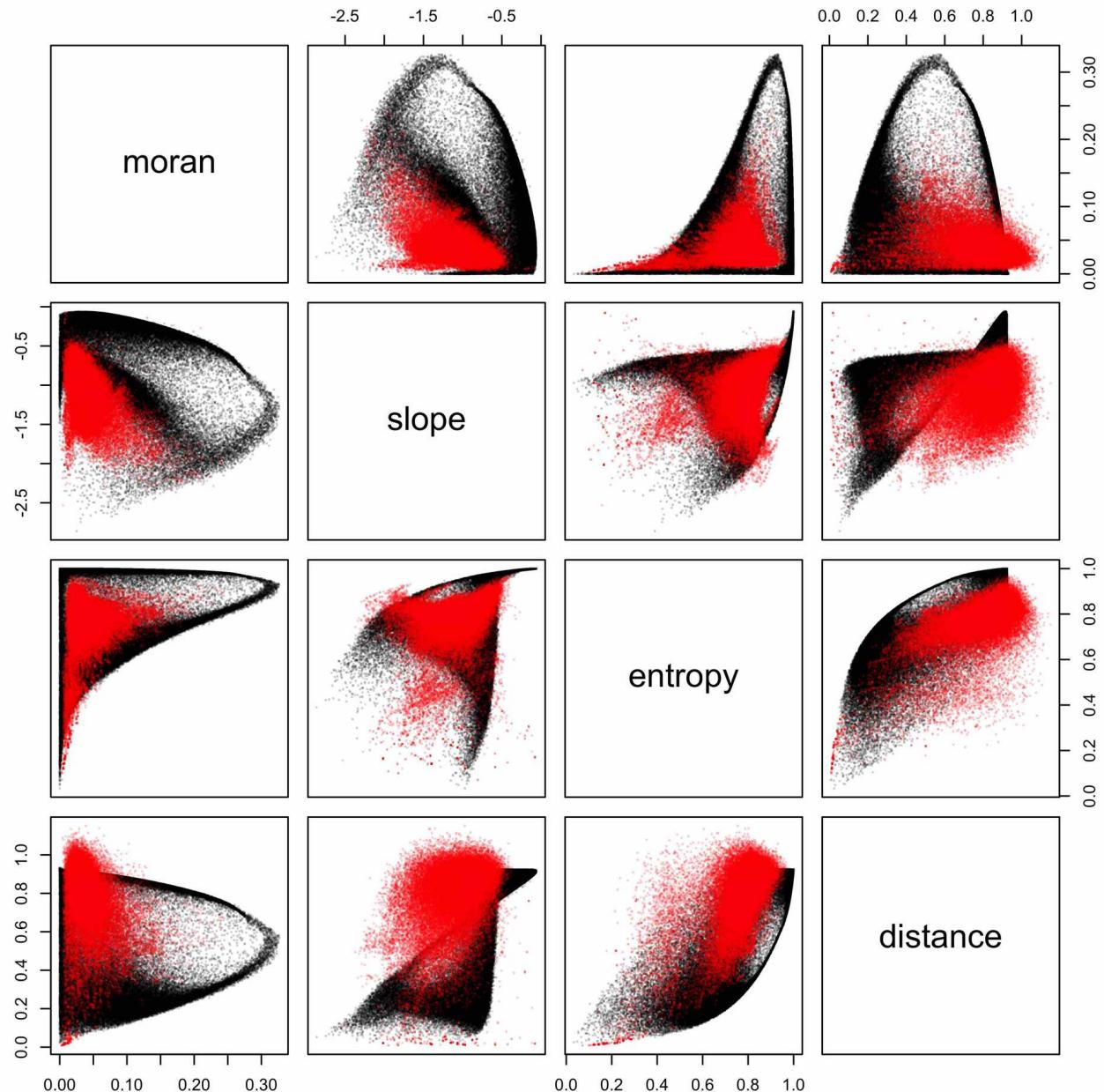


FIGURE 89: Nuages de points des indicateurs dans l'hypercube échantillonné de l'espace des paramètres. Les points rouges correspondent aux données réelles.

$$\tilde{p}(x, t) = p(x, t) + N_g \cdot \frac{p(x, t)^\alpha}{\sum_x p(x, t)^\alpha}$$

puisque toutes les unités de population sont ajoutées indépendamment. Si $\delta x \ll 1$ alors $\sum_x p^\alpha \simeq \int_x p(x, t)^\alpha dx$ et nous écrivons cette quantité $P_\alpha(t)$. Nous notons de plus $p = p(x, t)$ et $\tilde{p} = \tilde{p}(x, t)$ par la suite pour faciliter la lecture.

L'étape de diffusion est ensuite déterministe, et pour toute cellule qui n'est pas au bord ($0 < x < 1$), si δt est l'intervalle entre deux pas de temps, nous avons

$$\begin{aligned} p(x, t + \delta t) &= (1 - \beta) \cdot \tilde{p} + \frac{\beta}{2} [\tilde{p}(x - \delta x, t) + \tilde{p}(x + \delta x, t)] \\ &= \tilde{p} + \frac{\beta}{2} [(\tilde{p}(x + \delta x, t) - \tilde{p}) - (\tilde{p} - \tilde{p}(x - \delta x, t))] \end{aligned}$$

Sous l'hypothèse que les dérivées partielles existent, et comme $\delta x \ll 1$, nous faisons l'approximation $\tilde{p}(x + \delta x, t) - \tilde{p} \simeq \delta x \cdot \frac{\partial \tilde{p}}{\partial x}(x, t)$, ce qui donne

$$(\tilde{p}(x + \delta x, t) - \tilde{p}) - (\tilde{p} - \tilde{p}(x - \delta x, t)) = \delta x \cdot \left(\frac{\partial \tilde{p}}{\partial x}(x, t) - \frac{\partial \tilde{p}}{\partial x}(x - \delta x, t) \right)$$

et donc au second ordre

$$p(x, t + \delta t) = \tilde{p} + \frac{\beta \delta x^2}{2} \cdot \frac{\partial^2 \tilde{p}}{\partial x^2}$$

Le remplacement de \tilde{p} donne

$$\begin{aligned} \frac{\partial^2 \tilde{p}}{\partial x^2} &= \frac{\partial^2 p}{\partial x^2} + \frac{N_G}{P_\alpha} \cdot \frac{\partial}{\partial x} \left[\alpha \frac{\partial p}{\partial x} p^{\alpha-1} \right] \\ &= \frac{\partial^2 p}{\partial x^2} + \alpha \frac{N_G}{P_\alpha} \left[\frac{\partial^2 p}{\partial x^2} p^{\alpha-1} + (\alpha - 1) \left(\frac{\partial p}{\partial x} \right)^2 p^{\alpha-2} \right] \end{aligned}$$

En supposant que $\frac{\partial p}{\partial t}$ existe et que δt est petit, nous avons $p(x, t + \delta t) - p(x, t) \simeq \delta t \frac{\partial p}{\partial t}$, ce qui donne finalement, par combinaison des résultats ci-dessus, l'équation aux dérivées partielles

$$\delta t \cdot \frac{\partial p}{\partial t} = \frac{N_G \cdot p^\alpha}{P_\alpha(t)} + \frac{\alpha \beta (\alpha - 1) \delta x^2}{2} \cdot \frac{N_G \cdot p^{\alpha-2}}{P_\alpha(t)} \cdot \left(\frac{\partial p}{\partial x} \right)^2 + \frac{\beta \delta x^2}{2} \cdot \frac{\partial^2 p}{\partial x^2} \cdot \left[1 + \alpha \frac{N_G p^{\alpha-1}}{P_\alpha(t)} \right] \quad (22)$$

Les conditions initiales sont spécifiées par $p_0(x) = p(x, t_0)$. Pour obtenir un problème bien posé comme dans des formulations PDE plus classiques, nous devons supposer un domaine et des conditions au bord. Un support fini est traduit par $p(x, t) = 0$ pour tout t et x tel que $|x| > x_m$.

Solution stationnaire pour la densité

La non-linéarité et les termes intégraux rendant l'équation ci-dessus hors d'atteinte d'une résolution analytique, nous étudions son comportement de manière numérique pour certaines configurations. Prenant une condition initiale simple $p_0(0) = 1$ et $p_0(x) = 0$ pour $x \neq 0$, nous montrons que sur un domaine fini, la densité $d(x, t)$ converge toujours vers une solution stationnaire pour les grandes valeurs de t , pour un grand nombre de valeurs pour (α, β) avec $N_G = 10$ fixé ($\alpha \in [0.4, 1.5]$ variant avec un pas de 0.025 et $\log \beta \in [-1, -0.5]$ avec un pas de 0.1). Nous montrons en Fig. 90 les trajectoires correspondantes sur un sous-ensemble typique. La variation des distributions asymptotiques comme fonction de α et β ne sont pas directement observables, puisqu'elle dépendent des valeurs très faibles des flux sortants aux bords. Nous donnons en Fig. 91 leur comportement, en donnant la valeur du maximum de la distribution. Les valeurs faibles de β mènent à une inversion de l'effet de α , tandis que les fortes valeurs de β donnent des valeurs comparables pour tous les α .

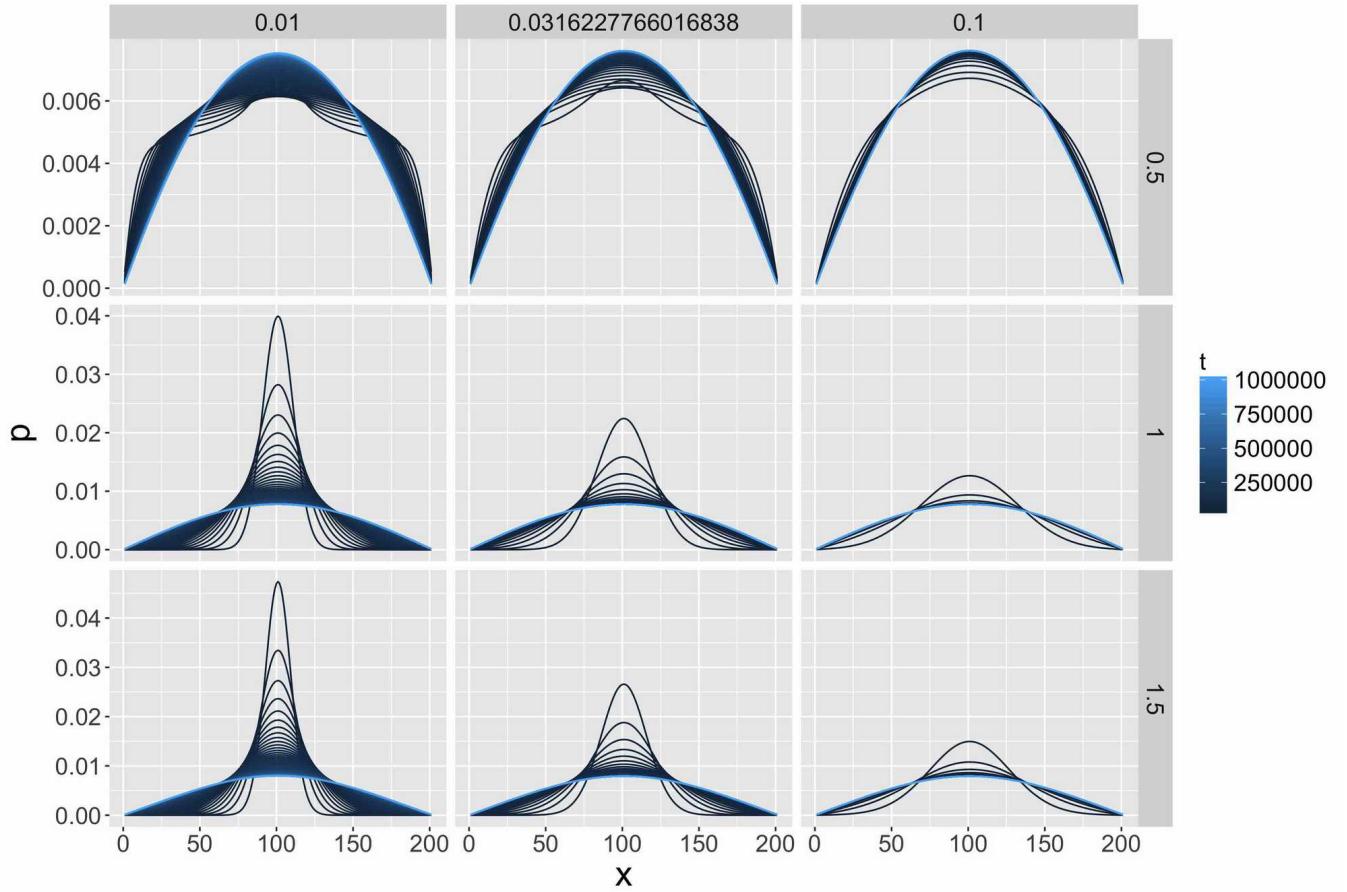


FIGURE 90: Trajectoires des densités en fonction de la coordonnée spatiale, pour β variable (colonnes) et α variable (lignes). Le niveau de couleur donne le temps.

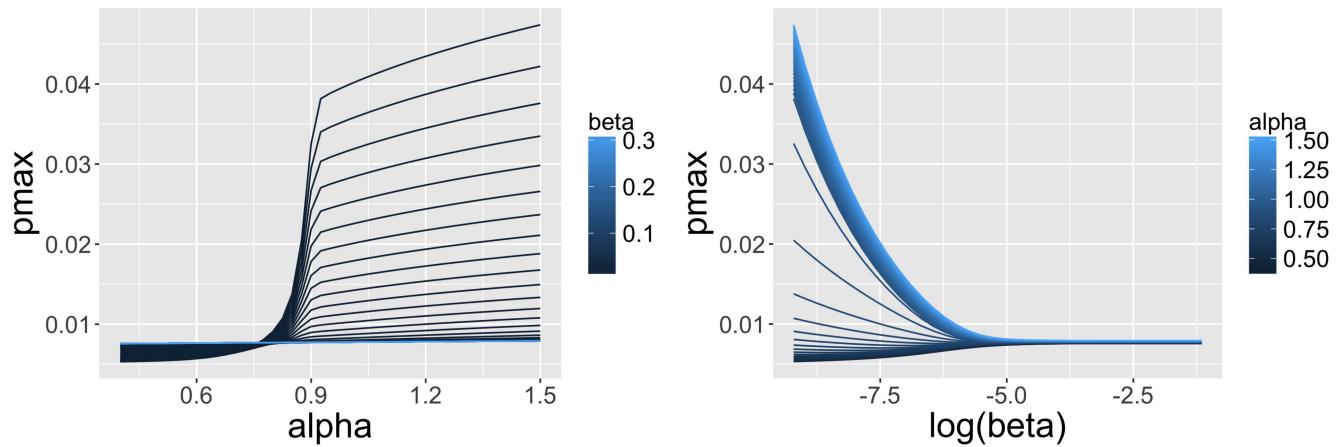


FIGURE 91: Dépendance de $\max d(t \rightarrow \infty)$ à α et β .

A.9 DONNÉES SYNTHÉTIQUES CORRÉLÉES

La Fig. 92 donne les erreurs sur les corrélations faisables ainsi que l'amplitude des corrélations.

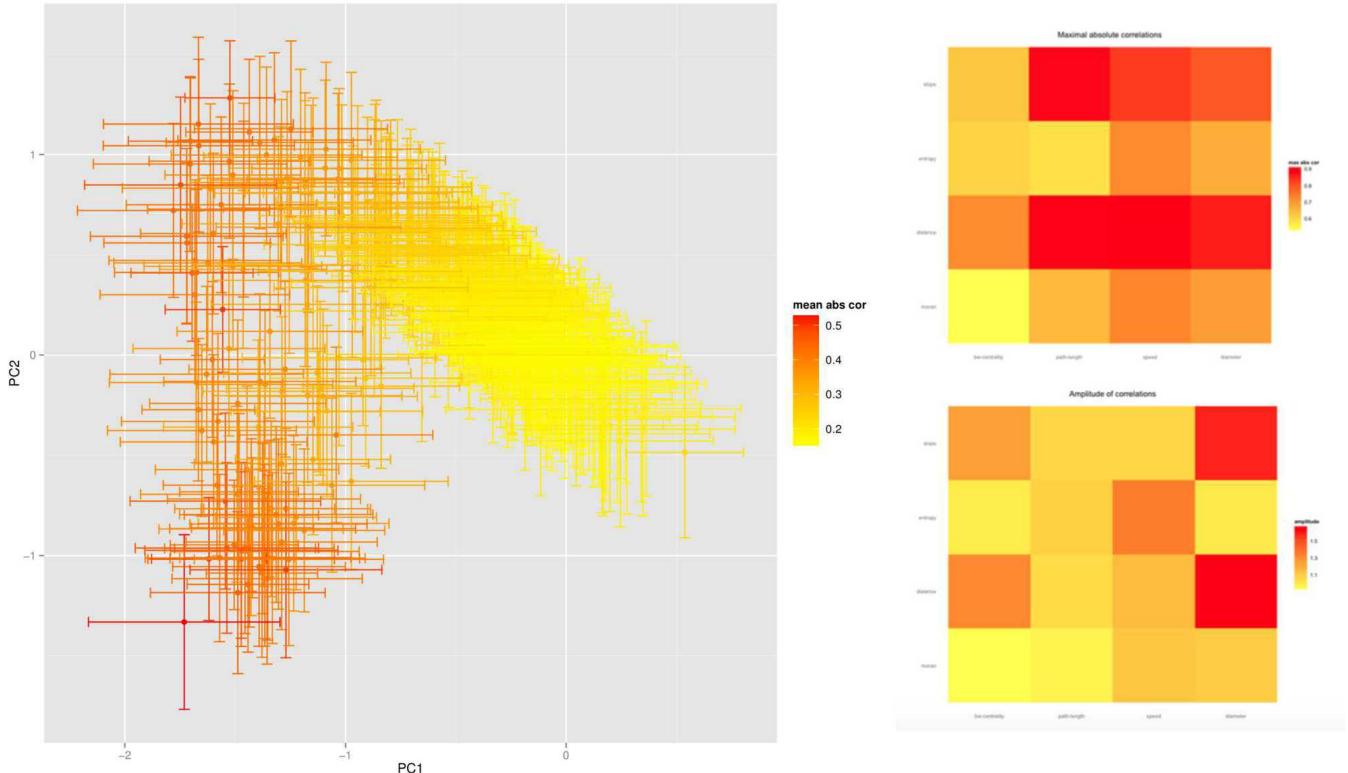


FIGURE 92: (*Gauche*) Projection des matrices de corrélations dans un plan principal obtenu par analyse en composantes principales sur la population des matrices (variances cumulées PC₁=38%, PC₂=68%, s'agissant de corrélations les données sont elles-mêmes corrélées d'où la structure du nuage de points); les barres d'erreur sont calculées initialement comme les intervalles de confiance à 95% sur chaque matrice (par méthode asymptotique de Fisher standard), et les bornes supérieures après transformation sont prises dans le plan principal; (*Droite*) Amplitude des corrélations, définie comme $a_{ij} = \max_k \rho_{ij}^{(k)} - \min_k \rho_{ij}^{(k)}$ et corrélation maximale absolue, définie comme $c_{ij} = \max_k |\rho_{ij}^{(k)}|$; l'échelle de couleur donne la corrélation moyenne absolue sur les matrices entières

A.10 EXPLORATION DU MODÈLE SIMPOPNET

A.11 MODÈLE DE CO-ÉVOLUTION MACROSCOPIQUE

A.11.1 *Données synthétiques*

A.11.2 *Données réelles*

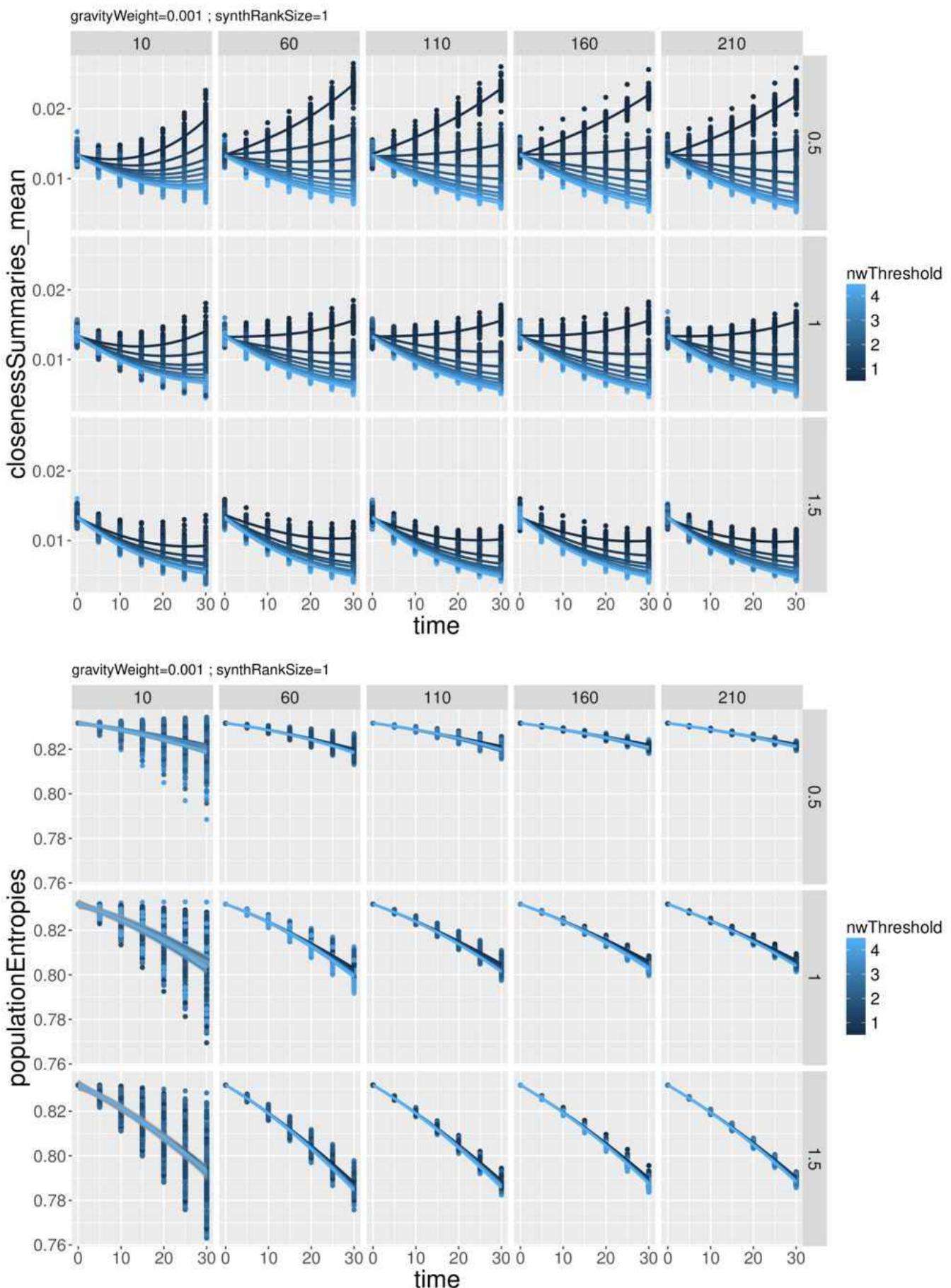


FIGURE 93: Comportement du modèle de co-evolution avec réseau abstrait sur un système de villes synthétique, pour $\alpha_S = 1$. (Haut) Moyenne des centralités de proximité en fonction du temps, pour d_G (colonnes), γ_G (lignes) et ϕ_0 (couleur) variables, à $w_G = 0.001$ fixé; (Bas) Entropie de populations, en fonction du temps, pour d_G (colonnes), γ_G (lignes) et ϕ_0 (couleur) variables, à $w_G = 0.001$ fixé Se référer au texte pour l'interprétation.

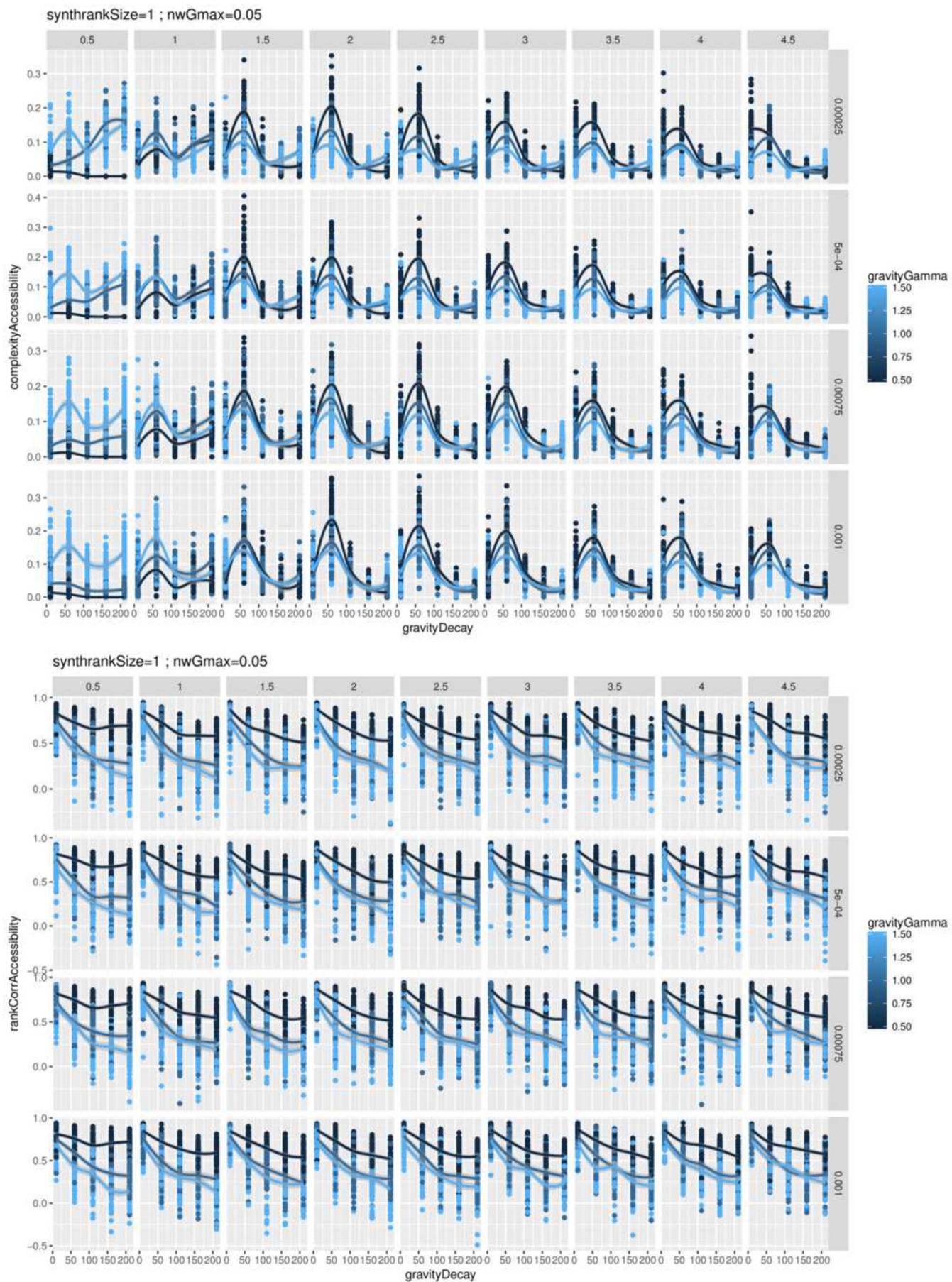


FIGURE 94: (Haut) Complexité des accessibilités, en fonction de d_G , pour ϕ_0 (colonnes), w_G (lignes) et γ_G (couleur) variables ; (Bas) Corrélations de rang des accessibilités, pour les mêmes paramètres.

[22 janvier 2018 at 12:17 – Thesis version 3.4.2]

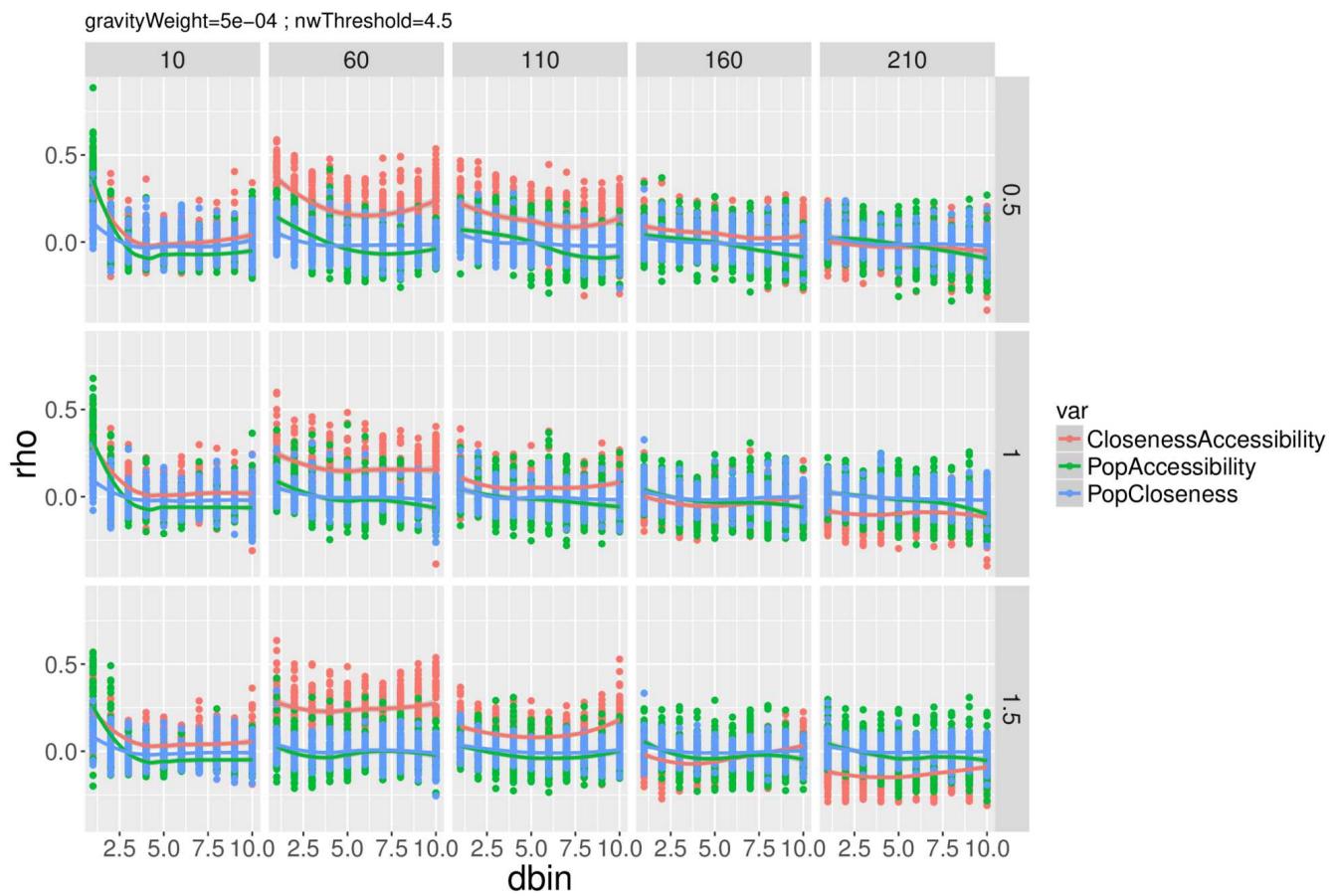


FIGURE 95: **Corrélations en fonction de la distance.** Correlation ρ_d entre couples de variables (donné par la couleur), en fonction de la distance d (discrétisée en déciles), pour d_G variable (colonnes) et γ_G variable (lignes), à $w_G = 5e - 4$ et $\phi_0 = 4.5$

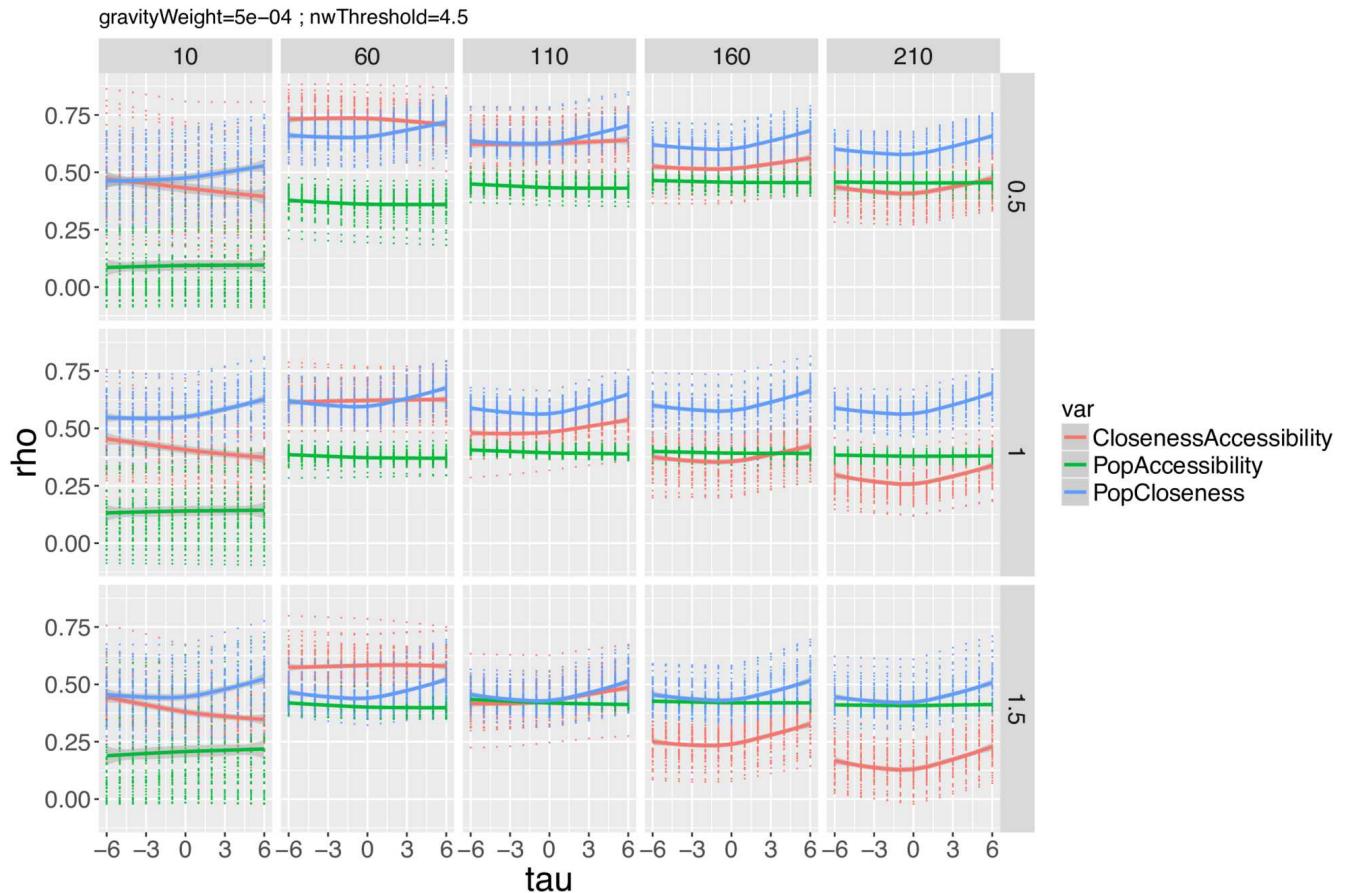


FIGURE 96: Corrélations retardées. Correlations retardées ρ_τ en fonction du retard τ , de manière similaire pour d_G variable (colonnes) et γ_G variable (lignes), à $w_G = 5e-4$ et $\phi_0 = 4.5$.

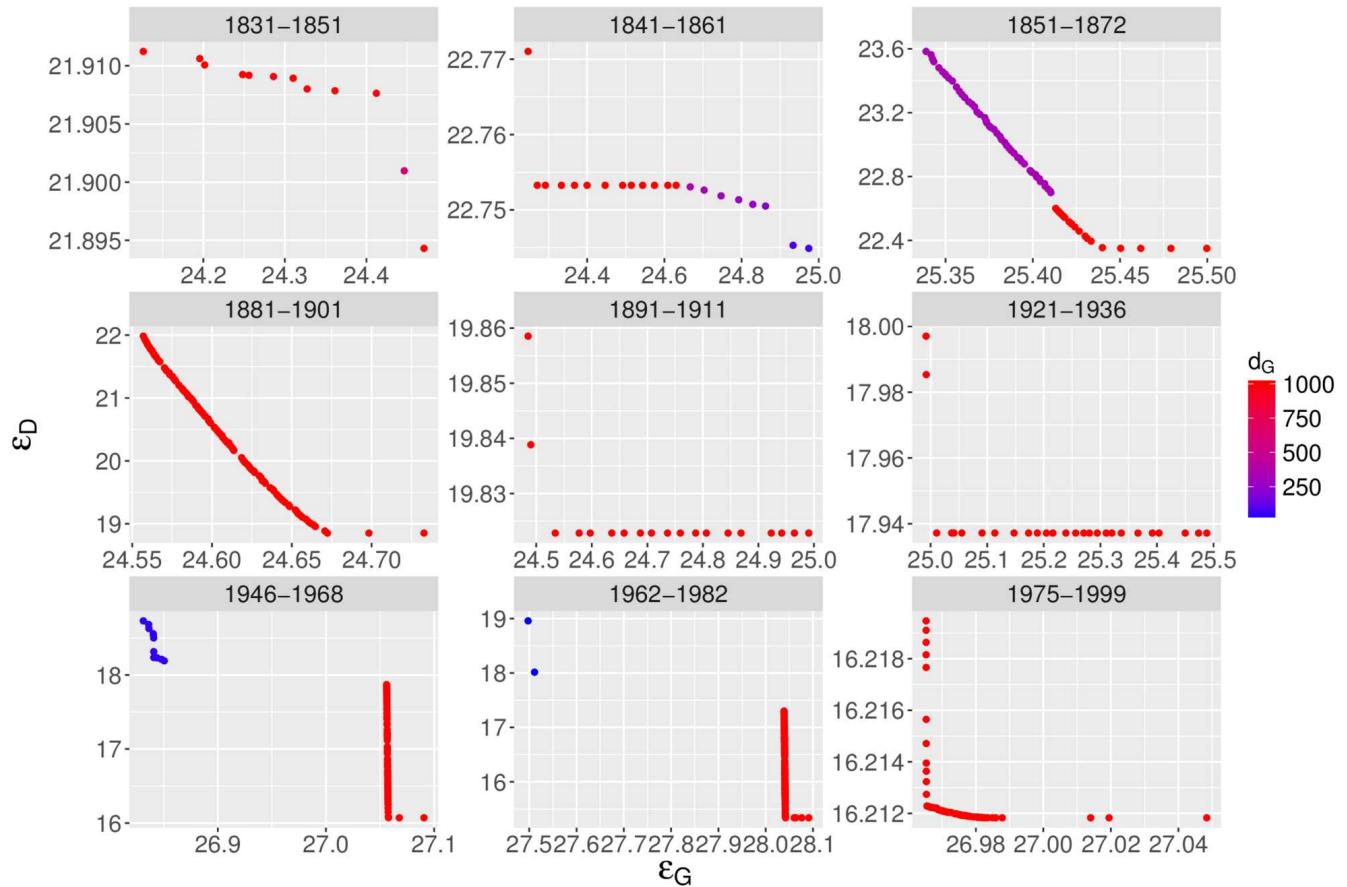


FIGURE 97: Fronts de Pareto pour la calibration bi-objectif population et distance. Les fronts sont donnés pour chaque période de calibration, et colorés en fonction de d_G (Bas).

A.12 HEURISTIQUES DE GÉNÉRATION DE RÉSEAU

A.12.1 Modèle de *slime mould*

Nous rappelons ici la procédure d'évolution du réseau biologique de type *slime mould*, à partir de [tero2007mathematical]. Le réseau est composé de noeuds caractérisés par leur pression p_i et de liens caractérisés par leur longueur L_{ij} , leur diamètre D_{ij} , une impédance Z_{ij} et le flux les traversant ϕ_{ij} . La relation analogue à la loi d'Ohm pour les liens s'écrit

$$\phi_{ij} = \frac{D_{ij}}{Z_{ij} \cdot L_{ij}} (p_i - p_j)$$

Par ailleurs, la conservation des flux à chaque noeud (loi de Kirchoff) impose

$$\sum_i \phi_{ij} = 0$$

pour tout j sauf pour la source et le puit, que nous supposons aux indices j_+ et j_- , tel que $\sum_i \phi_{ij_+} = I_0$ et $\sum_i \phi_{ij_-} = -I_0$ avec I_0 paramètre de flux initial.

La combinaison des contraintes ci-dessus donne pour tout j

$$\sum_i \frac{D_{ij}}{Z_{ij} \cdot L_{ij}} (p_i - p_j) = \mathbb{1}_{j=j_+} I_0 - \mathbb{1}_{j=j_-} I_0$$

ce qui se simplifie en une équation matricielle, en notant $\mathbf{Z} = \begin{pmatrix} \frac{D_{1j}}{Z_{1j} \cdot L_{1j}} & \dots & \frac{D_{nj}}{Z_{nj} \cdot L_{nj}} \end{pmatrix}_{ij}$, ainsi que $\vec{k} = \frac{\mathbb{1}_{j=j_+} I_0 - \mathbb{1}_{j=j_-} I_0}{\sum_i \frac{D_{ij}}{Z_{ij} \cdot L_{ij}}}$ et $\vec{p} = p_i$, qui se simplifie en

$$(Id - \mathbf{Z}) \vec{p} = \vec{k}$$

Le système admet une solution lorsque $(Id - \mathbf{Z})$ est inversible. L'espace des matrices inversible étant dense dans $M_n(\mathbb{R})$, par multilinéarité du déterminant, une perturbation infinitésimale de la position des noeuds permet d'inverser la matrice si celle-ci est effectivement singulière. On obtient donc les pressions p_i et par conséquent les flux ϕ_{ij} .

L'évolution du diamètre D_{ij} entre deux étapes d'équilibre est fonction du flux à l'équilibre, par l'équation

$$D_{ij}(t+1) - D_{ij} = \delta t \left[\frac{\phi_{ij}(t)^\gamma}{1 + \phi_{ij}(t)^\gamma} - D_{ij}(t) \right]$$

TABLE 23: Indicateurs morphologiques pour les centres des classes des grilles de densité initiales.

Classe	Moran I	distance \bar{d}	entropie \mathcal{E}	slope 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

Nous prenons pour simplifier $\gamma = 1.8$, suivant la configuration utilisée par [tero2010rules] pour génération d'un réseau dans une configuration réelle. Nous prenons par ailleurs $\delta t = 0.05$ et $I_0 = 10$.

La génération d'un réseau peut s'effectuer à partir d'un réseau initial, jusqu'à atteindre un critère de convergence, par exemple $\sum_{ij} \Delta D_{ij}(t) < \epsilon$ avec ϵ paramètre de seuil fixé. Nous utilisons ce modèle avec un critère de nombre d'itérations, et procédons à une itération pour obtenir des réseaux finaux avec un nombre raisonnable de liens.

A.12.2 Résultats

Dans l'expérience explorant la distance aux réseaux réels, l'initialisation de la densité est faite selon 50 grilles classées dans 5 classes morphologiques (10 grilles par classe). La Table 23 donne la composition des centres des classes en termes d'indicateurs morphologiques. Les classes peuvent être interprétées de la façon suivante :

- Classe 5 : plus bas Moran, distance, hiérarchie et entropie élevées ; nombreux foyers de peuplement localisés et dispersés.
- Classe 4 : plus fortes entropie et hiérarchie ; un petit nombre de foyers localisés.
- Classe 3 : plus basse distance et entropie ; population diffuse.
- Classe 2 : plus haut Moran ; un ou quelques centres de taille conséquente.
- Classe 1 : valeurs intermédiaires pour tous les indicateurs ; un certain nombre de centres de taille intermédiaire.

Les espaces topologiques des réseaux générés en 7.1 peuvent être conditionnés aux classes morphologiques pour la distribution de densité initiale. Ce conditionnement est montré en Fig. 98. Nous donnons également les espaces faisables avec les points réels. Les classe 1 et 5 semblent être celle pour laquelle le rapprochement aux points réels est le plus facile, en termes de points extrêmes.

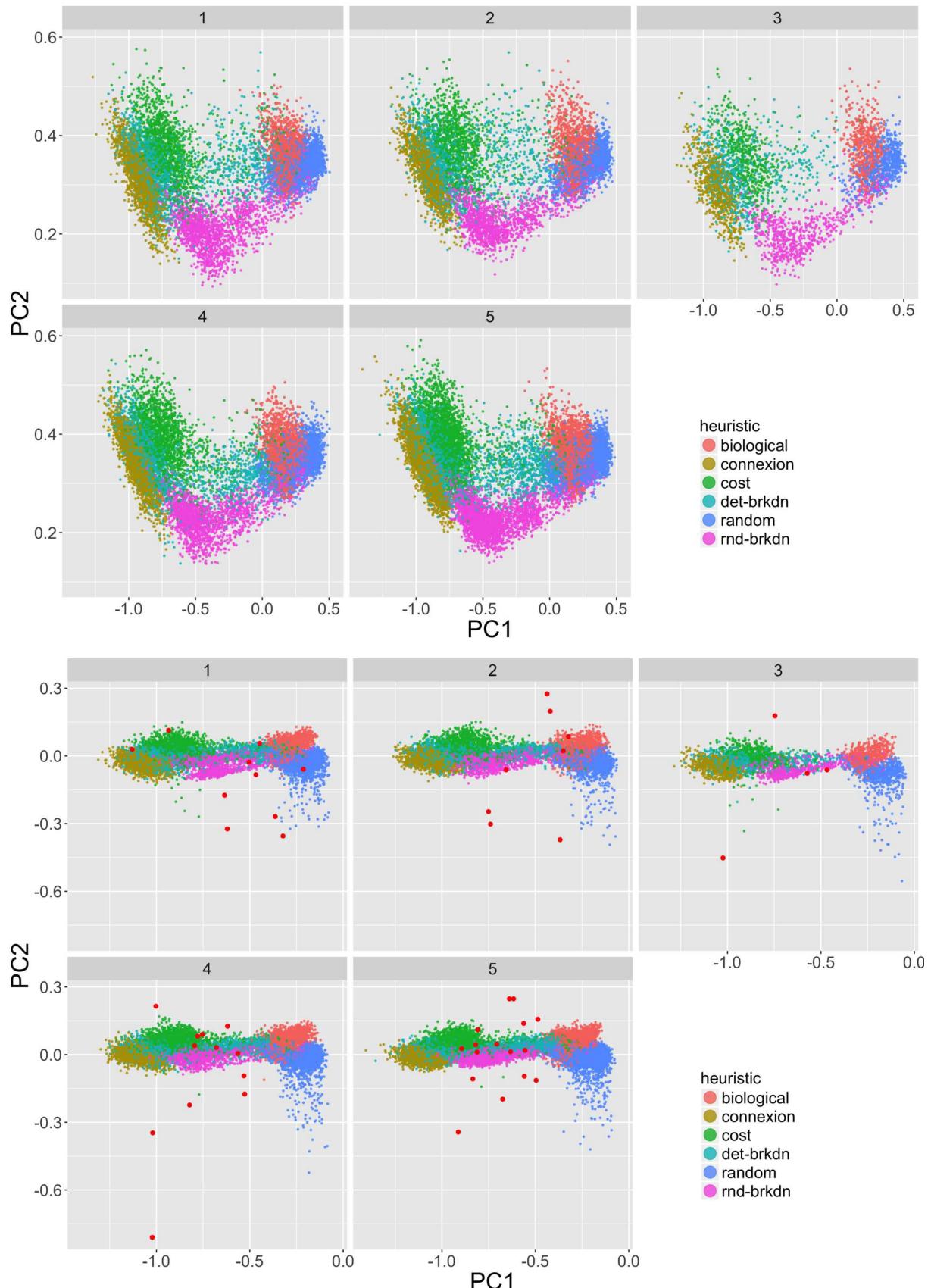


FIGURE 98: Conditionnement des résultats aux classes morphologiques pour la densité. (Haut) Espace topologique faisable pour les différentes heuristiques de génération, conditionné à la classe morphologique de densité. (Bas) Mêmes graphiques avec les points réels en rouge.

A.13 CO-ÉVOLUTION À L'ÉCHELLE MESOSCOPIQUE

A.13.1 Calibration

Afin de justifier l'agrégation des distances pour les indicateurs et pour les corrélations, nous avons contrôlé visuellement la forme des fronts de Pareto pour ces deux objectifs pour une vingtaine de points simulés. Un exemple pour deux points est donné en Fig. 99. Il apparaît que ces fronts sont quasi-inexistants, c'est à dire qu'il existe presque un optimum global.

Illustrons dans quelle mesure une agrégation linéaire à coefficient égaux peut être pertinente dans le cas d'un front de Pareto quasiment vertical/horizontal. La fonction

$$f_\alpha : x \mapsto \frac{1}{(x+1)^\alpha}$$

prend cette forme dans un voisinage de 0 lorsque α devient grand. Considérons alors les deux objectifs $o_1(x) = x$ et $o_2(x) = f_\alpha(x)$, qui peuvent soit être considérés pour une minimisation bi-objectifs, soit dans le cadre d'une agrégation linéaire par minimisation de $o(x) = \beta x + (1-\beta) \frac{1}{(x+1)^\alpha}$. Cette dernière est minimale en $x = \left(\frac{\beta}{\alpha(1-\beta)}\right)^{\frac{1}{\alpha+1}} - 1$, terme qui se développe en

$$x = \frac{\ln(\beta(1-\beta))}{\alpha+1} + \frac{\ln \alpha}{\alpha+1} + o\left(\frac{1}{\alpha}\right)$$

Par ailleurs, considérons que dans le cadre d'une optimisation bi-objectifs, on prenne le compromis auquel les variations de o_1 égalent celles de o_2 , ce qui revient à prendre x tel que $\frac{\partial f}{\partial x} = \frac{\partial f^{-1}}{\partial x}$. Cette équation conduit à $\frac{x^{\frac{1}{\alpha}}}{x+1} = \frac{1}{\alpha x^{\frac{2}{\alpha+1}}}$. On peut alors développer au second ordre de chaque côté pour obtenir

$$\frac{\ln x}{\alpha} = x \left[1 - 2 \frac{\ln \alpha}{\alpha+1} + o\left(\frac{1}{\alpha}\right) \right] - 2 \frac{\ln \alpha}{\alpha+1} + o\left(\frac{1}{\alpha}\right)$$

Or on a nécessairement $x \rightarrow_{\alpha \rightarrow \infty} 0$, puisque si $x \rightarrow K \neq 0$, on a une contradiction dans l'équation précédente car $1/(1+K) \neq 0$. Cela implique que $\frac{\ln x}{\alpha} = o\left(\frac{1}{\alpha}\right)$, et donc que

$$x = 2 \frac{\ln \alpha}{\alpha+1} + o\left(\frac{1}{\alpha}\right)$$

Pour avoir donc les mêmes ordres de grandeur pour les solutions aux deux approches, il faut éliminer le terme en $1/(\alpha+1)$ dans la première, ce qui revient à prendre $\ln(\beta(1-\beta)) = 0$ et donc $\beta = 1/2$.

Ainsi, il y a équivalence des ordres de grandeurs en α pour les deux approches si et seulement si $\beta = 1/2$. Vu la forme de nos fronts de Pareto, nous considérons la solution analogue et considérons ainsi la somme des deux distances.

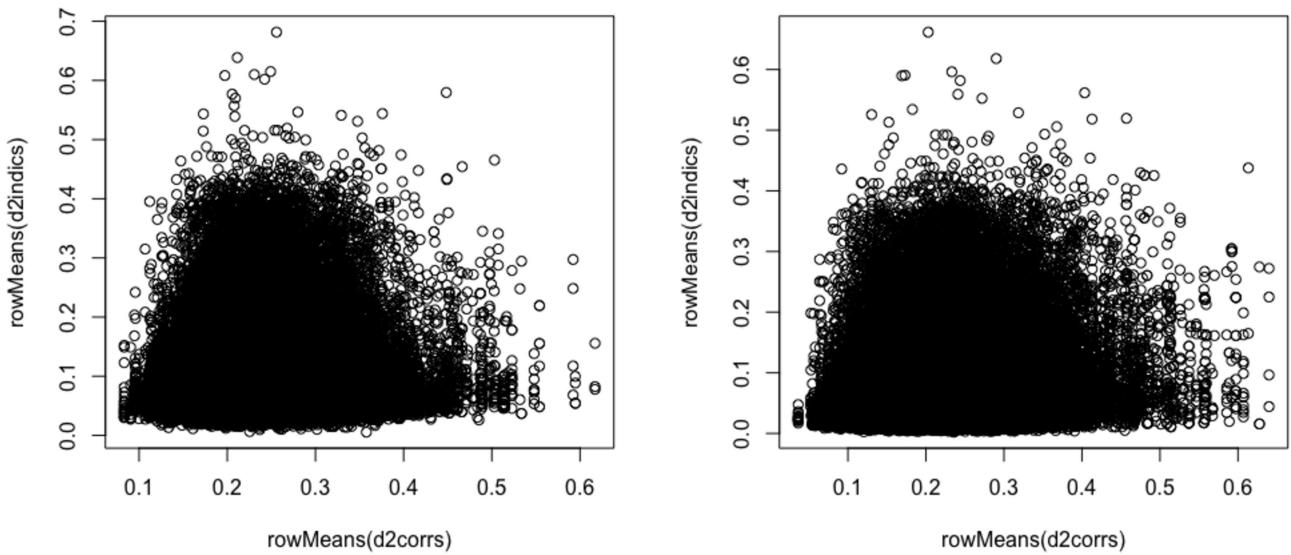


FIGURE 99: Exemples de fronts de Pareto pour la calibration au premier et au second ordre. Nous donnons pour deux points particuliers de simulation, les distances aux indicateurs d_I^2 et les distances aux corrélations d_C^2 pour l'ensemble des points réels.

A.14 MODÉLISATION DE LA GOUVERNANCE DU SYSTÈME DE TRANSPORT

A.14.1 Modèle d'usage du sol

Convergence

Nous étudions ici la question de la convergence dans le temps de la distribution des activités, à infrastructure fixe.

Considérons un cas très simple : en prenant $\lambda = 0$ on déspatialise le problème et en prenant $\gamma_A = 1$ on finit de découpler population et emplois. En posant $\beta' = \sum_j E_j \cdot \beta$ et $P_0 = \alpha \cdot \sum_i P_i$, l'existence d'un point fixe pour les populations se ramène à la résolution de

$$P_i = P_0 \cdot \frac{\exp(\beta' \cdot P_i)}{\sum \exp(\beta' \cdot P_i)}$$

La fonction est bien continue en les P_i et les plages de variations de la population sont $[0, \sum_i P_i]$, elle admet donc un point fixe par le Théorème du Point Fixe de Brouwer.

En fait, en toute généralité, si on écrit

$$(\vec{P}(t+1), \vec{E}(t+1)) = f(\vec{P}(t), \vec{E}(t))$$

pour des valeurs des paramètres arbitraires, la fonction f est également continue en chaque composante, et prend ses valeurs dans

un fermé borné (les emplois étant également limités) donc compact. De la même manière que [leurent2014user] l'établit pour un modèle de flux de traffic, on a aussi un point fixe dans notre cas, ce qui correspond à un point d'équilibre. L'unicité n'est cependant pas triviale et il n'y a pas de raison qu'elle soit vérifiée a priori. On vérifie empiriquement la convergence systématique à infrastructure fixe (voir ci-dessous l'exploration de l'espace des paramètres).

Exploration

Nous procédons à une exploration du comportement du modèle d'usage du sol seul, i.e. à infrastructure fixe, afin de comprendre l'influence des paramètres sur la forme urbaine. Nous fixons $\alpha = 1$ ici pour étudier le modèle dans un cas extrême.

Nous suivons les indicateurs de forme urbaine définis en 4.1, pour la distribution de la population et pour les emplois, dans le temps et jusqu'après convergence. Nous réduisons l'espace morphologique de la distribution spatiale des actifs dans un plan principal, tel que $PC_1 = -0.98 \cdot I - 0.13 \cdot \mathcal{E} + 0.05\bar{d} - 0.13 \cdot \gamma$ et $PC_2 = -0.19 \cdot I + 0.57 \cdot \mathcal{E} - 0.16\bar{d} + 0.77 \cdot \gamma$. La première composante exprime un niveau de dispersion et la seconde une agrégation hiérarchique.

La Fig. 100 donne des trajectoires temporelles dans le plan (PC_1, PC_2) pour $\gamma_A = 0.9, \gamma_E = 0.6, v_0 = 6$, pour différentes valeurs de λ et de β ainsi que pour différents réseaux initiaux. On constate qu'augmenter β a tendance à uniformiser les trajectoires. Pour $\beta = 1$, la forme du réseau conditionne fortement les trajectoires conjointement à λ : on passe par exemple d'une dispersion décroissante et d'une hiérarchie en cloche à une dispersion stable et une hiérarchie croissante pour les valeurs faibles de λ , entre aucun réseau et un réseau araignée.

La Fig. 101 donne la valeur de PC_1 pour la configuration finale sur l'ensemble de l'espace des paramètres exploré. Nous constatons ainsi la variabilité des formes (ici en termes de dispersion) en fonction de l'ensemble des paramètres : par exemple, pour des grandes valeurs de β , des diagrammes complexes émergent. Pour les faibles valeurs de β , on a une diagonale privilégiée pour la dispersion au sein de configurations concentrées.

Enfin, afin de comprendre l'influence des paramètres sur la mobilité totale au cours d'une trajectoire complète, nous étudions en Fig. 102 la variation cumulée des actifs donnée par $\tilde{\Delta} = \sum_t \sum_k |\Delta A_k(t)|$. On voit que des valeurs fortes de γ_A , pour β élevé, permettent de minimiser la quantité totale de relocalisation, qui ne dépend que très faiblement de γ_E . Il est ainsi possible d'optimiser, même à α fixé, la quantité totale d'étalement urbain.

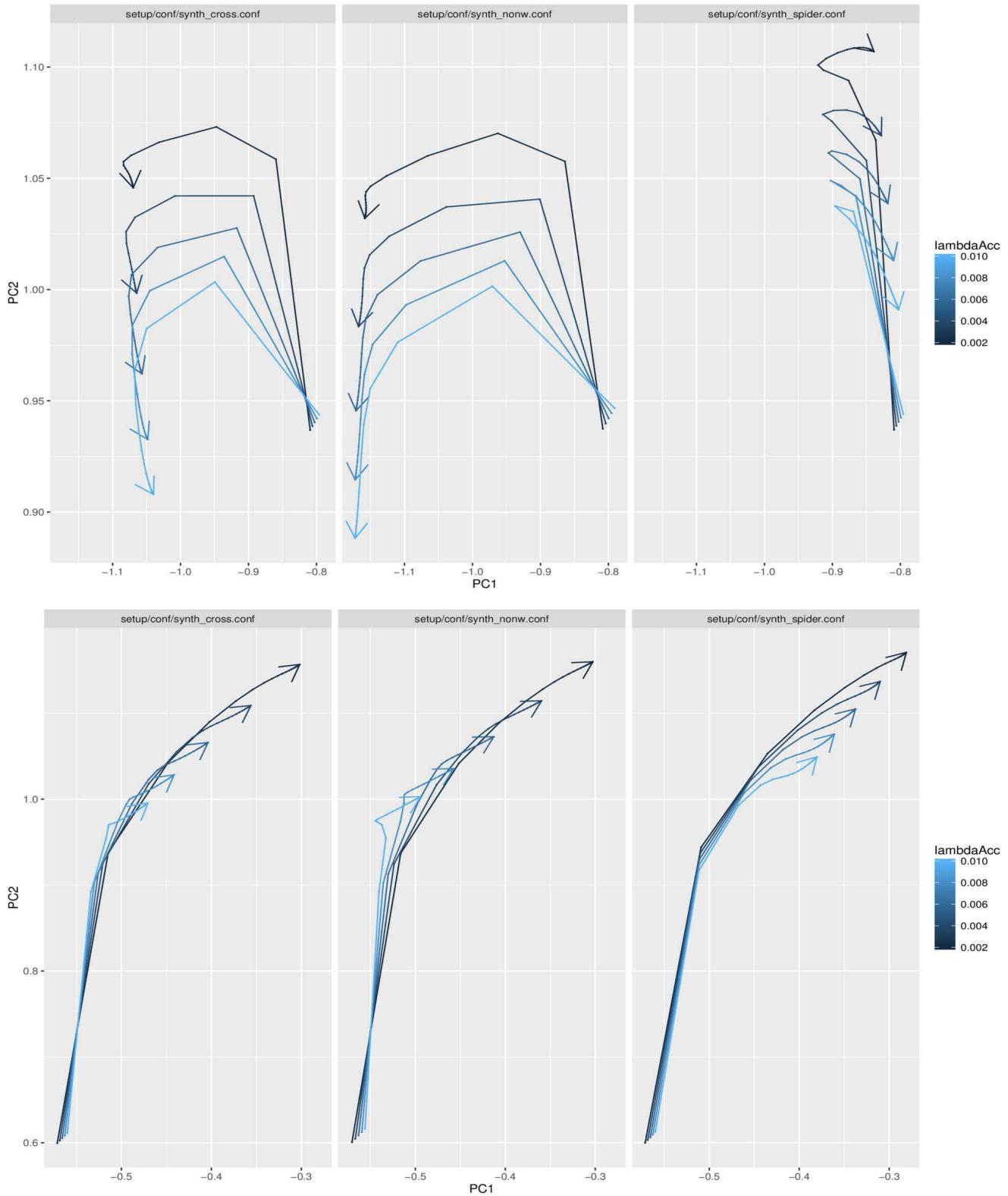


FIGURE 100: Trajectoires morphologiques pour la distribution de population. On fixe ici $\gamma_A = 0.9$ et $\gamma_E = 0.6$. (*Haut*) Trajectoires dans l'espace (PC_1, PC_2) pour $\beta = 1$, avec λ variable (couleur), et pour trois configurations de réseau différentes (colonnes) : réseau en croix, pas de réseau, réseau en croix avec ramifications (spider). (*Bas*) Mêmes graphes, pour $\beta = 2$.

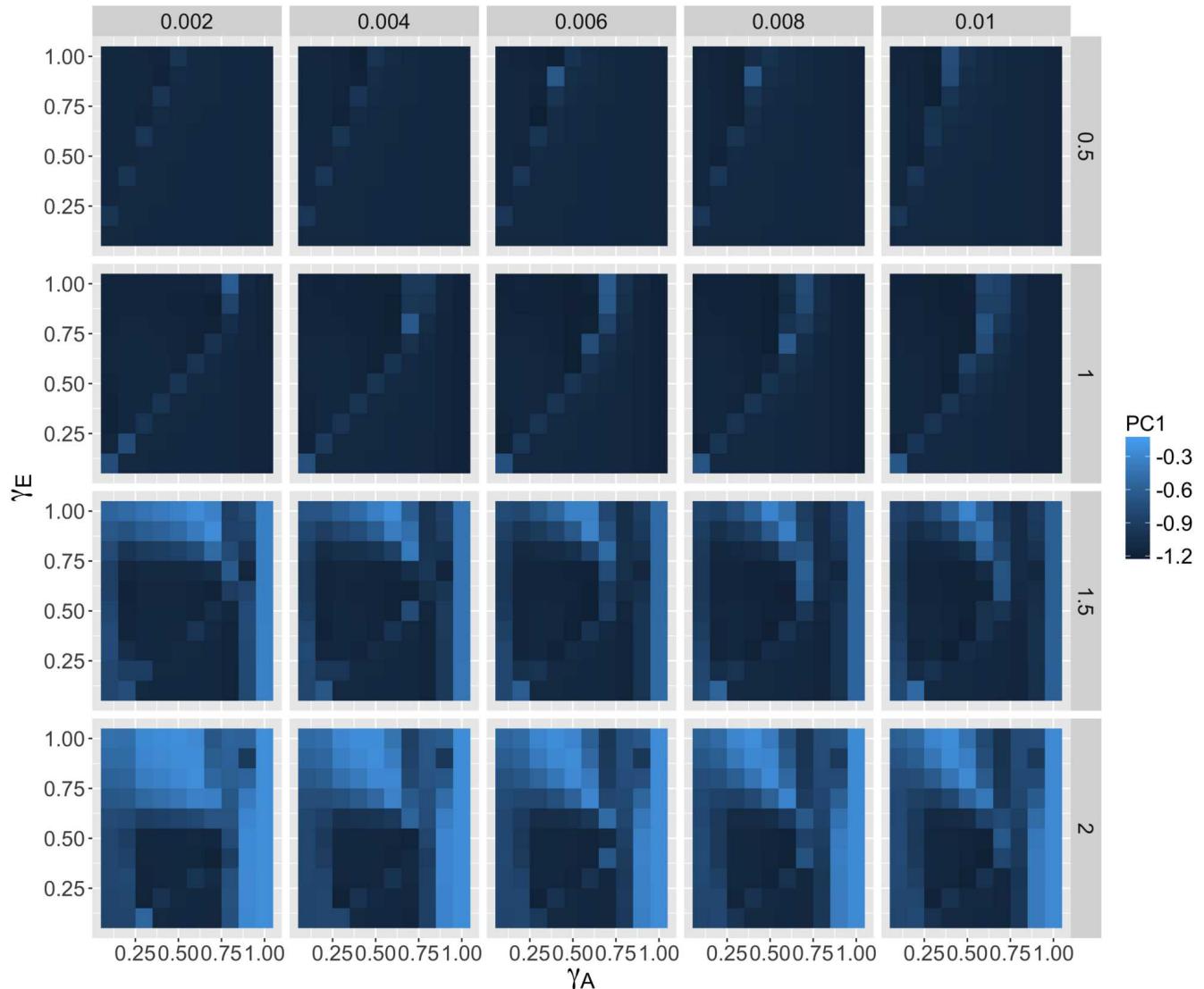


FIGURE 101: Sensibilité de la forme urbaine. Pour la distribution des populations, sans réseau initial, valeur de PC1 en fonction de (γ_A, γ_E) , avec λ variable (colonnes) et β variable (lignes).

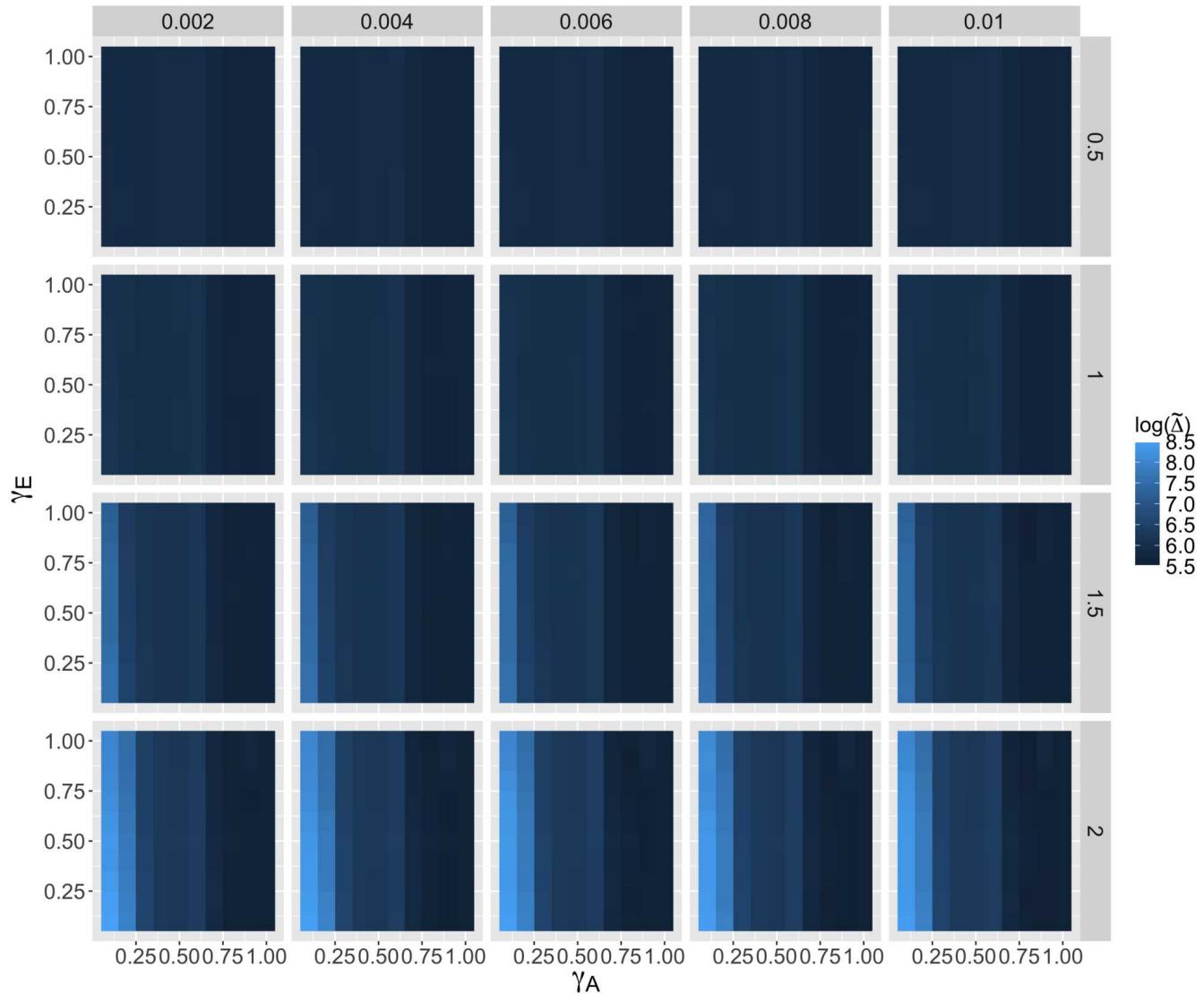


FIGURE 102: Variabilité cumulée des configurations urbaines. Valeur de $\ln \tilde{\Delta}$, sans réseau initial, en fonction de (γ_A, γ_E) , avec λ variable (colonnes) et β variable (lignes).

A.14.2 Module de transport

Nous n'avons pas pris en compte les flux de transport dans notre implémentation du modèle, supposant que les infrastructures construites sont de capacités suffisantes pour ne pas être significativement sensibles à la congestion.

Pour le calcul des flux entre cellules, l'opération est la suivante : les flux ϕ_{ij} sont calculés par résolution sur p_i, q_j par une méthode de point fixe (algorithme de Furness), du système des flux gravitaires :

$$\begin{cases} \phi_{ij} = p_i q_j A_i E_j \exp(-\lambda_{tr} d_{ij}) \\ \sum_k \phi_{kj} = E_j \\ \sum_k \phi_{ik} = A_i \\ p_i = \frac{1}{\sum_k q_k E_k \exp(-\lambda_{tr} d_{ik})} \\ q_j = \frac{1}{\sum_k p_k A_k \exp(-\lambda_{tr} d_{kj})} \end{cases}$$

où λ_{tr} est un paramètre donnant la portée spatiale des flux journaliers.

Pour implémenter l'étape de distribution des flux dans le réseau, une fois les flux entre cellules connus, il faudrait par exemple déterminer les flux de l'Équilibre Utilisateur Statique avec un algorithme approprié. Une affectation par plus courts chemins est implémentée avec le calcul des flux dans le modèle, mais nous désactivons ce processus pour simplifier l'étude du modèle.

La congestion peut être calculée comme un rapport à la capacité, comme c/c_{max} si c est le flux et c_{max} la capacité. La vitesse est obtenue par une fonction BPR sous la forme $v(c) = v_0 \left(1 - \frac{c}{c_{max}}\right)^{\gamma_c}$. Notre configuration revient à supposer une capacité infinie $c_{max} = \infty$.

A.14.3 Probabilités de coopération

L'hypothèse d'équilibre implique que les espérances conditionnelles de chaque joueur sont égales étant donné leur deux choix, i.e. que

$$\mathbb{E}[U_i | S_i = C] = \mathbb{E}[U_i | S_i = NC]$$

Cela revient en effet dans ce cas à maximiser $\mathbb{E}[U_i]$ par rapport à p_i , puisque en conditionnant on a $\mathbb{E}[U_i] = p_i \mathbb{E}[U_i | S_i = C] + (1 - p_i) \mathbb{E}[U_i | S_i = NC]$, et donc $\frac{\partial \mathbb{E}[U_i]}{\partial p_i} = \mathbb{E}[U_i | S_i = C] - \mathbb{E}[U_i | S_i = NC]$.

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.

Coordination par choix discrets

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

$$\frac{\partial f}{\partial x} = 2(\cosh u(x) - 1) + \beta^2 \Delta_i \Delta_{1-i} \frac{\exp(-\beta_{DC}(x\Delta_{1-i} - J))}{(1 + \exp(-\beta_{DC}(x\Delta_{1-i} - J)))^2}$$

$$\text{où } u(x) = -\beta_{DC} \left(\frac{\Delta_i}{1 + \exp(-\beta_{DC}(x\Delta_{1-i} - J))} - J \right).$$

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.14.4 Détails d'implémentation

MATRICE DES DISTANCES La matrice des distances est mise à jour de manière dynamique pour des questions de rapidité d'exécution (vu le nombre de mises à jour du réseau), de la façon suivante :

1. La matrice de distance euclidienne $d(i, j)$ est calculée analytiquement

2. Les plus courts chemins entre les intersections des liens (entre les cellules du réseau raster correspondant) sont mis à jour de manière dynamique (étape de complexité $O(N_{\text{inters}}^3)$):
 - Pour chaque nouvelle intersection, les plus courts chemins vers l'ensemble des autres intersections sont calculés par l'ancienne matrice et le nouveau lien.
 - Pour l'ensemble des anciens plus courts chemins, ils sont mis à jour si besoin après vérification des éventuels raccourcis par le nouveau lien.
 - La correspondance entre les cellules quelconques du réseau et les intersections est mise à jour.
3. Les composantes connexes et les distances entre celles-ci sont mises à jour (complexité en $O(N_{\text{nw}}^2)$)
4. Les distances par le réseau entre les cellules du réseau sont mises à jour, avec l'heuristique des connexions minimales uniquement (un lien unique le plus court entre chaque cluster) (complexité en $O(N_{\text{nw}}^2)$)
5. Les distances effectives entre l'ensemble des cellules (prenant la vitesse et la congestion en compte si celle-ci est implémentée) sont calculées comme le minimum entre la distance euclidienne et

$$\min_{C,C'} d(i, C) + d_{\text{nw}}(p_C(i), p'_C(j)) + d(C', j)$$

dont nous prenons une approximation avec \min_C uniquement dans l'implémentation, ce qui est consistant avec les portées d'interaction relativement faibles considérées. La complexité est en $O(N_{\text{clusters}}^2 \cdot N^2)$.

CROISSANCE DU RÉSEAU 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.14.5 Initialisation

Initialisation synthétique

Nous décrivons ici les détails de l'initialisation synthétique.

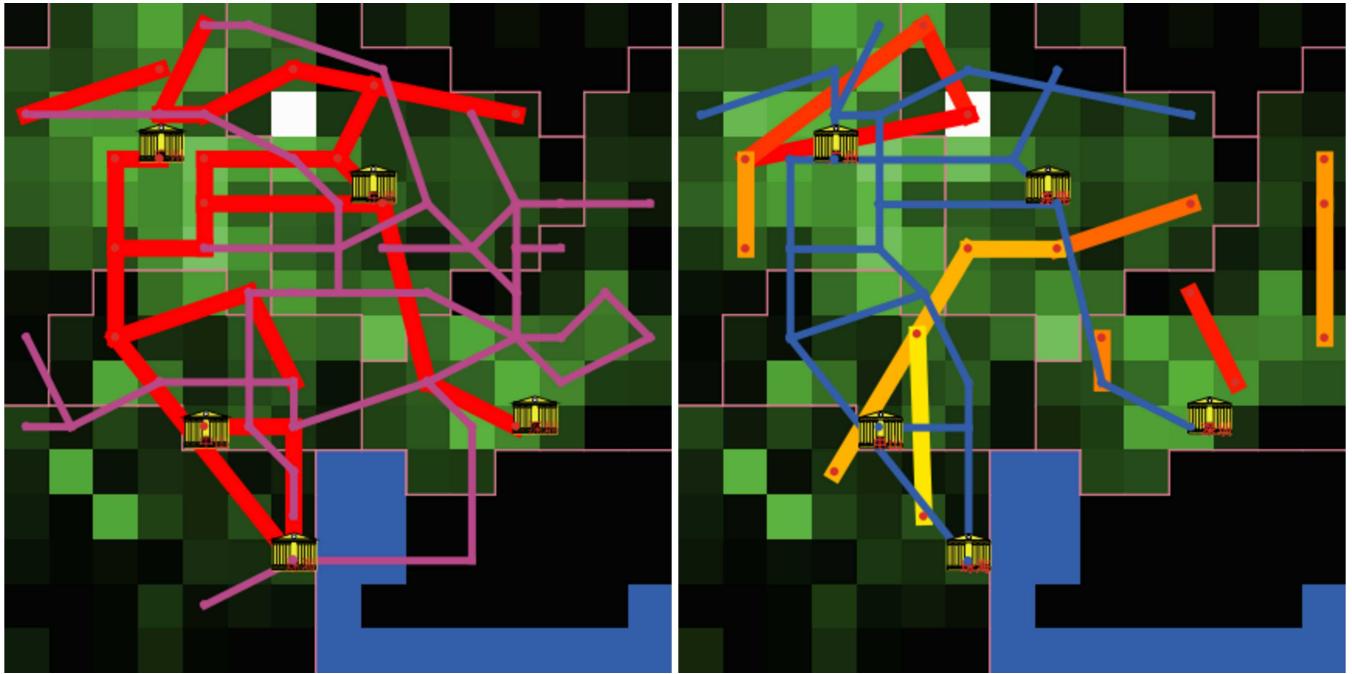


FIGURE 103: Initialisation sur données réelles utilisée lors de l’application. Pour des raisons de performances computationnelle, le nombre de cellules est ici diminué par rapport à l’illustration en texte principal. (*Gauche*) Réseaux à l’initialisation, en rouge le réseau initial correspondant au réseau en 2010, en violet fin le réseau cible pour la calibration, correspondant au réseau planifié. (*Droite*) Résultat obtenu avec $\alpha = 0$ à $t_f = 11$ après une initialisation sans réseau; en bleu le réseau cible, qui correspond au réseau de 2010.

Les distributions initiales des actifs et des emplois dans la configuration synthétique sont pris autour des centres de gouvernance (maires) aux positions \vec{x}_i avec des noyaux exponentiels par

$$A(\vec{x}) = A_{\max} \cdot \exp\left(-\frac{\|\vec{x} - \vec{x}_i\|}{r_A}\right); E(\vec{x}) = E_{\max} \cdot \exp\left(-\frac{\|\vec{x} - \vec{x}_i\|}{r_E}\right)$$

Initialisation sur configuration réelle

Nous montrons en Fig. 103 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.

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 [9.3](#).

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.2](#).

★ ★

★

B.1 MODÈLES STOCHASTIQUES DE CROISSANCE URBAINE

Les différents modèles stochastiques de croissance urbaine que nous avons développé suivent la même logique de règles autonomes pour reproduire les dynamiques des systèmes urbains. Nous proposons ici d'un point de vue méthodologique de mettre en valeur les liens entre les différents cadres, afin d'en formuler un cadre uniifié.

B.1.1 *Introduction*

Divers modèles stochastiques de croissance urbaine visant à reproduire des trajectoires de population, ou des faits stylisés sur celles-ci, souvent sur de longues échelles de temps et de grandes étendues spatiales (systèmes de villes) ont été proposé par la littérature dans des champs variés, de l'économie ou la physique à la géographie (voir par exemple 4.3 et 5.2 pour des revues à différentes échelles). Nous proposons ici une approche générale permettant de faire le liens entre plusieurs modèles existant, plus particulièrement les modèles de Gibrat, de Simon et d'attachement préférentiel.

Des modèles fondamentaux de croissance urbaine sont les modèles de Gibrat (voir 4.3) et le modèle de Simon [simon1955class] (qui a plus récemment été généralisé, voir par exemple [haran1973modified]). Diverses extensions on été données selon les disciplines. [benguigui2007dynamic] donne un modèle de système dynamique, tandis que [gabaix1999zipf] montre que le modèle de Gibrat produit la loi de Zipf pour la distribution de la taille des villes dans l'état stationnaire. Les approches en économie sont revues par [Gabaix20042341]. Un modèle inspiré par la Théorie Evolutive des Villes est décrit dans [favaro2011gibrat] et étend le modèle de Gibrat par l'addition de la propagation de l'innovation entre les villes. La question des échelles empiriques auxquelles ce type d'approche est pertinent a été traité dans le cas particulier de la France par [bretagnolle2002time], qui montre que de longues échelles de temps (supérieures à quelques décades) sont appropriées pour étudier la dynamique des systèmes urbains à une petite échelle spatiale.

B.1.2 *Cadre de Travail*

Le cadre que nous introduisons peut se comprendre comme un meta-modèle, au sens où chaque modèle peut être compris comme extension ou cas limite d'un autre modèle.

B.1.3 Dérivations

Généralisation de l'Attachement Préférentiel

[yamasaki2006preferential] donne une généralisation du modèle classique d'attachement préférentiel pour la croissance des réseaux, comme un modèle de vie et mort avec des entités évolutives. Plus précisément, les noeuds du réseau gagnent et perdent des unités de population à des probabilités fixes, et de nouveaux noeuds peuvent être ajoutés à un taux également fixe.

Lien entre Gibrat et Attachement Préférentiel

Considérons un modèle de croissance strictement positive de Gibrat donnée par $P_i(t) = R_i(t) \cdot P_i(t-1)$ avec $R_i(t) > 1$, $\mu_i(t) = \mathbb{E}[P_i(t)]$, $\lambda_i(t) = \mathbb{E}[R_i(t)]$ et $\sigma_i(t) = \mathbb{E}[R_i(t)^2]$. Les P_i sont les populations des villes tandis que R_i sont des taux de croissance aléatoires. D'autre part, soit un modèle simple d'attachement préférentiel, avec une probabilité d'attachement $\lambda \in [0, 1]$ et un nombre de nouveau arrivants $m > 0$, ce qui revient en espérance à $\mu_i(t+1) - \mu_i(t) = m \cdot \lambda$. Il est possible de dériver que le Gibrat est statistiquement équivalent à une limite de l'attachement préférentiel, sous l'hypothèse que toutes les fonctions génératrices des moments de $R_i(t)$ existent. Les distributions classiques qui peuvent être utilisées dans ce cas, e.g. une distribution normale ou log-normale, sont entièrement déterminées par leur deux premiers moments, ce qui rend cette hypothèse raisonnable.

Lemme 1 La limite d'un modèle d'attachement préférentiel quand $\lambda \ll 1$ est un modèle de croissance de Gibrat linéaire, avec le paramètres limites $\lambda_i(t) = 1 + \frac{\lambda}{m \cdot (t-1)}$.

Preuve S'intéressant au premier moment, nous notons $\bar{P}_i(t) = \mu_i(t) = \mathbb{E}[P_i(t)]$. L'indépendance entre les taux de croissance de Gibrat donne directement $\bar{P}_i(t) = \mathbb{E}[R_i(t)] \cdot \bar{P}_i(t-1)$. En partant du modèle d'attachement préférentiel, nous avons $\bar{P}_i(t) = \mathbb{E}[P_i(t)] = \sum_{k=0}^{+\infty} k \mathbb{P}[P_i(t) = k]$. Mais par ailleurs,

$$\{P_i(t) = k\} = \bigcup_{\delta=0}^{\infty} (\{P_i(t-1) = k-\delta\} \cap \{P_i \leftarrow P_i + 1\}^\delta)$$

où le second évènement correspond à la ville i étant augmentée δ fois entre $t-1$ et t (avec la convention que les évènements vides pour $\delta \geq k$). Ainsi, en prenant en compte la formulation conditionnelle de l'attachement préférentiel, qui postule que $\mathbb{P}[\{P_i \leftarrow P_i + 1\} | P_i(t-1) = p] =$

$\lambda \cdot \frac{p}{P(t-1)}$ (la population totale $P(t)$ étant déterministe), nous obtenons

$$\begin{aligned} \mathbb{P}[P_i \leftarrow P_i + 1] &= \sum_p \mathbb{P}[P_i \leftarrow P_i + 1 | P_i(t-1) = p] \cdot \mathbb{P}[P_i(t-1) = p] \\ &= \sum_p \lambda \cdot \frac{p}{P(t-1)} \mathbb{P}[P_i(t-1) = p] = \lambda \cdot \frac{\bar{P}_i(t-1)}{P(t-1)} \end{aligned}$$

Ce qui donne, sachant que $P(t-1) = P_0 + m \cdot (t-1)$ et en notant $q = \lambda \cdot \frac{\bar{P}_i(t-1)}{P_0 + m \cdot (t-1)}$

$$\begin{aligned} \bar{P}_i(t) &= \sum_{k=0}^{\infty} \sum_{\delta=0}^{\infty} k \cdot \left(\lambda \cdot \frac{\bar{P}_i(t-1)}{P_0 + m \cdot (t-1)} \right)^{\delta} \cdot \mathbb{P}[P_i(t-1) = k - \delta] \\ &= \sum_{\delta'=0}^{\infty} \sum_{k'=0}^{\infty} (k' + \delta') \cdot q^{\delta'} \cdot \mathbb{P}[P_i(t-1) = k'] \\ &= \sum_{\delta'=0}^{\infty} q^{\delta'} \cdot (\delta' + \bar{P}_i(t-1)) = \frac{q}{(1-q)^2} + \frac{\bar{P}_i(t-1)}{(1-q)} \\ &= \frac{\bar{P}_i(t-1)}{1-q} \left[1 + \frac{1}{\bar{P}_i(t-1)} \frac{q}{(1-q)} \right] \end{aligned}$$

On s'attend à ce que pour la majorité des villes, $\bar{P}_i(t) \ll P(t)$ (distributions fortement dissymétriques), la limite peut être prise pour λ uniquement. En prenant $\lambda \ll 1$, comme $0 < \bar{P}_i(t)/P(t) < 1$, nous obtenons $q = \lambda \cdot \frac{\bar{P}_i(t-1)}{P_0 + m \cdot (t-1)} \ll 1$, qui peut être développée au premier ordre en q . Cela donne finalement

$$\bar{P}_i(t) = \bar{P}_i(t-1) \cdot \left[1 + \left(1 + \frac{1}{\bar{P}_i(t-1)} \right) q + o(q) \right]$$

et donc

$$\bar{P}_i(t) \simeq \left[1 + \frac{\lambda}{P_0 + m \cdot (t-1)} \right] \cdot \bar{P}_i(t-1)$$

Cela signifie que cette limite est équivalente en espérance à un modèle de Gibrat avec $\mu_i(t) = \mu(t) = 1 + \frac{\lambda}{P_0 + m \cdot (t-1)}$.

Pour le second moment, on peut faire un calcul similaire. On a toujours

$$\mathbb{E}[P_i(t)^2] = \mathbb{E}[R_i(t)^2] \cdot \mathbb{E}[P_i(t-1)^2]$$

et

$$\mathbb{E}[P_i(t)^2] = \sum_{k=0}^{+\infty} k^2 \mathbb{P}[P_i(t) = k]$$

On obtient ainsi de la même façon

$$\begin{aligned}
 \mathbb{E}[P_i(t)^2] &= \sum_{\delta'=0}^{\infty} \sum_{k'=0}^{\infty} (k' + \delta')^2 \cdot q^{\delta'} \cdot \mathbb{P}[P_i(t-1) = k'] \\
 &= \sum_{\delta'=0}^{\infty} q^{\delta'} \cdot \left(\mathbb{E}[P_i(t-1)^2] + 2\delta' \bar{P}_i(t-1) + \delta'^2 \right) \\
 &= \frac{\mathbb{E}[P_i(t-1)^2]}{1-q} + \frac{2q \bar{P}_i(t-1)}{(1-q)^2} + \frac{q(q+1)}{(1-q)^3} \\
 &= \frac{\mathbb{E}[P_i(t-1)^2]}{1-q} \left[1 + \frac{q}{\mathbb{E}[P_i(t-1)^2]} \left(\frac{2\bar{P}_i(t-1)}{1-q} + \frac{(1+q)}{(1-q)^2} \right) \right]
 \end{aligned}$$

On a ainsi équivalence entre le modèle de Gibrat et une formulation continue de l'Attachement préférentiel (ou du modèle de Simon) dans la limite donnée ci-dessus. ■

Lien entre Simon et Attachement Préférentiel

Une reformulation du modèle de Simon le présente comme un cas particulier de l'attachement préférentiel généralisé, en particulier avec la probabilité de mort nulle.

Lien entre Favaro-Pumain et Gibrat

[**favaro2011gibrat**] généralise le modèle de Gibrat avec les dynamiques de propagation de l'innovation. En théorie, un équivalent microscopique devrait pouvoir être formulé si on considère l'ensemble des modèles dans une typologie par ontologie et par paradigme. Les modèles Marius [**cottineau2014evolution**] correspondent à un paradigme de Gibrat, et devraient aussi avoir leur contrepartie en termes de formulation microscopique.

B.2 SENSIBILITÉ DES LOIS D'ECHELLE URBAINES

Au centre de la théorie évolutive des villes se trouvent la hiérarchie et les lois d'échelle associées. Nous proposons ici un bref développement méthodologique sur la sensibilité des lois d'échelle à la définition de la ville.

Les lois d'échelle ont été montrées universelles des systèmes urbains à de nombreuses échelles et pour différents indicateurs socio-économiques ou techniques (PIB, éducation, emploi, crime, stock d'infrastructures, stock de logement). Des études récentes questionnent toutefois la cohérence de la détermination des exposants d'échelle, puisque leur valeur peut varier significativement selon les seuils utilisés pour définir les entités urbaines sur lesquelles les quantités urbaines sont intégrées, franchissant même dans certains cas la barrière qualitative de l'échelle linéaire, d'une loi infra-linéaire à une loi super-linéaire. Nous utilisons un modèle théorique simple de distribution spatiale des densités et des fonctions urbaines pour montrer analytiquement qu'un tel comportement peut être dérivé comme conséquence du type de distribution spatiale et de la méthode utilisée.

Les lois d'échelle pour les systèmes urbains, en commençant par la bien connue loi rang-taille de Zipf pour la distribution des tailles des villes [gabaix1999zipf], sont une caractéristique récurrente des systèmes urbains, à différentes échelles et pour différents types d'indicateurs. Elles reposent sur la constatation empirique que des indicateurs calculés sur des éléments du système urbain, qui peuvent être les villes dans le cas d'un système de villes, mais aussi des entités plus petites à une plus petite échelle, suivent relativement bien une distribution en loi de puissance en fonction de la taille de l'entité, i.e. pour l'entité i avec population P_i , on a pour une quantité intégrée A_i , la relation $A_i \simeq A_0 \cdot \left(\frac{P_i}{P_0}\right)^\alpha$. Les exposants d'échelle α peuvent être plus petits ou plus grands que 1, menant à des effets infra ou supra-linéaires. Diverses interprétations thématiques de ce phénomène ont été proposées, typiquement sous la forme d'analyse des processus. La littérature économique contient une production abondante sur le sujet (voir [Gabaix20042341] pour une revue), mais est généralement faiblement spatiale, donc de faible intérêt pour notre approche qui s'intéresse particulièrement à l'organisation spatiale. Des règles économiques simples comme un équilibre énergétique peut conduire à de simples lois d'échelles [bettencourt2008large] mais sont difficiles à ajuster empiriquement. Une proposition intéressante par PUMAIN est qu'elles sont intrinsèquement dues au caractère évolutionnaire des systèmes de villes, et que ces lois correspondent à différents niveaux de maturité dans les cycles d'innovation qui se diffusent hiérarchiquement dans les systèmes de villes [pumain2006evolutionary]. Même si un parallèle tentant peut être fait avec les systèmes biologiques ou physiques auto-organisés, [pumain2012urban] insiste sur le fait

que l'hypothèse d'ergodicité (voir 4.1) pour de tels systèmes n'est pas raisonnable dans le cas de système géographiques et que l'analogie est difficilement exploitable dans le cas des systèmes physiques. D'autres explications ont été proposées à d'autres échelles, comme le modèle de croissance urbaine à échelle mesoscopique (échelle de la ville) donné dans [2014arXiv1401.8200L] qui montre que la congestion dans les réseaux de transport pourrait être une raison de la forme des villes et des lois d'échelle correspondantes. On peut noter que les modèles "classiques" de croissance urbaine comme le modèle de Gibrat [favaro2011gibrat] fournissent une approximation au premier ordre des systèmes exhibant des lois d'échelle, mais que les interactions entre agents doivent être incorporées dans le modèle pour obtenir un résultat plus fidèle aux données réelles, comme le modèle de Favaro-Pumain pour la propagation des cycles d'innovation proposé dans [favaro2011gibrat], qui généralise un modèle de Gibrat pour la croissance des villes françaises avec une ontologie similaire à celle des modèles Simpop.

Cependant, l'application sans vergogne de l'estimation des exposants de lois d'échelle a été récemment rappelé comme pouvant mener à des interprétations divergentes, comme [2013arXiv1301.1674A] qui montre la variabilité des exposants calculés aux paramètres définissant les aires urbaines françaises, comme le seuil de densité. [] étudie empiriquement pour la France l'influence des 3 paramètres jouant un rôle dans la définition de la ville, qui sont un seuil de densité θ pour délimiter les limites d'une aire urbaine, un seuil du nombre de navetteurs θ_c qui correspond à la proportion de ceux-ci devant travailler dans la zone centrale pour que la zone considérée y soit associée, et un paramètre de cut-off P_c en dessous duquel les entités ne sont pas prises en compte pour la régression linéaire fournit l'exposant d'échelle. Un résultat significatif est que les exposants peuvent varier d'un comportement sous-linéaire à un comportement super-linéaire que les seuils varient. Une exploration systématique de l'espace des paramètres produit les diagrammes de phase des exposants pour diverses quantités. Une question qui est directement soulevée est la manière dont ces variations peuvent être expliquées par les caractéristiques de la distribution spatiale des variables. Résultent-elles de mécanismes intrinsèques au système ou peuvent-elles être expliquées simplement par le fait que le système est spatialisé d'une façon particulière ? Nous montrons avec un exemple analytique simplifié que même des distributions spatiales élémentaires induisent une variation significative des exposants le long d'une dimension des paramètres (seuil de densité), suggérant une réponse positive à la deuxième hypothèse.

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.

Formalisons le contexte théorique simple dans lequel nous dérivons la sensibilité des lois d'échelle à la définition de la ville. Considérons ainsi un système de villes polycentrique, dont la distribution spatiale des densité de population peut raisonnablement être estimé par la superposition de noyaux spatiaux rapidement décroissants, comme par exemple un modèle à mixture d'exponentielles [anas1998urban]. Prenant l'espace géographique comme \mathbb{R}^2 , nous prenons pour tout $\vec{x} \in \mathbb{R}^2$ la densité de population comme

$$d(\vec{x}) = \sum_{i=1}^N d_i(\vec{x}) = \sum_{i=1}^N d_i^0 \cdot \exp\left(\frac{-\|\vec{x} - \vec{x}_i\|}{r_i}\right)$$

où r_i sont les paramètres d'étendue des noyaux, d_i^0 les densités aux points d'origine et \vec{x}_i les positions des centres. Nous supposons de plus les contraintes suivantes :

1. Pour simplifier, chaque ville est supposée monocentrique, au sens que pour tout $i \neq j$, nous avons $\|\vec{x}_i - \vec{x}_j\| \gg r_i$.
2. Cela permet d'imposer une loi d'échelle "structurelle" au système urbain en contrignant les populations P_i . On obtient immédiatement par intégration que $P_i = 2\pi d_i^0 r_i^2$, ce qui donne par insertion dans l'hypothèse de loi d'échelle donnée par $\ln P_i = \ln P_{\max} - \alpha \ln i$, la relation suivante entre paramètres : $\ln [d_i^0 r_i^2] = K' - \alpha \ln i$.

Pour étudier les relations de lois d'échelles, nous considérons une variable aléatoire scalaire dans l'espace $a(\vec{x})$ représentant un aspect de la ville, qui peut être n'importe lequel mais a la dimension physique d'une densité spatiale, de telle façon que l'indicateur $A(D) = \mathbb{E}[\iint_D a(\vec{x}) d\vec{x}]$ représente la quantité espérée de a dans la zone D . Nous faisons l'hypothèse que $a \in \{0; 1\}$ (indicateur de comptage) et que sa loi est donnée par $\mathbb{P}[a(\vec{x}) = 1] = f(d(\vec{x}))$. Suivant le travail empirique fait par [], l'indicateur intégré sur la ville i en fonction de θ est donné par

$$A_i(\theta) = A(D(\vec{x}_i, \theta))$$

où $D(\vec{x}_i, \theta)$ est la zone centrée en \vec{x}_i telle que $d(\vec{x}) > \theta$. L'hypothèse 1 ci-dessus assure que les zones sont des cercles relativement disjoints. Nous considérons de plus une aménité qui suit également une loi d'échelle locale au sens que $f(d) = \lambda \cdot d^\beta$. Cette hypothèse semble raisonnable puisqu'il a été montré que de nombreuses variables urbaines suivent un comportement fractal à l'échelle intra-urbaine [keersmaecker2003using] ce qui implique une distribution en loi puissance [chen2010characterizing]. Nous faisons l'hypothèse supplémentaire que $r_i = r_0$ ne dépend pas de i , ce qui est raisonnable

si le système urbain est considéré à une petite échelle. L'exposant d'échelle estimé $\alpha(\theta)$ est alors pris comme le résultat de la régression logarithmique¹ de $(A_i(\theta))_i$ contre $(P_i(\theta))_i$ où $P_i(\theta) = \iint_{D(\vec{x}_i, \theta)} d$.

B.2.1 Dérivation Analytique de la Sensibilité

Avec les notations précédentes, dérivons l'expression de l'exposant estimé pour la quantité a en fonction du paramètre de seuil de densité θ . La quantité calculée pour une ville donnée i est, grâce à l'hypothèse monocentrique et dans une portée spatiale et des bornes pour θ telles que $\theta \gg \sum_{j \neq i} d_j(\vec{x})$, permettant d'approximer $d(\vec{x}) \simeq d_i(\vec{x})$ sur $D(\vec{x}_i, \theta)$, est donnée par

$$\begin{aligned} A_i(\theta) &= \lambda \cdot \iint_{D(\vec{x}_i, \theta)} d^\beta = 2\pi\lambda d_i^0 \beta \int_{r=0}^{r_0 \ln \frac{d_i^0}{\theta}} r \exp\left(-\frac{r\beta}{r_0}\right) dr \\ &= \frac{2\pi d_i^0 \beta r_0^2}{\beta^2} \left[1 + \beta \ln \frac{\theta}{d_i^0} \left(\frac{\theta}{d_i^0} \right)^\beta - \left(\frac{\theta}{d_i^0} \right)^\beta \right] \end{aligned}$$

Nous obtenons de la même manière l'expression de $P_i(\theta)$

$$P_i(\theta) = 2\pi d_i^0 r_0^2 \left[1 + \ln \left[\frac{\theta}{d_i^0} \right] \frac{\theta}{d_i^0} - \frac{\theta}{d_i^0} \right]$$

L'estimation par Moindres-carrés, pour résoudre le problème $\inf_{\alpha, C} \|(\ln A_i(\theta) - C - \alpha \ln P_i(\theta))_i\|^2$, donne la valeur $\alpha(\theta) = \frac{\text{Cov}[(\ln A_i(\theta))_i, (\ln P_i(\theta))_i]}{\text{Var}[(\ln P_i(\theta))_i]}$. Comme nous travaillons aux limites de la ville, le seuil est supposé être significativement plus petit que la densité au centre, i.e. $\theta/d_i^0 \ll 1$. Nous pouvons développer l'expression au premier ordre en θ/d_i^0 et utiliser la loi d'échelle globale pour les tailles des villes, ce qui donne

$$\ln A_i(\theta) \simeq K_A - \alpha \ln i + (\beta - 1) \ln d_i^0 + \beta \ln \frac{\theta}{d_i^0} \left(\frac{\theta}{d_i^0} \right)^\beta$$

et

$$\ln P_i(\theta) = K_P - \alpha \ln i + \ln \left[\frac{\theta}{d_i^0} \right] \frac{\theta}{d_i^0}$$

Le développement des variances et covariances donne finalement une expression de l'exposant d'échelle en fonction de θ , où k_j, k'_j sont des constantes obtenues dans le développement :

$$\alpha(\theta) = \frac{k_0 + k_1 \theta + k_2 \theta^\beta + k_3 \theta^{\beta+1} + k_4 \theta \ln \theta + k_5 \theta^\beta \ln \theta + k_6 \theta^\beta (\ln \theta)^2 + k_7 \theta^{\beta+1} (\ln \theta)^2 + k_8 \theta^{\beta+1} \ln \theta}{k'_0 + k'_1 \ln \theta + k'_2 \theta \ln \theta + k'_3 \theta^2 + k'_4 \theta^2 \ln \theta + k'_5 \theta^2 (\ln \theta)^2}$$

Cette fraction rationnelle en θ et $\ln \theta$ donne l'expression théorique de l'évolution des exposants d'échelle quand le seuil varie.

¹ Nous ne nous plaçons pas dans le cas d'une estimation raffinée des lois d'échelle, qui suppose un cut-off [1].

B.3 GÉNÉRATION DE DONNÉES SYNTHÉTIQUES CORRÉLÉES

Cette section correspond à l'introduction et la formalisation de [raimbault2016generation].

* * *

*

La génération de données synthétiques hybrides similaires à des données réelles présente des enjeux méthodologiques et thématiques pour la plupart des disciplines dont l'objet est l'étude de systèmes complexes. Comme l'interdépendance entre les éléments constitutifs d'un système, matérialisée par leur relations, conduit à l'émergence de ses propriétés macroscopiques, une possibilité de contrôle de l'intensité des dépendances dans un jeu de données synthétiques est un instrument de connaissance du comportement du système. Nous proposons une méthodologie de génération de données synthétiques hybrides sur lequel la structure de correlation est contrôlée. La méthode est illustrée sur des séries temporelles financières en C.5 et permet l'étude de l'interférence entre composantes à différentes fréquences sur la performance d'un modèle prédictif, en fonction des corrélations entre composantes à différentes échelles. La section 5.3 propose par ailleurs une application à un système géographique, dans laquelle le couplage faible d'un modèle de distribution de densité de population avec un modèle de génération de réseau permet la simulation de configurations territoriales. L'exploration intensive du modèle permet l'obtention d'un large spectre de valeurs pour la matrice de corrélation entre mesures morphologiques et mesures du réseau. On démontre ainsi les possibilités d'applications variées et les potentialités de la méthode.

B.3.1 Contexte

L'utilisation de données synthétiques, au sens de populations statistiques d'individus générées aléatoirement sous la contrainte de reproduire certaines caractéristiques du système étudié, est une pratique méthodologique largement répandue dans de nombreuses disciplines, et particulièrement pour des problématiques liées aux systèmes complexes, telles que par exemple l'évaluation thérapeutique [abadie2010synthetic], l'étude des systèmes territoriaux [moeckel2003creating ; pritchard2009advances], l'apprentissage statistique [bolon2013review] ou la bio-informatique [van2006syntren]. Il peut s'agir d'une désagrégation par création d'une population au niveau microscopique présentant des caractéristiques macroscopiques données, ou bien de la création de nouvelles populations au même

niveau d'agrégation qu'un échantillon donné avec un critère de ressemblance aux données réelles. Le niveau de ce critère peut dépendre des applications attendues et peut par exemple aller de la fidélité des distributions statistiques pour un certain nombre d'indicateurs agrégés, c'est à dire l'existence de motifs macroscopiques similaires. Dans le cas de systèmes chaotiques ou présentant de fortes caractéristiques d'émergence, une contrainte microscopique n'implique pas nécessairement le respect des motifs macroscopiques, et arriver à les reproduire est justement un des enjeux des pratiques de modélisation et simulation en sciences de la complexité. La donnée, qu'elle soit simulée, mesurée ou hybride est au cœur de l'étude des systèmes complexes de par la maturation de nouvelles approches computационnelles [arthur2015complexity], il est donc essentiel d'étudier des procédures d'extraction d'information des données (fouille de données) et de simulation d'une information similaire (génération de données synthétiques).

Si le premier ordre est de manière générale bien maîtrisé, il n'est pas systématique ni aisément de contrôler le second ordre, c'est à dire les structures de covariance entre les variables générées, même si des exemples spécifiques existent, comme dans [ye2011investigation] où la sensibilité des sorties de modèles de choix discrets à la forme des distributions des variables aléatoires ainsi qu'à leur structures de dépendance. Il est également possible d'interpréter les modèles de génération de réseaux complexes [newman2003structure] comme la création d'une structure d'interdépendance au sein d'un système, représentée par la topologie des liens. Nous proposons ici une méthode générique prenant en compte l'interdépendance lors de la génération de données synthétiques, sous la forme de correlations.

L'ensemble des méthodologies mentionnées en introduction sont trop variées pour être résumées par un même formalisme. Nous proposons ici une formulation générique ne dépendant pas du domaine d'application, ciblée sur le contrôle de la structure de correlation des données synthétiques.

B.3.2 Formalisation

Soit un processus stochastique multidimensionnel \tilde{X}_I (l'ensemble d'indexation pouvant être par exemple le temps dans le cas de séries temporelles, l'espace, ou une indexation quelconque). On se propose, à partir d'un jeu de réalisations $X = (X_{i,j})$, de générer une population statistique $\tilde{X} = \tilde{X}_{i,j}$ telle que

1. d'une part un certain critère de proximité aux données est vérifié, i.e. étant donné une précision ϵ et un indicateur f , $\|f(X) - f(\tilde{X})\| < \epsilon$

2. d'autre part le niveau de correlation est contrôlé, i.e. étant donné une matrice fixant une structure de covariance R, $\text{Var}[(\tilde{X}_i)] = R$, où la matrice de variance/covariance est estimée sur la population synthétique.

La satisfaction du deuxième point sera généralement conditionnée par la valeur de paramètres, dont dépendra la procédure de génération, qu'il s'agisse de modèles simples ou complexes. Formellement, les processus synthétiques sont des familles paramétriques $\tilde{X}_i[\vec{\alpha}]$. Nous proposons de décliner cette méthode sur deux exemples très différents mais tous deux typiques des systèmes complexes : des séries temporelles financières à haute fréquence, et les systèmes territoriaux. On illustre ainsi la flexibilité de la logique, ouvrant des portes interdisciplinaires par l'exportation de méthodes ou raisonnements par exemple. Dans le premier cas, la proximité aux données est l'égalité des signaux à une fréquence fondamentale, auxquels on superpose des composantes synthétiques dont il est facile de contrôler le niveau de correlation. On se place dans une logique de données hybrides, pour tester des hypothèses ou modèles dans un contexte plus proche de la réalité que sur des données purement synthétiques. Cet exemple est présenté en Appendice C.5. Dans le deuxième cas, la calibration morphologique d'un modèle de distribution de densité de peuplement permet de respecter le critère de proximité aux données. Les correlations de la forme urbaine avec celle d'un réseau de transport sont ensuite obtenues empiriquement par exploration du couplage avec un modèle de génération de réseau. Leur contrôle est dans ce cas indirect puisque constaté empiriquement.

B.4 ROBUSTESSE D'UNE ÉVALUATION MULTI-ATTRIBUTS

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 [raimbault2017discrepancy].
Elle est ici traduite et adaptée.*

* * *

*

Les évaluations multi-objectifs sont un aspect essentiel de la gestion de systèmes complexes, puisque la complexité intrinsèque d'un système est généralement étroitement liée au nombre d'objectifs d'optimisation potentiels. Cependant, une évaluation ne fait pas sens si sa robustesse, au sens de sa fiabilité, n'est pas donnée. Les méthodes statistiques usuelles fournissant une mesure de robustesse sont très dépendantes des modèles sous-jacents. Nous proposons une formulation d'un cadre indépendant du modèle, dans le cas d'indicateurs intégrés et agrégés (évaluation multi-attributs), qui permet de définir une mesure de robustesse relative prenant en compte la structure des données et les valeurs des indicateurs. La méthode est testée sur données urbaines synthétiques associées aux arrondissements de Paris, et à des données réelles de revenus pour l'évaluation de la ségrégation urbaine dans la région métropolitaine du Grand Paris. Les premiers résultats numériques montrent les potentialités de cette nouvelle mé-

thode. De plus, sa relative indépendance au type de système et au modèle pourrait la positionner comme une alternative aux méthodes statistiques classiques d'évaluation de la robustesse.

B.4.1 *Introduction*

Contexte Général

Les problèmes multi-objectifs sont organiquement liés à la complexité des systèmes sous-jacents. En effet, que ce soit dans le champ des *Systèmes Complexes Industriels*, dans le sens de systèmes conçus par ingénierie, où la construction de Systèmes de Systèmes (SoS) par couplage et intégration induit souvent des objectifs contradictoires [marler2004survey], ou dans le champ des *Systèmes Complexes Naturels*, au sens de systèmes non désignés, physiques, biologiques ou sociaux, qui présentent des propriétés d'émergence et d'auto-organisation, pour lesquels les objectifs peuvent e.g. être le résultat de l'interaction d'agents hétérogènes (voir [newman2011complex] pour une revue étendue des types de systèmes concernés par cette approche), l'optimisation multi-objectifs peut être explicitement introduite pour étudier ou désigner le système, mais régit généralement déjà implicitement les mécanismes internes du système. Le cas des Systèmes Complexes Sociaux-techniques est particulièrement intéressant puisque selon Haken [haken2003face], ils peuvent être vus comme des systèmes hybrides embarquant des agents sociaux dans des "artefacts techniques" (parfois jusqu'à un niveau inattendu, créant ce que PICON décrit comme *cyborgs* [picon2013smart]), et cumulent ainsi la potentialité d'être à l'origine de problèmes multi-objectifs². La notion récente d'*éco-quartier* [souami2012ecoquartiers] est un exemple typique pour lequel la durabilité implique des objectifs contradictoires. L'exemple des systèmes de transport, dont la conception a glissé durant la seconde moitié du 20ème siècle d'analyses coût-bénéfices à la price de décision multi-critères, est également typique de tels systèmes [bavoux2005geographie]. Les systèmes géographiques sont à présent bien étudiés d'un tel point de vue, en particulier grâce à l'intégration des cadres multi-objectifs au sein des Systèmes d'Information Géographiques [carver1991integrating]. Comme dans le cas microscopique des éco-quartiers, la planification et le design urbains mésoscopiques et macroscopiques peuvent être rendus durables grâce aux évaluations par indicateurs [jegou2012evaluation].

Un aspect crucial de l'évaluation est une certaine notion de sa fiabilité, que nous nommerons ici *robustesse*. Les méthodes statistiques incluent naturellement cette notion puisque la construction et l'esti-

² Nous désignons ici par *Evaluation Multi-objectifs* toutes les pratiques incluant le calcul de multiples indicateurs d'un système (il peut s'agir d'optimisation multi-objectif pour un design de système, une évaluation multi-objectif d'un système existant, une évaluation multi-attributs ; notre cadre particulier correspondra au dernier cas).

mation de modèles statistiques donne divers indicateurs de la consistance des résultats [launer2014robustness]. Le premier exemple venant à l'esprit est l'application de la loi des grands nombres pour obtenir la *p-valeur* d'une estimation de modèle, qui peut être interprété comme une mesure de confiance en les valeurs estimées. D'autre part, les intervalles de confiance et le *beta-power* sont d'autres indicateurs importants de robustesse statistique. L'inférence bayésienne fournit également des mesures de robustesse quand la distribution des paramètres est estimée de manière séquentielle. Concernant les optimisations multi-objectifs, en particulier par des algorithmes heuristiques (comme par exemple les algorithmes génétiques, ou les solveurs de recherche opérationnelle), la notion de robustesse d'une solution consiste plus en la stabilité de la solution dans l'espace des phases du système dynamique correspondant. Des progrès récents ont été faits vers une formulation unifiée de la robustesse pour les problèmes d'optimisation multi-objectifs, comme dans [deb2006introducing] où les fronts de Pareto robustes sont définis comme des solutions insensibles aux petites perturbations. Dans [1688537], la notion de degré de robustesse est introduite, formalisée comme une sorte de continuité des autres solutions dans des voisinages successifs d'une solution.

Cependant, il n'existe pas de méthode générique qui permettrait une évaluation de la robustesse de façon indépendante au modèle, i.e. qui serait extraite de la structure des données et des indicateurs mais ne dépendrait pas de la méthode utilisée. Un avantage serait par exemple une estimation *a priori* de la robustesse potentielle d'une évaluation et de décider ainsi si elle vaut la peine d'être faite. Nous proposons un cadre répondant à cette contrainte dans le cas particulier des évaluations multi-attributs, i.e. quand le problème est rendu unidimensionnel par agrégation des objectifs. Il est basé sur les données et non sur les modèles, au sens où l'estimation de la robustesse ne dépendra pas de la manière dont les indicateurs sont calculés, tant qu'ils respectent certaines hypothèses détaillées par la suite.

Approche Proposée

OBJECTIFS COMME INTÉGRALES SPATIALES Nous supposons que les objectifs peuvent être exprimés comme intégrales spatiales, ce qui devrait s'appliquer à tout système territorial, et nos cas d'application sont des systèmes urbains. Ce n'est pas si restrictif en terme d'indicateurs possibles si l'on utilise les bonnes variables et noyaux intégrés : de façon analogue à la méthode de Regression Géographique Pondérée [brunsdon1998geographically], toute variable spatiale peut être intégrée contre des noyaux réguliers de taille variable et le résultats sera une agrégation spatiale dont la signification dépendra de l'étendue du noyau. Les exemples utilisés par la suite comme des moyennes conditionnelles ou des sommes vérifient parfaitement cette hypothèse. Même un indicateur déjà agrégé dans l'espace peut être

interprété comme une intégrale spatiale en utilisant une distribution de Dirac au centroïde de la zone correspondante.

OBJECTIFSAGRÉGÉSLINÉAIREMENT Une seconde hypothèse que nous faisons est que l'évaluation multi-objectifs est effectuée par agrégation linéaire des objectif, c'est à dire qu'on se place dans le cadre d'un problème d'optimisation multi-attributs. Si $(q_i(\vec{x}))_i$ sont les valeurs des fonctions objectifs, on définit alors des poids $(w_i)_i$ afin de construire la fonction de prise de décision $q(\vec{x}) = \sum_i w_i q_i(\vec{x})$, dont la valeur détermine ensuite la performance d'une solution. Cette approche est analogue aux utilités agrégées en économie et est utilisée dans de nombreux domaines. La subtilité réside dans le choix des poids, i.e. de la forme de la fonction de projection, et différentes solutions ont été développées pour obtenir des poids selon la nature du problème. Récemment, [dobbie2013robustness] a proposé de comparer la robustesse des différentes techniques d'agrégation par une analyse de sensibilité, effectuée par simulations de Monte-Carlo pour produire des données synthétiques, ce qui permet d'obtenir la distribution des biais pour les différentes techniques, certaines étant significativement plus performantes que d'autres. Toutefois, la quantification de la robustesse dépend toujours des modèles utilisés dans ce travail.

Le reste de cette monographie est organisé de la façon suivante : la section 2 décrit intuitivement puis mathématiquement le cadre proposé ; la section 3 détaille ensuite l'implémentation, la collecte des données pour les cas d'étude et les résultats numériques pour une évaluation intra-urbaine synthétique et un cas réel métropolitain ; la section 4 discute finalement les limitations et les potentialités de la méthode.

B.4.2 *Description du Cadre*

Description Intuitive

Nous décrivons à présent le cadre proposé pour permettre théoriquement de comparer la robustesse d'évaluation de deux systèmes urbains différents. Ce cadre est une généralisation d'une méthode empirique proposée dans [ecodistrictReport] pour accompagner une étude dans un autre contexte effectuant une comparaison du sens et de la pertinence des indicateurs dans un contexte de durabilité. Intuitivement, la base empirique se base sur les principes suivants :

- Les systèmes urbains peuvent être vus selon l'information disponible, i.e. les données brutes décrivant le système. Dans une approche basée sur les données, celles-ci sont la base de notre cadre et la robustesse sera déterminée par leur structure.

- A partir des données sont capturés des indicateurs (fonctions objectifs). Nous supposons qu'un choix d'indicateurs est une intention particulière de traduire des aspects particuliers du système, i.e. de capturer une réalisation d'un "fait urbain" au sens de MANGIN [mangin1999projet] - une sorte de fait stylisé en terme de processus et de mécanismes, ayant différentes réalisations sur des systèmes distincts dans l'espace, dépendant de chaque contexte géographique précis.
- Etant donné plusieurs systèmes et indicateurs associés, un espace commun peut être construit pour les comparer. Dans cet espace, les données représentent plus ou moins bien le système réel, c'est à dire qu'elles sont imprécises en fonction de l'échelle initiale, de la précision effective des données. Nous proposons de capturer exactement ces différents aspects au travers de la notion de discrépance d'un nuage de points, qui est un outil mathématique provenant des théories d'échantillonnage, permettant d'exprimer la façon dont un jeu de données rempli l'espace dans lequel il s'insère [dick2010digital].

Synthétisant ces contraintes, nous proposons une notion de *Robustesse* d'une évaluation qui capture à la fois, en combinant la fiabilité des données à l'importance relative des indicateurs,

1. *Données manquantes* : une évaluation se basant sur des jeux de données plus raffinés sera naturellement plus robuste.
2. *Importance des indicateurs* : les indicateurs avec plus d'importance relative pèsent plus dans la robustesse totale.

Description Formelle

INDICATEURS Soit $(S_i)_{1 \leq i \leq N}$ un nombre fini de systèmes territoriaux, que nous supposons décrits par les données brutes et des indicateurs intermédiaires, donnés par $S_i = (X_i, Y_i) \in \mathcal{X}_i \times \mathcal{Y}_i$ avec $\mathcal{X}_i = \prod_k \mathcal{X}_{i,k}$ tel que chaque sous-espace contient des matrices réelles : $\mathcal{X}_{i,k} = \mathbb{R}^{n_{i,k}^X p_{i,k}^X}$ (de la même façon pour \mathcal{Y}_i). Nous définissons également une fonction d'indice ontologique $I_X(i, k)$ (resp. $I_Y(i, k)$) prenant des valeurs entières qui coïncident si et seulement si les deux variables ont même ontologie au sens de [livet2010], c'est à dire qu'elles sont supposées représenter le même objet réel. On distingue les "données brutes" X_i à partir desquelles les indicateurs sont calculés généralement par des fonctions déterministes explicites, des "indicateurs intermédiaires" Y_i qui sont déjà intégrés et peuvent être par exemple

les sorties de modèles élaborés simulant certains aspects du système urbain. Nous définissons l'espace caractéristique du "fait urbain" par

$$(X, Y) \underset{\text{def}}{=} \left(\prod \tilde{X}_c \right) \times \left(\prod \tilde{Y}_c \right) = \left(\prod_{x_{i,k} \in \mathcal{D}_X} \mathbb{R}^{p_{i,k}^X} \right) \times \left(\prod_{y_{i,k} \in \mathcal{D}_Y} \mathbb{R}^{p_{i,k}^Y} \right) \quad (23)$$

avec $\mathcal{D}_X = \{x_{i,k} | I(i, k) \text{ distincts, } n_{i,k}^X \text{ maximal}\}$ (de même pour Y_i). Il s'agit en fait de l'espace abstrait sur lequel les indicateurs sont intégrés. Les indices c introduit par définition correspondent aux différents indicateurs au sein des différents systèmes. Cette espace est l'espace minimal commun à tous les systèmes permettant une définition commune des indicateurs pour tous.

Soit $X_{i,c}$ les données projetées canoniquement sur le sous-espace correspondant, bien définies pour tout i et tout c . Nous faisons donc l'hypothèse clé que tous les indicateurs sont calculés par intégration contre un noyau donné, i.e. pour tout c il existe H_c espace de fonctions à valeurs réelles sur $(\tilde{X}_c, \tilde{Y}_c)$, tel que pour tout $h \in H_c$:

1. h est "suffisamment" régulière (distribution tempérée par exemple)
2. $q_c = \int_{(\tilde{X}_c, \tilde{Y}_c)} h$ est une fonction décrivant le "fait urbain" (l'indicateur en lui-même)

Des exemples typiques de noyaux peuvent être :

- Une moyenne des lignes de $X_{i,c}$ est calculée par $h(x) = x \cdot f_{i,c}(x)$ où $f_{i,c}$ est la densité de la distribution de la variable sous-jacente.
- Un taux d'éléments du jeu de données respectant une condition donnée C , $h(x) = f_{i,c}(x)\chi_C(x)$.
- Pour des variables déjà agrégées Y , une distribution de Dirac permet de les exprimer également comme des intégrales de noyaux.

AGRÉGATION La détermination des poids est en fait le point crucial des processus de prise de décision multi-attributs, et de nombreuses méthodes sont disponibles (voir [wang2009review] pour une revue dans le cas particulier de la gestion de l'énergie durable). Définissons les poids pour l'agrégation linéaire. Nous supposons les indicateurs normalisés, i.e. $h_c \in [0, 1]$, pour une construction plus simple des poids relatifs. Pour i, c et $h_c \in H_c$ donnés, le poids $w_{i,c}$ est simplement constitué par l'importance relative de l'indicateur $w_{i,c}^I = \frac{\hat{q}_{i,c}}{\sum_c \hat{q}_{i,c}}$ où $\hat{q}_{i,c}$ est un estimateur de q_c pour les données $X_{i,c}$ (i.e. la valeur calculée effectivement). On peut noter que cette étape n'est pas contrainte et que cela peut être étendu à tout ensemble d'attribution

de poids, en prenant par exemple $\tilde{w}_{i,c} = w_{i,c} \cdot w'_{i,c}$ si w' sont les poids fixés par le preneur de décisions. Nous nous concentrerons sur l'influence relative des attributs et pour cela choisissons cette forme simple pour les poids.

ESTIMATION DE LA ROBUSTESSE La scène est à présent apprêtée pour construire une estimation de la robustesse d'une évaluation faite par la fonction d'agrégation. Pour cela, nous appliquons un théorème d'approximation d'intégrale similaire au méthodes introduites dans [varet2010developpement], puisque la forme intégrée des indicateurs permet justement de bénéficier de tels résultats théoriquement puissant. Soit $\mathbf{X}_{i,c} = (\vec{X}_{i,c,l})_{1 \leq l \leq n_{i,c}}$ et $D_{i,c} = \text{Disc}_{\vec{X}_c, L^2}(\mathbf{X}_{i,c})$ le discrépance du jeu de données³ [niederreiter1972discrepancy]. Avec $h \in H_c$, on a la borne supérieure sur l'erreur d'approximation de l'intégrale

$$\left\| \int h_c - \frac{1}{n_{i,c}} \sum_l h_c(\vec{X}_{i,c,l}) \right\| \leq K \cdot \|h_c\| \cdot D_{i,c}$$

où K est une constante indépendante des points de données et des fonctions objectifs. Cela donne directement

$$\left\| \int \sum_c w_{i,c} h_c - \frac{1}{n_{i,c}} \sum_l w_{i,c} h_c(\vec{X}_{i,c,l}) \right\| \leq K \sum_c |w_{i,c}| \|h_c\| \cdot D_{i,c}$$

En supposant l'erreur réalisée de manière raisonnable (scénario du "pire de cas" pour la connaissance de la valeur théorique de la fonction agrégée), nous prenons cette borne supérieure comme une approximation de sa magnitude. De plus, la normalisation des indicateurs implique que $\|h_c\| = 1$. Nous proposons alors de comparer les bornes d'erreurs entre deux évaluations. Elle dépendent seulement de la distribution des données (équivalence à la *robustesse statistique*) et des indicateurs choisis (sorte de *robustesse ontologique*, i.e. est-ce que les indicateurs ont un sens réel dans le contexte choisi et est-ce que leur valeur fait sens), et sont un moyen de combiner ces deux types de robustesse dans une seule valeur.

Nous définissons ainsi un *ratio de robustesse* pour comparer la robustesse de deux évaluations par

$$R_{i,i'} = \frac{\sum_c w_{i,c} \cdot D_{i,c}}{\sum_c w_{i',c} \cdot D_{i',c}} \quad (24)$$

³ La discrépance est définie comme la norme-L2 de la discrépance locale qui est pour des points de données normalisés $\mathbf{X} = (x_{ij}) \in [0, 1]^d$, une fonction de $t \in [0, 1]^d$ comparant le nombre de points compris dans le volume de l'hypercube correspondant, donné par $\text{disc}(t) = \frac{1}{n} \sum_i \mathbb{1}_{\prod_j x_{ij} < t_j} - \prod_j t_j$. C'est une mesure de la manière dont le nuage de points couvre l'espace.

L’interprétation intuitive de cette définition est que l’on compare la robustesse des évaluations en comparant la plus grande erreur faite dans chaque cas selon la structure des données et l’importance relative.

En construisant une relation d’ordre sur les évaluations en comparant la position du ratio par rapport à un, il est clair qu’on obtient un ordre complet sur l’ensemble des évaluations possibles. Ce ratio devrait en théorie permettre de comparer n’importe quelle évaluation d’un système urbain. Afin de garder un sens ontologique à cela, il devrait être utilisé pour comparer des sous-systèmes disjoints avec une proportion raisonnable d’indicateurs en commun, ou le même sous-système avec des indicateurs différents. On peut noter que cela fournit un moyen de tester l’influence des indicateurs sur une évaluation, en analysant la sensibilité du ratio à leur suppression. Au contraire, la détermination d’un nombre “minimal” d’indicateurs faisant chacun varier le ratio fortement pourrait être un moyen d’isoler des paramètres essentiels régissant le sous-système.

B.4.3 Résultats

IMPLÉMENTATION Le pré-traitement des données géographiques est fait via QGIS [**qgis2011quantum**] pour des raisons d’ergonomie. L’implémentation du cœur est faite en R [**team200or**] pour la flexibilité de la gestion des données et du traitement statistique. De plus, le package DiceDesign [**franco20092**] conçu pour les expériences numériques et l’échantillonnage, permet un calcul efficient et direct des discordances. Enfin, tout aussi important, l’ensemble du code source est disponible de manière ouverte sur le dépôt git du projet⁴ pour permettre la reproductibilité et la réutilisation [**ram2013git**].

Implémentation sur Données Synthétiques

Nous proposons dans un premier temps d’illustrer l’implémentation par une application à des données et indicateurs synthétiques, pour des indicateurs de qualité de vie intra-urbaine pour la ville de Paris.

COLLECTE DES DONNÉES Le cas virtuel se base sur des données géographiques réelle, en particulier pour les arrondissements parisiens. Nous utilisons les données disponibles par le projet OpenStreetMap [**bennett2010openstreetmap**] qui fournit déjà des données précises en haute définition pour de nombreux aspects urbains. Nous utilisons le réseau de rues et la position des bâtiments dans la ville de Paris. Les limites des arrondissements, utilisées pour agréger et extraire les features lorsqu’on travaille sur un seul district, sont aussi pris de la même source. Nous utilisons les centroïdes des polygones

⁴ à <https://github.com/JusteRaimbault/RobustnessDiscrepancy>

des bâtiments et les segments du réseau de rues. Le jeu de données brutes consiste d'environ 200k bâtiments et 100k segments de rues.

CAS VIRTUEL Nous travaillons sur chaque arrondissement de Paris (du 1er au 20ème) comme un système urbain évalué. Des données synthétiques aléatoires sont associées aux features spatiales, chaque arrondissement pouvant alors être évalué de manière stochastique, et des répétitions permettent d'obtenir le comportement statistique moyen des indicateurs jouets et des ratios de robustesse. Les indicateurs choisis doivent être calculés comme des indicateurs résidentiels et du réseau de rues. Pour montrer différents exemples, nous implémentons deux kernels moyens et une moyenne conditionnelle, tous liés à la durabilité environnementale et la qualité de vie, chacun devant être maximisés. On peut noter que ces indicateurs ont un sens réel mais pas de raison particulière d'être agrégés, ils sont ici choisis pour l'aspect pratique du modèle jouet et de la génération de données synthétiques. Avec $a \in \{1 \dots 20\}$ le nombre d'arrondissements, $A(a)$ l'aire spatiale correspondante à chacun, $b \in B$ les coordonnées des bâtiments et $s \in S$ les segments de rues, nous prenons

- Le complémentaire de la distance journalière moyenne au travail en voiture par individu, approché par, avec $n_{cars}(b)$ nombre de voiture dans le bâtiment (généré aléatoirement en associant des voitures à bâtiments proportionnel au taux de motorisation attendu α_m 0.4 à Paris), d_w distance des individus à leur travail (généré à partir du bâtiment vers un point aléatoire distribué uniformément dans l'étendue spatiale du jeu de données), et d_{max} le diamètre de l'aire de Paris, $\bar{d}_w = 1 - \frac{1}{|B \in A(a)|} \cdot \sum_{b \in A(a)} n_{cars}(b) \cdot \frac{d_w}{d_{max}}$
- Le complémentaire des flots moyens de voitures des rues dans la zone, approché par, avec $\varphi(s)$ flot relatif dans le segment de rue s , généré par le minimum entre 1 et une distribution log-normale ajustée pour avoir 95% de masse plus petite que 1, ce qui mimique la distribution hiérarchique de l'utilisation des rues (qui correspond à la centralité de chemin), et $l(s)$ longueur du segment, $\bar{\varphi} = 1 - \frac{1}{|S \in A(a)|} \cdot \sum_{s \in A(a)} \varphi(s) \cdot \frac{l(s)}{\max(l(s))}$
- Longueur relative de rues piétonnes \bar{p} , calculé via une dummy variable aléatoire uniforme ajustée pour obtenir une proportion fixée de segments piédestre.

Comme les données synthétiques sont stochastiques, les simulations sont lancées pour chaque quartier $N = 50$ fois, ce qui était un compromis raisonnable entre convergence statistique et temps nécessaire au calcul. La table 1 montre les résultats (moyennes et déviations standard) des valeurs des indicateurs et le calcul du ratio de robustesse. Les déviations standard obtenues confirment que ce nombre de

TABLE 24: Résultats numériques des simulations pour chaque arrondissement avec $N = 50$ répétitions. Chaque valeur des indicateurs factice est donnée par sa moyenne sur les répétitions et la déviation standard associée. Le ratio de robustesse est calculé par rapport au premier arrondissement (choix arbitraire). Un ratio inférieur à 1 signifie que la borne de l'intégrale est plus petite pour le premier système, i.e. que l'évaluation est plus robuste pour celui-ci.

Arrdt	$\langle \bar{d}_w \rangle \pm \sigma(\bar{d}_w)$	$\langle \bar{\varphi} \rangle \pm \sigma(\bar{\varphi})$	$\langle \bar{p} \rangle \pm \sigma(\bar{p})$	$R_{i,1}$
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

simulations donnent des résultats constants. Les indicateurs obtenus en fixant un ratio fixe montre peu de variabilité, ce qui peut être une limite de cette approche jouet. On obtient toutefois le résultat intéressant que la majorité des arrondissements donne des évaluations plus robustes que le 1er arrondissement, ce qui était attendu par la taille et la fonction de ce quartier : il s'agit en effet d'un petit quartier avec de grand bâtiment administratifs, ce qui implique moins d'éléments spatiaux et pour cela une évaluation moins robuste selon la définition qu'on en a donnée.

Application à un cas réel : ségrégation métropolitaine

Le premier exemple avait pour but de montrer les potentialités de la méthode mais était purement synthétique, ne pouvant pour cela fournir pas de conclusion concrete ni d'implications pour la gouvernance. Nous proposons maintenant de l'appliquer à des données réelles dans le cas de la ségrégation métropolitaine.

DONNÉES Nous travaillons sur les données de revenus, disponible pour la France à un niveau intra-urbain (unités statistiques élémentaires IRIS) pour l'année 2011 sous la forme de résumé statistiques (déciles uniquement si la zone est peuplée suffisamment pour assurer l'anonymat), fournies par l'INSEE⁵. Les données sont associées à l'étendue géographique des unités statistiques, permettant le calcul d'indicateurs d'analyse spatiale.

INDICATEURS Nous utilisons ici trois indicateurs de ségrégation intégrés sur une zone géographique. Supposons la zone divisée en unités couvrantes S_i pour $1 \leq i \leq N$ avec pour centroïdes (x_i, y_i) . Chaque unité a des caractéristiques de population P_i et de revenu médian X_i . On définit des poids spatiaux utilisés pour quantifier l'intensité des interactions géographiques entre unités i, j , avec d_{ij} distance euclidienne entre centroïdes : $w_{ij} = \frac{P_i P_j}{(\sum_k P_k)^2} \cdot \frac{1}{d_{ij}}$ si $i \neq j$ et $w_{ii} = 0$. Les indicateurs normalisés sont les suivants

- Indice d'autocorrelation spatiale de Moran, défini comme la covariance pondérée normalisée du revenu médian par $\rho = \frac{N}{\sum_{ij} w_{ij}} \cdot \frac{\sum_{ij} w_{ij} (X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}$
- Indice de dissimilarité (proche du Moran mais intégrant les dissimilarités locales plutôt que les corrélations), donné par $d = \frac{1}{\sum_{ij} w_{ij}} \sum_{ij} w_{ij} |\tilde{X}_i - \tilde{X}_j|$
avec $\tilde{X}_i = \frac{X_i - \min(X_k)}{\max(X_k) - \min(X_k)}$
- Le complémentaire de l'entropie de la distribution des revenus, qui est une façon de capturer des inégalités globales $\varepsilon = 1 + \frac{1}{\log(N)} \sum_i \frac{X_i}{\sum_k X_k} \cdot \log \left(\frac{X_i}{\sum_k X_k} \right)$

De nombreuses mesures de ségrégation avec différentes signification à différentes échelles existent, comme par exemple à l'échelle d'une unité spatiale élémentaire par comparaison de la distribution de revenus empirique avec un modèle nul [louf2015patterns]. Le choix est ici arbitraire, afin d'illustrer la méthode avec un nombre raisonnable de dimensions.

⁵ <http://www.insee.fr>

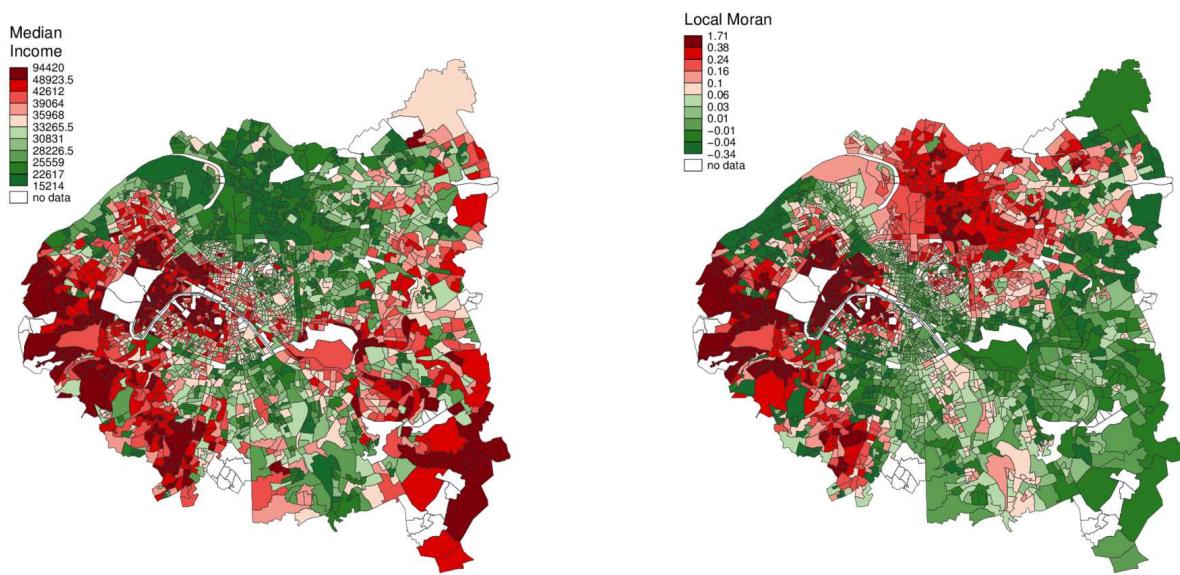


FIGURE 104: Cartes de ségrégation métropolitaine. Les cartes montrent le revenu annuel médian pour les unités statistiques élémentaires (IRIS) pour les trois départements correspondant globalement à la métropole du Grand Paris, et l'index local d'autocorrelation spatiale de Moran correspondant, défini pour l'unité i par $\rho_i = N / \sum_j w_{ij} \cdot \frac{\sum_j w_{ij} (X_j - \bar{X})(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2}$. Les zones les plus ségrégées coincident avec les plus riches et les plus pauvres, suggérant une augmentation de la ségrégation dans les cas extrêmes.

RÉSULTATS La méthode est appliquée avec ces indicateurs à la zone du Grand Paris, constitué de 4 département qui sont des niveaux administratifs intermédiaires. La création récente d'un nouveau système de gouvernance métropolitaine [gilli2009paris] met en évidence des interrogations sur sa pertinence, notamment sur ses capacités d'atténuer les inégalités spatiales. On peut voir en Fig. 104 les cartes de la distribution spatiale du revenu médian et de l'index local d'autocorrelation spatiale correspondant. La dichotomie bien connue entre est et ouest est retrouvée ainsi que la disparité des quartiers intra-muros, comme cela été présenté par diverses études, comme [guerois2009dynamique] à travers l'analyse des dynamiques des transactions immobilières. Notre cadre d'étude est ensuite appliqué à une question concrète ayant des implications pour la prise de décision : *dans quelle mesure une évaluation de la ségrégation au sein de différents territoires est sensible aux données manquantes?* Pour cela, on procède à des simulations de Monte-Carlo (75 répétitions) pour lesquelles une proportion fixe de données est supprimée aléatoirement, et l'indice de robustesse correspondant est évalué avec les indicateurs normalisés. Les simulations sont faites sur chaque département de façon indépendante, à chaque fois pour une robustesse relative à l'évaluation du Grand Paris complet. Les résultats sont présentés en Fig. ???. Toutes les zones ont une robustesse légèrement meilleure que la référence, ce qui pourrait être expliqué par une homogénéité locale et donc des indices de ségrégation plus fiables. Les implications pour la prise de décision qui peuvent être par exemple tirées sont des comparaisons directes entre les zones : une perte de 30% de l'information sur le 93 correspond à une perte de seulement 25% pour le 92. La première zone étant déjà défavorisée socio-économiquement, l'inégalité est augmentée par cette qualité moindre de l'information statistique. L'étude des déviations standard suggère des études plus approfondies comme différents régimes de réponse à la suppression de données semblent exister.

B.4.4 Discussion

Applicabilité à des situations réelles

IMPLICATIONS POUR LA PRISE DE DÉCISION L'application de notre méthode à des situations concrètes de prise de décision peu être pensée de différentes manières. Tout d'abord dans le cas d'un processus multi-attributs à but comparatif, comme la détermination d'un corridor pour une nouvelle infrastructure de transport, l'identification des territoires sur lesquels l'évaluation pourrait être biaisée (i.e. avec une mauvaise robustesse relative) devrait permettre une attention particulière pour ceux-ci, et l'adaptation des jeux de données ou la révision des points en conséquence. Dans tous les cas le processus total devrait être plus fiable. Une autre possibilité ressemble

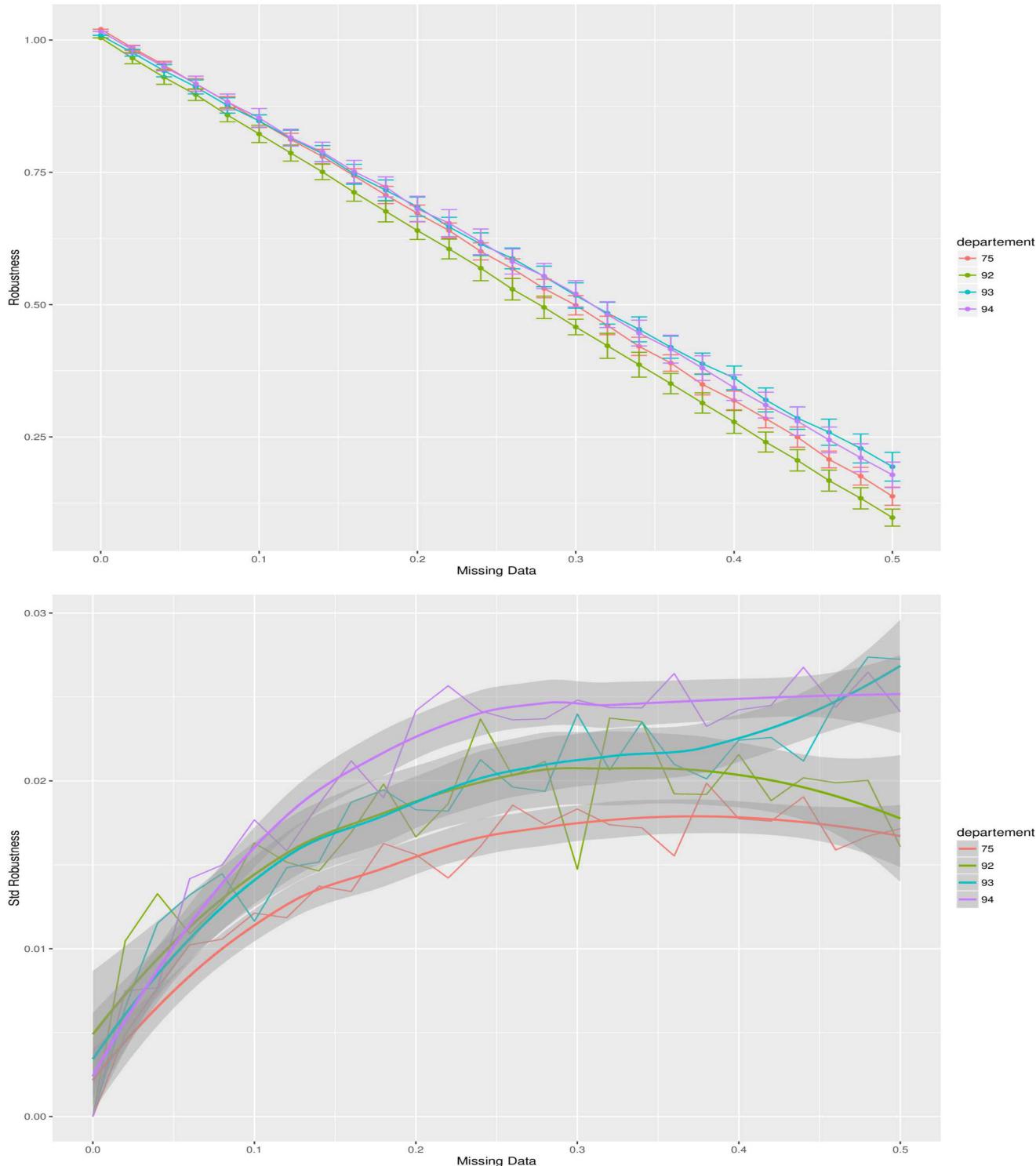


FIGURE 105: Sensibilité de la robustesse aux données manquantes. (Gauche) Pour chaque département, des simulations de Monte-Carlo ($N=75$ répétitions) sont utilisées pour déterminer l'impact des données manquantes sur la robustesse de l'évaluation de la ségrégation. Les ratios de robustesse sont tous calculés relativement à la région métropolitaine complète avec toutes les données disponibles. Le comportement quasi-linéaire traduit une décroissance approximativement linéaire de la discrépance en fonction de la taille des données. Les trajectoires similaires des départements les plus pauvres (93,94) suggère que la correction au comportement linéaire est fonction des motifs de ségrégation. (Droite) Déviations standard des ratios de robustesse. Les différents régimes (en particulier le 93 contre les autres) révèlent des transitions de phase à différents niveaux de données manquantes, signifiant que l'évaluation dans le 94 est de ce point de vue plus sensible aux données manquantes.

à l'application réelle que nous avons développé, i.e. la sensibilité de l'évaluation à divers paramètres comme les données manquantes. Si une décision paraît fiable car la taille de données est grande, mais que l'évaluation est très sensible à la suppression de données, il faudra être prudent pour l'interprétation des résultats et pour la prise de décision finale. Un travail approfondi et de test sera cependant nécessaire pour comprendre le comportement du cadre dans différents contextes et pouvoir piloter son application dans des situations réelles diverses.

INTÉGRATION AU SEIN DE CADRES EXISTANTS L'applicabilité de la méthode à des cas réels dépendra directement de son intégration potentielle dans des environnements existants. Au delà des difficultés techniques qui apparaissent nécessairement en essayant de coupler ou d'intégrer des implémentations existantes, des obstacles plus théoriques pourraient émerger, comme des formulations floues des fonctions ou des types de données, la cohérence des bases de données, etc. De tels cadres multi-critères sont nombreux. Un développement possible serait l'intégration dans un cadre open-source, comme par exemple celui décrit dans [tivadar2014oasis] qui calcule divers indices de ségrégation urbaine, comme on l'a déjà illustré pour l'application à la ségrégation métropolitaine.

DISPONIBILITÉ DES DONNÉES BRUTES De manière générale, des données sensibles comme des questionnaires de transport, ou des données de sondage à granularité très fine, ne sont pas disponibles de manière ouverte, mais fournis de manière déjà agrégée à un certain niveau (comme par exemple les données françaises de l'Insee sont disponibles publiquement au niveau des unités statistiques élémentaires ou pour des zones plus grandes selon les variables et des contraintes de population minimale, les données plus précises étant à accès restreint). Cela signifie que l'application de notre cadre peut impliquer une procédure de recherche de données laborieuse, l'avantage d'être flexible étant alors compensé par ces contraintes additionnelles.

Validité des hypothèses théoriques

Une limitation possible de notre approche est la validité de l'hypothèse qui formule les indicateurs comme des intégrales spatiales. En fait, de nombreux indicateurs socio-économiques ne dépendent pas nécessairement directement de l'espace, et essayer de les associer à des coordonnées peut entraîner sur une pente glissante (par exemple, associer des variables économiques individuelles à des coordonnées résidentielles aura un sens seulement si la variable a une relation à l'espace, autrement un devient un artefact superflu). Même des indicateurs qui ont une valeur spatiale peuvent dériver de variables non-spatiales, comme [kwan1998space] le souligne au sujet de l'accessi-

bilité, en opposant les mesures d'accessibilité intégrée aux mesures individu-centrées mais pas forcément basée sur l'espace (comme par exemple des décisions individuelles). Contraindre une représentation théorique d'un système pour le faire rentrer dans un cadre en changeant certaines de ses propriétés ontologiques (toujours dans le sens de la signification réelle des objets) peut être compris comme une violation d'une des règles pour la modélisation et la simulation en sciences sociales données par [banos2013HDR], car cela impliquerait qu'il pourrait exister un langage universel pour la modélisation, malgré qu'il ne puisse retranscrire certains systèmes, ayant pour conséquences des conclusions errantes à cause d'une rupture d'ontologie dans le cas d'une formulation sur-contrainte.

Généralité du Cadre

Nous soutenons qu'un des avantages fondamentaux de notre cadre est sa généralité et sa flexibilité, puisque la robustesse des évaluations est obtenue seulement par la structure des données si l'on relaxe les hypothèses sur les valeurs des poids. Des approfondissement pourraient inclure une formulation plus générale, en supprimant par exemple l'hypothèse d'agrégation linéaire. Des fonctions d'agrégation non-linéaires demanderaient toutefois de vérifier certaines propriétés regardant les inégalités intégrales. Par exemple, des résultats similaires pourraient être obtenus en s'orientant vers des inégalités intégrales pour fonctions Lipschitziennes, comme les résultats en une dimension de [dragomir1999ostrowski].

Conclusion

Nous avons proposé un cadre indépendant du modèle pour comparer la robustesse d'évaluations multi-attributs entre différents systèmes urbains. A partir de la discrépance des données, on fournit une définition générale de la robustesse relative sans aucune hypothèse de modèle pour le système, mais en supposant une agrégation linéaire des objectifs et des indicateurs exprimés comme des intégrales à noyaux. Nous proposons une première implémentation preuve de concept pour la ville de Paris pour laquelle les résultats numériques confirment la tendance générale attendue, et une implémentation sur des données réelles pour la ségrégation de revenus pour la région métropolitaine du Grand Paris, fournissant des réponses possibles à des questions de prise de décision plus concrètes. Des développements possibles peuvent inclure une analyse de sensibilité de la méthode, des applications à d'autres cas réels et une relaxation des hypothèses théoriques, c'est à dire de l'agrégation linéaire et de la formulation comme intégrale spatiale.

B.5 UN CADRE POUR LES SYSTÈMES SOCIO-TECHNIQUES

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 9.3 (sans encore inclure la structure algébrique pour l'opération sur les données).

B.5.1 Contexte

Contexte Scientifique

Les malentendus structurels entre les Sciences Sociales et Humanités d'une part, et les dénommées Sciences Exactes d'autre part, comme celui maintes fois évoqué déjà entre physiciens et géographes, loin d'être une règle nécessaire, semble toutefois avoir un impact conséquent sur la structure de la connaissance scientifique : [2015arXiv151103981H] montre comment la sociologie et la physique ont développé des méthodes d'analyse de réseau très similaire avec une inter-fertilisation faible. Ceux-ci peuvent être dus aux divergences épistémologiques qui elles-mêmes découlent de différences fondamentales dans les objets étudiés : les humains ne sont bien sûr pas des particules. Plus particulièrement, comme nous développons ici différents cadres théoriques, il est important de s'intéresser au rôle de celle-ci. La théorie, et en fait la signification elle-même du terme, a une place complètement différente dans l'élaboration de la connaissance, en partie à cause de différentes *complexités perçues*⁶ des objets étudiés. Par exemple, de nombreuses constructions mathématiques et par extension certaines en physique théorique sont *simples* au sens où elles sont résolubles de manière analytique (ou au moins semi-analytique)⁷, tandis que les sujets des Sciences Sociales tels les humains ou la société (pour prendre un exemple préconçu) sont *complexes* au sens de systèmes complexes. Cela implique un besoin accru d'une construction théorique (qui se base généralement sur l'empirique) pour identifier et définir qui sont nécessairement plus arbitraires dans la définition de leur limites, relations et processus, de par la multitude des points de vue possibles : PUMAIN suggère en effet dans [pumain2005cumulativite] une nouvelle approche de la complexité qui serait profondément ancrée dans les sciences sociales et qui serait "mesurée par la diversité des disciplines nécessaires pour élaborer une notion". Ces différences de fond

⁶ Nous utilisons le terme *perçu* car la plupart des systèmes étudiés en physique peuvent être décrits comme simple alors qu'ils sont intrinsèquement complexe et finalement mal compris [laughlin2006different].

⁷ nous prenons ici le parti que soluble analytiquement implique la simplicité, puisque le système n'exhibe alors pas d'émergence faible (voir 3.3).

sont naturellement bénéfiques pour la diversité scientifique, mais les choses peuvent se corser quand les terrains d'étude se chevauchent, typiquement dans le cas de problématiques liées aux systèmes complexes comme déjà détaillé, comme l'exemple géographique des systèmes urbains a récemment montré [[dupuy2015sciences](#)]. La Science des Systèmes Complexes⁸ est présentée par certains comme "un nouveau type de science" [[wolfram2002new](#)], et serait au moins symptomatique d'un changement de paradigme des pratiques, des approches analytiques "exactes" vers des approches computationnelles et *evidence-based* [[arthur2015complexity](#)], mais il est certain que cela permet de faire émerger, conjointement avec de nouvelles méthodologies, des nouveaux champs scientifiques au sens d'intérêts convergents de disciplines variées sur des questions transversales ou d'approches intégrées d'un champ particulier [[2009arXiv0907.2221B](#)]. Notre travail s'ancre particulièrement dans ce cadre et n'aurait pas de sens s'il était déconnecté de ces aspects notamment computationnels (voir [3.1](#)).

Objectifs

Dans ce contexte scientifique, l'étude de ce que nous désignons par *Systèmes socio-techniques*, que nous définissons de manière assez large comme des systèmes complexes hybrides qui incluent des agents ou objets sociaux qui interagissent avec des artefacts techniques et/ou un environnement naturel⁹, se situent précisément entre sciences sociales et sciences dures. L'exemple des systèmes urbains est relativement représentatif, puisque même avant l'arrivée de nouvelles approches prétendant être "plus exactes" que les approches des sciences sociales (typiquement par des physiciens, voir e.g. le positionnement de [[louf2014scaling](#)]), mais aussi par des chercheurs venant des sciences sociales comme BATTY [[batty2013new](#)]), une multitude d'aspects de l'étude des systèmes urbains étaient déjà traités dans des sciences dures très diverse, parmi lesquelles on peut citer sans hiérarchie particulière, l'hydrologie urbaine, la climatologie urbaine ou les aspects techniques des systèmes de transport, tandis que le centre de leur attention se reposait sur des sciences sociales comme la géographie, l'urbanisme, la sociologie, l'économie. D'où une place nécessaire de la théorie dans leur étude, vu son rôle comme domaine de connaissance pour la connaissance des systèmes complexes (voir le cadre introduit en [9.3](#)).

⁸ que nous appelons délibérément ainsi même si des débats existent sur le fait de considérer comme une science en elle-même ou comme une façon différente de faire de la science.

⁹ les systèmes géographiques au sens de [[dolfus1975some](#)] sont l'archetype de tels systèmes, mais cette définition peut couvrir d'autres types de systèmes comme un système de transport étendu, des systèmes sociaux pris dans un contexte environnemental, des systèmes industriels compliqués considérés avec leur utilisateurs, etc.

Nous proposons dans cette section de construire une théorie, ou plutôt un cadre théorique, pour faciliter certains aspects de l'étude de tels systèmes. De nombreuses théories existent déjà dans l'ensemble des champs liés à ce type de questionnement, et aussi à de plus haut niveaux d'abstraction concernant des méthodes comme e.g. la modélisation basée agent, mais il n'existe à notre connaissance pas de cadre théorique qui incluraient l'ensemble des points suivants que nous jugeons cruciaux (et qui peuvent être compris comme une base informelle de notre théorie) :

1. une définition précise et une emphase particulière sur la notion de couplage entre sous-systèmes, en particulier permettant de qualifier ou quantifier un certain niveau de couplage : dépendance, interdépendance, etc. entre composantes.
2. une précise définition de l'échelle, incluant l'échelle temporelle et l'échelle pour d'autres dimensions.
3. en conséquence des points précédents, une définition précise de ce qu'est un système.
4. la prise en compte de la notion d'émergence pour capturer les aspects multi-scalaires des systèmes.
5. une place centrale de l'ontologie dans la définition des systèmes, i.e. du sens dans le monde réel donné à ses objets¹⁰.
6. la prise en compte d'aspects hétérogènes du même système, qui peuvent être des composantes hétérogènes mais aussi différents points de vue sur le système qui se complètent.

La suite de cette section est organisée de la façon suivante : nous construisons la théorie dans la sous-section suivante en restant à un niveau abstrait, et proposons une première application à la question des sous-systèmes co-évolutifs. Nous discutons ensuite le positionnement au regard de théories existantes, ainsi que les développements possibles et des applications concrètes.

B.5.2 Construction de la Théorie

Perspectives et Ontologies

Le point de départ pour construire la théorie est une approche épistémologique perspectiviste des systèmes introduite par GIERE [[giere2010scientific](#)]. Pour résumer, cette position interprète toute démarche scientifique comme une perspective, au sein de laquelle chacun poursuit certains

¹⁰ comme déjà expliqué précédemment, ce positionnement combiné à l'importance de la structure pourrait être relié au *Réalisme Structurel Ontologique* dans des approfondissements.

objectifs et utilise ce qui est appelé *un modèle* pour les atteindre. Le modèle n'est alors rien de plus qu'un medium scientifique. VARENNE a développé [varenne2010framework] une typologie fonctionnelle des modèles qui peut être interprété comme un raffinement de cette théorie. Relâchons dans un premier temps cette précision potentielle et utilisons les perspectives comme des approximations des objets et concepts indéfinis. En effet, diverses visions du même objet (pouvant être complémentaires ou divergentes) ont la propriété de partager au moins l'objet lui-même, d'où notre proposition de définir les objets (et plus généralement les systèmes) à partir d'un ensemble de perspectives sur ceux-ci, qui vérifient certaines propriétés que nous formalisons par la suite.

Une perspective est définie dans notre cas comme une *Dataflow Machine M* au sens de [golden2012modeling], que nous considérons comme une boîte noire transformant un flux de données d'entrée en flux de sortie à une échelle de temps associée, et qui correspond au model comme medium. Celle-ci fournit un moyen adapté de représenter un modèle et d'y associer échelle de temps et données. On y associe un ontologie O au sens de [livet2010], i.e. un ensemble d'éléments qui correspondent à une entité (qui peut être un objet, un agent, un processus, un état, un concept, c'est à dire tout élément modulaire formalisable) du monde réel. Nous incluons seulement ces deux aspects (le modèle et les objets représentés) de la théorie de Giere, en faisant l'hypothèse que le but et le producteur de la perspective sont en fait contenus dans l'ontologie s'ils font sens pour l'étude du système : par exemple, dans le cas des sondages subjectifs en anthropologie ou sociologie, le sondeur est un élément clé et sera nécessairement inclus dans l'ontologie. De même pour l'objectif poursuivi, tout particulièrement en sciences humaines où la recherche n'est jamais neutre comme nous l'avons vu en 3. Formalisons cette définition :

Définition 2 *Une perspective sur un système est donnée par une Dataflow Machine M = (i, o, T) et une Ontologie associée O. Nous supposons que l'ontologie peut être décomposée de manière discrète en éléments atomiques O = (O_j)_j.*

Les éléments atomiques de l'ontologie peuvent être des constituants particuliers du systèmes, comme des agents ou des composantes, mais aussi des processus, interactions, états ou concepts par exemple. L'ontologie peut être vue comme la description exhaustive et rigoureuse du contenu de la perspective. L'hypothèse d'une *Dataflow Machine* implique que les entrées et sorties potentielles peuvent être quantifiées, ce qui n'est pas nécessairement restrictif aux perspectives quantitatives, puisque la plupart des approches qualitatives peuvent être traduites en variables discrètes à partir du moment où l'ensemble des possibles est connu ou supposé.

Nous définissons alors le système de manière “réciproque”, i.e. à partir d'un ensemble de perspectives sur ce qui constitue alors le système :

Définition 3 *Un système est un ensemble de perspectives sur un système : $S = (M_i, O_i)_{i \in I}$, où I n'est pas nécessairement fini.*

Nous désignons par $\mathcal{O} = (O_{j,i})_{j,i \in I}$ l'ensemble des éléments dans les ontologies.

Comme on part des perspectives sur un système pour définir le système dans son ensemble, il n'y a pas de contradiction. On peut noter qu'à ce stade de la construction, il n'existe pas nécessairement de cohérence structurelle, au sens d'une correspondance avec une structure réelle, sur ce qu'on appelle un système, puisque étant donné notre définition très large nous pourrions par exemple considérer un système comme une perspective sur un véhicule conjointement à une perspective sur un système de villes, ce qui ne fait pas raisonnablement sens. Des définitions approfondies et développements doivent permettre de se rapprocher des définitions classiques d'un système (entités en interaction, artefacts précisément définis, etc.). De la même manière, la définition d'un sous-système sera donnée plus loin. Les éléments de l'approche déjà introduits permettent jusqu'ici de répondre aux points trois, cinq et six des recommandations.

PRÉCISION SUR L'ASPECT RÉCURSIF DE LA THÉORIE Une conséquence directe de ces définitions doit être détaillée : le fait qu'elles peuvent être appliquées de manière récursive. En effet, on peut imaginer prendre comme perspective un système dans notre sens, c'est à dire un ensemble de perspectives sur un système, et le faire à tout ordre. Si on considère un système à n'importe quel sens classique, alors le premier ordre peut être interprété comme une épistémologie du système, i.e. l'étude de perspectives sur un système. Une ensemble de perspectives sur des systèmes en relation peut sous certaines conditions être un domaine ou un champ d'étude, et donc un ensemble de perspectives sur diverses perspectives l'épistémologie d'un champ. On peut proposer des analogies supplémentaires pour traduire l'idée derrière le caractère récursif de la théorie. C'est en effet crucial pour la signification et la cohérence de la théorie, notamment pour les raisons suivantes : (i) le choix des perspectives qui constituent un système est nécessairement subjectif et peut donc être compris comme une perspective en lui-même, et ainsi une perspective sur un système si l'on est en mesure de construire une ontologie générale; (ii) nous utiliserons des relations entre ontologies par la suite, dont la construction est basée sur l'émergence est également subjective et vue comme perspectives. Ces aspects de réflexivité sont fondamentaux, en écho à la discussion de 3.3 sur la production de connaissance et la nature de la complexité.

Graphé Ontologique

Nous proposons ensuite la structure du système en reliant les ontologies. Cette approche pourrait éventuellement être mise en perspective par rapport à un positionnement épistémologique de réalisme structurel [frigg2011everything], c'est à dire que les théories tendent à capturer une certaine structure existante du monde réel, puisqu'une connaissance du monde est ici partiellement contenue dans la structure des modèles, tout en gardant à l'esprit que notre position s'en éloigne en partie de par la conjugaison des perspectives qui induit un certain "degré de constructivisme" comme expliqué en 3.3. Pour cette raison, nous faisons le choix d'appuyer le rôle de l'émergence, suivant l'intuition qu'il pourrait s'agir d'un outil pratique minimaliste pour capturer de façon raisonnable la structure d'un système complexe¹¹. Nous prenons pour cet aspect le positionnement de BEDAU sur les différents types d'émergence déjà présenté plusieurs fois, en particulier sa définition de l'émergence faible donnée dans [bedau2002downward].

Rappelons brièvement les définitions que nous utiliserons par la suite. BEDAU commence par définir les propriétés émergentes puis étend le concept aux phénomènes, entités, etc. De la même manière, notre cadre n'est pas restreints aux objets ou propriétés et inclut ainsi les définitions généralisées comme lien entre ontologies. Nous appliquons la notion d'émergence sous les deux formes suivantes¹² :

- *Emergence nominale* : une ontologie O' est inclue dans une autre ontologie O mais l'aspect de O qui est dit nominalement émergent en rapport à O' ne dépend pas de O' .
- *Emergence faible* : une partie d'une ontologie O peut être dérivée de manière computationnelle par agrégation et interactions entre les éléments d'une ontologie O' .

Comme développé précédemment, la présence d'émergence, et spécifiquement d'émergence faible, constitue une perspective en soi. Elle peut être conceptuelle et postulée comme un axiome dans une théorie thématique, mais aussi expérimentale si des traces d'émergence faible sont effectivement mesurées entre objets. Dans tous les cas, la relation entre ontologies doit être encodée dans une ontologie, ce qui n'était pas nécessairement introduit dans la définition initiale d'un système. Ainsi pour simplifier, les perspectives permettent de décomposer le système en briques ontologiques spécifiant une description "complète".

¹¹ ce qui bien sûr ne peut être formulé comme une affirmation prouvable car cela dépendra de la définition d'un système, etc.

¹² la troisième forme rappelée par BEDAU, l'*émergence forte*, ne sera pas utilisée, car nous avons besoin de capturer rien de plus des relations de dépendance et d'autonomie, et l'émergence faible est plus adéquate en termes de systèmes complexes, puisqu'elle n'assume pas "des pouvoirs causaux irréductibles" aux objets des échelles supérieures à un niveau donné. L'émergence nominale est utilisée pour capturer des relations d'inclusion entre les ontologies.

Nous faisons pour cette raison l'hypothèse suivante importante par la suite :

Hypothèse 3 *Un système peut être partiellement structuré par son extension avec une ontologie qui contient (pas nécessairement uniquement) des relations entre les éléments des ontologies de ses perspectives. Nous la désignons ontologie de couplage et supposons son existence par la suite. Nous postulons de plus son atomicité, i.e. si O est en relation avec O' , alors tout sous-ensemble de O, O' ne peuvent être en relation, ce qui n'est pas contraignant puisqu'une décomposition en des sous-ensembles indépendants assurera cette propriété si elle n'est pas vérifiée initialement.*

Cette hypothèse revient concrètement qu'il est possible de coupler des perspectives, c'est à dire souvent des modèles en pratique, et que ce couplage peut être représenté de façon similaire. Notre expérience pratique du couplage tout au long de nos travaux nous pousse à faire cette hypothèse : tant que les systèmes considérés sont "raisonnables" (choisi raisonnablement l'un par rapport à l'autre, et donc choisi pour être couplés en quelque sorte), il est toujours possible de les coupler.

Cela nous permet d'exhiber des relations d'émergence pas seulement au sein d'une perspective elle-même, mais également entre les éléments de différentes perspectives. Nous définissons ensuite des relations de pré-ordre entre les sous-ensemble des ontologies :

Proposition 3 *Les relations binaires suivantes sont des pré-ordres sur $\mathcal{P}(\mathcal{O})$:*

- *Emergence (basée sur l'émergence faible) : $O' \preceq O$ si et seulement si O émerge faiblement de O' .*
- *Inclusion (basée sur l'émergence nominale) : $O' \Subset O$ si et seulement si O émerge nominalement de O' .*

Avec la convention qu'il peut être admis qu'un objet émerge de lui-même, on a réflexivité (si une telle convention paraît absurde, on peut définir les relations comme O émerge de O' ou $O = O'$). La transitivité est clairement contenue dans la définition de l'émergence.

Notons que la relation d'inclusion est plus général qu'une inclusion entre ensembles, puisqu'elle traduit une inclusion "au sein" des éléments de l'ontologie. Par exemple, une ontologie peut supposer un couplage fort non-décomposable (qui serait une hypothèse de la perspective en elle-même), et une autre perspective contenir l'un des éléments de ce couplage. Nous allons voir que ces relations d'ordre vont nous permettre de définir un graphe par l'algorithme de réduction qui suit.

Définition 4 *Le graphe ontologique est construit par induction de la manière suivante :*

1. Un graphe est construit, avec pour noeuds des éléments de $\mathcal{P}(\mathcal{O})$ et des liens de deux types : $E_W = \{(O, O') | O' \preccurlyeq O\}$ et $E_N = \{(O, O') | O' \sqsubseteq O\}$
2. Les noeuds sont réduits¹³ par : si $o \in O, O'$ et $(O' \preccurlyeq O \text{ ou } O' \sqsubseteq O)$ mais pas $(O \preccurlyeq O' \text{ or } O \sqsubseteq O')$, alors $O' \leftarrow O' \setminus o$
3. Les noeuds avec des ensemble se recouplant sont fusionnés, en gardant les liens liant des noeuds fusionnés. Cette étape assure des noeuds ne se recouplant pas.

Arbre Ontologique Minimal

La structure topologique du graphe, qui contient en un sens la *structure du système*, peut être réduite en un arbre minimal qui capture la structure hiérarchique essentielle pour la théorie.

Nous devons d'abord donner cohérence au système :

Définition 5 Une partie cohérente du graphe ontologique est une composante du graphe faiblement connectée au sens d'un graphe dirigé. Nous assumons pour la suite travailler sur une partie cohérente.

La notion de système cohérent, ainsi que de sous-système ou d'échelle de temps des noeuds qui seront définies par la suite, nécessite de reconstruire des perspectives à partir des éléments ontologiques, i.e. l'opération inverse de ce qui a été fait dans notre procédure qui peut être vue comme une deconstruction.

Hypothèse 4 Il existe $\mathcal{O}' \subset \mathcal{P}(\mathcal{O})$ tel que pour tout $O \subset \mathcal{O}'$, il existe une Dataflow Machine M correspondante telle que la perspective correspondante est cohérente avec les éléments initiaux du système (i.e. les machine sont équivalentes sur les parties communes des ontologies). Si $\Phi : M \mapsto O$ est la correspondance initiale, nous notons cette construction réciproque étendue par $M' = \Phi^{<-1>}(O)$.

REMARQUE Cette hypothèse pourrait éventuellement être changée en une proposition prouvable, en supposant que l'ontologie de couplage correspond effectivement à une perspective de couplage, dont la composante *Dataflow Machine* est cohérente avec les entités couplées. Ainsi, le postulat de décomposition de [golden2012modeling] devrait permettre d'identifier des composantes de base correspondantes à chaque élément de l'ontologie, et construire ainsi la nouvelle perspective par induction. Nous trouvons toutefois ces hypothèses trop restrictives, puisque par exemple divers éléments de l'arbre ontologique peuvent être modélisés par la même machine irréductible, à l'image d'une équation différentielle aux variables agrégées. Nous

¹³ la procédure de réduction vise à supprimer la redondance, gardant une entité au plus haut niveau où elle existe.

préférions être moins restrictifs et postuler l'existence de la correspondance inverse sur certaines sous-ontologies, qui devraient être en pratique celles sur lesquelles le couplage peut effectivement être modélisé.

Grace à l'hypothèse ci-dessus, on peut définir le système cohérent comme l'image réciproque de la partie cohérente du graphe ontologique. Cela permet la connectivité du système qui est un pré-requis pour la construction de l'arbre.

Proposition 4 *La décomposition arborescente du graphe ontologique dans laquelle les noeuds contiennent les composantes fortement connexes est unique. L'arbre réduit, qui correspond au graphe ontologique les composantes fortement connexes ont été fusionnées et les liens gardés, est nommé Arbre Ontologique Minimal.*

Preuve (esquisse) L'unicité découle de la définition univoque puisque les noeuds sont fixés comme les composantes fortement connexes. Il s'agit trivialement d'une décomposition en arbre puisque dans un graphe dirigé, les composantes fortement connexes ne se recoupent pas, d'où la cohérence de la décomposition.

Toute boucle $O \rightarrow O' \rightarrow \dots \rightarrow O$ dans le graphe ontologique suppose que tous ses éléments sont équivalents au sens de \preccurlyeq . Ces boucles d'équivalence devrait aider à définir la notion de couplage fort comme une application de la théorie, avec cependant un caractère qualitatif dans la nature du couplage, ne permettant pas une définition fine de la force de couplage par exemple.

L'Arbre Minimal Ontologique (MOT) est un arbre au sens non-dirigé, mais une forêt au sens dirigé. Sa topologie contient une représentation des hiérarchies du système. Les sous-systèmes cohérents sont définis à partir de l'ensemble \mathcal{B} des branches de la forêt, comme $(\Phi^{<-1>}(\mathcal{B}), \mathcal{B})$. L'échelle de temps d'un noeud, et par extension d'un sous-système, est l'union est échelles de temps des machines correspondantes. Les niveaux de l'arbre sont définis à partir des noeuds racine, et les relations d'émergence entre les noeuds implique une inclusion verticale entre échelles de temps.

Action sur des Données

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 9.3, impliquerait le couplage de ce cadre avec le cadre de connaissances qui y est introduit.

Echelles

Enfin, nous proposons de définir les échelles associées à un système. Suivant [manson2008does], un continuum épistémologique de visions sur l'échelle est une conséquence des différences propres à chaque discipline, comme nous avons développé en introduction. Cette proposition est en fait compatible avec notre cadre, puisque la construction d'échelles pour chaque niveau de l'arbre ontologique résulte en une grande variété d'échelles.

Soit (M, O) un sous-système et T l'échelle de temps correspondante. Nous proposons de définir "l'échelle thématique" (par exemple l'échelle spatiale) en supposant un théorème de représentation, i.e. qu'un aspect (aspect thématique) de la machine peut être représenté par une variable d'état dynamique $\vec{X}(t)$. Etant donné un opérateur d'échelle¹⁴ $\|\cdot\|_s$ et que la variable d'état est différentiable à un certain niveau, l'échelle thématique pour cet aspect, c'est à dire l'échelle typique à laquelle les agents ou processus correspondants opèrent (pouvant être multiple si l'opérateur est multidimensionnel), est définie par $\|(d^k \vec{X}(t))_k\|_s$.

B.5.3 Applications et discussion

Le cas particulier des systèmes géographiques

Dans [dolfus1975some], DURAND-DASTÈS introduit une définition des systèmes et structures géographiques, la structure étant le contenu spatial des systèmes vus comme des systèmes complexes ouverts en interaction (donné par ses éléments et leur attributs, les relations entre éléments et les entrée/sorties avec le monde extérieur). Pour un système donné, sa définition est une perspective, complété par la structure pour avoir un système selon notre sens. Selon la manière dont les relations sont définies, cela peut être plus ou moins aisément d'extraire la structure ontologique.

Modularité et sous-systèmes en co-évolution

Pour l'exemple des systèmes urbains, la théorie évolutive des villes entre dans ce cadre en utilisant notre théorie thématique développée dans la section précédente. La décomposition en sous-systèmes décorrélés fournit précisément des composantes fortement couplées comme des composantes en co-évolution. La corrélation entre sous-systèmes devrait d'une certaine façon être corrélée à la distance topologique dans l'arbre. Si on définit les éléments d'un noeud avant réduction comme *éléments fortement couplés*, dans le cas d'ontologies

¹⁴ qui peut être de nature variée : étendue, étendue probabiliste, échelles spectrales, échelles de stationnarité, etc.

dynamiques, cela fournit une définition de la *co-évolution* et de sous-systèmes en co-évolution, équivalente à la définition thématique.

Discussion

LIEN AVEC DES CADRES EXISTANTS Un lien avec le cadre de Cottineau-Chapron pour la multi-modélisation [10.1371/journal.pone.0138212] pourrait être fait dans le cas où ils ajouteraient la couche bibliographique, qui correspondrait à la reconstruction des perspectives. [reymond2013logique] propose la notion de “couplage interdisciplinaire” qui est proche de notre notion de coupler des perspectives. Une correspondance avec les approches de Système de Systèmes (voir e.g. [luzeaux2015formal] pour un cadre récent englobant la modélisation et la description des systèmes) pourrait être également possible puisque nos perspectives sont construites comme des *Dataflow Machines*, mais avec la différence cruciale que la notion d’émérgence est centrale dans notre cas.

CONTRIBUTION À L’ÉTUDE DES SYSTÈMES COMPLEXES Nous ne prétendons pas exhiber une théorie des systèmes (il faut généralement se méfier de la cybernétique, la systémique etc. qui ne peuvent pas tout modéliser), mais plutôt un cadre majoritairement axiomatique et la structure associée pour guider les questions de recherche (e.g. dans notre cas les conséquences directes sont les études d’épistémologie quantitative qui vient de la construction des systèmes comme perspectives ; les études empiriques pour construire des ontologies robustes pour les perspectives ; des études thématiques ciblées pour révéler des relations causales ou l’émérgence pour la construction des réseaux ontologiques ; l’étude des couplages comme processus contenant possiblement de la co-évolution ; l’étude des échelles ; etc.). Cela peut être compris comme une meta-théorie dont l’application donne une théorie, la théorie thématique qui précède étant une implémentation aux systèmes territoriaux en réseau. Nous appuyons la notion de système socio-technique, croisant une approche des systèmes sociaux complexes (ontologies) avec une description des artefacts techniques (*Dataflow Machines*), prenant “le meilleur des deux mondes”.

Réflexivité

Nous pouvons tirer de l’application de ce cadre à notre travail, c’est à dire d’une réflexivité, une clarification des directions de recherche menées jusqu’ici, et donc de la co-construction des réponses à ces questions avec les différents cadres théoriques.

1. L’approche perspectiviste implique une compréhension large des perspectives existantes sur un système, et des possibilités de couplage entre celle-ci ; d’où une emphase sur l’épistémologie quantitative qui inclue la revue systématique algorithmique (exploration de l’espace des connaissances), la cartographie des

connaissances (extraction de sa structure) et de possibilités de fouille de contenu (raffinement au niveau atomique de la connaissance scientifique) qui correspondent au travail de 2.2.

2. A un niveau plus fin de particularité, la connaissance des perspectives signifie une connaissance des faits stylisés empiriques, comme par exemple ceux pour le traffic routier 8.2, les prix des carburants 8.3, les formes urbaines et de réseau 4.1.

★ ★

★

B.6 EXPLORATION D'UN PAYSAGE SCIENTIFIQUE INTERDISCIPLINAIRE

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.2 pour la production commune), initié pour les 20 ans de la revue en mai 2016. Les résultats préliminaires ont été présentés comme [raimbault2016indirect] à la conférence anniversaire, et le texte de cette annexe est extrait et traduit de [raimbault2017exploration].

* * *

*

Les motifs d'interdisciplinarité en science peuvent être quantifiés au travers de diverses dimensions complémentaires. Cette monographie étudie comme cas d'étude l'environnement scientifique d'un journal généraliste en géographie, *Cybergeo*, afin d'introduire une nouvelle méthodologie qui combine analyse du réseau de citation et analyse sémantique. Nous construisons un corpus massif d'environ 200,000 articles avec leur résumés et le réseau de citation correspondant qui fournit une première classification par citations. Les mots-clés pertinents sont extraits pour chaque article par analyse textuelle, permettant la construction d'une classification sémantique. Nous étudions les motifs qualitatifs de relations entre disciplines endogènes au sein de chaque classification, et montrons finalement la complémentarité des classifications et des mesures d'interdisciplinarité associées. Les outils développés en conséquence sont ouverts et réutilisables pour des études similaires à grande échelle d'environnements scientifiques.

B.6.1 *Introduction*

Cette section développe un cas d'étude qui couple exploration et analyse du réseau de citation avec analyse textuelle, dans le but de cartographier le paysage scientifique du voisinage d'un journal particulier. Nous choisissons d'étudier un journal électronique en Géographie, appelé *Cybergeo*¹⁵, qui publie des articles dans l'ensemble des champs de la Géographie et est de cette façon multi-disciplinaire. Le choix est initialement dû à la disponibilité des données, mais répond à différentes contraintes le rendant particulièrement pertinent dans le contexte décrit précédemment. Tout d'abord, la "discipline" de la Géographie est très large et par essence interdisciplinaire [**bracken2016interdisciplinarity**] : le spectre s'étend de la Géographie Humaine et Critique à la Géographie Physique et la géomorphologie, et les interactions entre ces sous-champs sont nombreuses. Dans un second temps, les données bibliographiques sont difficiles à obtenir, soulevant la difficulté de la façon dont un paysage scientifique est perçu peut être influencée par les acteurs de la dissémination et donc loin d'être objectif, ce qui rend ainsi des solutions techniques comme celle que nous développerons en conséquence ici des outils cruciaux pour une science ouverte et neutre. Enfin, il s'agit d'un cas d'étude particulièrement intéressant puisque la politique éditoriale est généraliste et préoccupée par des questions de science ouverte comme la transparence de l'éthique de revue par les pairs [[10.1371/journal.pone.0147913](https://doi.org/10.1371/journal.pone.0147913)], les pratiques d'ouverture des données et des modèles, comme rappelé par [**pumain2015adapting**], et ce travail contribue à celles-ci en encourageant l'ouverture de la reflexivité.

Le contexte méthodologique dans lequel ce travail se situe est développé en 2.2. Cette contribution est originale et significative sur au moins deux aspects :

1. nous combinons des classifications endogènes dans l'idée d'un réseau multi-couches, en utilisant l'information sémantique ;
2. un jeu de données massif est construit à partir de rien pour étudier un journal non référencé dans les bases principales, répondant ainsi à des problèmes à la fois liés à la collecte de données et au traitement de données massives.

Le reste de cette section est organisée de la façon suivante : nous décrivons d'abord le jeu de données utilisé et la procédure de collecte des données. Nous étudions ensuite les propriétés du réseau de citation et décrivons la procédure pour construire la classification sémantique par analyse textuelle. Nous étudions finalement des mesures complémentaires d'interdisciplinarité obtenues par les différentes classifications.

¹⁵ <http://cybergeo.revues.org/>

b.6.2 Construction de la base de données

Notre approche impose des contraintes sur le jeu de données utilisé, à savoir : (i) couvrir un certain voisinage du journal étudié dans le réseau de citation pour avoir une vue cohérente sur le paysage scientifique ; (ii) avoir au moins une description textuelle pour chaque noeud. Pour satisfaire celles-ci, nous devons rassembler et compiler des données de sources hétérogènes. Nous utilisons pour cela une application spécifiquement conçue, dont l'architecture générale est donnée en Fig. 106. Le code source de l'application et l'ensemble des scripts utilisés ici sont disponibles sur le dépôt git ouvert du projet¹⁶. Les données brutes et traitées sont également disponibles de manière ouverte sur Dataverse¹⁷. Nous rappelons qu'une importante contribution de ce travail est la construction d'un tel jeu de données hybride à partir de sources hétérogènes, et le développement d'outils associés qui peuvent être utilisés et étendus dans des cadres similaires.

Corpus initial

La base de production de *Cybergeo* (snapshot pris en février 2016, fournit par l'équipe éditoriale), permet d'obtenir après pré-traitement la base initiale des articles, avec information élémentaire (titre, résumé, année de publication, auteurs). La version traitée que l'on utilise est disponible en même temps que l'ensemble des données construites, comme dump mysql, à l'adresse donnée précédemment. Cette base fournit également les enregistrements de la bibliographie des articles, qui fournissent l'ensemble des références citées par la base initiale (citations *données* par le corpus initial).

Données de citation

Les données de citation sont collectées à partir de Google Scholar, qui est la seule source pour les citations entrantes [noruzi2005google] dans notre cas puisque le journal n'est pas référencé dans les autres bases¹⁸. Nous sommes conscients des possibles biais de l'utilisation de cette source unique (voir par exemple [bohannon2014scientific])¹⁹, mais ces critiques sont plutôt dirigées vers les résultats de recherche ou de possibles manipulations ciblées que envers la structure globale du réseau de citation. La collecte automatique demande l'utilisation d'un logiciel de collecte pour transférer les requêtes, à savoir TorPool [torpool] qui fournit une API Java permettant une intégration aisée au sein de notre application de collecte. Un crawler peut ainsi récupérer les pages html et obtenir les citations inverses, c'est à

¹⁶ à <https://github.com/JusteRaimbault/HyperNetwork>

¹⁷ à <http://dx.doi.org/10.7910/DVN/VU2XKT>

¹⁸ ou vient d'y être ajouté comme dans le cas du *Web of Science* qui n'indexe *Cybergeo* que seulement depuis mai 2016

¹⁹ ou <http://iscpif.fr/blog/2016/02/the-strange-arithmetic-of-google-scholars>

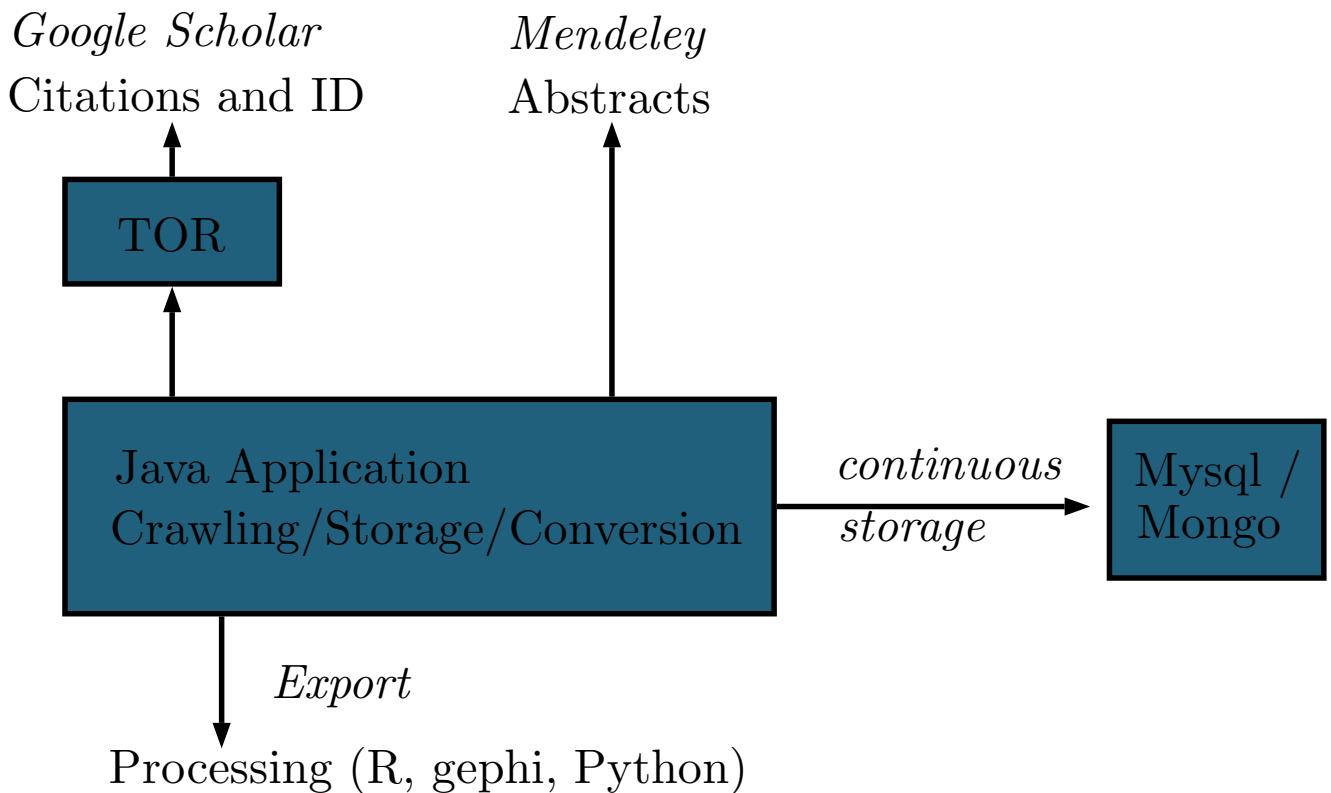


FIGURE 106: Collecte et traitement de données bibliographiques hétérogènes. Architecture de l'application pour la récolte des données de contenu (données sémantiques), des métadonnées et des données de citation. L'hétérogénéité des tâches requière l'utilisation de multiples langages : la collecte et la gestion des données sont faites en Java, et les données sont stockées dans des bases de données (Mysql et MongoDB) ; le traitement des données est effectué en python pour le traitement du langage naturel et en R pour les analyses statistiques et de réseaux ; les visualisations de graphes sont effectuées avec le logiciel Gephi.

dire l'ensemble des articles citant un article initial donné. Nous récupérons de cette manière deux sous-corpus : les références citant des articles dans *Cybergeo* et les références citant celles citées par *Cybergeo*. A ce stade, le corpus complet contient autour de $4 \cdot 10^5$ références.

Pour simplifier, nous appellerons *référence* toute production scientifique standard qui peut être citée par une autre (article de journal, livre, chapitre de livre, article de conférence, communication, etc.) et contient des champs basiques (titre, résumé, auteurs, année de publication). Nous travaillons par la suite sur le réseau de références, reliées par des citations.

Données textuelles

Une description textuelle pour l'ensemble des références est nécessaire pour une analyse sémantique complète. Nous utilisons pour cela une autre source de données, qui est le catalogue en ligne du logiciel de gestion bibliographique *Mendeley* [[mendeley](#)]. Celui fournit une API gratuite permettant de récupérer divers champs sous un format structuré. Même s'il n'est pas complet, le catalogue fournit une couverture raisonnable dans notre cas, autour de 55% du réseau de citation complet. Cela correspond à un corpus final avec résumés complets de taille $2.1 \cdot 10^5$. La structure et les statistiques descriptives du réseau de citation correspondant sont rappelés en Fig. [107](#).

B.6.3 Méthodes et Résultats

Propriétés du réseau de citation

PROPRIÉTÉS Comme détaillé précédemment, nous sommes en mesure de récupérer autour de $4 \cdot 10^5$ références par reconstruction du réseau de citation à profondeur ± 1 à partir des 927 références initiales du journal, parmi lesquelles $2.1 \cdot 10^5$ ont un texte de résumé permettant une analyse sémantique. Un premier regard sur les propriétés du réseau de citation fournit des informations utiles. Le degré moyen entrant (qui peut être interprété comme un facteur d'impact stationnaire intégré) pour les références pour lesquelles il peut être défini a une valeur de $\bar{d} = 121.6$, tandis que pour les articles de *Cybergeo* nous avons $\bar{d} = 3.18$. Cette différence suggère une variété de status de références, de travaux anciens classiques (le plus cité a 1051 citations entrantes) à des travaux plus récents moins influents.

Cette diversité est confirmée par l'organisation hiérarchique examinée en Fig. [108](#) qui révèle trois régimes superposés. Plus précisément, nous nous intéressons à la courbe rang-taille, donnée par le logarithme du nombre de citations reçues comme fonction du logarithme du rang de l'article. Nous obtenons, comme attendu [[redner1998popular](#)], des comportements de loi puissance localisés. Un premier ensemble d'environ 150 références présente une hiérarchie très faible (exposant

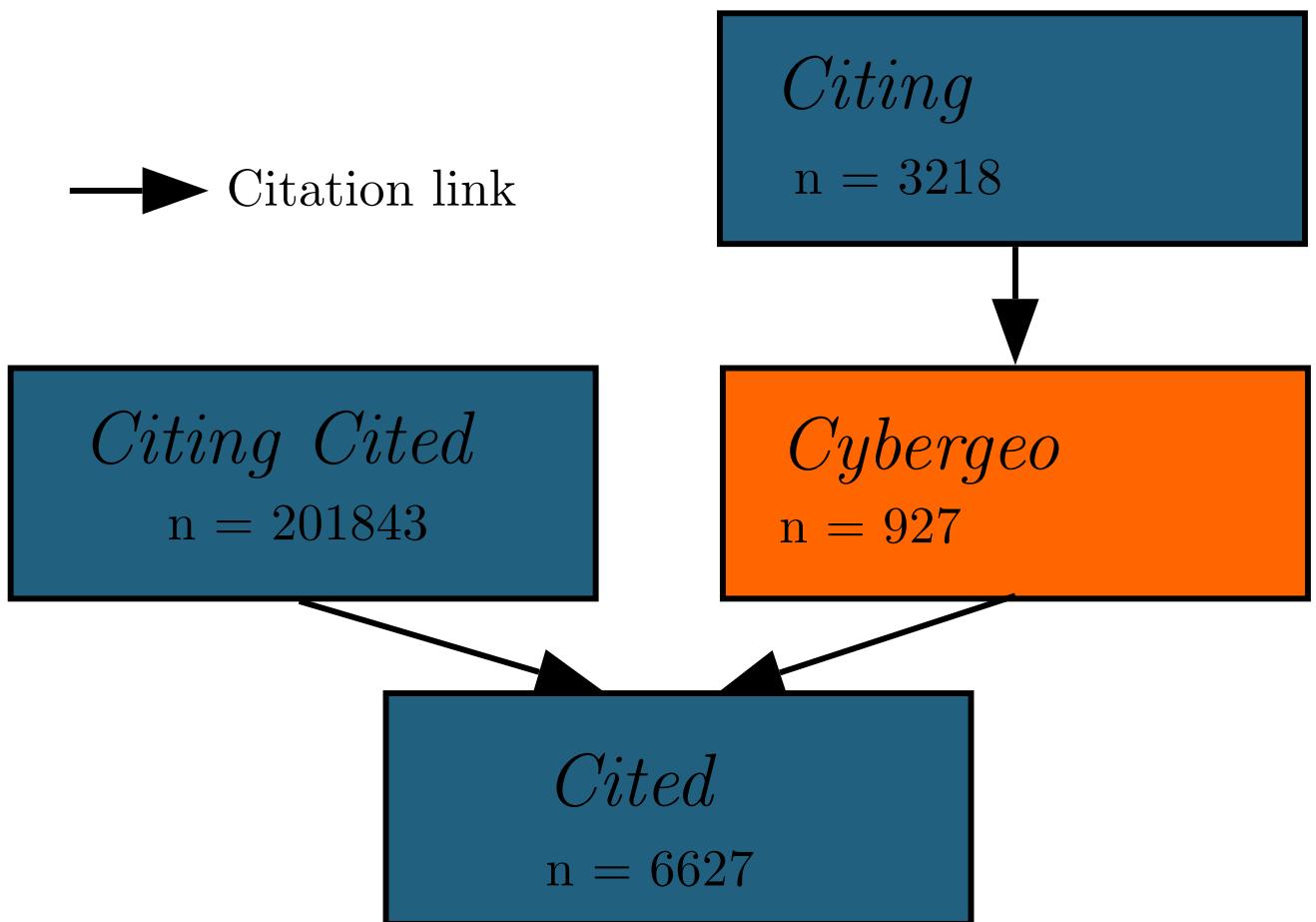


FIGURE 107: **Structure et contenu du réseau de citation.** Le corpus initial de *Cybergeo* est composé de 927 articles, eux-mêmes cités par un corpus légèrement plus grand (donnant un facteur d'impact stationnaire autour de 3.18), cite \simeq 6600 références, elles-mêmes co-citées par plus de $2 \cdot 10^5$ travaux pour lesquels nous avons une description textuelle.

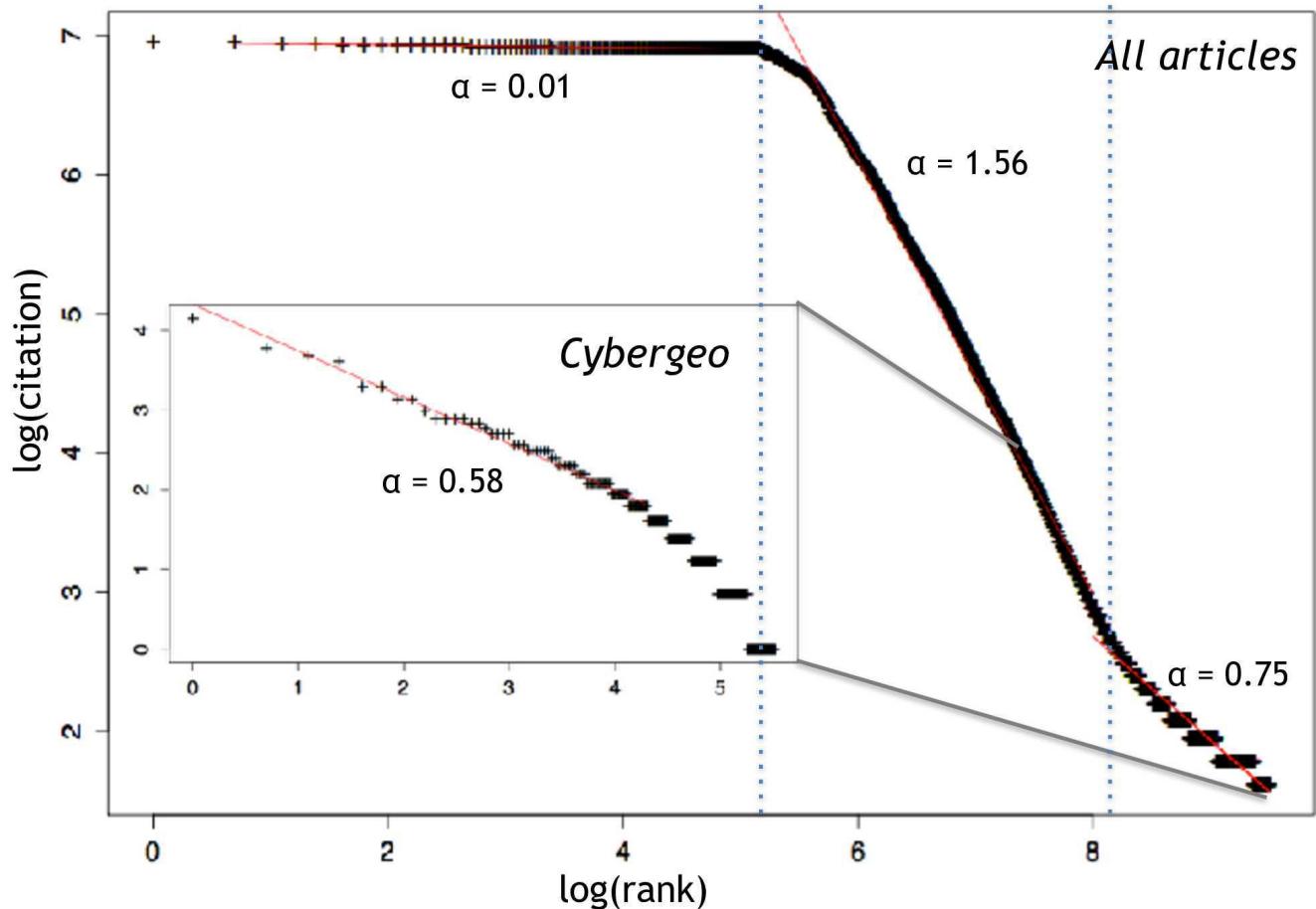


FIGURE 108: **Graphe rang-taille des citations reçues.** La courbe dévoile trois régimes de citation superposés, correspondant à des lois puissance avec différents niveaux de hiérarchie. Les références dans *Cybergeo* (graphe en insert) sont elles-mêmes dans la queue et moins hiérarchiques.

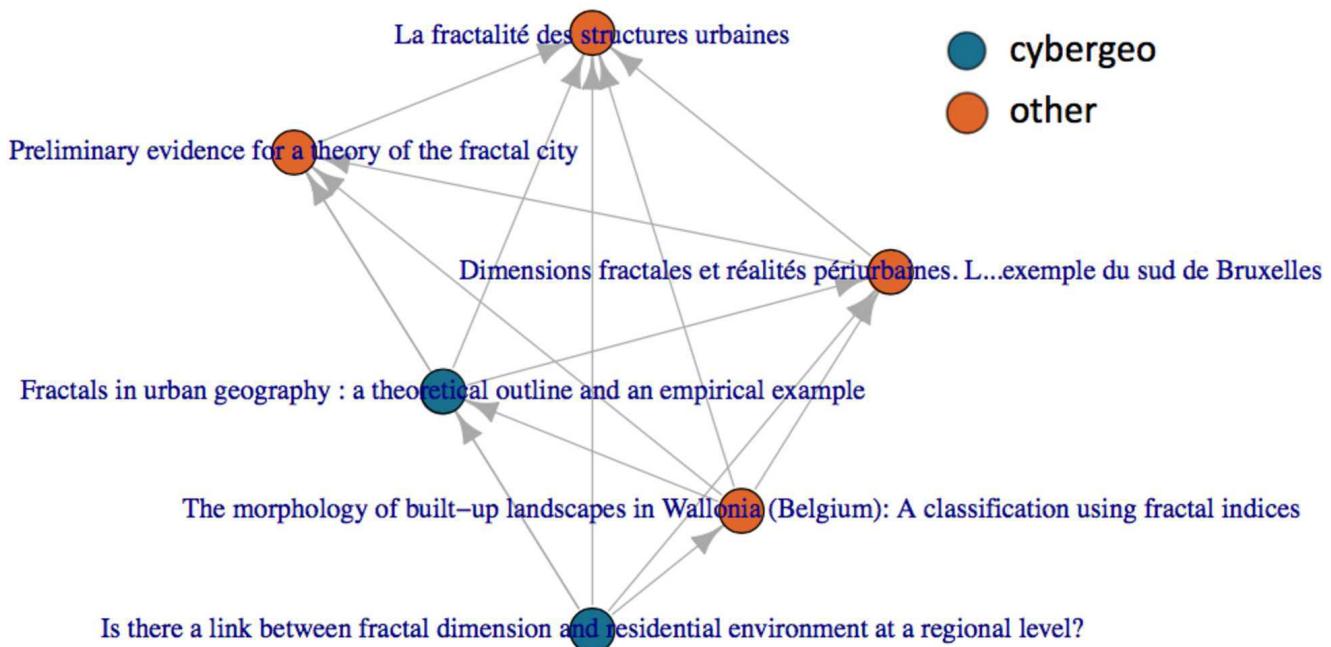


FIGURE 109: Exemple d'une clique maximale dans le réseau de citation. Les articles de *Cybergeo* sont en bleu. Une telle structure topologique révèle certaines pratiques de citation comme ici une citation systématique des travaux précédents dans la niche de recherche.

rang-taille $\alpha = 0.01$) et correspond aux références classiques dans différentes disciplines. Un second régime ($\alpha = 1.56$) est bien plus hiérarchisé, suivi par un dernier régime moins hiérarchique ($\alpha = 0.75$) contenant des articles plus récents (année moyenne de publication mi-2005, contre mi-1998 pour le second et 1983 pour le premier).

D'autres propriétés topologiques révèlent des motifs typiques de pratiques de citation : par exemple, l'existence de cliques (sous-graphes complets) de fort ordre implique des pratiques de citation dont la compatibilité avec la nature cumulative de la connaissance peut être remise en question [pumain2005cumulativite], puisque celles-ci doivent toujours trouver la source de la production de connaissance dans les travaux les plus récents. Un exemple de telle clique est montré en Fig. 109.

COMMUNAUTÉS DE CITATION Le réseau de citation est une première opportunité pour construire des disciplines endogènes, par extraction des communautés de citation. Plus précisément, cette étape vise à identifier des motifs récurrents dans les citations, qui définiraient une discipline par ses pratiques de citation. Afin de rester cohérent avec la structure particulière de données que nous avons (citations entrantes manquantes pour les sous-corpus à profondeur maximale), nous filtrons le réseau en supprimant tous les noeuds de degré inférieur à 1. Cela assure que les noeuds restants sont soit au moins

cités par un autre noeud (et donc il n'y a pas de liens manquants pour ces noeuds) ou citent au moins deux autres noeuds, ce qui peut faire des "ponts" entre sous-communautés. Le réseau obtenu a une taille de $|V| = 107164$ noeuds et $|E| = 309778$ liens. Celui-ci est visualisé en Fig. 110.

Nous utilisons un algorithme standard d'optimisation de modularité pour identifier des communautés (**blondel2008fast**) dans ce réseau de citation. Il fournit 29 communautés avec une modularité de 0.71. En comparaison, un bootstrap de 100 tirages aléatoires des liens dans le réseau donne une modularité moyenne de $-1.0 \cdot 10^{-4} \pm 4.4 \cdot 10^{-4}$ ce qui signifie une forte significativité des communautés.

Nous nommons les communautés par inspection des titres des références les plus citées dans chaque. Les 14 communautés qui ont une taille plus grande que 2.5% du réseau sont : *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*. Ces catégories ne correspondent pas directement à des disciplines bien définies, puisque certaines correspondent plus à des méthodes (ABMS), des objets d'étude (*Urban Studies*), ou des paradigmes (*Complex Networks*). Certains sont des "spécialisations" d'autres : la plupart des travaux en *Urban Studies* peuvent aussi être classifiés comme géographie critique ou sociale. De cette façon, nous construisons des disciplines endogènes qui correspondent à des *pratiques scientifiques* (ce qui est cité) plus qu'à leur représentations (les disciplines "officielles"). Le positionnement relatif des communautés en Fig. 110, obtenu par un algorithme de Force-Atlas, en dit long sur leurs relations respectives : par exemple la géographie sociale fait une pont entre les *Urban Studies* et l'économie géographique, tandis que la connection entre socio-écologie et les simulations urbaines est fait par le GIS (ce qui pouvait être attendu car la géomatique est un champ interdisciplinaire). Le GIS sépare également et connecte deux sous-champs de l'écologie, d'une part des études plus thématiques sur les habitats écologiques, et d'autre part des méthodes statistiques. Ces relations informeront déjà qualitativement des motifs d'interdisciplinarité, au sens de mesures d'intégration. Nous allons par la suite utiliser ces communautés pour situer la classification sémantique.

Construction des communautés sémantiques

Nous présentons à présent les détails méthodologiques de la construction de la classification sémantique. Cette étape adopte la méthodologie décrite par [bergeaud2017classifying], qui construit une classification sémantique sur données de brevets.

EXTRACTION DES MOTS-CLÉS PERTINENTS Nous rappelons que notre corpus avec des textes disponibles consiste en environ $2 \cdot 10^5$

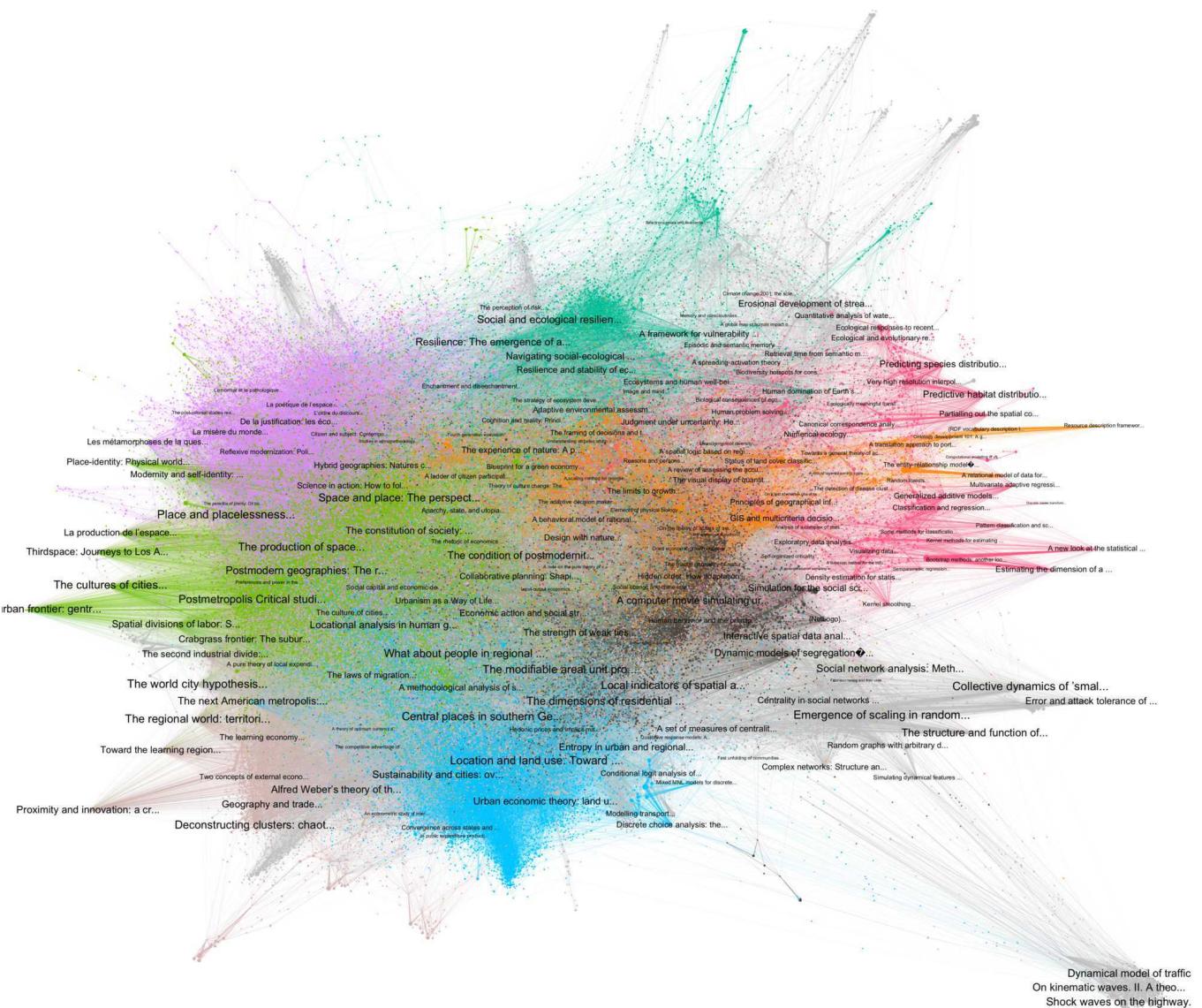


FIGURE 110: Réseau de citations. Nous visualisons uniquement le “coeur” du réseau de citation, composé des références avec un degré plus grand que 1 ($|V| = 107164$ et $|E| = 309778$). L’algorithme de détection de communautés fournit 29 communautés avec une modularité de 0.71. Les couleurs des noeuds et des liens donnent les communautés principales (par exemple l’écologie en magenta, GIS en orange, la Socio-écologie en turquoise, la géographie sociale en vert, l’analyse spatiale en bleu). Les labels des noeuds donnent les titres raccourcis des références les plus citées, la taille étant à l’échelle de leur degré entrant. Le graphe est spatialisé par l’utilisation d’un algorithme Force-Atlas.

résumés des publications à une distance topologique plus petite que 2 du journal *Cybergeo* dans le réseau de citation. La première étape importante est l'extraction de mots-clés pertinents à partir des résumés. L'analyse textuelle est effectuée avec la bibliothèque python `nltk` [**bird2006nltk**]. Nous ajoutons un traitement particulier à la méthode de [**bergeaud2017classifying**], puisque notre corpus est multilingue : la détection de langue est faite par technique des *stop-words* (**baldwin2010language**). Nous utilisons également un tagger spécifique (la fonction permettant l'attribution de fonctions grammaticales aux mots), TreeTagger (**schmid1994probabilistic**) pour les langues autres que l'anglais.

En résumé, le flux d'extraction des mots-clés suit les étapes suivantes :

1. La détection de la langue est faite par utilisation des *stop-words*
2. Le *pos-tagging* (détection des fonctions des mots) et le *stemming* (extraction du *stem*) sont effectués différemment selon la langue :
 - Anglais : *pos-tagger* intégré à `nltk`, combiné à un *PorterStemmer*
 - Français ou autre : utilisation de TreeTagger (**schmid1994probabilistic**)
3. Sélection de *n-grams* potentiels (mots-clés de longueur *n* avec $1 \leq n \leq 4$) suivant les règles grammaticales données : pour l'anglais $\cap\{\text{NN} \cup \text{VBG} \cup \text{JJ}\}$, et pour le français $\cap\{\text{NOM} \cup \text{ADJ}\}$. Les autres langues sont une proportion négligeable du corpus et ne sont pas pris en compte.
4. Estimation de la pertinence des *n-grams*, par attribution d'un score suivant la déviation des distribution statistiques des co-occurrences à une distribution aléatoire.

RÉSEAU SÉMANTIQUE Nous conservons à cette étape un nombre fixé K_W de *n-grams*, en se basant sur leur score de pertinence, qui seront désignés comme mots-clés pertinents. Nous observons que pour des grandes valeurs de K_W , les résultats ne sont pas sensibles au nombre total de mots-clés, et prenons ainsi une grande valeur raisonnable pour la performance computationnelle, $K_W = 50,000$. Nous construisons la matrice de co-occurrence des mots-clés pertinents. Cette matrice de co-occurrence fournit le réseau sémantique comme sa matrice d'adjacence : les noeuds sont les mots-clés, et ils sont reliés en fonction de leurs co-occurrences.

ANALYSE DE SENSIBILITÉ Nous observons un phénomène similaire à celui observé par [**bergeaud2017classifying**], qui est la présence de noeuds avec un grand degré mais non spécifiques à un champ particulier : par exemple `model` et `space` sont utilisés dans la majorité des sous-champs de la Géographie. Nous adaptons également la procédure de filtration originale, puisque nous ne disposons

pas ici d'information exogène pour calibrer les paramètres. Nous supposons que les termes de plus haut degré ne portent pas d'information sur des classes en particulier et peuvent ainsi être filtrés étant donné un degré maximal k_{\max} . Nous gardons le second filtre sur un seuil de poids minimal des liens θ_w . Nous ajoutons la contrainte supplémentaires que les mots-clés sont aussi filtrés par leur fréquence d'apparition dans les documents sur une fenêtre $[f_{\min}, f_{\max}]$ (nombre de références dans lesquelles ils apparaissent), ce qui est légèrement différent de la filtration du réseau.

Une analyse de sensibilité de la topologie du réseau résultant à ces quatre paramètres est présentée en Fig. 111. Etant donné un réseau filtré, nous détectons les communautés en utilisant une optimisation de la modularité comme ci-dessus pour le réseau de citation. Diverses propriétés du réseau peuvent être optimisées, et nous nous intéressons en particulier à sa taille (nombre de mots-clés après filtrage), la modularité optimale, le nombre de communautés, et l'équilibre entre leurs tailles (défini comme un indice de concentration $\sum_k s_k^2 / (\sum_k s_k)^2$). Ce problème d'optimisation multi-objectif ne possède pas de solution unique car les objectifs sont contradictoires de manière complexes, et un point de compromis doit être choisi. Nous prenons un point compromis entre modularité et taille du réseau, avec un fort équilibre et un nombre raisonnable de communautés, donné par $k_{\max} = 1200, \theta_w = 100, f_{\min} = 50, f_{\max} = 10000$. Ces valeurs donnent un réseau de taille 2868, avec 18 communautés et une modularité de 0.57.

Notons que la petite proportion de mots-clés en français est toujours séparée du reste du réseau puisqu'ils ne peuvent pas coïncider avec des mots-clés anglais, et qu'avec ces valeurs des paramètres aucun mot-clé français n'est conservé. L'ensemble des communautés décrites par la suite ne contient pour cette raison que des mots-clés en anglais.

COMMUNAUTÉS SÉMANTIQUES Nous obtenons ainsi des communautés dans le réseau sémantique avec les paramètres de filtration optimisés. A l'exception d'une petite proportion s'apparentant apparemment à du bruit (représentant moins de 10 mots-clés dans 3 communautés que nous supprimons, i.e. 0.33% des mots-clés), les communautés correspondent à des champs scientifiques, domaines, ou approches bien définis. La dénomination est également faite par inspection des mots-clés les plus pertinents dans chaque communauté, afin de se tenir ici à un certain niveau de supervision.

La Table 112 résume les communautés, donnant leurs noms, tailles, et mots-clés correspondants. La communauté la plus importante est en rapport avec des questions de Sciences politiques et de Géographie critique, ce qui pouvait être attendu puisque plusieurs communautés de citation obtenues précédemment (*Social Geography, Urban Studies*)

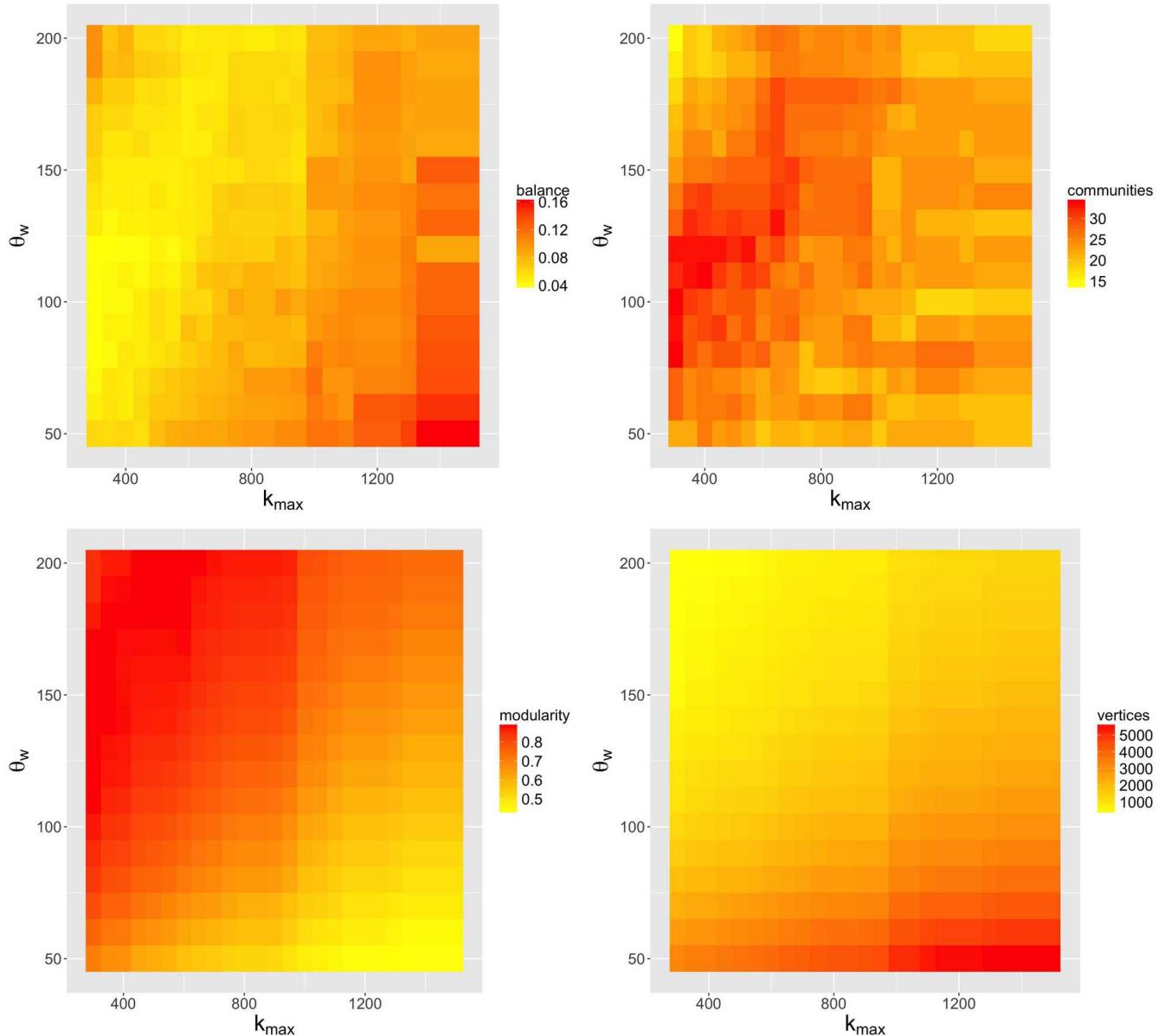


FIGURE 111: Analyse de sensibilité des indicateurs de réseau aux paramètres de filtrage. Nous donnons ici 4 indicateurs (équilibre entre les tailles des communautés, modularité de la décomposition, nombre de communautés, nombre de noeuds), comme fonction des paramètres k_{\max} et θ_w , à des valeurs fixées $f_{\min} = 50$, $f_{\max} = 10000$. Des valeurs proches pour ces deux derniers paramètres (dans des bornes raisonnables) donne un comportement similaire.

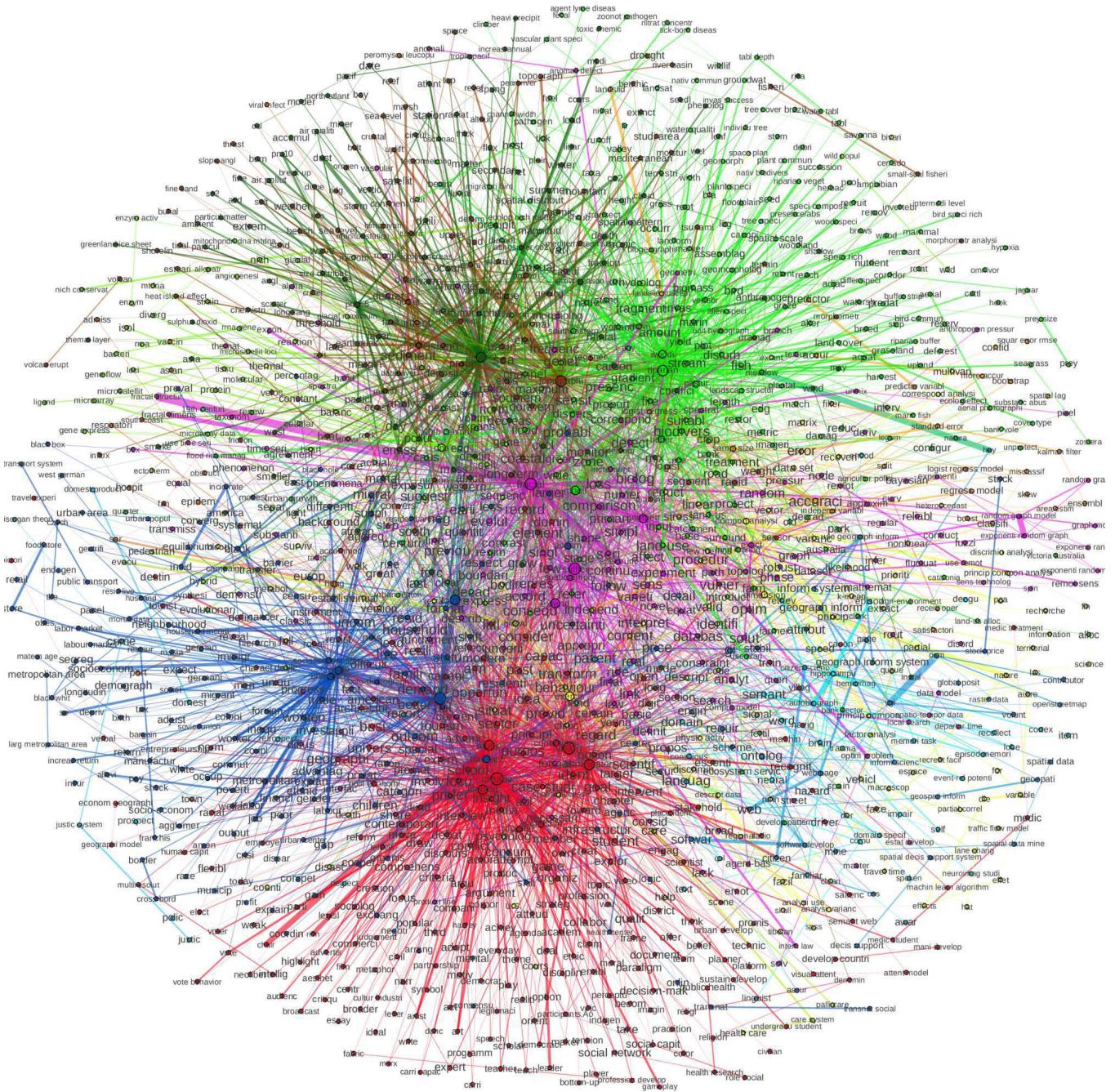


FIGURE 113: Visualisation du réseau sémantique. Le réseau est construit par co-occurrences des mots-clés les plus pertinents. Les paramètres de filtrage sont pris ici selon l'optimisation multi-objectifs faite en Fig. 111, i.e. ($k_{\max} = 1200$, $\theta_w = 100$, $f_{\min} = 50$, $f_{\max} = 10000$). L'algorithme de spatialisation du graphe (Fruchterman-Reingold), malgré son caractère stochastique et dépendant au chemin, révèle de l'information sur le positionnement relatif des communautés.

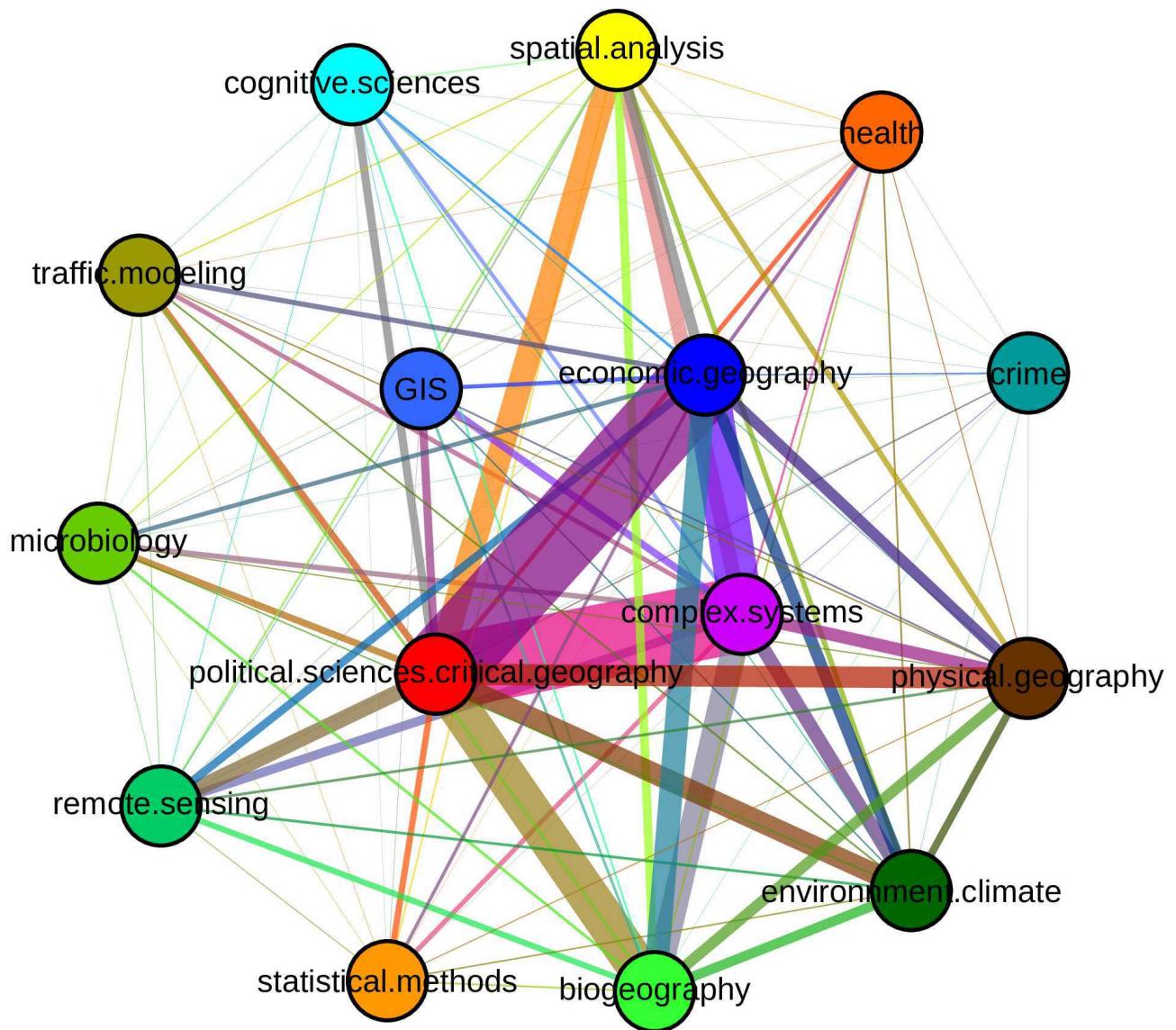


FIGURE 114: Synthèse des communautés sémantiques et de leurs liens. Les poids des liens sont calculés comme les probabilités de co-occurrence des mots-clés correspondants au sein des références.

FIGURE 112: Composition des communautés sémantiques. Elles sont construites par détection de communautés dans le réseau sémantique.

Name	Size	Keywords
Political sciences/critical geography	535	decision-mak, polit ideolog, democraci, stakehold, neoliber
Biogeography	394	plant densiti, wood, wetland, riparian veget
Economic geography	343	popul growth, transact cost, socio-econom, household incom
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, fac
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 geograph inf
Traffic modeling	63	simul model, lane chang, traffic flow, crowd behavior
Health	52	epidem, vaccin strategi, acut respiratori syndrom, hospit
Remote sensing	48	land-cov, landsat imag, lulc
Crime	17	crimin justic system, social disorgan, crime

s'occupent de ces questions. Nous obtenons ensuite un groupement conséquent en termes de biogéographie, qui doit correspondre aux publications en écologie et socio-écologie identifiées précédemment, avec une communauté en environnement et climat.

De manière similaire au communautés de citation, mais plus prononcée ici, nous obtenons des “disciplines” endogènes qui peuvent correspondre à des vraies disciplines, à des méthodologies, à des objets d’étude. Cette classification révèle pour cela également des *pratiques scientifiques effectives*, ici en termes de contenu sémantique. Une classe ici en rapport avec les Systèmes Complexes peut être associées à un paradigme et à différentes approches qui étaient séparées dans les communautés de citation : les modèles basés-agent et les réseaux complexes par exemple. Au contraire, certaines études qui étaient rassemblées dans de larges domaines précédemment peuvent être différencier précisément dans le réseau sémantique, comme la microbiologie et la santé ici qui sont utilisées par des études en relation à l’écologie et la socio-écologie dans le réseau de citation. Certains domaines très spécifiques apparaissent ici puisqu’ils ont très peu de connections avec les autres dans leur contenu sémantique : par exemple, la géographie du crime est très précise et déconnectée des autres communautés.

Nous montrons en Fig. 113 une visualisation du réseau sémantique, dans lequel le positionnement des communautés, induit par un algorithme de Fruchterman-Reingold (que nous utilisons ici pour avoir un positionnement plus précis dans le positionnement relatif en comparaison du Force-Atlas ([jacomy2014forceatlas2](#))). Les ponts entre disciplines distantes est effectué différemment en comparaison du réseau de citation, et révèle ainsi qualitativement une autre dimension de l'interdisciplinarité, i.e. la sémantique partagée par les disciplines. Ici, les communautés correspondant à l'Economie Géographique (bleu) et à la Géographie critique (rouge) sont proches dans le réseau de citation, mais sont liés à l'écologie et la géomorphologie (vert et marron) par les Systèmes Complexes (magenta), bien que ceux-ci n'étaient pas présents comme communauté dans le réseau de citation. Les méthodologies de la complexité comme les fractales, les loi d'échelles [[west2017scale](#)], ou les réseaux [[newman2003structure](#)] sont en effet largement utilisés à la fois en sciences sociales et en physique ou biologie. L'analyse sémantique montre ainsi que des disciplines très distantes, qui sont distantes dans leur motifs de citation, sont finalement proches en termes de contenu observé.

En termes de chevauchements entre les communautés, au sens des co-occurrences des mots-clés correspondants dans les textes des références, nous montrons une synthèse des liens entre communautés sémantiques en Fig. 114. Nous voyons que les communautés comme la Géographie critique et la biogéographie ne sont pas totalement déconnectées et partagent finalement un certain nombre de co-occurrences. Des communautés plus isolées peuvent être identifiées comme les géographies de la Santé et du Crime. De manière surprenante, les méthodes statistiques ne partagent pas de liens forts avec d'autres communautés, ce qui pourrait signifier que des articles traitant de questions méthodologiques dans ce champ sont plutôt déconnectées du champ d'application, ou au moins ne le décrivent pas en détail. Au contraire, les méthodes en Systèmes Complexes sont organiquement intégrées avec les questions thématiques qu'ils traitent.

Composition sémantique des communautés de citation

Nous pouvons à présent nous tourner vers l'étude des relations entre classifications. Tout d'abord, une façon simple de les relier est d'étudier le contenu sémantique des communautés de citations. Chaque référence a une proportion donnée de mots-clés dans chaque classe sémantique, et une composition moyenne en termes de classes sémantiques pour chaque classe de citation peut ainsi être calculée. Nous montrons ces compositions en Fig. 115. Des résultats attendus sont obtenus, comme *Complex Networks* (citation) ayant la proportion la plus forte en *Complex Systems* (sémantique), ou le GIS (citation) le plus fort en GIS (sémantique), et de même pour l'économie géographique.

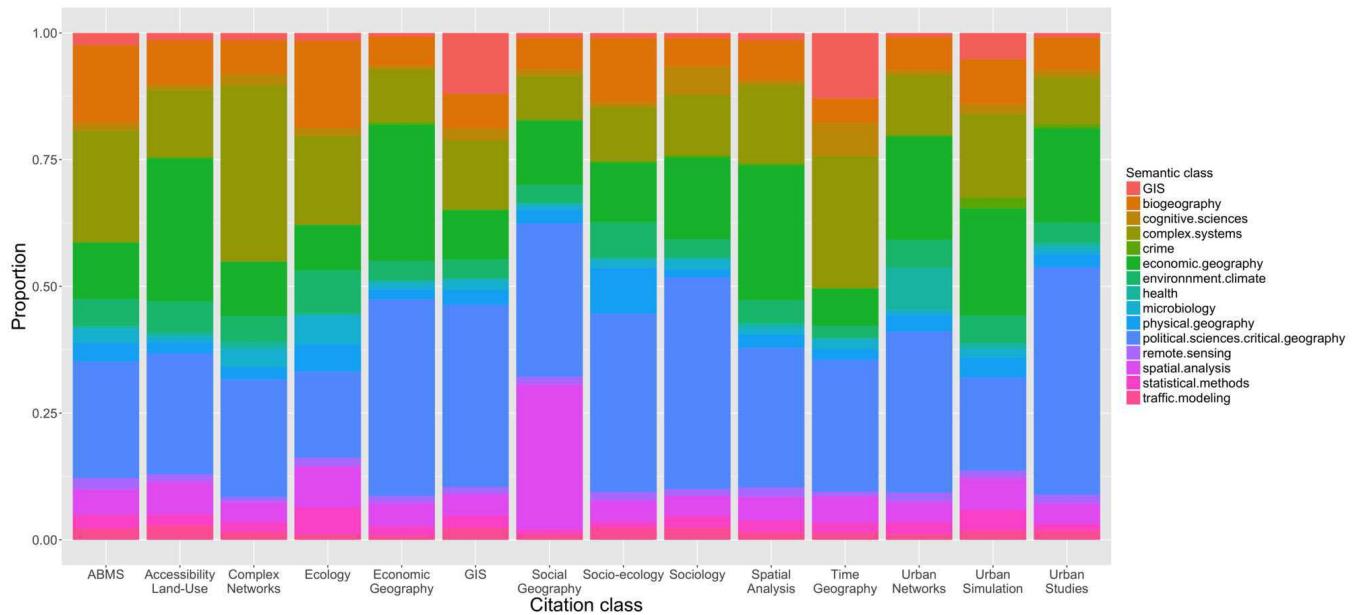


FIGURE 115: Composition des communautés de citation en termes de contenu sémantique. Pour chaque classe de citation (horizontalement), la barre est décomposée entre les proportions de chaque classe sémantique (données par la couleur).

Mais l'étude de motifs qui auraient pu ne pas être attendus est très informatif, et dévoile des pratiques d'interdisciplinarité. Par exemple, la *Time Geography* (citation) utilise quasiment autant de GIS (sémantique) que le GIS (citation), ce qui signifie qu'ils doivent utiliser les méthodes et outils correspondants pour étudier la question thématique des trajectoires spatio-temporelles des agents géographiques. Le plus important en termes de sciences politiques (sémantique) sont les *Urban Studies*, ce qui suggère une convergence de la Ville comme objet d'étude et des disciplines de Sciences politiques et de Géographie critique. Egalement de manière intéressante, les communautés de citation utilisant le plus la biogéographie sont l'écologie (ce qui pouvait être attendu) et les ABMS, confirmant ainsi le rôle de l'application thématique dans les méthodologies des Systèmes Complexes.

Mesure de l'interdisciplinarité

Nous avons eu jusqu'à présent une vue qualitative sur les motifs d'interdisciplinarité, en s'intéressant à la localisation relative des communautés au sein des classifications de citation et sémantique, et la relation entre les classifications. Nous proposons à présent de regarder des mesures quantitatives de l'interdisciplinarité, pour chaque classification.

Plus précisément, pour une classification donnée $C \in \{\text{Citation}, \text{Semantic}\}$ une référence i peut être représentée par un vecteur de probabilités $(p_{ij}^{(C)})_j$ sur les classes j qui donne pour chaque classe la probabilité

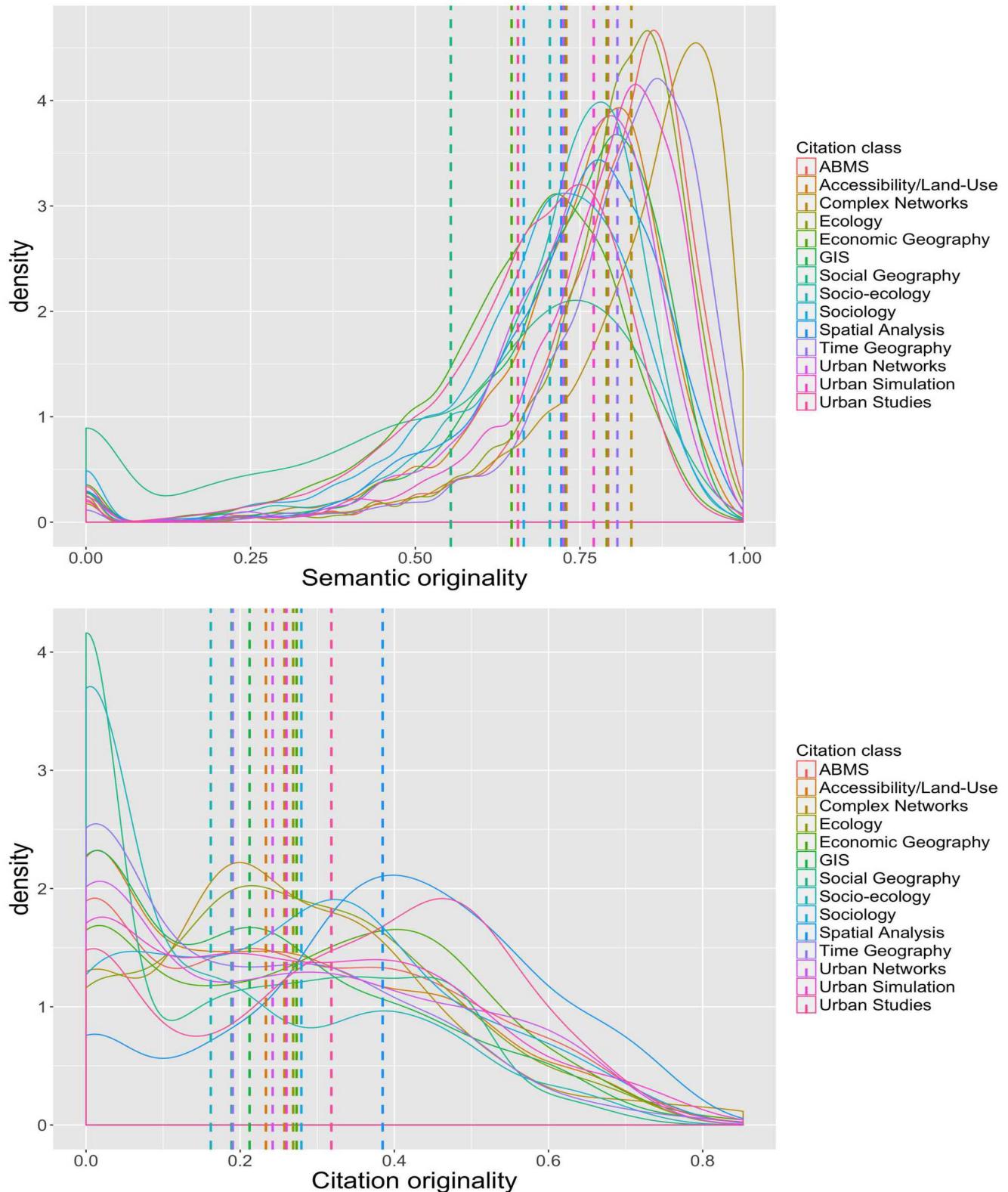


FIGURE 116: **Distribution statistique des originalités.** Nous donnons les densités de probabilité listées des indices d'originalité, par classe de citation (donnée par la couleur), pour l'originalité sémantique $o^{(\text{Semantic})}$ (*Haut*) et pour l'originalité de citation $o^{(\text{Citation})}$ (*Bas*). Les lignes pointillées donnent la moyenne pour chaque distribution, avec la couleur correspondante.

d'y appartenir. Etant donné cette configuration, nous mesurons l'interdisciplinarité d'une référence en utilisant une concentration d'Herfindhal [**porter2009science**], qui peut aussi être appelée indice d'originalité. Nous définissons l'originalité par

$$o_i^{(C)} = 1 - \sum_j p_{ij}^{(C)}{}^2$$

Pour la classification sémantique, les probabilités sont définies comme la proportion de mots-clés du résumé dans chaque classe sémantique. Avec la classification de citation déterministe, chaque référence a une classe unique et l'index d'originalité est toujours nul. Pour cette raison, afin de pouvoir comparer les deux classifications, nous associons une probabilité à chaque classe de citation pour chaque article comme la proportion de citations reçues de cette classe. L'indice induit est original, et mesure l'interdisciplinarité comme la façon dont une référence *est utilisée* par différentes disciplines pendant sa vie.

Nous montrons en Fig. 116 la distribution statistique des deux indices $o^{(\text{Semantic})}$ et $o^{(\text{Citation})}$, stratifiées par classe de citation. Cela permet une comparaison directe entre les deux et également une comparaison indirecte par la variation des distributions sémantiques selon les classes de citations. Pour la distribution des originalités sémantiques, l'ensemble des classes de citation présentent un motif similaire, qui correspond à un pic autour d'une grande valeur et un pic plus faible à zéro. Cela signifie que soit les références sont fortement spécialisées et n'ont des mots-clés que dans une seule classe, ou bien elles utilisent des mots-clés de différentes classes de façon relativement équilibrée (pour comparaison, un résumé avec moitié de mots dans une classe et moitié dans une autre donne une originalité de 0.5). La classe de citation la plus originale, i.e. la plus mélangée, est *Complex Networks*, avec une distribution clairement détachée des autres, ce qui confirmerait leur utilisation comme une méthode pour un certain nombre de problèmes différents. La Géographie sociale est de très loin la moins originale, avec un grand nombre de références à classe unique, et une moyenne bien plus basse que celle des autres classes, ce qui signifierait une présence accrue d'isolation dans les disciplines associées.

En termes d'indice d'originalité de citation, le comportement général est fondamentalement différent, les indices d'originalité moyens étant tous inférieurs à 0.4 et la plupart des distributions ont leur mode en 0, ce qui signifie que la majorité des références sont citées uniquement par leur propre classe de citation. A nouveau, la Géographie sociale est la moins originale, confirmant un comportement similaire en termes de pratiques de citation qu'en termes de contenu de la recherche. Les classes les plus originales en moyenne, avec un pic dans les grandes valeurs, sont *Spatial Analysis* et *Urban Simulation* : cela correspond au fait que ces classes contiennent des méthodes relativement génériques qui peuvent être appliquées dans différents

champs et sont citées de manière appropriée. Les *Complex Networks* n’atteignent pas le même niveau, mais présentent cependant un pic autour de 0.2 et pas de pic en 0, de la même manière que l’écologie, suggérant des disciplines ayant toujours un impact significatif sur les autres disciplines.

En résumé, nous montrons (i) différents motifs d’interdisciplinarité, selon les disciplines, en termes de contenu scientifique (sémantique) et d’impact scientifique (citation); et (ii) une forte différence qualitative de comportement des originalités entre les deux classifications, ce qui suggère leur complémentarité.

Corrélation entre classifications

Afin de renforcer l’idée d’une complémentarité des classifications, qui captureraient chacune différentes dimensions des processus de production de connaissance, nous examinons finalement la matrice des corrélations entre les classifications. Nous utilisons cette fois les probabilités de classes effectives pour la classification de citation, i.e. un vecteur de zeros à l’exception d’un un à l’index de la classe de la référence. Nous calculons un coefficient de corrélation de Pearson entre les classes k (sémantique) et la classe k' (citation) comme

$$\rho_{k,k'} = \frac{\text{Cov} \left[(p_{ik}^{(\text{Sem})})_i, (p_{ik'}^{(\text{Cit})})_i \right]}{\sqrt{\sigma \left[(p_{ik}^{(\text{Sem})})_i \right] \sigma \left[(p_{ik'}^{(\text{Sem})})_i \right]}}$$

où la covariance est estimée avec l’estimateur non biaisé.

La structure de la matrice de corrélation rejoint les conclusions obtenues lors de l’étude de la composition sémantique des communautés de citation, comme le GIS étant fortement corrélé au GIS ($\rho = 0.26$), ou la Sociologie à la Science politique ($\rho = 0.16$). Plus importantes pour notre questions sont les statistiques de synthèse de la matrice entière. Elle a un minimum de -0.16 (*Ecology* (citation) contre les sciences politiques (sémantique)), une moyenne de -0.002 et un maximum de 0.33 (*Social geography* (citation) et *Spatial Analysis* (sémantique)). Les valeurs “fortes” sont fortement isolées, comme le premier décile est à -0.06 et le dernier à 0.09 , ce qui signifie que 80% des coefficients tombent dans cet intervalle, correspondant à des corrélations faibles. En résumé, les classifications sont cohérentes puisque les corrélations les plus fortes sont observées où on peut les attendre, mais la majorité des classes ne sont pas corrélées, ce qui signifie que les classifications sont relativement orthogonales et ainsi complémentaires.

b.6.4 *Discussion*

Nous avons ainsi montré la complémentarité des classifications dans les motifs qualitatifs qu’elles dévoilent, mais aussi quantitativement

en termes de mesures d'interdisciplinarité et quantitativement en termes de corrélations. Notre travail peut être étendu selon différents aspects, desquels nous donnons certaines suggestions par la suite.

Développements

Un premier développement consiste en la comparaison de journaux. Le point de départ pour la construction de l'environnement scientifique, le journal *Cybergeo*, était le point d'entrée mais pas le sujet principal de notre étude. Un développement se concentrant plus sur les journaux, essayant par exemple de répondre à des questions comparatives, ou de classifier les journaux selon leur niveau effectif d'interdisciplinarité selon différentes dimensions, serait potentiellement intéressant. La collection de données précises sur l'origine des références est cependant un premier point qui doit d'abord être résolu.

La performance de la classification sémantique n'a également pas été quantifiée ici. Une validation approfondie de la pertinence en utilisant l'information complémentaire contenue dans les deux classifications pourrait être menée par l'analyse des modularités dans le réseau de citation, comme fait dans [bergeaud2017classifying]. Cela nécessiterait cependant une classification de référence pour comparaison, qui n'est pas disponible avec le type de données que nous utilisons. Des archives ouvertes comme arXiv (principalement pour la physique) ou Repec (pour l'économie) fournissent des API pour accéder aux métadonnées incluant les résumés, et pourraient être des points de départ pour de telles études ciblées.

Applications

Une première application potentielle de notre méthodologie se base sur le fait que les deux classifications dévoilent à la fois des domaines thématiques (objets d'étude), des disciplines classiques, des communautés méthodologiques. Ces différents types de communautés peuvent en effet être comprises comme différents *Domaines de Connaissance*. [raimbault2017applied] postule des Domaines de Connaissance en co-évolution dans tout processus de production de connaissance scientifique, qui sont les domaines Théorique, Empirique, des Modèles, Méthodologique, des Outils et des Données. La majorité sont nécessaire pour tout processus, et les investigations dans l'un conditionne les avances dans les autres. Un raffinement des classifications, associé à une classification supervisée pour associer des domaines de connaissance à des communautés (en utilisant potentiellement les textes complets pour avoir une information plus précise sur la proportion de chaque domaine de connaissance impliqué dans chaque), permettrait de quantifier les relations entre domaines. De plus, l'utilisation de données temporelles avec les dates des publications, fournirait une quantification effective de la *co-évolution* des do-

maines au sens des motifs de corrélations temporelles (e.g. causalité de Granger).

Une autre direction intéressante est l'application de nos classifications à la quantification de la diffusion spatiale de la connaissance, comme [maisonobe2013diffusion] fait pour la diffusion d'un question spécifique en génétique. Il n'est pas évident si des dimensions différentes de la connaissance se diffusent de la même façon : par exemple les pratiques de citation peuvent être corrélées aux réseaux sociaux et peuvent ainsi montrer différents motifs que les contenus effectifs de la recherche. Ainsi, notre travail permettrait d'étudier de telles questions de points de vue complémentaires.

Enfin, nous affirmons que les outils que nous avons développé peuvent contribuer à une autonomisation accrue des auteurs et au développement de pratiques de science ouverte. Au sein des différentes visions de la Science Ouverte [fecher2014open], l'ouverture des données est toujours un aspect important, en même temps d'un développement de la réflexivité dans toutes les disciplines, au delà des seules sciences sociales auxquelles elle est classiquement associée. Le premier point est traité par nos outils ouverts pour la construction de jeux de données, tandis que le second est impliqué par la connaissance nouvelle des différentes dimensions de l'environnement scientifique que nous avons étudié.

b.6.5 Conclusion

Nous avons introduit une approche multi-dimensionnelle pour la compréhension de l'interdisciplinarité, basée sur les analyses du réseau de citation et du réseau sémantique. A partir d'un journal généraliste en Géographie, nous construisons un vaste corpus du voisinage de citation, duquel nous extrayons les mots-clés pertinents pour élaborer une classification sémantique. Nous montrons ensuite qualitativement et quantitativement la complémentarité des classifications. La méthodologie et les outils associés sont ouverts et peuvent être réutilisés dans des études similaires pour lesquelles les données sont difficiles à obtenir ou faiblement référencées dans les bases classiques.

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 se rapportent à des questions épistémologiques.

1. 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 appliquée.
2. Celui-ci est également illustré 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.
3. 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.

Les deux développements thématiques suivants traitent des questions d'outils et de méthodes à partir de contextes thématiques précis.

4. 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.
5. 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.

Enfin, la dernière section est importante quant à des question de modélisation.

6. Un modèle multi-échelles de dynamiques de migrations résidentielles à l'échelle métropolitaine est présenté avec les premiers résultats issus de son exploration.

* * *

*

Les publications ou communications correspondant au contenu de ces annexes sont détaillées pour chacune, avec le détail des contributions des différents collaborateurs.

C.1 PONTS ENTRE GÉOGRAPHIE ET ECONOMIE

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 10 ci-dessous présente l'appel à communication.

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.

ENCADRÉ 10: ECTQG 2017 Special Session : bridges between economics and geography

Contributions

Les contributions à la session

Synthèse des débats

* * *

*

C.2 CYBERGEONETWORKS : UNE ANALYSE BIBLIOMÉTRIQUE MULTIDIMENSIONNELLE SPATIALISÉE

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 Big Data & Society) est ici traduit et adapté.

* * *

*

La biométrie est devenue monnaie courante et largement utilisée par les auteurs et les journaux pour suivre, évaluer et identifier le lectorat dans un contexte de publications toujours accrues. Cette contribution vise à se détacher des comptages en temps réel pour s'intéresser aux proximités sémantiques et l'évolution des articles publiés dans le journal en ligne *Cybergeo* depuis sa création en 1996. Nous comparons trois stratégies pour construire des réseaux sémantiques, en utilisant les mots-clés (thématiques auto-déclarées), les citations (aires de recherche utilisant les articles publiés dans *Cybergeo*) et les textes complets (thèmes dérivés des mots utilisés dans l'écriture). Nous interprétons ces réseaux et les proximités sémantiques selon leur évolution temporelle ainsi que leur inscription spatiale, en considérant les pays étudiés dans les articles considérés. Enfin, nous comparons les trois méthodes et concluons que leur complémentarité peut contribuer à dépasser les simples statistiques pour mieux comprendre l'évolution épistémologique d'une communauté scientifique et de l'audience visée du journal.

C.2.1 Introduction

Depuis le travail fondateur de KUHN au début des années 1960, le développement des études de la science s'est basée sur trois pilier disciplinaires : l'histoire des sciences, la philosophie de la science et la sociologie de la science. Dans les années 1980, les sciences politiques ont pris une importance grandissante en étudiant les liens entre la production de la connaissance et l'utilisation de la connaissance. Ce "tournant politique" a commencé avec la création du journal *Knowledge* en 1979. Depuis la fin des années 1990, les études de la science ont pris un "tournant spatial" et ont vu l'émergence d'une géographie de la science [livingston_spaces_1995 ; livingston_science_2003 ; livingston_geography_2005 ; withers_place_2009]. Ce travail se positionne dans cette tendance : cet article propose une approche de bibliométrie spatialisée.

Faisant face à un nombre croissant d'articles, de journaux et de canaux de publication utilisés par les chercheurs dans un monde digital et d'accès ouvert, les journaux ont besoin d'identifier leur lectorat et les auteurs ont besoin de cette information pour mieux atteindre leur audience cible, en adaptant les mots-clés, le vocabulaire et les citations adaptés. Ce travail fournit un ensemble d'outils digitaux complémentaires qui satisfont trois pré-requis : 1) d'aller au delà les métriques de citation classiques et de proposer des analyses sémantiques et de réseaux extraits directement des contenus scientifiques des articles ; 2) de situer la position des ensemble d'articles selon les champs sémantiques de leurs thèmes ; 3) d'identifier les variations significatives dans les thématiques de recherche qui peuvent être reliées à l'origine géographique des auteurs ou aux pays étudiés. Ce dernier point est particulièrement intéressant dans notre cas comme notre cas d'étude est un journal en géographie.

L'anniversaire des 20 ans du premier journal exclusivement digital en sciences sociales, Cybergeo, a été l'occasion d'analyser un corpus conséquent de plus de 700 articles publiés dans 7 langages, selon la géographie des auteurs et du lectorat. Nous effectuons une analyse en épistémologie quantitative des articles scientifiques publiés depuis 1996 pour mesurer leur similarité selon trois types d'indicateurs textuels : leur mots-clés (la façon dont les auteurs situent leur recherche), leur réseau de citation (la façon dont l'article est utilisé par d'autres champs et disciplines), ou leur textes complets (le vocabulaire utilisé pour écrire l'article et présenter la recherche).

Ces analyses sont complémentaires et montrent l'évolution d'un journal vers des thèmes de recherche émergents. Elles montrent aussi le besoin pour Cybergeo d'étendre sa base d'auteurs au delà de la communauté francophone, afin de remplir son ambition de journal Européen en géographie. Cette contribution consiste en ces conclusions épistémologiques spécifiques, mais aussi en une entrée métho-

dologique et technique plus large pour gérer de manière interactive des corpus scientifiques hétérogènes à grande échelle. Nous montrons dans quelle mesure le couplage de vues complémentaires peut créer une connaissance au second ordre : la contextualisation spatiale des trois méthodes de classification révèle des motifs inattendus. De plus, les outils dédiés que nous avons construit sont disponibles comme un logiciel open source, qui peut être utilisé par les journaux pour une réflexivité scientifique accrue, mais aussi par les institutions et les scientifiques eux-mêmes pour une autonomisation bottom-up de la Science Ouverte.

Le reste de cet article est organisé de la façon suivante : nous revoyons d'abord les approches similaires s'intéressant à la bibliométrie de manière hétérogène ou multidimensionnelle, et décrivons l'étude de cas sur laquelle nous travaillerons. Nous développons ensuite les détails techniques des différentes méthodes utilisées, et la manière dont celles-ci sont couplées par exploration interactive de données spatiales ; puis nous décrivons les résultats au premier ordre (chaque méthode) et au second ordre (obtenus par couplage) ; et nous discutons finalement des implications plus larges pour l'épistémologie et la réflexivité en Science Ouverte.

Contexte bibliographique

Les études en bibliométrie ayant pour objet principal la complémentarité de différentes approches sont plutôt rares. [[2016arXiv160106075O](#)] montre que la prise en compte des données de citation et de disciplines dans un réseau multicouches permet de comprendre les motifs d'interdisciplinarité. [[cronin2014beyond](#)] est une tentative d'un aperçu de la nature complexe de la mesure des publications scientifiques et de la nature multidimensionnelle de la production de connaissance. Il fournit à la fois des contributions techniques récentes et des approches critiques, et insiste sur la nature “à-tête-de-Janus” des métriques, confirmant que la reduction de la production de connaissance à peu de dimensions n'est pas seulement trompeur mais aussi dangereux pour la science. La dimension géographique de la science a été étudiée par de nombreuses études de cas, comme [[maisonobe2013diffusion](#)] qui étudie la diffusion de questions et pratiques spécifiques en biologie moléculaire autour du monde.

Cybergeo comme cas d'étude

Cybergeo a été fondé en 1996 comme une journal européen de Géographie entièrement digital. Depuis, 737 articles scientifiques ont été publiés (jusqu'en mai 2016) par 1351 auteurs de 51 pays. Ces articles sont à l'origine de 2710 citations au total sur les 20 dernières années, ce qui correspond à la moitié des autres articles cités dans Cybergeo (5545).

La majorité des contributions proviennent d'une institution française (561), bien que des pays francophones (35 articles comprennent un auteur affilié au Canada, 21 en Suisse) et des pays voisins (23 contributions pour le Royaume-Uni, 18 pour l'Italie) soient également bien représentés (Fig. 117). Les sujets géographiques des articles eux-mêmes présentent une plus grande diversité, comme le monde est quasiment entièrement couvert (Fig. 117). Toutefois, la France et des pays voisins comme l'Espagne ou l'Allemagne sont le sujet principal de la majorité des articles, bien que les Etats-unis soient le 5ème pays le plus étudié. En reliant les auteurs à leur sujet géographique (Fig. 118), différentes tendances peuvent être mises en valeur :

- Des pays européens et nord-américains s'étudient mutuellement au travers des articles de Cybergeo ;
- Les pays d'Amérique sont étudiés par des auteurs affiliés en Europe et Amérique du Nord ;
- Les pays asiatiques et africains sont principalement étudiés par les européens, et de façon marginale par les américains et eux-mêmes ;
- La Russie et l'Australie sont étudiés par des auteurs occidentaux et étudient leur propre territoire.

Enfin, l'évolution temporelle montre une croissance accélérée du nombre d'auteurs - même si le nombre d'articles sur des périodes de 5 ans reste stable, une extension de la couverture géographique, avec plus d'articles publiés sur les pays émergents et les territoires extra-européens, ainsi que des connexions croissantes dans les réseaux de citation. Il existe un biais de renforcement en faveur d'auteurs francophone, révélé par l'origine des auteurs ainsi que la part des articles publiés en français.

C.2.2 Méthodes

Un aspect principal de cette contribution est la combinaison complémentaire de différentes méthodologies, chacune ayant ses potentialités et limitations, mais aussi des questions et des objets d'étude spécifiques. Nous détaillons dans cette section les différentes méthodes et comme celles-ci sont couplées pour produire une nouvelle connaissance.

Réseau sémantique interne

La première méthode d'exploration est basée sur l'ensemble des mots-clés déclarés par les auteurs des articles du journal *Cybergeo*. Deux réseaux peuvent être construits à partir des articles et des mots-clés : un réseau d'articles reliés par les mots-clés communs, et un réseau

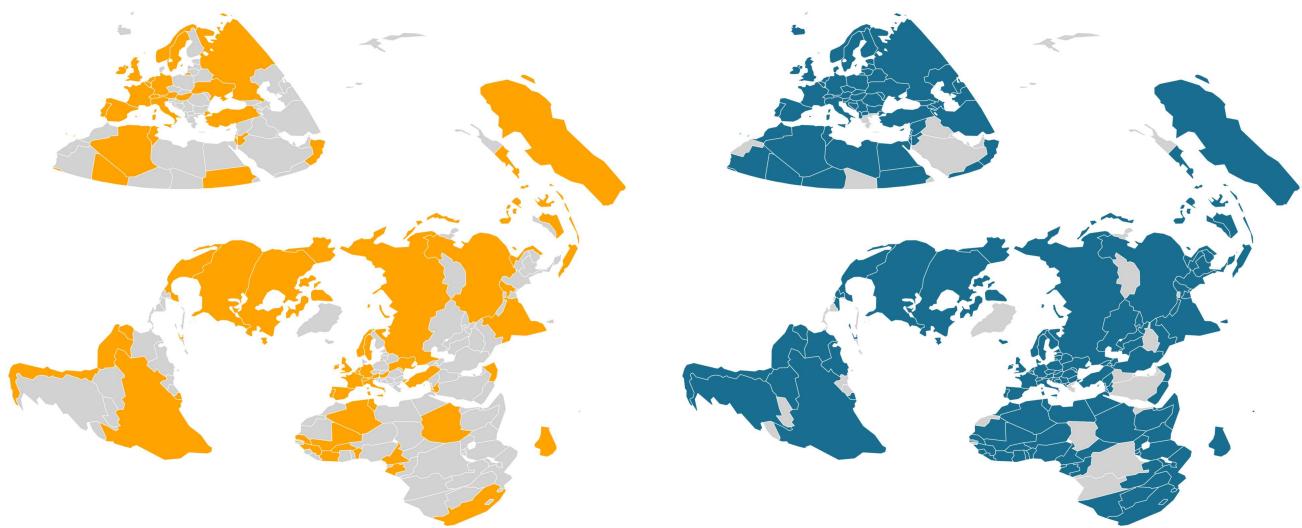


FIGURE 117: Pays avec au moins un auteur, 1996-2015 (Gauche); Pays étudiés au moins une fois, 1996-2015 (Droite)



FIGURE 118: Origine géographique et destination des articles, 1996-2015

de mots-clés reliés par les articles communs [roth_social_2010]. Ce dernier sera le réseau sémantique.

Les noeuds possèdent une variable de fréquence correspondant au nombre d'articles dans lequel ils apparaissent, ainsi que leur degré défini par leur nombre de liens. A partir des degrés, on peut calculer la probabilité de deux noeuds d'être connectés, ce qui donne un poids attendu au lien correspondant. Le poids observé étant le nombre d'articles citant les deux mots-clés, on définit le *poids modal* comme le ratio entre le poids observé et la racine du poids attendu. Ce poids modal peut être utilisé comme une mesure d'attachement préférentiel.

A partir de cette mesure d'attachement préférentiel deux types de visualisations sont utilisées : les champs sémantiques et les communautés. Le champ sémantique représente un mot-clé donné au centre d'un graphe avec l'ensemble de ses voisins à une distance inversement proportionnelle au poids modal. Les communautés sont obtenues avec l'algorithme de Louvain [blondel2008fast]. Cette méthode de détection de communauté est choisie en particulier car elle est basée sur une mesure de modularité similaire au poids modal défini ci-dessus.

Réseau sémantique externe

Le deuxième développement méthodologique se concentre sur la combinaison de l'exploration du réseau de citations avec l'analyse du réseau sémantique. La méthode appliquée pour ce développement est décrite en détails par [raimbault2017exploration]. Les réseaux de citation ont été largement utilisés dans les études de la science, par exemple comme outil prédictif pour le succès d'un article [2013arXiv1310.8220N], ou pour révéler des fronts de recherche émergents [shibata2008detecting].

En effet, la bibliographie d'un article contient d'une certaine façon un positionnement scientifique, comme un héritage auquel il cherche à contribuer et les champs sur lesquels il se base. Dans l'autre sens, les citations inverses, i.e. les contributions citant un article donné, jusqu'à un certain niveau, montre dans quelle mesure la connaissance produite a été comprise, interprétée et utilisée, et en particulier par quel champ (sur ce point l'exemple intéressant de [jacobs2016death], cité en masse aujourd'hui par les études quantitatives de la ville par les physiciens, montre comment le type d'audience peut être inattendu).

Nous définissons le voisinage de citation de notre corpus comme tous les articles citant les articles publiés dans *Cybergeo*, tous les articles citant ceux cités par *Cybergeo*, et tous les articles citant ceux-ci (obtenant ainsi un réseau à profondeur 2). Les données de citation sont collectées en utilisant une collecte automatiques de données de façon similaire à [raimbault2017exploration].

Après avoir construit ce voisinage de citation, nous introduisons une méthode pour analyser son contenu par analyse textuelle. Plus

précisément, nous nous intéressons aux mots-clés *pertinents* des résultats, en un sens précis tel qu'introduit par [chavalias2013phylomemetic] pour étudier l'évolution des champs scientifiques, et plus tard raffiné et étendu à des données massives pour une base de données de brevets par [bergeaud2017classifying]. En utilisant les co-occurrences des n-grams (mots-clés avec de multiples composantes, obtenus après un premier filtrage et nettoyage de texte), la déviation à une distribution uniforme entre les textes par un test du chi-deux donne une mesure de la pertinence des mots-clés, sur laquelle un nombre fixe $N_k = 50,000$ est filtré. Le réseau de co-occurrence pondéré correspondant entre les mots-clés pertinents capture leur relation au second ordre et nous supposons que sa topologie contient une information sur la structure des disciplines présentes dans le réseau de citation. Une analyse de sensibilité de la structure des communautés aux paramètres de filtrage du réseau (poids minimal sur les liens, degré maximal pour les noeuds, fréquence minimal et maximale de l'occurrence dans les documents) fournit un réseau robuste avec une structure de communautés optimale, qui permet d'associer à chaque article une liste de mots-clés et de disciplines correspondantes. Ceux-ci sont complémentaires aux mots-clés déclarés et aux thèmes des textes complets présentés ci-dessous, comme ils révèlent la manière dont les auteurs se positionnent dans le paysage sémantique associé au voisinage de citation, ou ce qu'est leur "bagage culturel".

Attribution de thématiques avec les textes complets

La troisième et dernière méthode d'exploration détaille la construction de thèmes à partir des textes complets, et se révèle ainsi complémentaires aux analyses précédentes. L'extraction de thèmes de documents textuels est un champ de recherche intense. Un thème est défini comme un ensemble de mots fréquemment utilisés conjointement dans les documents, et les documents sont des mélanges de thèmes. Les premières méthodes se basaient sur une pondération des mots, en particulier la méthode *term frequency inverse document frequency (tf-idf)* introduite par [salton_introduction_1986]. [hofmann1999probabilistic] a plus récemment proposé les premiers modèles probabilistes généralisés, qui ont conduit à la méthode de la *Latent Dirichlet Allocation (LDA)* [blei2003latent].

La méthode LDA déstructure les textes et considère les articles comme des ensembles de mots. Afin de garder un certain niveau de structure, nous utilisons ici le *part-of-speech tagging* développé par [schmid_probabilistic_1994] qui fournit la fonction des mots dans les phrases et extrait leur racine. Les noms, pronoms et verbes sont filtrés et pondérés par leur statistique *tf-idf*. On utilise la méthode LDA pour alors produire la composition des documents en termes de thèmes et la composition des thèmes en termes de mots-clés. Etant donné la matrice a priori β de composition des thèmes en termes de

mots-clés, les documents sont générés suivant diverses lois de probabilités. De manière itérative, les paramètres des distributions, incluant β sont estimés, par l'utilisation d'un échantillonnage de Gibbs [geman_stochastic_1984]. On peut alors analyser β pour expliquer les thèmes présents dans le corpus. Le nombre de thèmes est un paramètre fixé. Un nombre optimal peut être obtenu par minimalisation de la perplexité et maximisation de l'entropie des thèmes dans le corpus, comme proposé par [blei2003latent].

Aggregation géographique des profils sémantiques

Pour pouvoir produire les cartes des Fig.117 et Fig.118 ainsi que les analyses au niveau du pays, les articles ont été géocodés de deux façons. Dans un premier temps, la pays d'affiliation de ou des auteur(s) a été codé par les identifiants ISO à deux lettres. Dans un second temps, les articles ont été lus un par un pour extraire les sujets géographiques principaux. Les articles ont alors été codés par un pays si le pays ou une sous-région de celui-ci constituaient l'objet principal de l'étude. Dans le cas des pays européens, différents ensembles de pays ont été associés à la publication, selon le périmètre du sujet (par exemple EU15, EU25, espace Schengen, EuroMed, etc.). Etant donné une caractérisation sémantiques des articles (en utilisant les mots-clés, les citations ou les textes complets), il est ensuite possible de déterminer deux profils sémantiques pour les pays : l'un utilisant les pays comme origine des auteurs, l'autre comme la destination des sujets. Le profil sémantique d'un pays est constitué de la part moyenne des thèmes présent dans les articles dont l'auteur en provient ou étudiant celui-ci. Au total, étant donné les trois caractérisations sémantiques des articles et les deux attributions géographiques, chaque pays a au maximum six profils sémantiques distincts. Ces profils peuvent être utilisés pour regrouper les pays. La méthode de clustering utilisée est un algorithme de classification ascendante hiérarchique avec le critère de maximisation de distance de Ward. En analysant les clusters des profils d'auteurs, on considère les groupes de pays dans lesquels une certaine géographie est développée et écrite. Cet aspect est intéressant du point de vue de la réflexivité mais en pratique relativement aléatoire à cause de la forte concentration des émissions et par conséquent du faible nombre de pays écrivant. Pour cette raison, les résultats présentés seront basés sur les pays étudiés. Pour cela, nous considérons la manière dont un certain groupe de territoires est étudiés, quel mots-clés les auteurs utilisent pour les désigner et dans quel domaine de recherche les articles les concernant sont utilisés.

Open Data + interactivité = reproductibilité & transparence

Enfin, notre contribution méthodologique est également étroitement liée aux questions de réflexivité, transparence et reproductibilité dans

le processus de production de connaissance. Il s'agit d'une idée acceptée que ces aspects sont en interaction et que leur couplage participe à un cercle vertueux encourageant et accélérant la production de connaissances, comme on peut le voir dans les différentes approches de la Science Ouverte ([fecher2014open](#)). Par exemple, la revue par les pairs ouverte émerge progressivement comme une pratique alternative aux canons rigides et lents de la communication scientifique : [[10.12688/f1000research.11369.1](#)] procède à une revue systématique des approches de la notion pour en donner une définition unifiée et comprendre ses bénéfices ou défauts potentiels. Dans le domaine des sciences computationnelles, les outils pour assurer une réproductibilité et une transparence sont nombreux mais requièrent une discipline stricte d'utilisation et ne sont pas toujours facilement accessibles ([wilson2017good](#)). La Science Ouverte suggère une transparence du processus de production de connaissance lui-même, mais aussi des motifs de communication des connaissances : sur ce point nous appuyons l'idée que l'exploration interactive des motifs épistémologiques quantitatifs sont nécessaires. Nous construisons pour cela une application interactive pour permettre l'exploration de corpus scientifiques hétérogènes.

L'application web est disponible en ligne à <http://shiny.parisgeo.cnrs.fr/CybergeoNetworks/>. Le code source et les données, à la fois pour les analyses et pour l'application web, sont disponibles sur le dépôt git ouvert du projet à <https://github.com/AnonymousAuthor3/cybergeo20>.

c.2.3 Résultats

Réseau sémantique interne

COMMUNAUTÉS ET CHAMPS SÉMANTIQUES L'algorithme de détection de communauté trouve une modularité optimale qui donne 10 clusters : mobilité et transports; télédétection et SIG; climat et environnement; histoire et épistémologie; soutenabilité, risques, planification; géographie économique; territoires et populations; dynamiques urbaines; statistiques et modélisation; géographie des émotions. Certains clusters concentrent un grand nombre de mots-clés et d'article, comme "télédétection et SIG" ou "statistiques et modélisation". Ce résultat était prévisible vu le but et la portée originaux du journal. A côté des clusters principaux et d'un ensemble de clusters de taille moyenne, deux clusters de petite taille et totalement inattendus émergent : "géographie des émotions" et "environnement et climat". L'application *CybergeoNetworks* propose un ensemble de paramètres de visualisation pour tracer les communautés, comme présenté en Fig. 119, comme la taille des noeuds et des liens pouvant être variée selon différentes variables (degré, fréquence, poids modal).

Les métriques de poids modal expliquée précédemment peuvent être utilisée pour tracer les champs sémantiques. L'application *CybergeoNetworks* propose la liste complète des mots-clés. L'utilisateur choisit un mot-clé dans cette liste, celui-ci est placé au centre du graphe, et l'ensemble de ses voisins sont placés à une distance inversement proportionnelle à l'attachement préférentiel (poids modal). L'application propose des paramètres de visualisation comme la taille des caractères en fonction du poids des mots-clés (fréquence dans les articles ou degré dans le réseau), comme montré en Fig. 119. Certaines proximités sont attendues ("urban" est étroitement relié à "city"), d'autres sont attendus sachant l'étendue originale du journal dans le champ de la géographie théorique et quantitative ("model" ou "spatial statistics" sont reliés à "city"). Certaines sont inattendues, comme pour "city" l'attachement préférentiel de mots-clés comme "movie", "web", "virtual".

COMMUNAUTÉS SPATIALISÉES En utilisant les distributions des mots-clés pour déterminer les profils sémantiques des 128 pays étudiés dans les articles de *Cybergeo*, nous obtenons un clustering en 4 groupes représentant 16.5% de l'inertie initiale. Sa distribution géographique est montrée en Fig. 120 avec le profil moyen de chaque groupe.

Les pays sont premièrement différenciés par le fait ou non que les pays les étudiant déclarent également des mots-clés en lien avec la mobilité et le transport, l'histoire et l'épistémologie, les systèmes urbains, la géographie émotionnelle. En effet, le premier groupe de 83 pays (en bleu, Fig. 120) est défini par ces thèmes. Les pays correspondants sont les territoires les plus développés et les plus riches du monde, incluant les pays émergents comme les BRICS. Les mots-clés utilisés pour mettre en valeur les articles sur ceux-ci suivent les modes de la géographie, avec des mentions des émotions et de la mobilité par exemple.

Les pays des autres groupes sur-représentent les mots-clés en relation avec :

- les méthodes (en orange) comme les statistiques et la modélisation. Les pays associés à ces mots-clés sont tous situés en Afrique centrale et du sud, à l'exception du Laos. Ces pays sont étudiés par un petit nombre d'articles qui se concentrent sur les approches méthodologiques. Par exemple, le seul article étudiant le Rwanda [**querria2004localisation**] est en relation avec un problème de localisation optimale tandis que [**vallee2009disparites**] utilise le mot-clé "multilevel modelling" pour le seul article traitant du Laos.
- La soutenabilité et les risques (en jaune). C'est le cas des articles sur l'Indonésie par exemple, qui sont tous en relation avec les aléas et la vulnérabilité : des tsunamis [**ozer2005tsunami**], aux

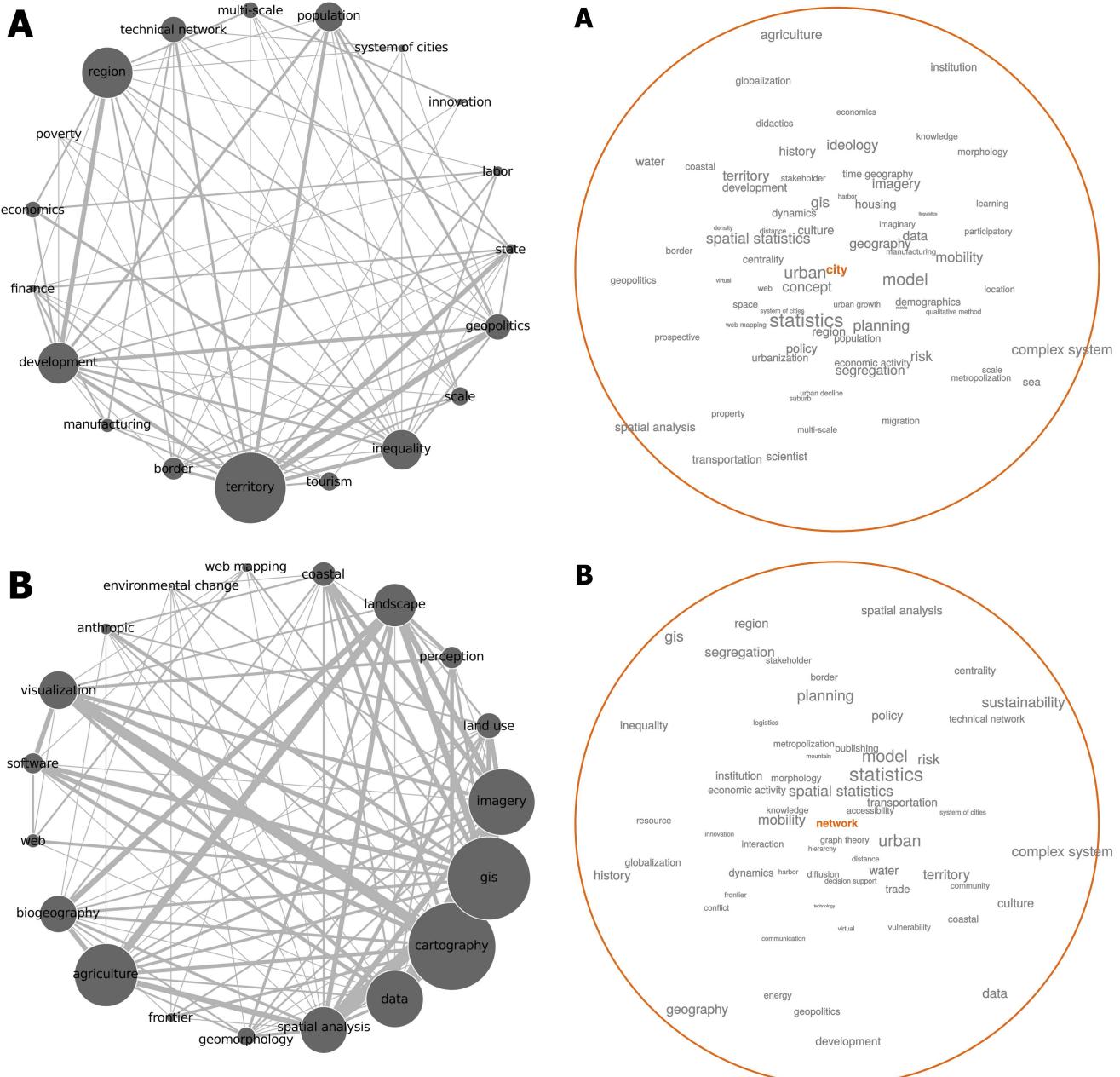


FIGURE 119: Structure de communauté du réseau sémantique interne : A-Territory, B-Imagery & GIS (Gauche); Champs sémantiques : A-City, B-Network (Droite)

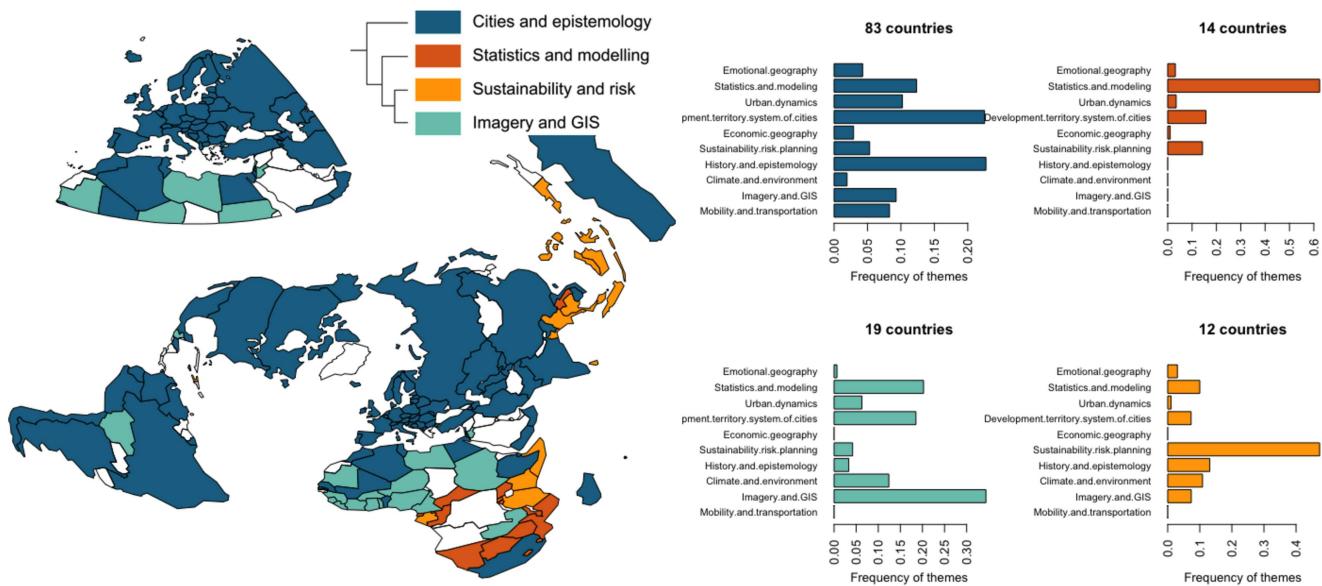


FIGURE 120: Communautés géographiques d'intérêt déclaré (*Gauche*); Profil sémantique des groupes correspondants (*Droite*).

volcans [[belizal2011quand](#)] et à la rareté des ressources en eau [[putra2009einfluss](#)].

- Enfin, 19 pays sont associés à des mots-clés en relation avec la télédétection et les SIG (en turquoise). Ils sont situés principalement en Afrique saharienne. Dans de nombreux cas, il s'agit d'articles qui présente une méthodologie qui exploite les images aériennes ou satellites pour extrapoler des données socio-économiques absentes [[ackermann2003analysis](#); [devaux2007extraction](#)].

Ainsi, en construisant les communautés d'intérêt déclaré, nous trouvons une dichotomie intéressante entre les pays riches d'une part, qui sont étudiés de manière intensive dans la littérature et pour lesquels les auteurs utilisent des mots-clés à la mode pour se détacher des travaux existants ou concurrents; et les pays en développement d'autre part, qui sont associés à des mots-clés plus techniques témoignant un spectre des domaines plus étroit et des problématiques spécifiques aux données.

Réseau sémantique externe

L'application permet l'exploration du voisinage de citation d'un article choisi, en termes de contenu sémantique (la visualisation des réseaux complets n'est pas faisable techniquement comme le corpus complet contient autour de 200,000 articles). Les nuages de mots donnent le contenu de l'article et le contenu des articles dans le voisinage, chaque mot étant associé aux communautés sémantiques. L'utilisateur peut ainsi situer un travail dans son contexte sémantique, et

nous nous attendons à ce que des liens non anticipés puissent être faits avec ces outils, comme les auteurs ne sont pas forcément au courant de travaux similaires dans des disciplines étrangères.

STRUCTURE DES COMMUNAUTÉS Comme expliqué précédemment, la réseau sémantique brut est optimisé pour la modularité et la taille, en prenant un compromis entre ces deux objectifs opposés, tandis que les paramètres de filtrage des liens et des noeuds varient. Cela permet d'obtenir 12 communautés, qui peuvent correspondre à des disciplines existantes, à des questions méthodologiques, ou à des sujets thématiques très précis. Les communautés sont, par ordre d'importance en termes de proportion de l'ensemble des mots-clés : Science politiques/communication ; Biogéographie ; Géographie Economique et Sociale ; Climat ; Géographie physique ; commerce ; analyse spatiale ; microbiologie ; neurosciences ; SIG ; agriculture ; santé. Cette méthode a la caractéristique de grouper les mots-clés par co-occurrences, révélant ainsi la structure effective du contenu des résumés : il s'agit simultanément d'un avantage en révélant des liens comme pour le champ très large de la Géographie Economique et Sociale, mais peut aussi bruiter l'information en groupant des communautés plus détaillées. Des communautés de taille modeste et très précises apparaissent, comme elles sont très isolées du reste des communautés. Cette structure est particulière, et témoigne d'une dimension de la connaissance qui par exemple n'est pas révélée par les analyses de citation classiques.

COMMUNAUTÉS SPATIALISÉES A l'aide des précédents réseaux pour construire le profil sémantique des 128 pays étudiés dans un article de Cybergeo, nous obtenons une classification en 5 groupes qui représentent 19,3% de l'inertie initiale. Sa distribution géographique est montrée en Fig. 121 ainsi que le profil moyen de chaque groupe.

Le plus grand groupe de pays recoupe largement le cluster le plus large des communautés de mots-clés présentées dans la section précédente. En effet, des pays riches et émergents sont étudiés dans des articles utilisés de manière similaire dans les réseaux de citation. Il existe des sous-divisions au sein de ce groupe. Un premier sous-groupe de pays (en bleu) est étudié par les articles de Cybergeo cités préférentiellement dans les champs du commerce et des analyses socio-économiques et politiques. Ceux-ci correspondent principalement à des articles en Economie et Sciences Sociales.

Le second sous-groupe de pays (en orange) comprend l'Australie, l'Azerbaïjan, l'Iran, le Laos, les Philippines et l'Islande. Il correspond à des pays traités par des articles cités préférentiellement dans des champs méthodologiques (analyse spatiale et GIS). En effet, le seul article à propos de l'Iran présente un système collaboratif d'aide à la décision ([jelokhani2012web](#)) tandis que le seul article

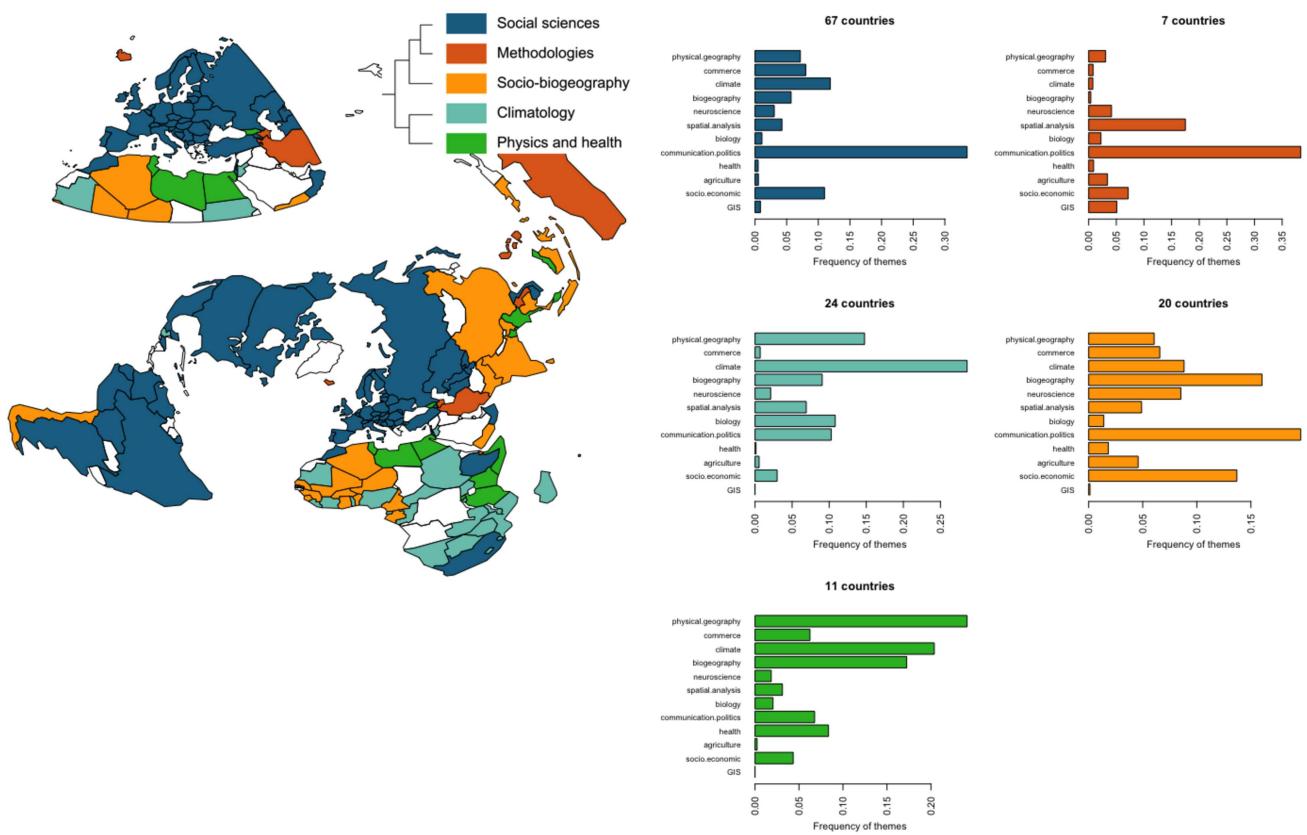


FIGURE 121: Communautés géographiques d'usage bibliographique (*Gauche*); Profil sémantique correspondant. (*Droite*).

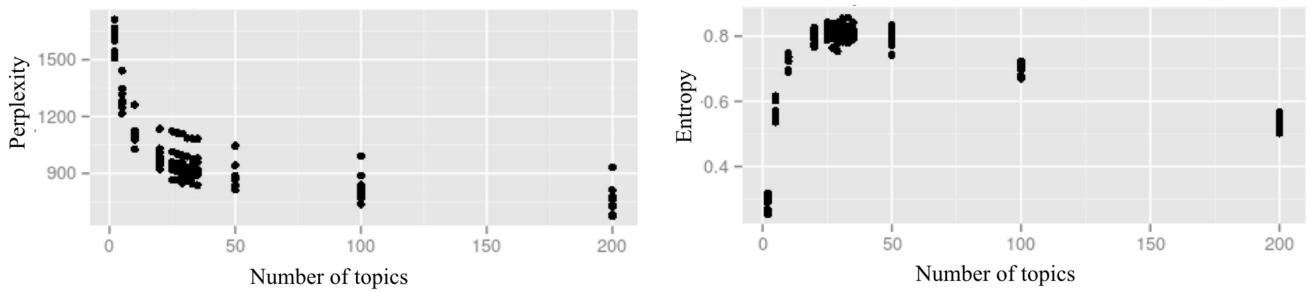


FIGURE 122: Perplexité et entropie du modèle LDA par nombre de thèmes.

sur l’Australie est une revue de produits cartographiques en ligne ([escobar2000distribution](#)). Ce type d’article tend ensuite à rester dans la clique de citation de la géomatique.

Le troisième sous-groupe correspond à des pays d’Asie du Sud-est, d’Afrique occidentale, le Yemen et le Chili. Les articles les étudiants sont cités en préférence dans les champs de la biogéographie et des études socio-économiques, bien qu’ils correspondent au profil moyen.

Dans le second groupe de pays, nous trouvons un premier sous-groupe de pays sub-Saharien (en turquoise) associés à des articles cités dans la communauté de citation de la climatologie. Le second sous-groupe est composé de pays d’Afrique de l’Est, du Nord et d’Asie du Sud-est (en vert) associés à des articles cités dans les champs de la géographie physique et de la santé.

Ainsi, en construisant les communautés d’usage bibliographique, nous trouvons une dichotomie intéressante entre les pays riches d’une part, qui sont associés à des articles cités dans des communautés très larges, incluant des champs thématiques et méthodologiques ; et des pays pauvres ou en développement d’autre part, qui sont associés à des articles cités principalement en relation dans la littérature aux catastrophes naturelles, la santé et les risques.

Thèmes avec textes complets

EVOLUTION DES THÈMES DU CORPUS La déstructuration des documents et le filtrage donne un dictionnaire d’environ $1.4 \cdot 10^5$ mots. Les paramètres LDA sont estimés pour un nombre de thèmes variant de 2 à 200, avec différentes résolutions notamment entre 20 et 40. La stochasticité est prise en compte en répétant 10 fois chaque estimation pour un nombre donné de thèmes. Comme montré en Fig. 122, le nombre de 20 thèmes est optimal au regard des indicateurs de perplexité et d’entropie.

Les 20 thèmes obtenus, classés par ordre d’importance (en terme de fréquence d’occurrence dans l’ensemble des documents) peuvent être synthétisés comme correspondant à : logement et voisinages ; mobilité et accessibilité ; télédétection ; planification et gouvernance ;

risques et vulnérabilité; santé; mots de liaison; modélisation des systèmes complexes; ville; ressources en eau; cartographie; histoire de la géographie; éducation; agglomérations urbaines; géopolitique; géographie électorale; administration; géomorphologie; paysage; géographie maritime. Nous pouvons observer l'utilisation des thèmes par année, comme présenté en Fig. 123. Différents profils d'évolution se distinguent : décroissant, ponctuel, constant et croissant. Les articles sur la cartographie (11) sont en nombre décroissant. Les articles sur la télédétection (3) ont été majoritairement produits en 2000, de la même manière que les articles sur les ressources en eau (10) en 2004 et 2011. Les articles sur les agglomérations urbaines sont produits de façon régulière. Des thèmes comme le voisinage (1) et la mobilité (2) ont tendance à augmenter.

COMMUNAUTÉS DE TEXTE COMPLET SPATIALISÉES En utilisant les textes complets pour construire les profils sémantiques des 128 pays étudiés dans les articles de *Cybergeo*, nous obtenons une classification en 4 groupes qui représentent 13.4% de l'inertie totale. Sa distribution géographique est montrée en Fig. 124 avec le profil moyen de chaque groupe.

Dans cette analyse de classification, nous ne retrouvons pas les dichotomies des pays selon leur richesse et leur niveau de développement économique. Le lien entre la proximité sémantique et géographique est également moins évidente au niveau mondial, bien qu'une région soit particulièrement révélée : les limites institutionnelles de l'Europe. Le groupe de pays qui contient l'EU27 et les USA, le Brésil et le Chili (en jaune) apparaissent fortement similaires en termes de vocabulaire utilisé pour les décrire. En particulier, les thèmes en relation avec les questions de frontières administratives ("communes") et la planification régionale ("amenagement") décrivent bien ces pays (par exemple : [santamaria2009schema ; lusso2009musees ; le2011consommation]). Deux sous-groupes sont voisins de ce cluster dans l'arbre de classification. Le premier inclut les pays étudiés par les articles écrits en Anglais. Le second sous-groupe inclut des pays de tous les continents et correspond aux articles qui utilisent préférentiellement des mots comme "eau" et "entreprise". Enfin, 59 pays sont distants de ces groupes en ce que les mots utilisés pour les décrire réfèrent aux villages et aux frontières ("frontière"), dans des contextes aussi divers que le Canada, l'Equateur, la Malaisie ou le Zimbabwe. Les communautés de vocabulaire et de pratiques d'écriture apparaissent ainsi moins évidentes et moins liées à la proximité géographique. Le résultat principal consiste en le fait qu'il existe un ensemble spécifique de mots pour écrire à propos de l'Union Européenne, une sorte de novlangue de l'EU27 composée de mots comme "Eurovision", "subventions" ou "Perspectives de développement spatial".

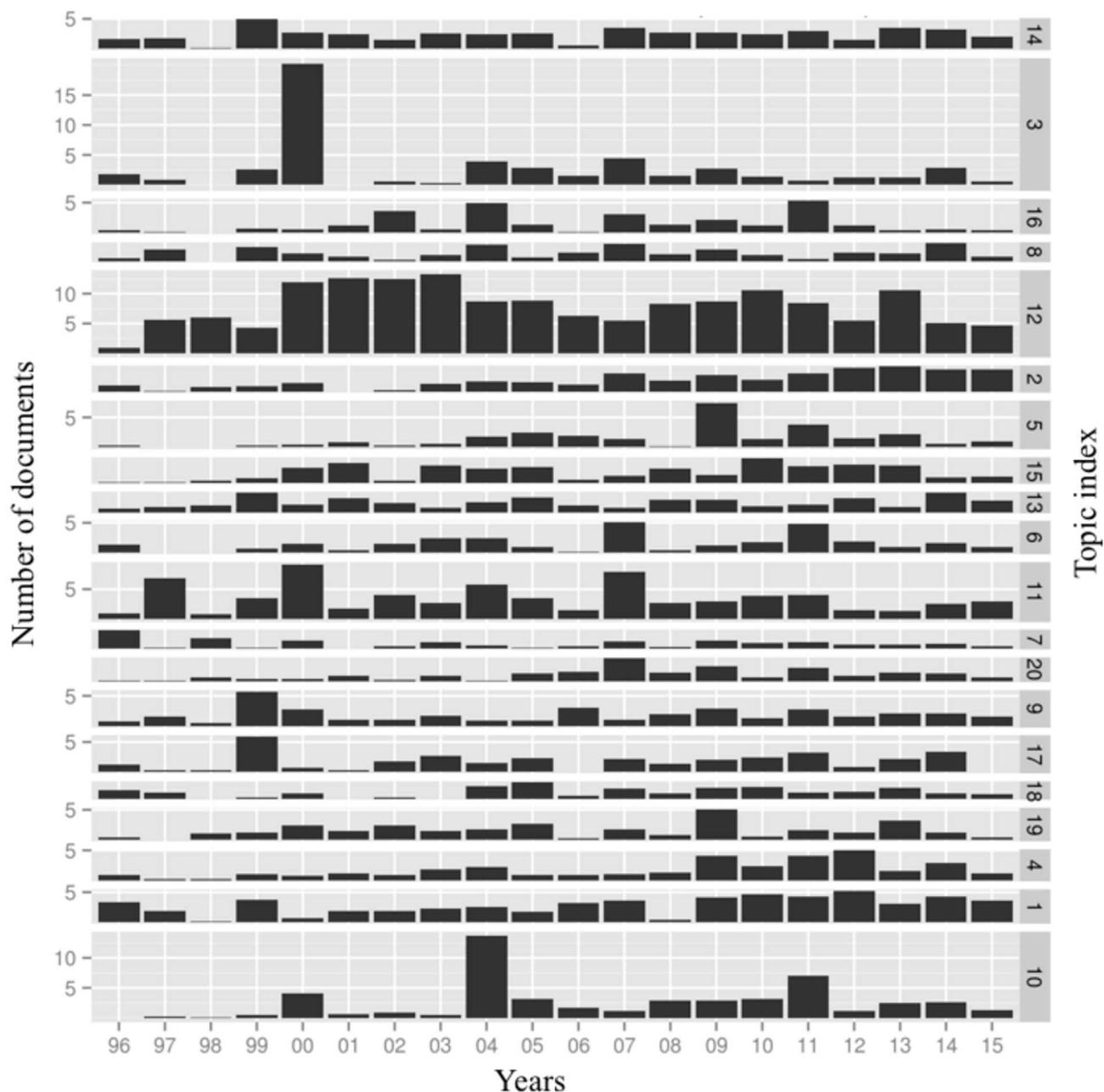


FIGURE 123: Nombre de documents traitant d'un thème par années, entre 1996 et 2015.

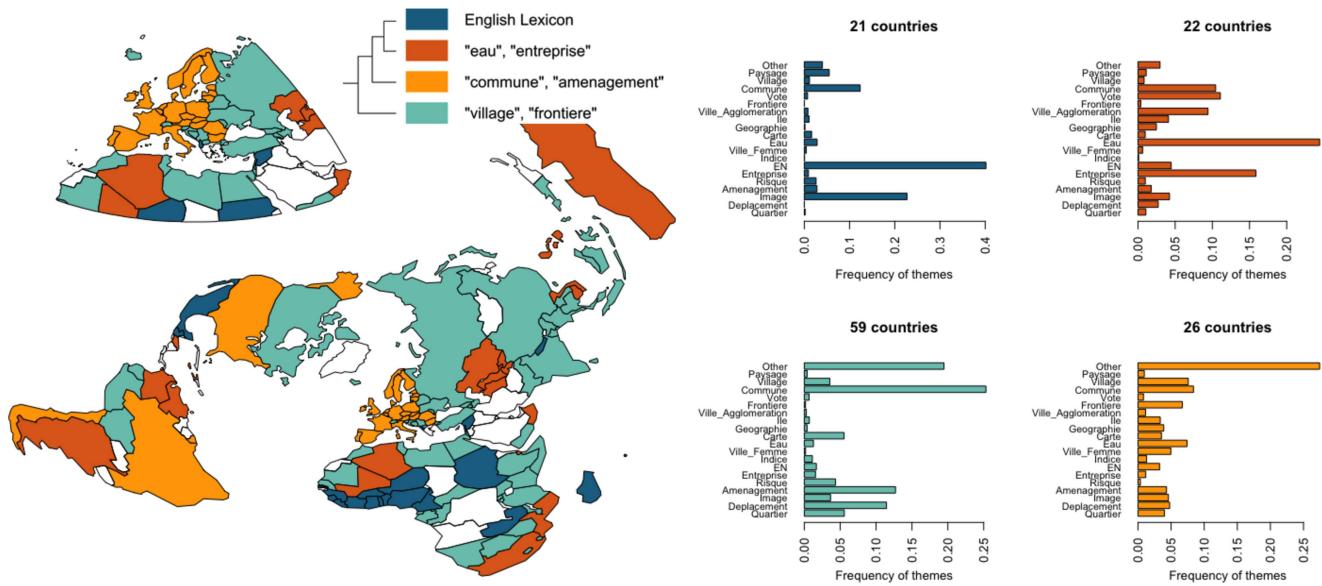


FIGURE 124: Communautés géographiques des pratiques d'écriture (Gauche); Profile sémantique des groupes correspondants (Droite)

C.2.4 Discussion

Evaluation de la complémentarité des approches

Cette section supporte les comparaisons qualitatives précédentes des approches par leur spatialisation par des mesures quantitatives de leur complémentarité. Bien que nous aillons vu que les communautés obtenues par les trois méthodes différentes sont distinctes sémantiquement et géographiquement, nous ne savons pas précisément comment elles se complètent. Une analyse de recouvrement est rendu difficile par le fait que les articles appartiennent simultanément à différents clusters pour chaque classification. Pour cela, nous comparons les méthodes 2 à 2 en calculant la part des articles classifiés simultanément dans chaque paire possible de clusters des deux méthodes. En d'autres termes, si une méthode M_1 (par exemple basée sur les communautés de citation) est composée de n catégories et une méthode M_2 (par exemple basée sur les communautés de mots-clés) est composée de m catégories, nous calculons pour chaque article $n \cdot m$ produits de co-occurrences et sommes ensuite ces produits en flux sur l'ensemble du corpus de *Cybergeo*. Si les méthodes étaient des moyens équivalents pour décrire et classifier les articles, nous devrions obtenir uniquement des flux de la forme $1 : 1, n : 1$ ou $1 : n$, comme les méthodes ne donnent pas les mêmes nombres de clusters. Si les méthodes étaient complètement orthogonales, chaque flux devrait être proportionnel à la taille du cluster d'origine et à celle du cluster de destination. Le fait que nous trouvions $n : n$ flux et qu'ils ne sont

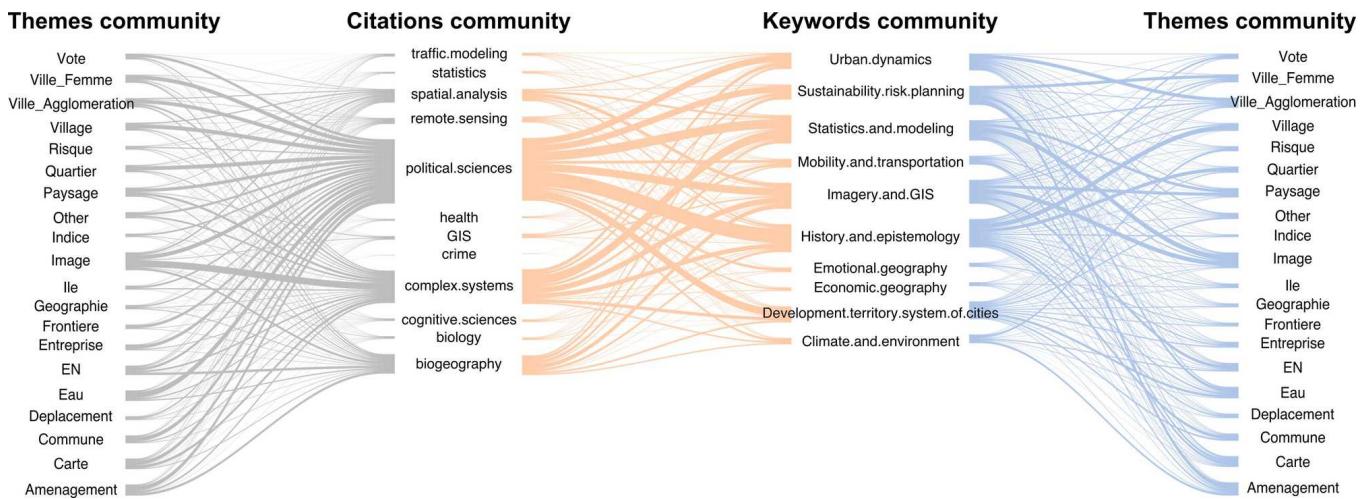


FIGURE 125: Recouvrement entre les communautés sémantiques.

pas entièrement déterminés par les tailles des clusters d'origine et de destination signifie que nos trois méthodes de classification sont ni équivalentes ni orthogonales (Fig. 125). Au contraire, ils donnent différents éclairages sur le corpus du journal.

Par exemple, il existe clairement des relations préférentielles positives et négatives entre certaines communautés de citation et communautés de mots-clés (Fig. 125). D'une part, 35% des articles de Cybergeo dans le cluster GIS des citations sont caractérisés par des mots-clés identifiés comme "télédétection et SIG". D'autre part, il n'existe pas d'article dans la communauté de citation "crime" qui ont des mots-clés dans la communauté de mots-clés "Climate and Environment". Ces relations font sens, puisque la manière dont un papier est promu dans ses mots-clés est l'un des premiers éléments indiquant au lecteur potentiel si le papier est pertinent ou non. De façon intéressante, la communauté de citation "complex systems" est caractérisée par une variété de communautés de mots-clés (27% des articles sont marqués dans "statistics and modelling", 17% dans "Imagery and GIS", 13% dans "History and Epistemology", 11% dans "Urban Dynamics"). Cela suggère que les champs des systèmes complexes, unifiés par des méthodes plutôt que par des objets de recherche, sont plus ouverts à des thèmes divers en comparaison aux autres communautés de citation. Cela peut également signifier qu'au sein de Cybergeo, les auteurs des articles pertinents pour le champ des systèmes complexes présentent leurs articles avec des mots-clés de la discipline géographique plutôt que les méthodes afin d'attirer des lecteurs thématiques également.

S'intéressant aux relations entre les communautés de mots-clés et celles de thèmes, nous trouvons que certains thèmes nécessitent des mots spécifiques pour les désigner. Par exemple, les articles étiquetés comme "Imagery and GIS" utilisent plus de mots de la catégorie thé-

matique "EN", qui correspond aux mots en anglais (plutôt qu'en français). Les études urbaines se distinguent entre leur côté quantitatif (présenté par des mots-clés autour de "urban dynamics" et utilisant des mots comme "agglomeration") et son côté qualitatif (présenté par des mots-clés autour de "sustainability, risk and planning" et utilisant des mots comme "femme"). De manière intéressante, les mots comme "risque" sont eux-mêmes utilisés plus fréquemment dans des articles étiquetés comme "Climate and Environment" plutôt que "sustainability, risk and planning". Enfin, les flux entre les communautés thématiques et les communautés de citation apparaissent globalement proportionnels aux tailles des clusters à l'origine et à la destination, suggérant que les citations sont relativement indépendantes du vocabulaire utilisé dans les articles. Cela se reflète dans l'analyse quantitative ci-dessous (Fig.127), cette paire ayant la corrélation absolue moyenne la plus faible. En résumé, les mots qui importent dans la stratégie de citation sont plus les mots-clés que le contenu effectif de l'article.

Nous synthétisons les relations de flux entre classifications en examinant leur structure de covariance de manière agrégée. Plus précisément, étant donné les matrices de probabilité $(p_{ki}) = (P_i)$ et $(p_{kj}) = (P_j)$ résumant deux classifications, dans lesquelles les articles sont indexés par la ligne, nous estimons la matrice de corrélation entre leurs colonnes $\rho_{ij} = \hat{\rho}[P_i, P_j]$ en utilisant un estimateur de corrélation de Pearson standard. Nous regardons ensuite des mesures agrégées, qui sont la corrélation minimale, la corrélation maximale et la corrélation absolue moyenne. Afin de disposer d'une référence pour interpréter les valeurs de ces corrélations, nous les comparons à deux modèles nul obtenus en par un bootstrap de corpus aléatoires. L'estimation pour le modèle nul inférieur (ρ_0) doit minimiser la corrélation et est obtenu par permutation de l'ensemble des lignes d'une des deux matrices, ce qui est fait successivement sur chaque pour assurer la symétrie. Le modèle nul supérieur (ρ_+) est construit est construit par calcul des corrélations entre une matrice et l'identique dont une proportion fixée de lignes ont été permutées. Nous fixons cette proportion à 50%, ce qui est un assez haut niveau de similitude, et calculons le modèle pour chaque matrice à chaque fois. Les moyennes et deviation standard sont calculées sur $b = 10000$ répétitions de bootstrap. La Table 126 résume les résultats. Nous trouvons que la corrélation maximale pour le corpus Cybergeo, qui peut être interprétée comme un recouvrement maximal entre les approches de classification sémantique, est toujours significativement plus petite (autour de $5 \cdot \sigma$) que pour le modèle nul supérieur. Cela confirme que nos trois classifications sont fortement indépendantes l'une de l'autre dans leurs composantes principales. Il est intéressant de relever que pour la relation Mots-Clés/Thèmes, la corrélation absolue moyenne est dans la déviation standard de la corrélation absolue

FIGURE 126: Corrélations entre les 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 ± 0.070	0.69 ± 0.070	0.059 ± 0.0021	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 ± 0.068	0.64 ± 0.068	0.070 ± 0.0026	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 ± 0.075	0.66 ± 0.075	0.091 ± 0.0022	0.040 ± 0.0022	0.080 ± 0.020

Notes : Pour chaque paire de classification et chaque mesure, nous donnons également la moyenne et la déviation standard pour les modèles nuls inférieur (ρ_0) et supérieur (ρ_+), obtenus par bootstrap de $b = 10000$ corpus aléatoires.

moyenne du modèle nul supérieur, suggérant que celles-ci doivent être plutôt proches sur les faibles recouvrements. Elles sont effectivement plus proches qu'avec Citation pour l'ensemble des indicateurs. Nous confirmons aussi que Thèmes/Citation ont le plus bas recouvrement absolu moyen.

Pour rendre ces conclusions plus robustes, nous complémentons l'analyse par une analyse de modularité de réseau, qui est une méthode largement appliquée pour évaluer la pertinence d'une classification dans un réseau. Pour être en mesure de comparer deux classifications, comme le réseau de citations est trop creux pour toute analyse comme déjà mentionné, nous évaluons la modularité d'une classification au sein du réseau induit par l'autre. Plus précisément, étant donné un seuil de distance θ et deux documents donnés par leur probabilités dans une classification $\vec{p}_i^{(c)}, \vec{p}_j^{(c)}$, nous considérons le réseau avec les documents comme noeuds, liés si et seulement si $d(\vec{p}_i^{(c)}, \vec{p}_j^{(c)}) < \theta$ avec d distance euclidienne. On peut ensuite calculer la modularité multi-classes de l'autre classification au sens de [nicosia2009extending]. Nous montrons en Fig. 127, pour différents seuils, les modularités normalisées par la modularité de la classification du réseau dans son propre réseau. Le plus proche de 1 est la mesure, le plus proches sont les classifications. La plupart des couples ont de faibles valeurs pour de larges plages de θ , confirmant les conclusions précédentes d'orthogonalité. De plus, les différents comportements en fonction de θ (croissant ou décroissant) suggère différentes *structures internes* des classifications, ce qui est cohérent

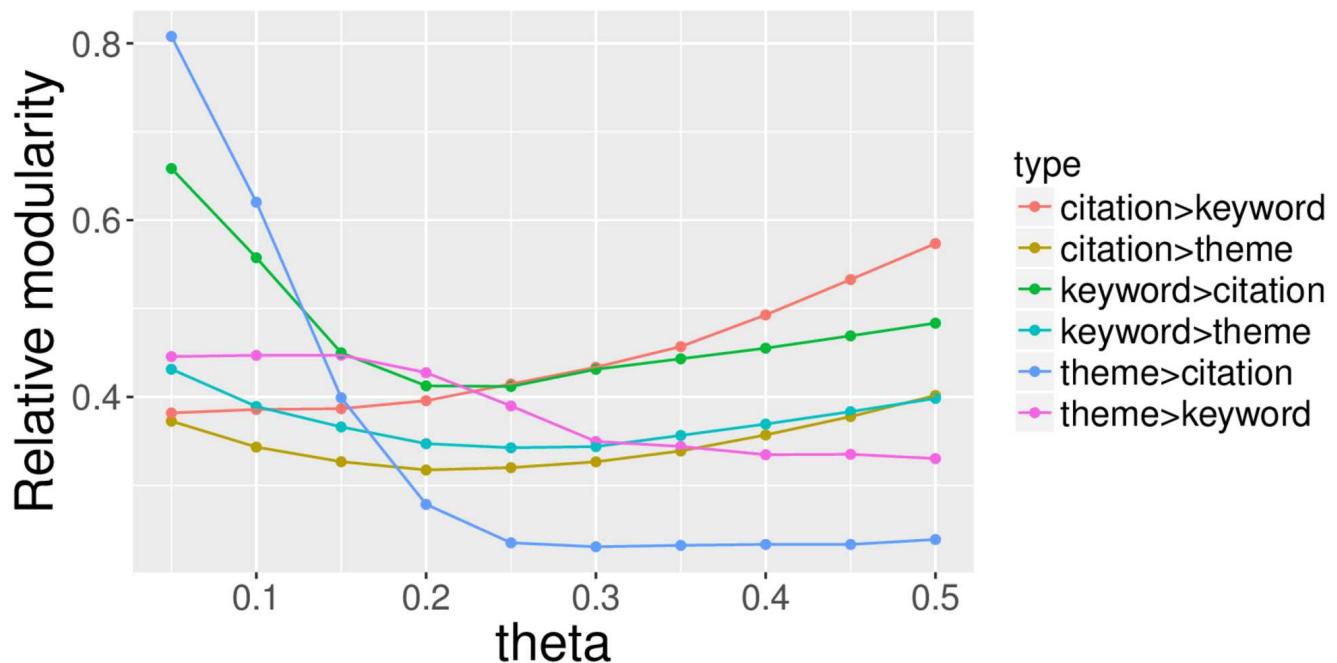


FIGURE 127: Evaluation de la complémentarité des classifications par la modularités des réseaux.
Le graphe donne la modularité relative de la première classification dans le réseau induit par la seconde avec le seuil θ (voir texte), pour chaque couple de classifications (couleurs).

avec le fait qu'elles reposent sur des processus différents pour classifier les données.

En combinaison aux diagrammes visuels, ces analyses montrent la complémentarité des classifications dans l'exploration de la diversité sémantique des publications dans un journal qui fête ses 20 ans.

Perspectivisme appliqué

Notre approche peut être comprise comme un “perspectivisme appliqué”, que nous postulons comme une manière de contribuer à la création d'une connaissance au second ordre et d'assurer une réflexivité. Le perspectivisme est une position épistémologique défendue par [giere2010scientific], qui vise à sortir par le haut des débats entre constructivisme et réalisme. En se concentrant sur les agents scientifiques comme supports de la création de connaissances, toute entreprise scientifique est une certaine *perspective* sur le monde, prise par l'agent dans un certain but et grâce à un certain medium qui est considéré comme étant son *modèle*. Les perspectives sont généralement complémentaires puisqu'elles résultent d'approches différentes sur les mêmes objets, bien que les définitions des objets et les questions de recherche ne seront pas nécessairement les mêmes. Le couplage de perspectives devrait ainsi être une caractéristique essentielle de l'interdisciplinarité. Nous positionnons notre travail comme une tentative délibérée de coupler des perspectives complémentaires sur

le même corpus. [varenne2017theories] rappelle que l'une des nombreuses fonctions des modèles est de favoriser le couplage entre théories par le couplage des modèles eux-mêmes, permettant la création d'une connaissance nouvelle au sein de la spirale vertueuse entre disciplinarité et interdisciplinarité mise en valeur par [banos2013pour]. Notre travail vise précisément à accélérer et améliorer de tels processus.

Encourager la Science ouverte et la Réflexivité

Les outils ouverts et les logiciels que nous produisons participent à un effort plus large de conception d'outils de réflexivité dans le contexte de la Science Ouverte. Ils visent à être complémentaires aux plateformes existantes, comme le *Community Explorer* développé pour la communauté des Systèmes Complexes développé par l'ISCPIF¹ qui fournit une visualisation interactive des réseaux sociaux de la recherche en combinaison aux réseaux sémantiques basés sur les mots-clés auto-déclarés donnés par les chercheurs. Un autre exemple plus proche de ce que nous avons développé est l'outil Gargantext² qui fournit des fonctions d'exploration de corpus. Linkage³ est un outil similaire avec des méthodes différentes, qui utilise un allocation latente de thèmes pour les réseaux avec annotations textuelles (**bouveyron2016stochastic**). Nous nous différencions de ceux-ci en explorant simultanément de multiples dimensions de classification sémantique et de façon plus importante en y ajoutant la composante géographique. De plus, en comparaison aux outils variés que les éditeurs privés commencent à introduire, la nature ouverte et collaborative de notre travail est essentielle. Par exemple, [bohannon2014google] suggère qu'il faut rester précautionneux quant à l'utilisation des résultats de recherche issus d'un engin académique de requêtes populaire, comme les mécanismes de l'algorithme de classement et ainsi les biais multiples restent inconnus. La comparaison est similaire avec les services payant de fouille de texte fournis par des entreprises privées, puisque nous suggérons qu'une synergie subtile entre le contenu de la connaissance et les processus de production de connaissance (qui est permise par des outils ouverts de manière privilégiée) peut être plus bénéfique aux deux.

c.2.5 Conclusion

Nous avons étudié un corpus scientifique d'un journal en Géographie, combinant des points de vue multiples par leur plongement dans l'espace géographique. Ce travail est pour cela en lui-même réflexif, illustrant le type d'approche nouvelle à la science qu'il cherche

¹ disponible à <https://communityexplorer.org>

² <https://gargantext.org/>

³ <https://linkage.fr/>

à promouvoir. Nous sommes convaincus que les outils ouverts que nous développons dans ce contexte contribueront à la prise d'autonomie des auteurs dans la Science Ouverte.

C.3 CLASSIFICATION SÉMANTIQUE DES BREVETS

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 [10.1371/journal.pone.0176310]. Elle est ici traduite et adaptée.

* * *

*

Ainsi, nous étendons ici les techniques de classification usuelles par une approche d'analyse de réseau et de fouille de données à grande échelle. Ces nouvelles techniques, qui sont conçues particulièrement pour être adaptées aux données massives, sont utilisées pour construire une base consolidée ouverte à partir de données brutes de 4 millions de brevets issus de l'US Patent Office (USPTO) depuis 1976. Pour construire le réseau, nous utilisons les titres, mais aussi les résumés complets desquels nous extrayons les mots-clés pertinents. Nous

désignons cette classification comme *approche sémantique*, en comparaison avec la plus classique *approche technologique* qui consiste à considérer la topologie induite par les classes technologiques de l'USPTO. De plus, nous obtenons des mesures topologiques fortement différentes pour les deux approches. Cela suggère que cette méthode est un outil puissant pour extraire une information endogène.

Introduction

L'innovation et le changement technologique ont été présentés par de nombreux chercheurs comme des moteurs principaux de la croissance économique comme dans [aghionhowitt1992] et [romer1990]. [RePEc:nbr:nberwo:3301] a proposé l'utilisation des brevets comme un indicateur économique et comme un bon proxy pour l'innovation. Par conséquent, la disponibilité accrue de bases de données plus complètes et détaillées des brevets et l'augmentation du nombre d'études qui ont permis une utilisation plus efficace de ces données (par exemple [Hall2001]) ont ouvert la voie à des analyses très variées. La majorité des statistiques issues des bases de données sur les brevets se basent sur un petit nombre de caractéristiques essentielles : l'identité de l'inventeur, le type et l'identité du possesseur des droits, les citations faites par le brevet à des travaux antérieurs et les classes technologiques attribuées par le bureau des brevets après une revue du contenu du brevet. La combinaison des ces informations est particulièrement pertinent pour tenter de capturer la diffusion des connaissances et les interactions entre champs technologiques comme étudié dans [Youn:2015fk]. Par des méthodes comme la modélisation dynamique des citations, étudiée par [2013arXiv1310.8220N], ou les analyses de réseaux de co-auteurs dans [2014arXiv1402.7268S], une partie conséquente de la littérature comme [sorenson2006complexity] ou [kay2014patent] s'est intéressée aux réseaux de citations de brevets pour comprendre les processus à l'origine de l'innovation technologique, de la diffusion et de la naissance de clusters technologiques. Enfin, [bruck2016recognition] étudie par exemple la dynamique des citations depuis différentes classes pour montrer que la technologique de l'imprimante laser jet d'encre est le résultat de la recombinaison de deux technologies existantes.

Par conséquent, la classification technologique combinée à d'autres caractéristiques des brevets peut être un outil efficace pour l'étude des technologies à travers l'histoire et pour prédire des innovations futures en se basant sur la connaissance passée et les interactions entre secteurs et technologies. Mais il est également crucial pour les entreprises qui font face à une structure toujours changeante de la demande et qui ont besoin d'anticiper les tendances technologiques futures et les convergences (voir par exemple [currant2011patent]) pour s'adapter à la compétition accrue qui en ressort, discutée dans

[**Katz1996remarks**] et pour maintenir une part de marché. De manière surprenante, et malgré le grand nombre d'études qui analysent les interactions entre technologies [**Furman2011shoulders**], la connaissance du "réseau d'innovation" [**AAKnetwork2016**] sous-jacent est relativement faible.

Nous proposons ici une classification alternative basée sur l'analyse sémantique des résumés des brevets et explorons la nouvelle information qui en émerge. En opposition avec la classification technologique classique qui est produite par les choix du relecteur du brevet, la classification sémantique est opérée automatiquement en se basant sur le contenu du résumé du brevet. Bien que les *patent officers* (experts chargés de la classification) soit experts dans leur champs, la pertinence de la classification existante est limitée par le fait qu'elle est basée sur l'état de la technologie au moment où le brevet est accordé, et ne peut pas anticiper la naissance de nouveau champs. Pour corriger cela, l'USPTO change régulièrement ses classifications afin de les adapter au changement technologique (par exemple, la classe "nanotechnologie" (977) a été établie en 2004 et de manière retroactive à tous les brevets antérieurs pour lesquels elle était pertinente). Au contraire, ce problème n'est pas présent dans l'approche sémantique. Les liens sémantiques peuvent être des indices d'une technologie s'inspirant d'une autre et de bons prédicteurs de futures convergences technologiques (par exemple [**preschitschek2013**] étudie des similarités sémantiques dans l'ensemble du texte de 326 brevets sur les *phytosterols* et montre que l'analyse sémantique a un bon pouvoir prédictif d'une convergence technologique future). On peut par exemple considérer le cas du mot *optic*. Jusqu'à récemment, ce mot était souvent associé aux technologies comme la photographie et la chirurgie de l'oeil, alors qu'il est maintenant presque exclusivement utilisé dans le contexte de la conception des semi-transistors et de l'optique électronique. Cette dérive sémantique ne s'est pas opérée par hasard mais contient de l'information sur le fait que l'électronique moderne utilise de manière intensive des technologies qui ont été initialement développées en optique.

Des travaux existants ont déjà proposé d'utiliser les réseaux sémantiques pour étudier les domaines technologiques et détecter la nouveauté. [**yoon2004text**] était parmi les premiers à utiliser cette approche dans l'idée de visualiser les réseaux de mots-clés, illustrée sur un champ technologique restreint. La même approche peut être utilisée pour aider les entreprises à identifier l'état de l'art dans leur domaine et éviter les violations de brevets, comme dans [**park2014semantic**] et [**yoon2011detecting**]. Plus proche de notre méthodologie, [**gerken2012new**] développe une méthode basée sur l'analyse sémantique des brevets pour appuyer l'idée que cette approche est plus performante pour suivre la technologie et pour l'identification de la nouveauté dans les innovations. L'analyse sémantique a déjà montré ses capacités dans

des champs variés, comme l'étude des technologies (par exemple [[choi2014patent](#)] et [[fattori2003text](#)]) et en sciences politiques [[2015arXiv151003797G](#)].

En se basant sur ces travaux, nous faisons diverses contributions en dépassant certaines limitations des études existantes, comme par exemple l'utilisation de mots-clés simples sélectionnés par leur fréquence. Tout d'abord, nous développons et implémentons une nouvelle méthodologie totalement automatisée pour classifier les brevets selon le contenu sémantique de leur résumés, qui est à notre connaissance la première de ce type. Elle inclut les raffinements suivants pour lesquels des détails peuvent être trouvés dans la section suivante : (i) utilisation de racines multiples comme mots-clés potentiels ; (ii) filtration des mots-clés selon une mesure de pertinence au second ordre (cooccurrences) et sur une mesure externe indépendante (dispersion technologique) ; (iii) optimisation multi-objectifs de la modularité et de la taille du réseau sémantique. L'utilisation de l'ensemble de ces techniques dans le contexte de la classification sémantique est nouveau et essentiel dans une perspective appliquée.

De plus, la majorité des études existantes se basent sur un échantillon des données de brevet, alors que nous implémentons ici notre méthode sur l'ensemble de la base USPTO de 1976 à 2013. De cette manière, une structure générale de l'innovation technologique peut être étudiée. Nous tirons de cette application divers faits stylisés qualitatifs, comme un changement qualitatif de régime autour de la fin des années 1990, et une amélioration significative de la modularité de citation pour la classification sémantique en comparaison à la classification technologique. Ces conclusions thématiques valident notre méthode comme un outil utile pour extraire de l'information endogène, de manière complémentaire à la classification technologique.

Grace à cette complémentarité, nous postulons que les *patent officers* pourraient bénéficier significativement de la considération du réseau sémantique lors de la considération des citations potentielles pour un brevet en revue.

Ce travail est organisé de la façon suivante : nous présentons d'abord les données, la classification existante, et donnons des détails sur la procédure de collection des données. Nous expliquons ensuite la procédure de construction des classes sémantiques. La pertinence de celles-ci est alors testée par des analyses exploratoires. Enfin, nous discutons des développements potentiels en conclusion.

Contexte

Dans l'analyse, nous considérons l'ensemble des brevets accordés par l'USPTO de 1976 à 2013. Une définition plus précise d'un brevet d'utilité (*utility patent*) est donnée en Annexe de [[10.1371/journal.pone.0176310](#)]. De plus, des informations supplémentaires sur la manière d'exploiter

correctement les données de brevet est donnée dans [Hall2001] et [lerner2015use].

Une classification existante : le système USPC

Chaque brevet USPTO est associé avec un ensemble non vide de classes et de sous-classes technologiques. Il existe actuellement autour de 440 classes et plus de 150000 sous-classes, qui constituent le système *United States Patent Classification* (USPC). Alors qu'une classe technologique correspond au champ technologique couvert par le brevet, une sous-classe correspond à une technologie spécifique ou une méthode utilisée dans l'invention. Un brevet peut avoir de multiples classes technologiques : en moyenne dans nos données un brevet a 1.8 classes différentes et 3.9 paires de classe/sous-classe. A cette étape, deux caractéristiques de ce système valent la peine d'être mentionnées : (i) la classe et la sous-classe ne sont pas choisies par l'inventeur du brevet mais par l'examinateur pendant le processus de revue, en fonction du contenu du brevet ; (ii) la classification a évolué au cours du temps et continue à évoluer afin de s'adapter aux nouvelles technologies par la création ou l'édition de classes. Lorsqu'un changement a lieu, l'USPTO revoit l'ensemble des brevets précédents afin de créer une classification cohérente.

Un réseau bibliographique entre brevets : les citations

De la même manière que les publications scientifiques, les brevets doivent faire référence à l'ensemble des brevet précédent qui correspondent à une connaissance antérieure nécessaire. Les citations indiquent donc la connaissance passée en relation avec l'invention brevetée. Toutefois, au contraire des publications scientifiques, elles ont aussi un important rôle légal puisqu'elles sont utilisées pour délimiter la portée des droit propriétaires donnés par le brevet. On peut consulter [oecdpatentmanual] pour plus de détail sur cette procédure. Un manque de références aux brevets passés peut mener à une invalidation du brevet [martin2015]. Une autre différence cruciale est que la majorité des citations est en fait choisie par les examinateurs et non les inventeurs eux-mêmes. A partir de l'USPTO, nous rassemblons l'information de l'ensemble des citations faites par chaque brevet (citations données) et toutes les citations reçues par chaque brevet à la fin de 2013 (citation reçues). Nous pouvons ainsi construire un réseau complet de citations qui sera utilisé par la suite dans l'analyse.

En se tournant vers la structure du délai entre les brevets citants et les brevets cités en termes de date de soumission, nous constatons que la moyenne de ce délai est de 8.5 ans et la médiane 7 ans. Cette distribution est fortement dissymétrique, le 95^{ème} centile étant 21 ans. Nous relevons également 164000 citations avec un délai négatif. Cela est du au fait que des citations peuvent être ajoutées pendant

le processus de revue et que des brevets requièrent plus de temps que d'autres pour être accordés.

Dans la suite, nous choisissons de restreindre notre attention aux paires de citations avec un délai inférieur ou égal à 5 ans. Cette restriction est imposée pour deux raisons. D'abord, le nombre de citations reçues présente un pic 4-5 ans après l'application. Ensuite, la structure du délai de citation est nécessairement biaisée par la restriction de l'échantillon : les brevets les plus récents reçoivent mécaniquement moins de citations que les plus anciens. Comme nous nous restreignons aux citations reçues en moins de 5 ans après la date de soumission, cet effet n'affectera que les brevets avec une date de soumission postérieure à 2007.

Collecte des données et description élémentaire

Chaque brevet contient un résumé et un texte qui décrit l'invention. Pour voir à quoi ressemble un brevet en pratique, on peut se référer à la base USPTO des textes complets <http://patft.uspto.gov/netahtml/PTO/index.html> ou à Google Patent qui publie les brevets USPTO au format pdf à <https://patents.google.com>. Même si l'inclusion des textes complets serait naturelle et probablement très utile dans une approche d'analyse textuelle systématique comme faite dans [tseng2007text], ceux-ci sont trop longs pour être inclus ici et nous ne considérons donc que les résumés pour l'analyse. En effet, l'analyse sémantique porte sur plus de 4 million de brevets, avec des résumés correspondants avec une longueur moyenne de 120.8 mots (et une déviation standard de 62.4), une taille qui est déjà un défi en termes de charge computationnelle et de taille des données. De plus, les résumés visent à synthétiser le but et le contenu des brevets et doivent pour cela être un objet d'étude pertinent (voir [Adams2010text]). L'USPTO définit une ligne directrice qui précise que le résumé doit être "un résumé de l'information comme contenue dans la description, les positions et les figures ; le résumé doit également indiquer le champ technique dans lequel l'invention se situe et doit être rédigé d'une manière permettant une compréhension claire du problème technique, le cœur de la solution au problème apporté par l'invention et le ou les usages principaux de l'invention" (PCT Règle 8).

Nous construisons à partir des données brutes une base unifiée. Les données sont récupérées à partir de la page des téléchargement du *redbook* des brevets de l'USPTO, qui fournit sous format brut (formats dat ou xml spécifiques) l'ensemble des informations des brevets, à partir de 1976. La procédure détaillée de collection des données, de parsing et de consolidation sont disponible en Annexe de [[10.1371/journal.pone.0176310](#)]. L'image la plus récente de la base est disponible au format MongoDB à <http://dx.doi.org/10.7910/DVN/BW3ACK>. La collection et l'homogénéisation des données en une base directement utilisable avec les informations de base et les résumés est

une étape importante, puisque des formats USPTO bruts sont impliqués et changent fréquemment.

Nous dénombrons 4,666,365 brevets d'utilité avec un résumé obtenus entre 1976 et 2013. Un très faible nombre de brevet ont un résumé manquant, ils correspondent aux brevets qui ont été annulés et nous ne les considérons pas dans l'analyse. Le nombre de brevets accordés chaque année de environ 70000 en 1976 à environ 278000 en 2013. Lorsqu'ils sont distribués par année d'application, le schéma est légèrement différent. Le nombre de brevets s'accroît à taux constant de 1976 à 2000 et reste constant à environ 200,000 par an de 2000 à 2007. En restreignant l'échantillon aux brevets dont la date de soumission est entre 1976 et 2007, il reste 3,949,615 brevets. Ces brevets en citent 38,756,292 autres avec le délai empirique qui a été étudié de manière détaillée par [Hall2001]. Conditionnellement à être cité au moins une fois, un brevet reçoit en moyenne 13.5 citations sur une fenêtre de 5 ans. 270,877 brevets ne reçoivent aucune citation pendant les 5 ans qui suivent la soumission, 10% des brevets ne reçoivent qu'une seule citation et 1% reçoit plus de 100 citations. Une citation interne à une classe est définie comme une citation entre deux brevets partageant au moins une classe technologique. Suivant cette définition, 84% des citations sont internes aux classes. 14% des citations sont entre des brevets qui partagent exactement le même ensemble de classes technologiques.

Vers une classification complémentaire

La potentialité des techniques de fouille textuelle comme un moyen alternatif pour analyser et classifier les brevets est documenté dans [tseng2007text]. L'argument principal de l'auteur, en support d'un outil de classification automatique pour les brevets, est la réduction de la quantité de travail humain considérable nécessaire pour classer toutes les soumissions. Le travail conduit dans le champ du traitement naturel du langage et/ou de l'analyse textuelle a été développé afin d'améliorer la performance des recherches dans les bases de brevets, construire des cartes des technologies ou investiguer de potentiels risque de violation de brevets avant le développement d'une nouvelle technologie (voir [abbas2014literature] pour une revue). La fouille de texte des documents de brevets est également largement utilisée comme outil pour construire des réseaux qui contiennent une information complémentaire au modèles simplistes de connection bibliographique comme argumenté dans [yoon2004text]. D'après les auteurs, l'utilisation de fouille de texte comme moyen de construction d'une classification globale des brevets reste cependant largement sous explorée. Une exception notable peut être trouvée dans [preschitschek2013] où une classification sémantique est montrée plus performante que la classification standard dans la prédiction de la convergence technologique même sur de petits échantillons. L'analyse sémantique se révèle

plus flexible et plus rapidement adaptable à l'apparition de nouveaux clusters technologiques. En effet, comme décrit dans [preschitschek2013], avant que deux technologies distinctes commencent à converger clairement, on peut s'attendre à ce que des termes similaires soient utilisés dans des brevets des deux technologies.

Enfin, une classification sémantique dans laquelle les brevets sont rassemblés en se basant sur le fait qu'ils partagent des mots-clés significatifs similaires a l'avantage d'inclure une caractéristique du réseau qui ne se retrouve pas dans le cas de l'USPC, c'est à dire que chaque brevet est associé à un vecteur de probabilités d'appartenir à chacune des classes sémantiques (plus de détails sur cette caractéristique sont donnés plus loin). En utilisant les co-occurrences des mots-clés, il est ensuite possible de construire un réseau de brevets et d'étudier l'influence de caractéristiques topologiques clés. Comme rappelé précédemment, l'utilisation des co-occurrences est la manière usuelle de construire un réseau sémantique. D'autres techniques hybrides comme les réseaux bipartites auteur/sémantique n'ont pas la caractéristique agréable de se baser uniquement sur l'information sémantique endogène contenue dans les données.

Construction de la classification sémantique

Dans cette section, nous décrivons la méthode et les analyses empiriques menant à la construction du réseau sémantique et la classification correspondante.

Extraction des mots-clés

Soit \mathcal{P} l'ensemble des brevets. Nous assignons pour commencer à un brevet $p \in \mathcal{P}$ un ensemble de mots-clés potentiellement significatifs $K(p)$ à partir de son texte $A(p)$ (qui correspond à la concaténation de son propre titre et de son résumé). Les éléments de $K(p)$ sont extraits selon une procédure similaire à celle détaillée dans [chavaliarias2013phylomemetic] :

1. Parsing du texte et mise en tokens : nous transformons les textes bruts en un ensemble de mots et phrases, qui sont lues (parsing) et séparés en entités élémentaires (mots organisés en phrases).
2. Attribution des tags *part-of-speech* : attribution d'une fonction grammaticale à chacun des tokens définis précédemment.
3. Extraction des racines (*stems*) : les familles de mots sont généralement dérivées d'une unique racine appelée *stem* (par exemple *compute*, *computer*, *computation* dérivent tous de la même racine *comput*) que nous extrayons des tokens. A ce point, le texte du résumé est réduit à un ensemble de *stems* et leur fonctions grammaticales.

4. Construction des multi-stems : Ce seront les unités sémantiques de base utilisées dans les analyses par la suite. Ils sont construits comme groupes de stems successifs dans une phrase, qui satisfont une règle de fonction grammaticale simple. La longueur du groupe doit être entre 1 et 3 et ses éléments être soit des noms, soit des verbes attributifs, soit des adjectifs. Nous choisissons d'extraire la sémantique à partir de tels groupes nominaux à la vue de la nature technique des textes, qui ont peu de probabilité de contenir des nuances subtiles dans les combinaisons entre verbes et groupes nominaux.

Les opérations d'analyse textuelle sont implémentées en python afin d'utiliser les fonctions intégrées à la bibliothèque nltk [nltk] pour la majorité des opérations ci-dessus. Cette bibliothèque supporte la plupart des opérations de traitement du langage naturel au niveau de l'état de l'art. Le code source est disponible de manière ouverte sur le dépôt du projet à <https://github.com/JusteRaimbault/PatentsMining>.

Estimation de la pertinence des mots-clés

ESTIMATION DE LA PERTINENCE Suivant l'heuristique de [chavalarias2013phylomem] nous estimons un score de pertinence pour filtrer les multi-stems. Le choix du nombre total de mots-clés extraits, que nous notons K_w est important, puisque des valeurs trop faibles donnent des structures de réseau similaires mais qui contiennent moins d'information, tandis que des très grandes valeurs tendent à inclure trop de mots-clés non pertinents. Nous choisissons de fixer ce paramètre à $K_w = 100,000$. Nous considérons pour commencer une filtration de $k \cdot K_w$ (avec $k = 4$) mots, pour garder un grand nombre de mots-clés potentiels mais calculer un nombre raisonnable de co-occurrences. Cette étape n'a que des effets marginaux sur la nature des mots-clés finaux mais est nécessaire pour des raisons computationnelles. La filtration est faite sur la mesure *unithood* u_i , définie pour le mot-clé i par $u_i = f_i \cdot \log(1 + l_i)$ où f_i est le nombre d'apparitions du multi-stem dans l'ensemble du corpus et l_i sa longueur en nombre de mots. Une seconde filtration de K_w mots-clés est faite sur la *termhood* t_i , dont la définition formelle est donnée en Eq. 25. Elle est calculée comme un score du chi-deux sur la distribution des co-occurrences du stem et comparée à une distribution uniforme sur l'ensemble du corpus. Intuitivement, les termes distribués uniformément seront identifiés comme du langage courant, et ne sont donc pas pertinents pour la classification. Plus précisément, nous calculons la matrice de co-occurrence (M_{ij}), où M_{ij} est défini comme le nombre de brevets où les stems i et j apparaissent simultanément. Le score de *termhood* t_i est alors défini par

$$t_i = \sum_{j \neq i} \frac{(M_{ij} - \sum_k M_{ik} \sum_k M_{jk})^2}{\sum_k M_{ik} \sum_k M_{jk}}. \quad (25)$$

ESTIMATION SUR FENÊTRE GLISSANTE Les scores précédents sont estimés sur une fenêtre glissante de longueur temporelle fixe, suivant l'idée que la pertinence présente est donnée par le contexte le plus récent et que l'influence diminue lorsqu'on remonte dans le passé. Par conséquent, la matrice de co-occurrence est construite pour l'année t en se restreignant aux brevets qui ont été soumis pendant la fenêtre temporelle $[t - T_0; t]$. Notons que la propriété causale de la fenêtre est essentielle puisque le futur ne peut jouer aucun rôle dans l'état courant des mots-clés et des brevets. De cette manière, nous obtiendrons des classes sémantiques exploitables sur une durée T_0 . Par exemple, cela nous permettra plus loin de calculer la modularité des classes dans le réseau de citations. Dans la suite, nous prenons $T_0 = 4$ (ce qui correspond à une fenêtre temporelle de 5 ans), en cohérence avec le choix du délai maximal pour les citations fait précédemment. Nous présentons l'analyse de sensibilité effectuée pour $T_0 = 2$ en Annexe de [[10.1371/journal.pone.0176310](#)].

Construction du réseau sémantique

Nous conservons l'ensemble des \mathcal{K}_W mots-clés les plus pertinents et obtenons leur matrice de co-occurrence comme défini précédemment. Cette matrice peut directement être interprétée comme une matrice d'adjacence pondérée du réseau sémantique. A cette étape, la topologie du réseau brut ne permet pas l'extraction de communautés claires. Cela est partiellement dû à la présence de hubs qui correspondent à des termes fréquents communs à de nombreux champs (par exemple `method`, `apparat`) qui sont considérés pertinents mais sont des faux positifs. Nous introduisons donc une mesure supplémentaire pour corriger la topologie du réseau : la concentration des mots-clés au sein des classes technologiques, définie par

$$c_{tech}(s) = \sum_{j=1}^{N^{(tec)}} \frac{k_j(s)^2}{(\sum_i k_i(s))^2},$$

où $k_j(s)$ est le nombre d'occurrences du mot-clé s dans la classe technologique j prise parmi les $N^{(tec)}$ classes USPC. Plus c_{tech} est grand, plus le noeud est spécifique à une classe technologique. Par exemple, le terme `semiconductor` est largement utilisé en électronique et ne contient pas d'information significative au regard de ce champ. Nous utilisons un paramètre de seuil, défini comme θ_c , et conservons les noeuds tels que $c_{tech}(s) > \theta_c$. De la même manière, les liens avec des poids faibles correspondent à des co-occurrences rares et sont

considérées comme du bruit. Pour prendre cela en compte, nous définissons le paramètre θ_w pour les liens, et nous filtrons les liens qui ont un poids en dessous de θ_w , suivant la logique que deux mots-clés ne sont pas connectés “par chance” s’ils apparaissent simultanément un nombre minimal de fois. Pour contrôler l’effet de taille, nous normalisons en prenant $\theta_w = \theta_w^{(0)} \cdot N_P$ où N_P est le nombre de brevets dans le corpus ($N_P = |\mathcal{P}|$). $\theta_w^{(0)}$ est ainsi un paramètre variable qui peut s’interpréter comme un seuil de bruit *par brevet*. Les communautés sont ensuite extraites par l’utilisation d’une procédure standard de maximisation de modularité comme décrite par [clauset2004finding], à laquelle nous ajoutons les deux contraintes capturées par θ_w et θ_c , c’est à dire que les liens doivent avoir un poids plus grand que θ_w et les noeuds une concentration plus grande que θ_c . A cette étape, les deux paramètres θ_c et $\theta_w^{(0)}$ ne sont pas contraints et leur choix n’est pas direct. En effet, différents objectifs d’optimisation sont possibles, comme la modularité, la taille du réseau ou le nombre de communautés. Nous trouvons que la modularité est maximisée à une valeur relativement stable de θ_w pour les différentes valeurs de θ_c pour chaque année, ce qui correspond à une valeur stable de $\theta_w^{(0)}$ au cours des années, et ce qui amène le choix $\theta_w^{(0)} = 4.1 \cdot 10^{-5}$. Ensuite, pour le choix de θ_c , différents points candidats se trouvent sur un front de Pareto pour l’optimisation bi-objectif sur le nombre de communautés et la taille du réseau. Il n’y a a priori pas de raisons de choisir un point spécifique entre les différents optima. Par conséquent, nous avons testé l’analyse pour toutes les valeurs candidates pour θ_c et trouvé que les résultats sont les plus raisonnables avec $\theta_c = 0.06$ (voir Fig. 128). Nous montrons en Fig. 129 un exemple de visualisation du réseau sémantique.

Characteristics of Semantic Classes

Pour chaque année t , nous définissons comme $N_t^{(sem)}$ le nombre de classes sémantiques qui ont été calculées par la détection de communautés pour les brevets soumis sur la période $[t - T_0, t]$ (nous rappelons que nous avons pris $T_0 = 4$). Chaque classe sémantique $k = 1, \dots, N_t^{(sem)}$ est caractérisée par un ensemble de mots-clés $K(k, t)$ qui est un sous-ensemble de \mathcal{K}_W sélectionné comme décrit dans les précédentes sections. La distribution des cardinaux de $K(k, t)$ pour l’ensemble des classes sémantiques k a une très longue queue avec quelques classes sémantiques contenant plus de 1000 mots-clés, la plupart avec approximativement le même nombre de mots-clés. A l’opposé, il existe de nombreuses classes qui ne contiennent que deux mots. Il y a en moyenne 30 mots-clés par classe sémantique et la médiane est 2 pour tout t . La Fig. 130 montre que le nombre moyen de mots-clés est relativement stable de 1976 à 1992 et ensuite présente un pic autour de 1996 avant de redescendre à nouveau.

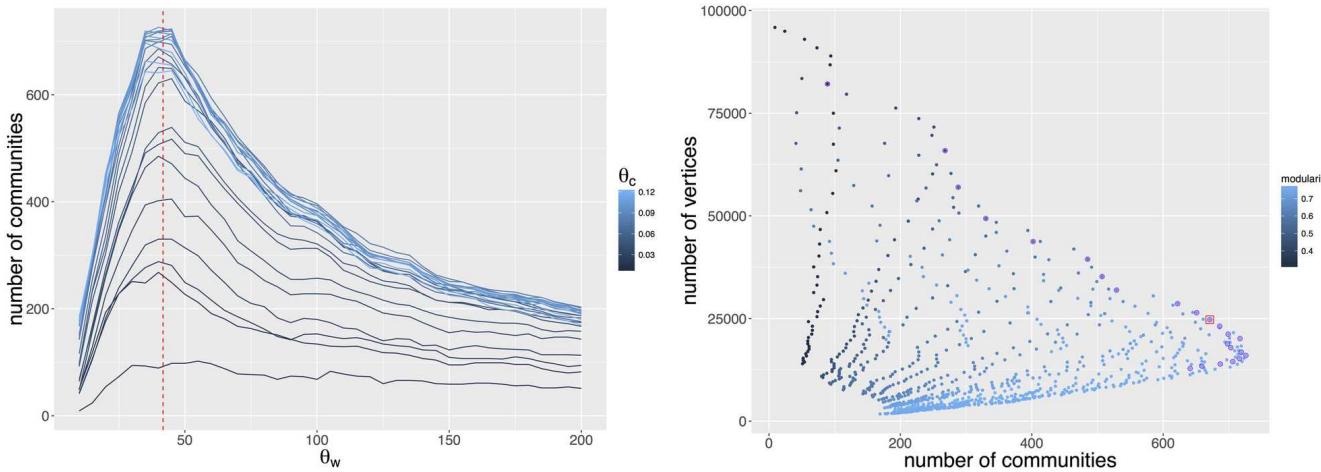


FIGURE 128: Analyse de sensibilité de la structure des communautés du réseau aux paramètres de filtrage. Nous considérons une fenêtre spécifique 2000-2004. Les graphes obtenus sont typiques. (*Gauche*) Le graphe donne le nombre de communautés en fonction du paramètre de seuil pour les liens θ_w pour différentes valeurs du paramètre de seuil des noeuds θ_c . Le maximum est globalement stable pour les différents θ_c (ligne rouge pointillée). (*Droite*) Pour choisir θ_c , nous faisons une optimisation de Pareto sur les communautés et la taille du réseau : le point de compromis (surligné en rouge) sur le front de Pareto (surligné en violet : les choix possibles après avoir fixé $\theta_w^{(0)}$; le niveau de bleu donne la modularité) correspond à $\theta_c = 0.06$.

TITRE DES CLASSES SÉMANTIQUES Les classes technologiques USPC sont définies par un titre et une définition très précise qui aide à retrouver les brevets facilement. Le titre peut être un simple mot (par exemple, la classe 101 : "Printing") ou plus compliqué (par exemple, classe 218 : "High-voltage switches with arc preventing or extinguishing devices"). Notre but étant de produire une base exhaustive dans laquelle chaque brevet est associé à un ensemble de classes sémantiques, il est nécessaire de donner un aperçu de ce que ces classes représentent en leur associant une courte description ou un titre comme dans [tseng2007text]. Dans notre cas, cette description est prise comme un sous-ensemble de mots-clés pris dans $K(k, t)$. Pour la grande majorité des classes sémantiques qui ont moins de 5 mots-clés, nous décidons de garder l'ensemble de ces mots-clés comme description. Pour les classes restantes qui ont en moyenne une cinquantaine de mots-clés, nous nous reposons sur les propriétés topologiques du réseau sémantique. [yang2000improving] suggère de ne garder que les termes les plus fréquents dans $K(k, t)$. Une autre possibilité est de sélectionner 5 mots-clés en se basant sur leur centralité dans le réseau, en suivant l'idée que les mots-clés très centraux sont les meilleurs candidats pour décrire le thème général capturé par une communauté. Par exemple, la plus grande classe sémantique en 2003-2007 est caractérisée par les mots-clés : Support Packet; Tree Network; Network Wide; Voic Stream; Code Symbol Reader.

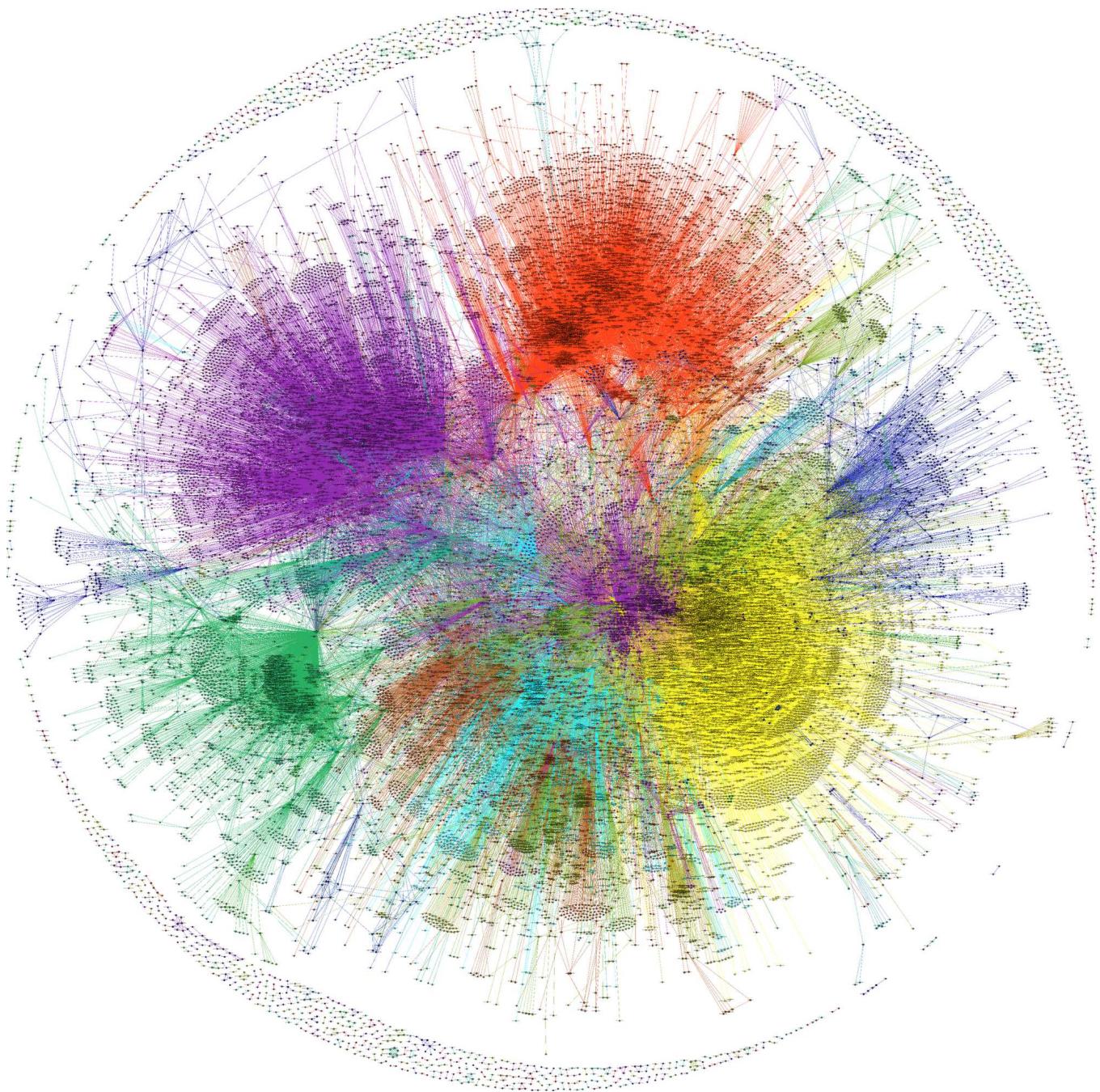


FIGURE 129: Un exemple de visualisation du réseau sémantique. Nous montrons le réseau obtenu pour la fenêtre 2000-2004, avec les paramètres $\theta_c = 0.06$ et $\theta_w = \theta_w^{(0)} \cdot N_p = 4.5e^{-5} \cdot 9.1e^5$. Le fichier correspondant sous un format vectoriel (.svg), qui peut être zoomé et exploré, est disponible en Annexe de [[10.1371/journal.pone.0176310](#)].

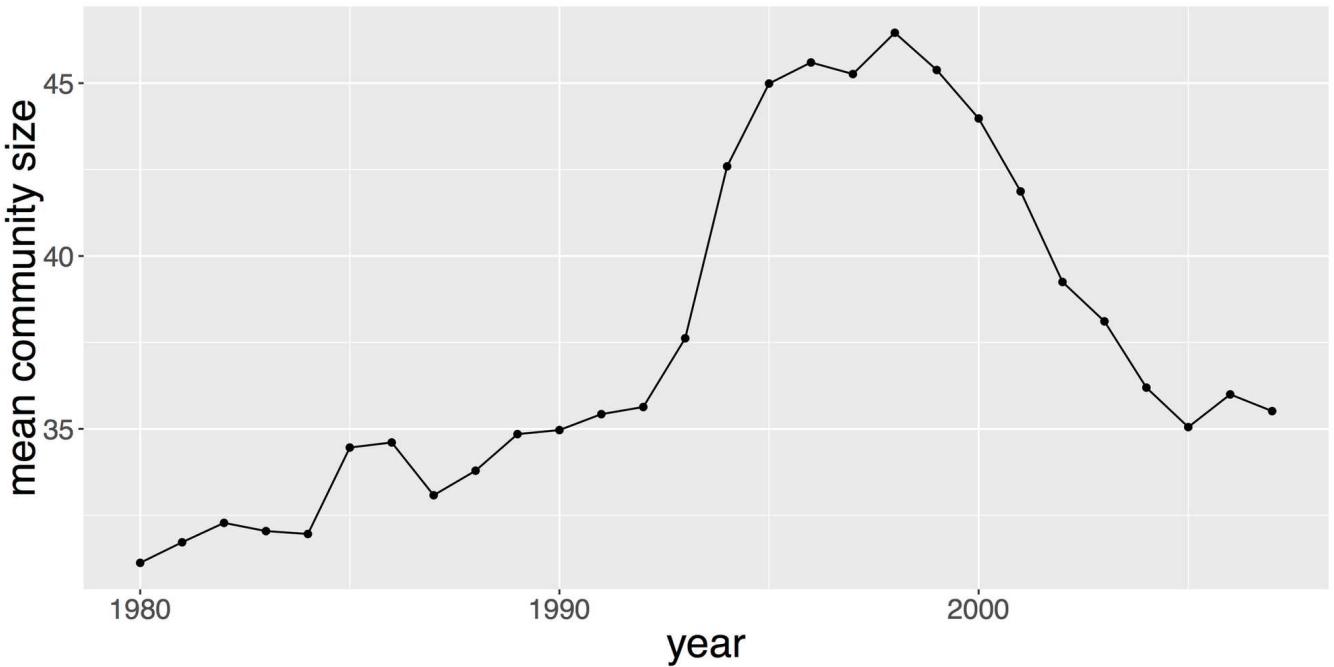


FIGURE 130: Cette figure montre le nombre moyen de mots-clés par classe sémantique pour chaque fenêtre $[t-4; t]$ de $t = 1980$ à $t = 2007$.

TAILLES DES CLASSES SÉMANTIQUES ET TECHNOLOGIQUES Nous considérons une fenêtre spécifique d’observations (par exemple 2000-2004), et nous définissons Z le nombre de brevets qui apparaissent dans cette fenêtre temporelle. Pour chaque brevet $i = 1, \dots, Z$, nous lui associons un vecteur de probabilités où chaque composante $p_{ij}^{(sem)} \in [0, 1]$, avec $j = 1, \dots, N(sem)$ et où

$$\sum_{j=1}^{N(sem)} p_{ij}^{(sem)} = 1$$

(lorsqu’il n’y a pas risque de confusion, nous oublions l’indice t dans $N_t^{(sem)}$). En moyenne sur l’ensemble des fenêtres temporelles, un brevet est associé à 1.8 classes sémantiques avec une probabilité strictement positive. Nous définissons alors la taille d’une classe sémantique comme

$$S_j^{(sem)} = \sum_{i=1}^Z p_{ij}^{(sem)}.$$

De manière correspondante, nous proposons d’introduire une définition cohérente pour les classes technologiques. Pour cela, nous suivons la méthode dite du “compte fractionnel”, qui a été introduite par l’USPTO et consiste en la division égale des brevets entre l’ensemble des classes auxquelles ils appartiennent. Formellement, nous définissons le nombre de classes technologiques comme $N^{(tec)}$ (qui

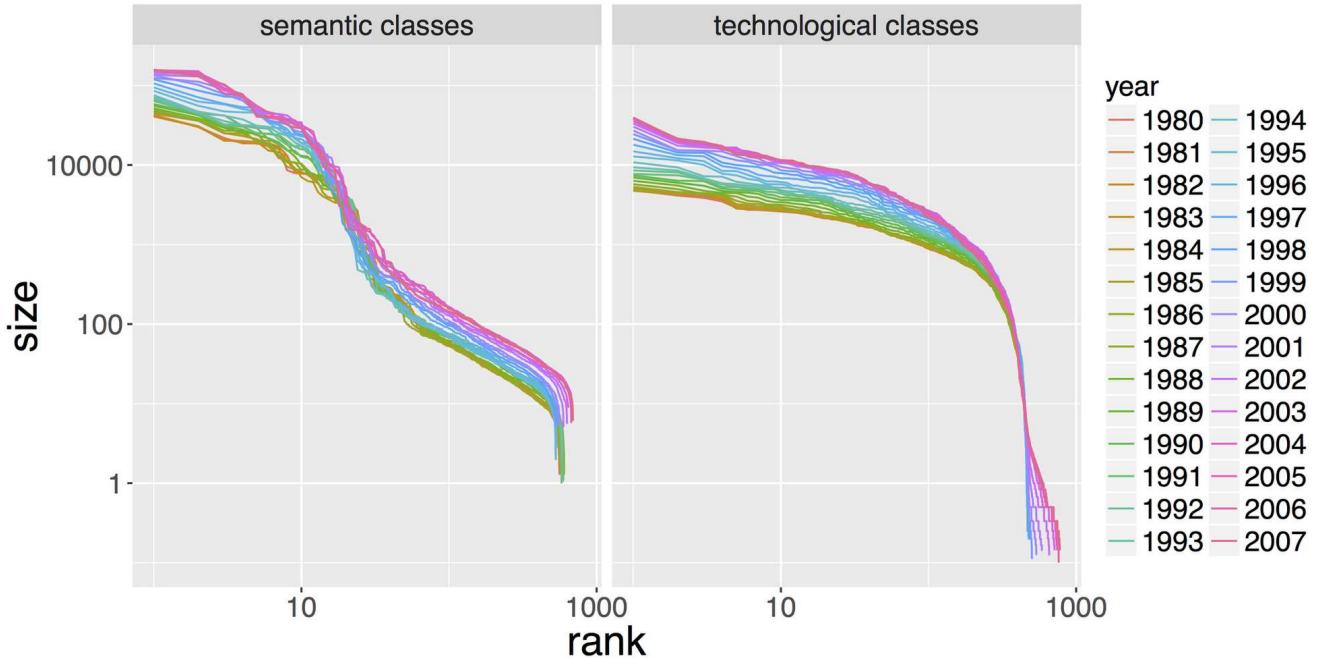


FIGURE 131: Tailles des classes. Pour chaque année de $t = 1980$ à $t = 2007$, nous montrons la taille des classes sémantiques (*Gauche*) et des classes technologiques (*Droite*) sur la fenêtre temporelle correspondante $[t - 4, t]$, de la plus grande à la plus petite. La définition formelle des tailles est donnée est section C.3. Chaque couleur correspond à une année donnée. Les classes sémantique annuelles et les classes technologiques présente une structure hiérarchique similaire, ce qui confirme la comparabilité des deux classifications. Dans le temps, les courbes sont translatées et le niveau de hiérarchie reste globalement constant.

ne dépend pas du temps contrairement au cas sémantique) et pour $j = 1, \dots, N^{(tec)}$ la matrice de probabilité correspondante définie par

$$p_{ij}^{(tec)} = \frac{B_{ij}}{\sum_{k=1}^{N^{(tec)}} B_{ik}},$$

où B_{ij} est égal à 1 si le i ème brevet appartient à la j ème classe technologique et 0 sinon. Quand il n'y a pas risque de confusion, nous oublions l'exposant et écrivons uniquement p_{ij} pour référer soit à la matrice sémantique soit à la matrice technologique. Empiriquement, nous obtenons que les deux classifications présentent une structure hiérarchique similaire, ayant toutes les deux une distribution de type loi puissance pour la taille des classes comme montré en Fig. ???. Cette caractéristique est importante, puisqu'elle suggère qu'une classification basée sur le contenu sémantique des brevets a un certain pouvoir de séparation au sens qu'elle ne divise pas les brevets en uniquement une ou deux communautés.

Extensions possibles à la méthode

Notre classification sémantique pourrait être améliorée en la combinant à d'autres techniques comme l'Allocation de Dirichlet Latente qui est une méthode répandue en détection de thèmes (voir par exemple [**blei2003latent**]), déjà utilisée sur les données de brevets comme dans [**kaplan2015double**] où elle fournit une mesure de la nouveauté d'une idée et le fait stylisé contre-intuitif que les innovations de rupture ont plus de chance d'émerger de recherches locales dans un champ plutôt que de recombinations de technologies distantes. L'utilisation de cette approche devrait dans un premier temps permettre d'évaluer plus précisément la robustesse de nos conclusions qualitatives (validation externe). Selon le niveau d'orthogonalité avec notre classification, cela pourrait également potentiellement constituer une caractéristique supplémentaire pour la caractérisation des brevets, dans l'esprit de la multimodélisation où les modèles voisins sont combinés pour tirer parti de l'ensemble des points de vue sur le système.

Notre utilisation d'analyse de réseau pourrait aussi être étendue en utilisant des techniques d'analyse par hyper-réseau récemment développées. En effet, les brevets et les mots clés peuvent par exemple être les noeuds d'un réseau bipartite, ou les brevets être les liens d'un hyper-réseau, au sens de multiples couches avec différents liens de classification et de citation. La combinaison de la modélisation du réseau de citations par de la Modélisation Stochastique par Blocs, à la modélisation des thèmes, a été étudiée dans le cas des articles scientifiques par [**zhu2013scalable**], ayant une performance supérieure au algorithmes de prédiction de liens précédents. [**iacovacci2015mesoscopic**] propose une méthode pour comparer les structures macroscopiques des différentes couches d'un réseau multi-couches, qui pourrait être appliquée comme un raffinement des analyses de l'intersection et de la modularité que nous faisons ici. De plus, il a récemment été montré que les mesures calculées sur des projections des réseaux multi-couches induisent une perte d'information non-négligeable en comparaison au mesures correspondantes généralisées [**de2015ranking**], ce qui confirme la pertinence de tels développements que nous laissons pour des recherches futures.

Un autre développement de recherche potentiel serait une exploitation approfondie de la nature temporelle du jeu de données. En effet, des progrès conséquents ont récemment été faits en analyse de réseau de données de séries temporelles (voir [**gao2017complex**] pour une revue). Par exemple, [**gao2015multiscale**] développe une méthode pour construire des réseaux multi-échelles à partir de séries temporelles, ce qui pourrait dans notre cas être une solution pour identifier des structures dans les trajectoires des brevets à différents niveaux, et être une alternative à l'analyse de modularité à une seule échelle que nous utilisons.

Résultats

Dans cette partie, nous présentons les caractéristiques principales de notre classification sémantique obtenue, qui présente à la fois une complémentarité et des différences avec la classification technologique. Nous présentons d'abord diverses mesures dérivées de cette classification sémantique au niveau du brevet : diversité, originalité, généralité, et recouvrement pour les classes. Nous montrons ensuite des les deux classifications présentent des mesures topologiques fondamentalement différentes.

Mesures pour les brevets

Etant donné un système de classification (classes sémantiques ou technologiques), et les probabilités associées p_{ij} pour chaque brevet i d'appartenir à la classe j (comme définies précédemment), on peut définir une mesure de diversité au niveau du brevet comme le complémentaire dans un de l'indice de concentration de Herfindhal sur les p_{ij} par

$$D_i^{(z)} = 1 - \sum_{j=1}^{N^{(z)}} p_{ij}^2, \text{ with } z \in \{\text{tec, sem}\}.$$

Nous montrons en Fig. 132 la distribution dans le temps de la diversité sémantique et technologique avec les séries temporelles correspondantes pour les moyennes. L'analyse est faite pour deux configurations différentes, c'est à dire en incluant ou non les brevets avec diversité nulle (les brevets n'appartenant qu'à une seule classe). Nous désignons les autres brevets comme "brevets compliqués" par la suite. Tout d'abord, la présence de masse dans les faibles probabilités pour la diversité sémantique mais non celle technologique confirme que la classification sémantique contient des brevet qui recouvrent un grand nombre de classes. D'autre part, une décroissance générale de la diversité pour les brevets compliqués, à la fois pour les systèmes de classification sémantique et technologique, peut être interprété comme une augmentation de la spécialisation des inventions. Il s'agit d'un fait stylisé bien connu comme documenté dans [ARCHIBUGI199279]. De plus, un changement de régime qualitatif sur la classification s'opère autour de 1996. Celui-ci s'observe que l'on inclue ou non les brevets à diversité nulle. La diversité des brevets compliqués se stabilise après une décroissance constante, et la diversité globale commence à décroître fortement. Cela signifie d'une part que le nombre de brevet n'ayant qu'une seule classe commence à augmenter et d'autre part les brevets compliqués ne changent pas en diversité. Ce phénomène peut s'interpréter comme un changement dans le régime de spécialisation, le nouveau régime étant causé par une augmentation du nombre de brevets à classe unique.

Plus classiques dans la littérature sont les mesures d'originalité et de généralité. Ces mesures correspondent à la même idée que la diversité définie précédemment pour quantifier la diversité des classes (qu'elles soient technologiques ou sémantiques) associées à un brevet. Mais plutôt que de considérer les classes du brevet, celles-ci considèrent les classes des brevets qui sont cités ou citants. Formellement, l'originalité $O_i^{(z)}$ et la généralité $G_i^{(z)}$ d'un brevet i sont définis par

$$O_i^{(z)} = 1 - \sum_{j=1}^{N^{(z)}} \left(\frac{\sum_{i' \in I_i} p_{i'j}}{\sum_{k=1}^{N^{(z)}} \sum_{i' \in I_i} p_{i'k}} \right)^2 \quad \text{and} \quad G_i^{(z)} = 1 - \sum_{j=1}^{N^{(z)}} \left(\frac{\sum_{i' \in \tilde{I}_i} p_{i'j}}{\sum_{k=1}^{N^{(z)}} \sum_{i' \in \tilde{I}_i} p_{i'k}} \right)^2,$$

où $z \in \{\text{tec}, \text{sem}\}$, I_i dénote l'ensemble des brevets qui sont cités par le i ème brevet sur une fenêtre temporelle de cinq ans (i.e. si le i ème brevet apparaît en année t , alors nous considérons les brevets sur $[t - T_0, t]$) pour calculer l'originalité et \tilde{I}_i l'ensemble des brevets qui citent le brevet i après moins de cinq ans (i.e. nous considérons les brevets sur $[t, t + T_0]$) dans le cas de la généralité. Il est important de noter que la mesure de généralité est anticipative au sens où $G_i^{(z)}$ utilise l'information qui ne sera disponible que 5 ans après la soumission du brevet. Les deux mesures sont inférieures à la moyenne lorsqu'on se base sur la classification sémantique par rapport à la classification technologique. Fig. 133 donne les valeurs moyennes de $O_i^{(\text{sem})}$, $O_i^{(\text{tec})}$, $G_i^{(\text{sem})}$ and $G_i^{(\text{tec})}$.

Intersection des classes

Une mesure de proximité entre deux classes peut être définie comme leur recouvrement en nombre de brevets. De telles mesures peuvent par exemple être utilisées pour construire une métrique entre les classes sémantique. Intuitivement, des classes qui ont un très fort recouvrement seront très proche en terme de contenu technologique et on peut alors les utiliser pour mesurer la distance entre entreprises en terme de technologie développée comme fait dans [Bloom2005distance]. Formellement, en rappelant la définition de (p_{ij}) comme la probabilité du i ème brevet d'appartenir à la j ème classe et N_p le nombre de brevets, le recouvrement est donné par

$$\text{Overlap}_{jk} = \frac{1}{N_p} \cdot \sum_{i=1}^{N_p} p_{ij} p_{ik}. \quad (26)$$

Le recouvrement est normalisé par le nombre de brevets pour prendre en compte l'effet de la taille du corpus : par convention, nous supposons que le recouvrement est maximal quand il y a une unique classe

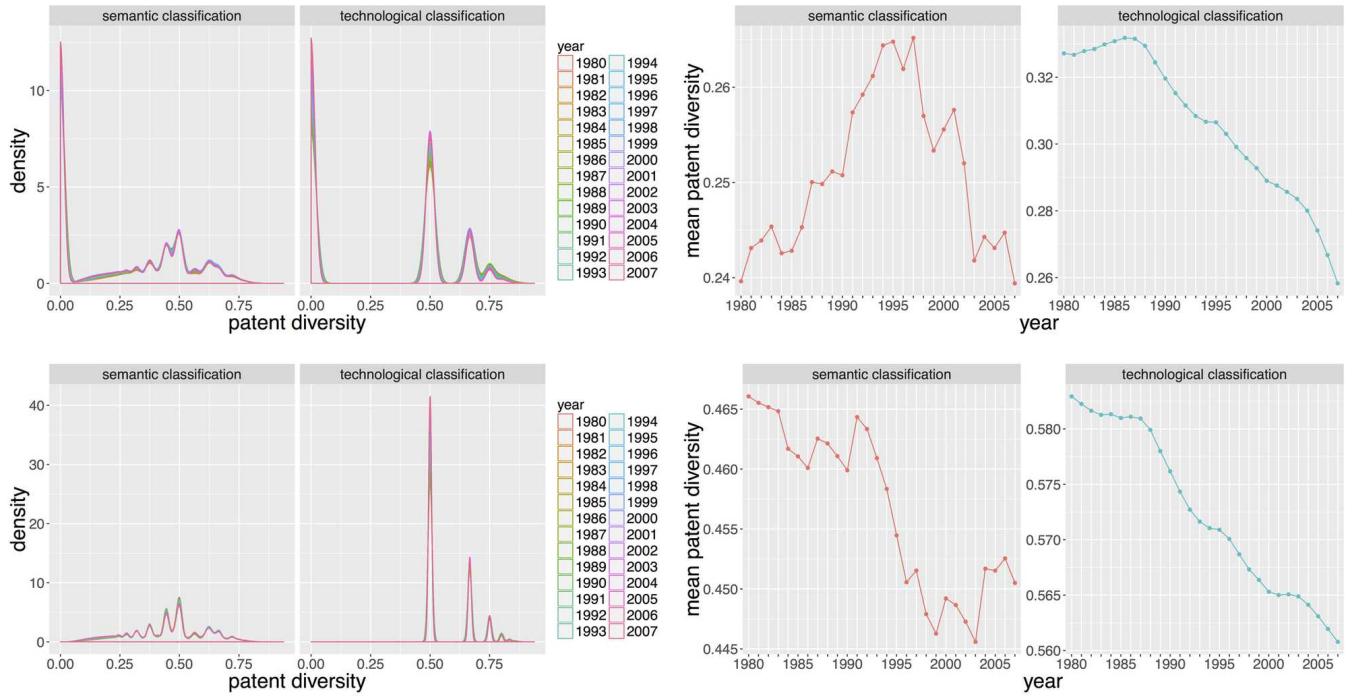


FIGURE 132: Diversité au niveau du brevet. Distribution des diversités (Colonne de gauche) et séries temporelles correspondantes pour leur moyenne (Colonne de droite) pour $t = 1980$ à $t = 2007$ (avec la fenêtre temporelle correspondante $[t - 4, t]$). La première ligne inclut l'ensemble des brevets classifiés, tandis que la deuxième ligne inclut uniquement les brevets avec plus d'une classe (c'est à dire les brevets avec une diversité supérieure à 0).

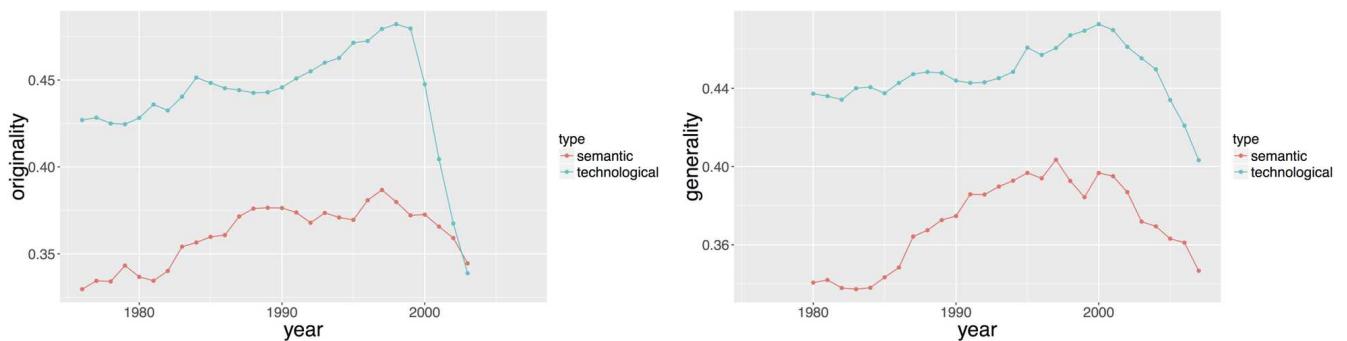


FIGURE 133: Originalité (Gauche) et généralité (Droite) au niveau du brevet de $t = 1980$ à $t = 2007$ (avec la fenêtre temporelle correspondante $[t - 4, t]$) comme défini ci-dessus.

dans le corpus. Un recouvrement relatif correspondant est calculé comme une mesure de similarité d'ensemble sur le nombre de brevets communs à deux classes A et B, donné par $\text{o}(A, B) = 2 \cdot \frac{|A \cap B|}{|A| + |B|}$

RECOUVREMENT INTRA-CLASSIFICATIONS L'étude des distributions des recouvrements à l'intérieur de chaque classification, c'est à dire entre les classes technologiques et les classes sémantiques séparément, révèle des différences structurelles entre les deux méthodes de classification, suggérant leur nature complémentaire. Leur évolution dans le temps donne de plus des indications sur les tendances de spécialisation. Nous montrons en Fig. 134 les distributions et les séries temporelles moyennes des recouvrements pour les deux classifications. La classification technologique suit globalement toujours une tendance décroissante, correspondant à des classes de plus en plus isolées, i.e. des inventions spécialisées, confirmant le fait stylisé obtenu précédemment. Pour les classes sémantiques, la dynamique est en quelque sorte plus intrigante et soutient l'hypothèse d'un changement de régime qualitatif suggérée précédemment. Même si celle-ci décroît globalement comme pour le recouvrement technologique, le recouvrement normalisé (respectivement relatif) moyen présente un maximum (plus clair pour le recouvrement normalisé) correspondant à l'année 1996 (resp. 1999). En étudiant les recouvrements normalisés, on constate que la structure de classification a été relativement stable jusqu'en 1990, puis a fortement augmenté pour culminer en 1996 puis décroître à une allure similaire jusqu'à aujourd'hui. Les technologies ont commencé par partager de plus en plus jusqu'à un point de rupture quand une isolation croissante est devenue à nouveau la règle. Une perspective évolutionnaire sur l'évolution technologique [ziman2003technological] pourrait éclaircir sur de possibles interprétations de ce changement de régime : quand une espèce évolue, l'environnement de fitness aurait d'abord été localement favorable à des inséminations réciproques, jusqu'à ce que chaque fitness dépasse un seuil au dessus duquel l'auto-spécialisation devient le chemin optimal. Ce phénomène est très comparable à l'établissement d'une niche écologique [holland2012signals], la forte interdépendance qui a son origine ici dans l'insémination mutuelle, résultant dans une situation finale très fortement dépendante au chemin.

CORRESPONDANCE ENTRE LES CLASSIFICATIONS Les recouvrements *entre* les classifications sont définis comme précédemment, mais avec j désignant la j ème classe technologique et k la k ème classe sémantique : p_{ij} sont les probabilités technologiques et p_{ik} les probabilités sémantiques. Ils décrivent la correspondance relative entre les deux classifications et sont un bon indicateur pour détecter des changements relatifs, comme montré en Fig. 135. Le recouvrement inter-classifications moyen présente clairement deux tendances linéaires,

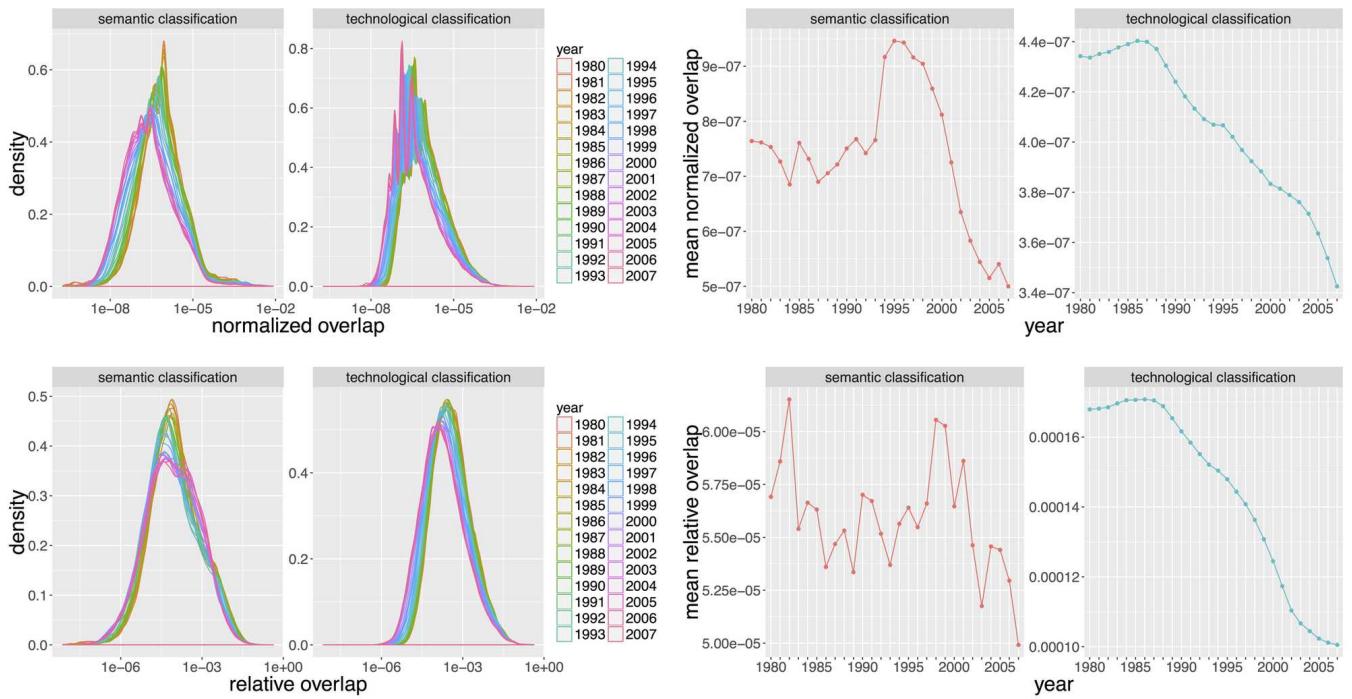


FIGURE 134: Recouvrements intra-classification. (Colonne de gauche) Distribution des recouvrements O_{ij} pour tous $i \neq j$ (les valeurs nulles sont supprimé par l'échelle logarithmique). (Colonne de droite) Séries temporelles moyennes correspondantes. (Première ligne) Recouvrements normalisés. (Deuxième ligne) Recouvrements relatifs.

la première constante de 1980 à 1996, suivie par une décroissance constante. Même s'il est difficile à interpréter directement, ce fait stylisé dévoile clairement un changement dans la *nature* des inventions, ou au moins dans la relations entre le contenu des inventions et la classification technologique. Comme le point de rupture est à la même date que ceux observés précédemment et comme les indicateurs sont différents, il est peu probable qu'il s'agisse d'une pure coïncidence. Ainsi, ces observations pourraient être des marqueurs d'un changement structurel sous-jacent caché des processus.

Modularité de citation

Une source exogène d'information concernant la pertinence des classifications est le réseau de citations décrit précédemment. La correspondance entre les liens de citation et les classes devrait fournir une mesure de la précision des classifications, au sens d'une validation externe puisqu'il est bien connu que l'homophilie de citation doit avoir de fortes valeurs (voir par exemple [AAKnetwork2016]). Cette section étudie empiriquement les modularités du réseau de citation pour les différentes classifications. La modularité est une mesure simple de la manière dont la partition des noeuds d'un réseau correspond plus ou moins bien à de plus fréquentes connexions internes aux commu-

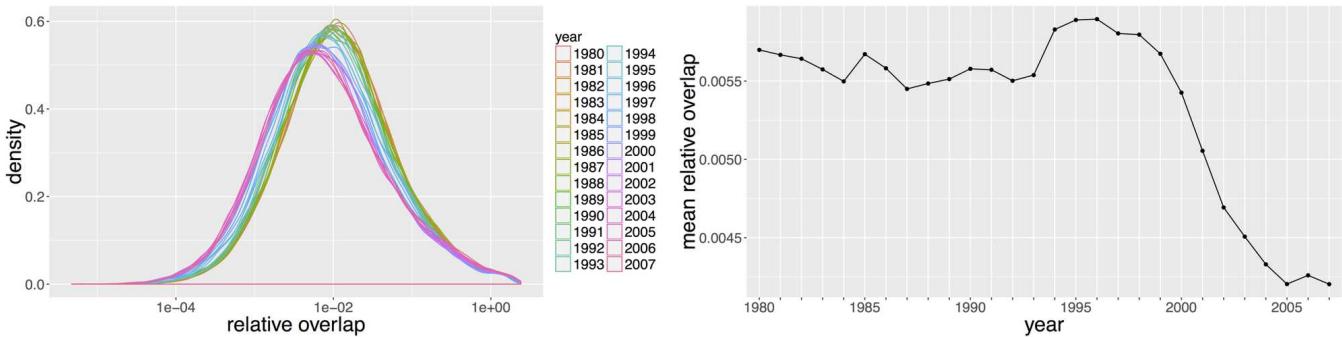


FIGURE 135: Distribution des recouvrements relatifs entre les classifications. (Gauche) Distributions des recouvrements à chaque date; (Droite) Série temporelle moyenne correspondante. La tendance décroissante commençant autour de 1996 confirme un changement qualitatif de régime à cette période.

nautés (voir [clauset2004finding] pour la définition formelle). Bien qu'initialement conçue pour des classifications univoques, cette mesure peut être étendue au cas où les noeuds peuvent appartenir simultanément à plusieurs classes, dans notre cas avec différentes probabilités comme introduit dans [nicosia2009extending]. La modularité dirigée simple est donnée dans notre cas par

$$Q_d^{(z)} = \frac{1}{N_p} \sum_{1 \leq i, j \leq N_p} \left[A_{ij} - \frac{k_i^{in} k_j^{out}}{N_p} \right] \delta(c_i, c_j),$$

avec A_{ij} la matrice d'adjacence du réseau de citations (i.e. $A_{ij} = 1$ s'il existe une citation du i ème brevet vers le j ème brevet, et $A_{ij} = 0$ sinon), $k_i^{in} = |I_i|$ (resp. $k_i^{out} = |I_i|$) degré entrant (resp. degré sortant) des brevets (i.e. le nombre de citations faites par le i ème brevet à des autres et le nombre de citations reçues par le i ème brevet). Q_d peut être définie pour chacun des systèmes de classification : $z \in \{\text{tec}, \text{sem}\}$. Si $z = \text{tec}$, c_i est défini comme la classe principale du brevet, qui est prise comme la première classe tandis que si $z = \text{sem}$, c_i est la classe avec la probabilité la plus forte.

La modularité multi-classes est quant à elle donnée par

$$Q_{ov}^{(z)} = \frac{1}{N_p} \sum_{c=1}^{N^{(z)}} \sum_{1 \leq i, j \leq N_p} \left[F(p_{ic}, p_{jc}) A_{ij} - \frac{\beta_{i,c}^{out} k_i^{out} \beta_{j,c}^{in} k_j^{in}}{N_p} \right],$$

où

$$\beta_{i,c}^{out} = \frac{1}{N_p} \sum_j F(p_{ic}, p_{jc}) \text{ and } \beta_{j,c}^{in} = \frac{1}{N_p} \sum_i F(p_{ic}, p_{jc}).$$

Nous prenons $F(p_{ic}, p_{jc}) = p_{ic} \cdot p_{jc}$ comme suggéré par [nicosia2009extending]. La modularité est une mesure agrégée de la façon dont le réseau

dévie d'un modèle nul où les liens sont attribués de manière aléatoire en respectant les degrés. En d'autres termes elle capture la tendance qu'on les liens d'être à l'intérieur des classes. La modularité recouvrante étend naturellement la modularité simple en prenant en compte le fait que les noeuds peuvent appartenir simultanément à plusieurs classes.

Nous donnons en Fig. 136 à la fois les modularités simple et multi-classes dans le temps. Pour la modularité simple, $Q_d^{(tec)}$ est bas et stable dans le temps tandis que $Q_d^{(sem)}$ est légèrement supérieure et s'accroît. Ces valeurs sont cependant faibles et suggèrent que les classes uniques ne sont pas suffisantes pour capturer l'homophilie de citation. Les modularités multi-classes donnent des résultats différents. Tout d'abord, les modularités pour les deux classifications ont une claire tendance croissante, signifiant qu'elles deviennent de plus en plus adéquates au réseau de citation. Les spécialisations dévoilées à la fois par les diversités au niveau du brevet et les recouvrements des classes sont une explication potentielle pour l'accroissement de ces modularités. Ensuite, la modularité sémantique est plus grande que la modularité séquentielle par un ordre de grandeur (par exemple 0.0094 pour la technologique contre 0.0853 pour la séquentielle en 2007) à chaque date. Cette différence a une forte signification qualitative. Notre classification séquentielle correspond mieux au réseau de citations avec des classes multiples. Comme les technologies peuvent être comprises comme une combinaison de différentes composantes comme montré par [Youn:2015fk], cette nature hétérogène est probablement mieux prise en compte par notre classification séquentielle multi-classes.

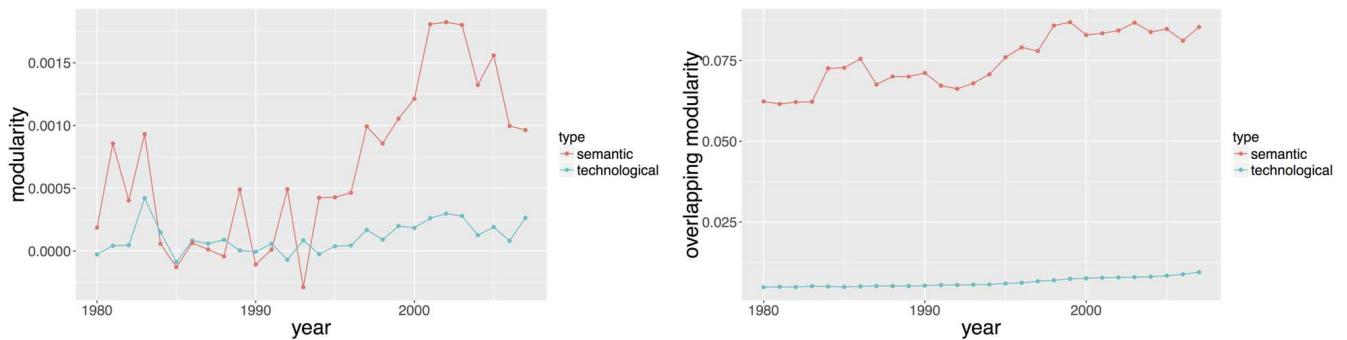


FIGURE 136: Evolution temporelle des modularités séquentielle et technologique du réseau de citation. (Gauche) Modularité dirigée simple, calculée avec les classes principales des brevets (classe technologique principale et classe séquentielle avec la plus grande probabilité). (Droite) Modularité multi-classes, calculée selon [nicosia2009extending].

Perspectives

La contribution principale de cette étude repose sur deux points. Tout d'abord nous avons précisé comment construire un réseau de brevets produisant une classification qui utilise l'information sémantique des résumés. Nous avons montré que cette classification partage des similarités avec la classification technologique traditionnelle, mais qu'elles ont aussi des caractéristiques distinctes. Deuxièmement, nous fournissons de manière ouverte les données résultantes de notre analyse, qui incluent : (i) une base de données reliant chaque brevet avec son ensemble de classes sémantiques et les probabilités associées ; (ii) une liste de ces classes sémantiques avec une description basée sur les mots clés les plus pertinents ; (iii) une liste des brevets avec leurs propriétés topologiques dans le réseau sémantique (centralité, fréquence, degré, etc.). La disponibilité de ces données suggère de nouvelles directions pour la recherche future. La jointure de notre base avec des bases ouvertes existantes peut mener à des développements potentiellement prometteurs. Par exemple, son utilisation simultanée avec la base des inventeurs désambiguisee fournie par [li2014disambiguation] pourrait être une manière d'étudier les profils sémantiques des inventeurs, ou de villes puisque les adresses des inventeurs sont fournies. L'étude de la diffusion spatiale de l'innovation entre les villes, qui est une composante cruciale de la Théorie Evolutive des Villes de PUMAIN [pumain2010theorie], serait rendue possible.

Une première application potentielle est l'utilisation des mesures topologiques des brevets qu'ils héritent de leurs mots-clés pertinents. La fait que ces mesures remontent dans le passé et sont immédiatement disponibles après la publication de l'information du brevet est un point important. Il serait par exemple très intéressant de tester leur pouvoir prédictif pour juger la qualité d'une innovation, en utilisant le nombre de citations reçues par un brevet, et par conséquent l'effet futur sur la valeur de marché de l'entreprise.

Au sujet de la stratégie d'innovation des entreprises, une deuxième extension serait d'étudier les trajectoires des entreprises dans les deux réseaux : technologique et sémantique. La combinaison de ces informations avec des données sur la valeur de marché des entreprises peut apporter beaucoup d'information sur les stratégies d'innovation les plus efficaces, sur l'importance de la convergence des technologies ou sur l'acquisition de petites entreprises innovantes. Cela permettra aussi d'observer les motifs d'innovation sur l'ensemble du cycle de vie d'une entreprise et comment ceux-ci diffèrent selon les champs technologiques.

Une dernière extension serait d'étudier plus en profondeur l'histoire de l'innovation. Les données des brevets USPTO ont été digitalisées depuis le premier brevet en juillet 1790. Cependant, ils ne

contiennent pas tous un texte qui est directement exploitable. Nous considérons que la qualité des images des brevets est assez fiable pour permettre l'utilisation de techniques d'*Optical Character Recognition* pour récupérer les textes jusqu'au moins 1920. Avec de telles données, il serait possible d'étendre notre analyse en remontant plus loin dans le temps et d'étudier comment le progrès technologique se produit et se combine dans le temps. [akcigit2013mechanics] procède à un travail similaire en étudiant les recombinaisons et l'apparition des sous-classes technologiques. En se basant sur le fait que les communautés sont construites chaque année, on peut construire une mesure de proximité entre deux classes successives. Cela pourrait fournir une vision plus claire sur la manière dont des technologies ont convergé dans le temps et quand d'autres sont devenues obsolètes et remplacées par des nouvelles méthodes.

C.4 COMMUNICATION SCIENTIFIQUE PAR LA GAMIFICATION

La question de la communication scientifique, notamment entre les agents producteurs de connaissance, a été un thème récurrent de notre travail. Celle-ci intervient également dans le contact avec le public comme médiation scientifique, et le développement de celui-ci peut en retour informer les entreprises d'interdisciplinarité. Nous développons ici deux modèles sous forme de jeux, ayant un objectif similaire de transmettre des concepts d'éologie d'eau douce. Cela renforce l'idée du modèle comme instrument crucial de la médiation scientifique.

* * *

*

Cette section est le fruit d'une collaboration interdisciplinaire avec l'éco-toxicologue DR. HÉLÈNE SERRA (Université de Bordeaux et Ineris) et a été présentée à la conférence SETAC 2016 comme [serra:halshs-01322860].

* * *

*

C.4.1 *Introduction*

L'attente de prise de conscience et d'implication pour le public concernant les questions environnementales est croissante. Toutefois, une connaissance experte est souvent nécessaire pour comprendre le enjeux sous-jacents à la plupart de ces problèmes. L'un des défis de la science aujourd'hui réside dans le fait d'expliquer des questions complexes de façon simple et compréhensible à une audience non-spécialisée. Les jeux apparaissent comme un medium pertinent pour la vulgarisation scientifique. En effet, la première forme d'apprentissage est en général par le jeu. Les jeux sont très populaires et présentent divers avantages d'un point de vue éducatif. Ceux-ci sont dynamiques et interactifs. Ainsi, l'engagement du joueur est augmenté, ainsi que sa rétention de connaissances. De plus, le joueur est immergé dans un monde nouveau et découvre un environnement virtuel où il doit développer des stratégies et identifier les processus fondamentaux. Ces caractéristiques peuvent être aisément utili-

sées pour transmettre des concepts scientifiques, et la gamification a déjà été proposée comme un outil pour une meilleure propagation de la pensée scientifique [**morris2013gaming**] comme en pharmacologie [**cain2015serious**] ou les géosciences [**reynard2015application**]. Dans ce contexte, ce projet vise à développer des outils basés sur les jeux pour transmettre des concepts basiques en écologie d'eau douce. Nous nous intéressons à un jeu de plateau classique et à un jeu informatique car ceux-ci sont complémentaire dans l'audience visée (joueurs en groupe et joueurs en ligne) et dans les possibilités offertes, en particulier concernant les interactions entre joueurs et les dynamiques du système.

c.4.2 Méthodologie

La méthodologie pour la conception des deux types de jeux est divisée de manière similaire en 5 étapes : (1) sélection des espèces ; (2) définition des instructions (objets, environnement du jeu, règles) ; (3) inclusion des stress environnementaux (biotiques et abiotiques) ; (4) conception et construction des interfaces (plateau et implémentation informatique) ; (5) test avec des joueurs. L'ensemble des étapes sont interdépendantes et sont menées en parallèle pendant le développement des jeux.

Tandis que le jeu de plateau est inspiré d'expériences de joueurs, le jeu informatique se base sur un modèle de simulation de l'écosystème. De manière à introduire les notions d'équilibre et ses perturbations qui surviennent à une échelle de temps plus longue que celle du jeu de plateau, nous proposons d'implémenter un modèle basé-agent (ABM) et de coupler sa dynamique avec des actions de jeu. Les ABM sont déjà largement utilisés en écologie [**grimm2005pattern**]. Ainsi, nous choisissons un modèle dynamique de chaîne trophique (modèle proie-prédateur étendu) qui est capable d'inclure des règles comportementales pour les poissons et un environnement spatial hétérogène. Un tel modèle est particulièrement adapté pour l'implémentation du jeu : les comportements des poissons sont influencés par les joueurs tandis que l'écosystème est perturbé par des événements extérieurs.

c.4.3 Résultats

Les deux jeux sont basés sur les mêmes règles générales, même si des adaptations sont nécessaires selon le type. L'objectif du jeu est de garantir la stabilité d'une communauté écologique dans un lac.

Jeu de plateau

Jeu pour ordinateur

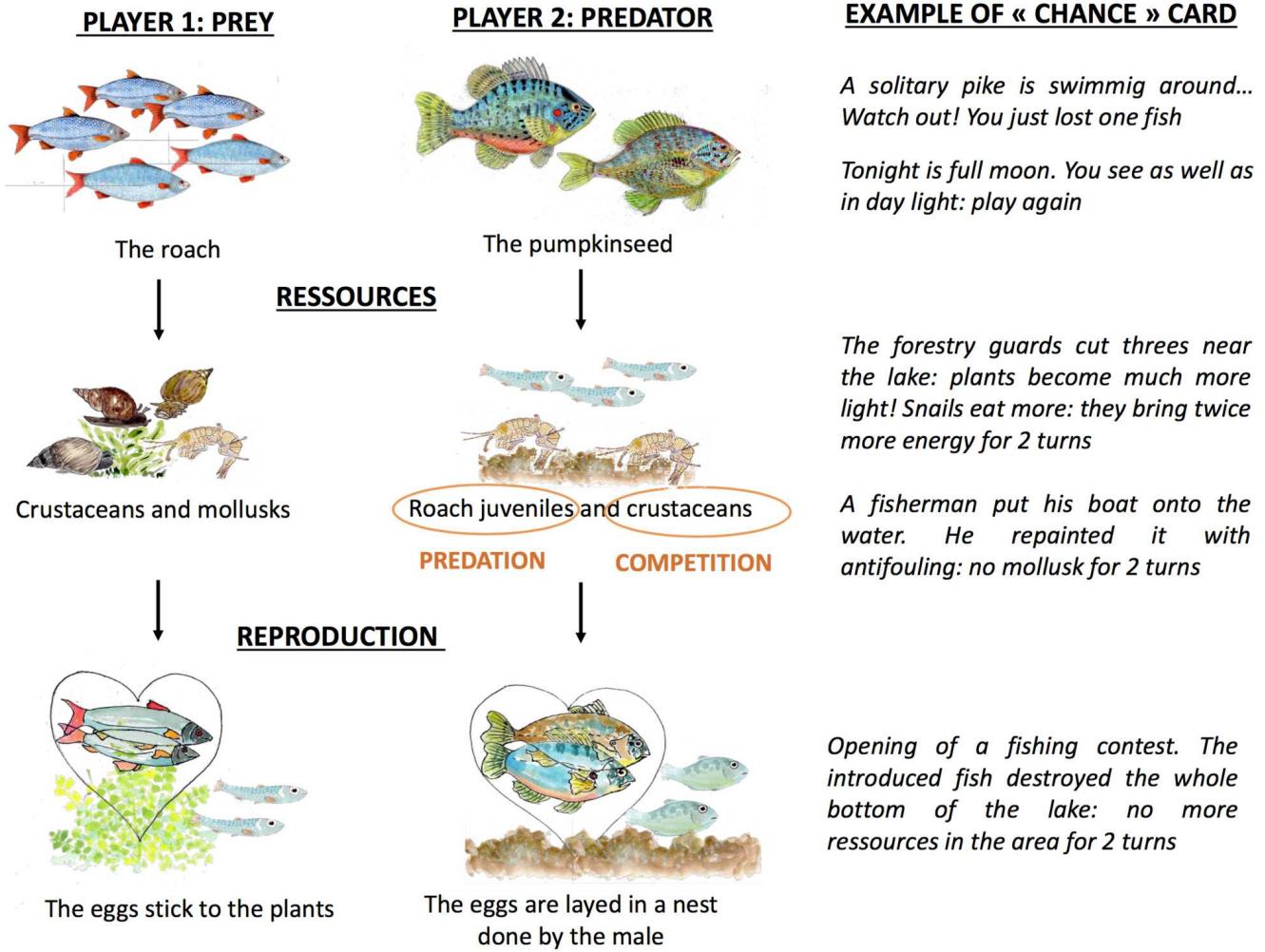


FIGURE 137: Principes du jeu de plateau.

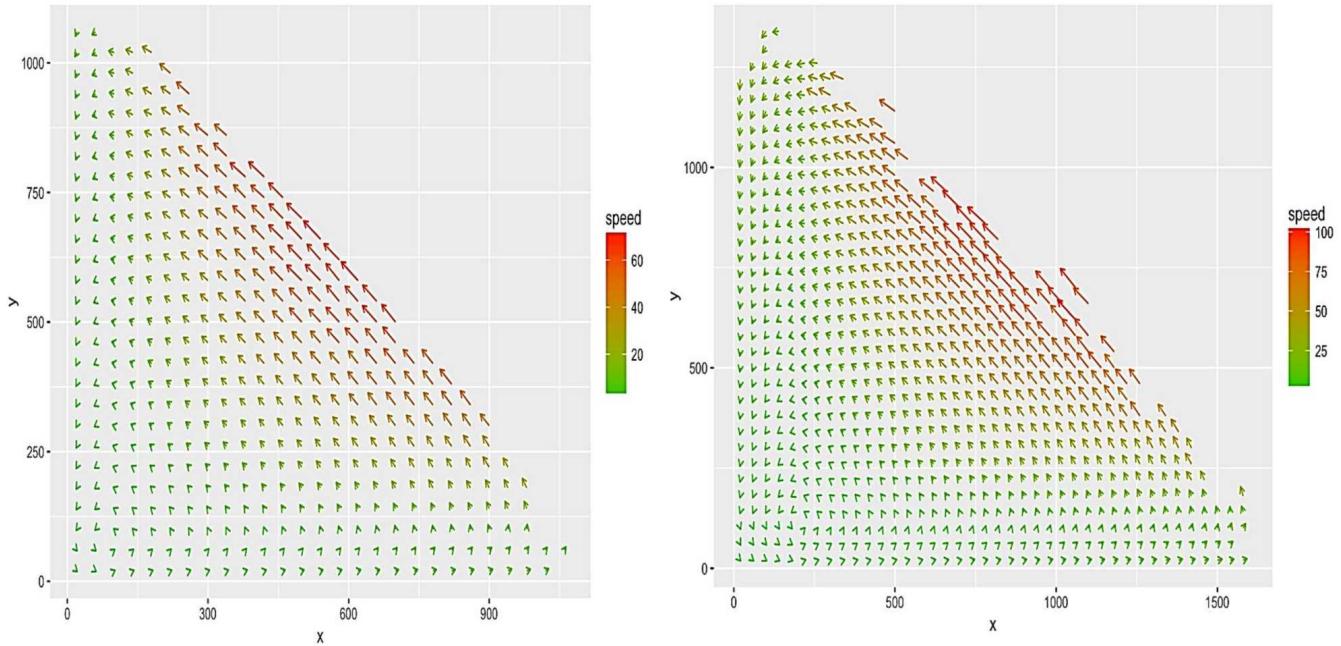


FIGURE 138: Exemples de diagrammes de phase du modèle proie-prédateur.

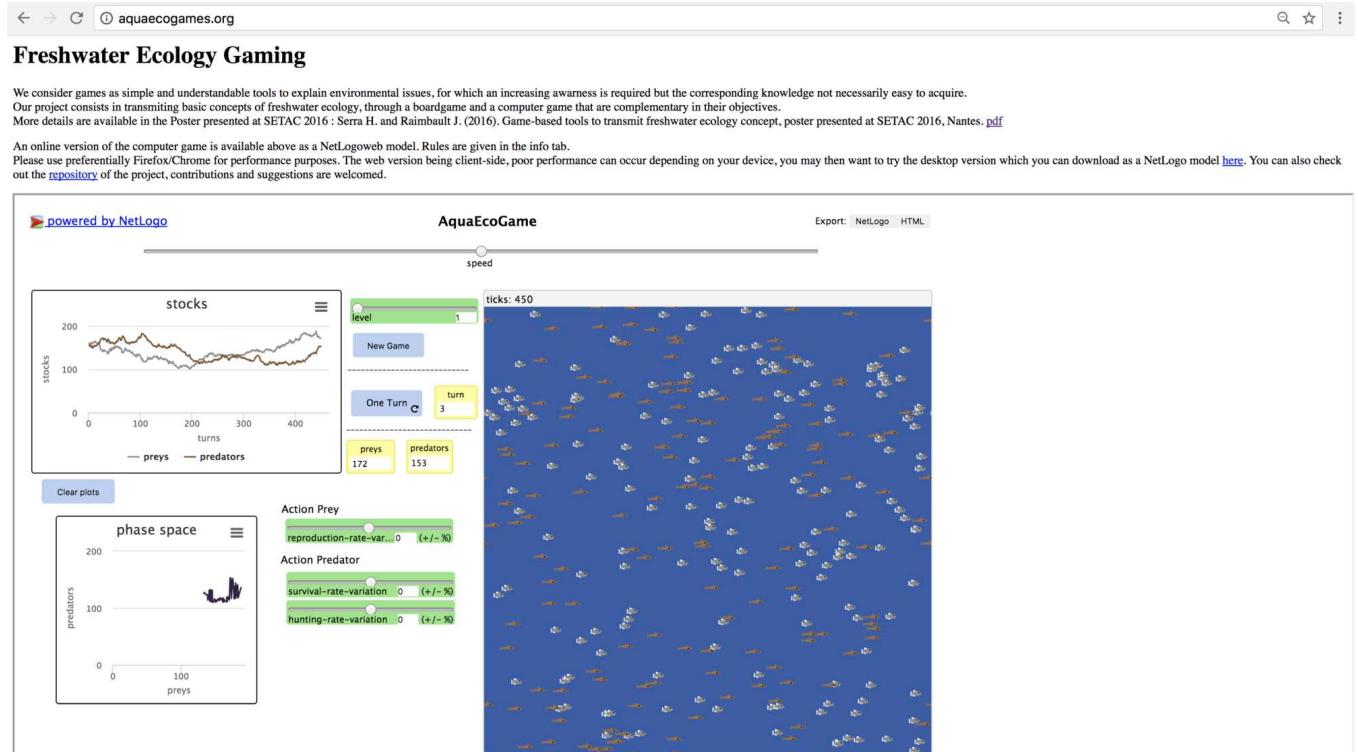


FIGURE 139: Capture de l'application web qui implémente le jeu informatique.

c.4.4 *Discussion*

C.5 DONNÉES SYNTHÉTIQUES CORRÉLÉES : SÉRIES TEMPORELLES FINANCIÈRES

Contexte

Un domaine d'application proposé pour la méthode de données synthétiques présentée en [B.3](#) est celui des séries temporelles financières, signaux typiques de systèmes complexes hétérogènes et multiscalaires [[mantegna2000introduction](#)] et pour lesquels les corrélations ont fait l'objet d'abondants travaux. Ainsi, l'application de la théorie des matrices aléatoires peut permettre de débruiter, ou du moins d'estimer la part de signal noyée dans le bruit, une matrice de corrélations pour un grand nombre d'actifs échantillonnés à faible fréquence (retours journaliers par exemple) [[2009arXiv0910.1205B](#)]. De même, l'analyse de réseaux complexes construits à partir des corrélations, selon des méthodes type arbre couvrant minimal [[2001PhyA..299...16B](#)] ou des extensions raffinées pour cette application précise [[tumminello2005tool](#)], ont permis d'obtenir des résultats prometteurs, tels la reconstruction de la structure économique des secteurs d'activités. A haute fréquence, l'estimation précise de paramètres d'interdépendance dans le cadre d'hypothèses fixées sur la dynamique, fait l'objet d'importants travaux théoriques dans un but de raffinement des modèles et des estimateurs [[barndorff2011multivariate](#)]. Les résultats théoriques doivent alors être testés sur des jeux de données synthétiques, qui permettent de contrôler un certain nombre de paramètres et de s'assurer qu'un effet prédit par la théorie est bien observable *toutes choses égales par ailleurs*. Par exemple, [[potiron2015estimation](#)] dérive une correction du biais de l'estimateur de *Hayashi-Yoshida* qui est un estimateur de la covariance de deux browniens corrélés à haute fréquence dans le cas de temps d'observation asynchrones, par démonstration d'un théorème de la limite centrale pour un modèle généralisé endogénisant les temps d'observations. La confirmation empirique de l'amélioration de l'estimateur est alors obtenue sur un jeu de données synthétiques à un niveau de corrélation fixé.

Formalisation

Cadre

Considérons un réseau d'actifs $(X_i(t))_{1 \leq i \leq N}$ échantillonnés à haute fréquence (typiquement 1s). On se place dans un cadre multi-scalaire (utilisé par exemple dans les approches par ondelettes [[ramsey2002wavelets](#)] ou analyses multifractales du signal [[bouchaud2000apparent](#)]) pour interpréter les signaux observés comme la superposition de composantes à des multiples échelles temporelles : $X_i = \sum_{\omega} X_i^{\omega}$. On notera $T_i^{\omega} = \sum_{\omega' \leq \omega} X_i^{\omega'}$ le signal filtré à une fréquence ω donnée. Prédire l'évolution d'une composante à une échelle donnée est alors un pro-

blème caractéristique de l'étude des systèmes complexes, pour lequel l'enjeu est l'identification de régularités et leur distinction des composantes considérées comme stochastiques en comparaison⁴. Dans un souci de simplicité, on représente un tel processus par un modèle de prédiction de tendance à une échelle temporelle ω_1 donnée, formellement un estimateur $M_{\omega_1} : (T_i^{\omega_1}(t'))_{t' < t} \mapsto \hat{T}_i^{\omega_1}(t)$ dont l'objectif est la minimisation de l'erreur sur la tendance réelle $\|T_i^{\omega_1} - \hat{T}_i^{\omega_1}\|$. Dans le cas d'estimateurs auto-regressifs multivariés, la performance dépendra entre autre des correlations respectives entre actifs et il est alors intéressant d'utiliser la méthode pour évaluer celle-ci en fonction de niveaux de correlation à plusieurs échelles. On assume une dynamique de Black-Scholes [jarrow1999honor] pour les actifs, i.e. $dX = \sigma \cdot dW$ avec W processus de Wiener, ce qui permettra d'obtenir facilement des niveaux de correlation voulus.

Génération des données

Il est alors aisément de générer \tilde{X}_i tel que $\text{Var}[\tilde{X}_i^{\omega_1}] = \Sigma R$ (Σ variance estimée et R matrice de corrélation fixée), par la simulation de processus de Wiener au niveau de corrélation fixé et tel que $X_i^{\omega \leq \omega_0} = \tilde{X}_i^{\omega \leq \omega_0}$ (critère de proximité au données : les composantes à plus basse fréquence qu'une fréquence fondamentale $\omega_0 < \omega_1$ sont identiques). En effet, si $dW_1 \perp\!\!\!\perp dW_1^{\perp\!\!\!\perp}$ (et $\sigma_1 < \sigma_2$ pour fixer les idées, quitte à échanger les actifs), alors $W_2 = \rho_{12}W_1 + \sqrt{1 - \frac{\sigma_1^2}{\sigma_2^2} \cdot \rho_{12}^2} W_1^{\perp\!\!\!\perp}$ est tel que $\rho(dW_1, dW_2) = \rho_{12}$. Les signaux suivants sont construits de la même manière par orthonormalisation de Gram. On isole alors la composante à la fréquence ω_1 voulue par filtrage, c'est à dire $\tilde{X}_i^{\omega_1} = W_i - \mathcal{F}_{\omega_0}[W_i]$ (avec \mathcal{F}_{ω_0} filtre passe-bas à fréquence de coupure ω_0), puis on reconstruit les signaux synthétiques par $\tilde{X}_i = T_i^{\omega_0} + \tilde{X}_i^{\omega_1}$.

Résultats

Méthodologie

La méthode est testée sur un exemple de deux actifs du marché des devises (EUR/USD et EUR/GBP), sur une période de 6 mois de juin 2015 à novembre 2015. Le nettoyage des données⁵, originellement échantillonées à l'ordre de la seconde, consiste dans un premier temps à la détermination du support temporel commun maximal (les séquences manquantes étant alors ignorées, par translation verticale des séries, i.e. $S(t) := S(t) \cdot \frac{S(t_n)}{S(t_{n-1})}$ lorsque t_{n-1}, t_n sont les extrémi-

⁴ voir [gell1995quark] pour une discussion étendue sur la construction de *schema* pour l'étude de systèmes complexes adaptatifs (par des systèmes complexes adaptatifs).

⁵ obtenues depuis <http://www.histdata.com/>, sans licence spécifiée, les données nettoyées et filtrées à ω_m uniquement sont mises en accessibilité pour respect du copyright.

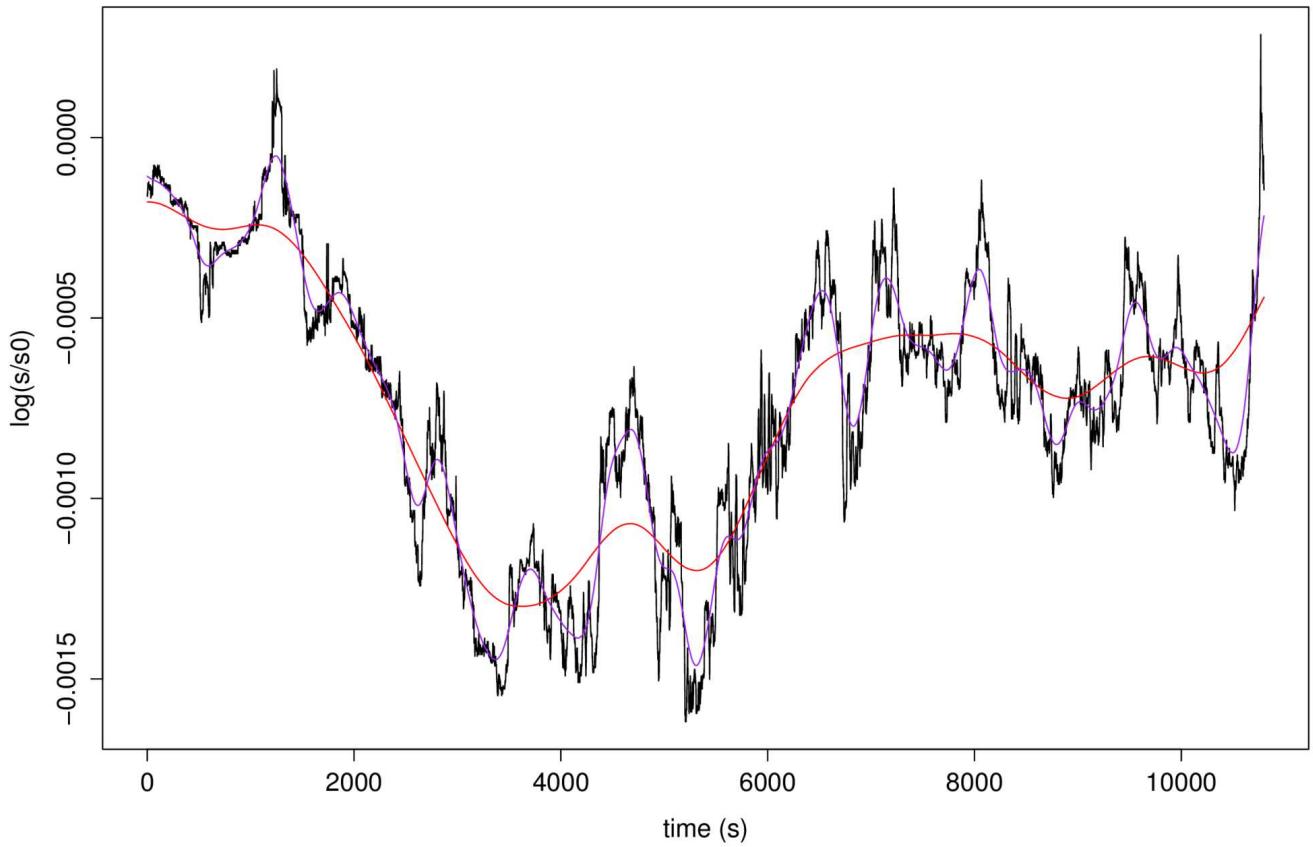
EUR/USD, 1st November 2015

FIGURE 140: Exemple de la nature multi-scalaire des signaux. Celle-ci est à la base de la construction des signaux synthétiques. Les *log-prices* sont représentés sur une fenêtre d'environ 3h le 1er novembre 2015 pour l'actif EUR/USD, avec en violet la tendance à 10min et en rouge à 30min.

tés du “trou” et $S(t)$ la valeur de l'actif, ce qui revient à garder la contrainte d'avoir des retours à pas de temps similaires entre actifs). On étudie alors les *log-price* et *log-returns*, définis par $X(t) := \log \frac{S(t)}{S_0}$ et $\Delta X(t) = X(t) - X(t-1)$. Les données brutes sont filtrées à une fréquence $\omega_m = 10\text{min}$ (qui sera la fréquence maximale d'étude) pour un souci de performance computationnelle. On utilise un filtre gaussien non causal de largeur totale ω . On fixe $\omega_0 = 24\text{h}$ et on se propose de construire des données synthétiques aux fréquences $\omega_1 = 30\text{min}, 1\text{h}, 2\text{h}$. Voir la figure 140 pour un exemple de la structure du signal à ces différentes échelles.

Il est crucial de noter l'interférence entre les fréquences ω_0 et ω_1 dans le signal construit : la corrélation effectivement estimée est

$$\rho_e = \rho [\Delta \tilde{X}_1, \Delta \tilde{X}_2] = \rho [\Delta T_1^{\omega_0} + \Delta \tilde{X}_1^\omega, \Delta T_2^{\omega_0} + \Delta \tilde{X}_2^\omega]$$

ce qui conduit à dériver dans la limite raisonnable $\sigma_1 \gg \sigma_0$ (fréquence fondamentale suffisamment basse), lorsque $\text{Cov} [\Delta \tilde{X}_i^{\omega_1}, \Delta X_j^\omega] =$

0 pour tous $i, j, \omega_1 > \omega$, et les retours d'espérance nulle à toutes échelles, en notant $\rho_0 = \rho [\Delta T_1^{\omega_0}, \Delta T_2^{\omega_0}]$, $\rho = \rho [\Delta \tilde{X}_1^{\omega_1}, \Delta \tilde{X}_2^{\omega_1}]$, et $\varepsilon_i = \frac{\sigma(\Delta T_i^{\omega_0})}{\sigma(\Delta \tilde{X}_i^{\omega_1})}$, la correction sur la corrélation effective due aux interférences : la corrélation effective est alors au premier ordre

$$\rho_e = [\varepsilon_1 \varepsilon_2 \rho_0 + \rho] \cdot \left[1 - \frac{1}{2} (\varepsilon_1^2 + \varepsilon_2^2) \right] \quad (27)$$

ce qui donne l'expression de la corrélation que l'on pourra effectivement simuler dans les données synthétiques.

La corrélation est estimée par méthode de Pearson, avec l'estimateur de la covariance au biais corrigé, c'est à dire

$$\hat{\rho}[X1, X2] = \frac{\hat{C}[X1, X2]}{\sqrt{\text{Var}[X1]\text{Var}[X2]}}$$

, où $\hat{C}[X1, X2] = \frac{1}{(T-1)} \sum_t X_1(t)X_2(t) - \frac{1}{T(T-1)} \sum_t X_1(t) \sum_t X_2(t)$ et $\text{Var}[X] = \frac{1}{T} \sum_t X^2(t) - \left(\frac{1}{T} \sum_t X(t) \right)^2$.

Le modèle de prédiction M_{ω_1} testé est simplement un modèle ARMA pour lequel on fixe les paramètres $p = 2, q = 0$ (on ne crée pas de corrélation retardée, on ne s'attend donc pas à de grand ordre d'auto-régression, les signaux originaux étant à mémoire relativement courte ; de plus le lissage n'est pas nécessaire puisqu'on travaille sur des données filtrées), appliqué de manière adaptative⁶. Plus précisément, étant donné une fenêtre temporelle T_W , on estime pour tout t le modèle sur $[t - T_W + 1, t]$ afin de prédire les signaux à $t + 1$.

IMPLÉMENTATION L'implémentation est faite en langage R, utilisant en particulier la bibliothèque MTS [Tsay:2015xy] pour les modèles de séries temporelles. Les données nettoyées et le code source sont disponibles de manière ouverte sur le dépôt git du projet⁷.

RÉSULTATS La figure 141 donne les corrélations effectives calculées sur les données synthétiques. Pour des valeurs standard des paramètres (par exemple pour $\omega_0 = 24h$, $\omega_1 = 2h$ et $\rho = -0.5$), on a $\rho_0 \simeq 0.71$ et $\varepsilon_i \simeq 0.3$ et ainsi $|\rho_e - \rho| \simeq 0.05$. On constate dans l'intervalle $\rho \in [-0.5, 0.5]$ un bon accord entre la valeur ρ_e prédite par ?? et les valeurs observées, et une déviation pour de plus grandes valeurs absolues, d'autant plus grande que ω_1 est petit : cela confirme

⁶ il s'agit d'un niveau d'adaptation relativement faible, les paramètres T_W, p, q et même le type de modèle restant fixés. On se place ainsi dans le cadre de [potiron2016estimating] qui suppose une dynamique localement paramétrique, mais pour lequel on fixe les métaparamètres de la dynamique. On pourrait imaginer estimer un T_W variable qui s'adapterait pour une meilleure estimation locale, à l'image de l'estimation de paramètres en traitement du signal Bayesien effectuée via augmentation de l'état par les paramètres.

⁷ at <https://github.com/JusteRaimbault/SynthAsset>

l'intuition que lorsque la fréquence descend et se rapproche de ω_0 , les interférences entre les deux composantes vont devenir non négligeables et invalider les hypothèses d'indépendance par exemple.

On applique ensuite le modèle prédictif décrit ci-dessus aux données synthétiques, afin d'étudier sa performance moyenne en fonction du niveau de corrélation des données. Les résultats pour $\omega_1 = 1\text{h}, 1\text{h}30, 2\text{h}$ sont présentés en figure 142. Le résultat a priori contre-intuitif d'une performance maximale à corrélation nulle pour l'un des actifs confirme l'intérêt d'une génération de données hybrides : l'étude des corrélations décalées (*lagged correlations*) montre une dissymétrie présente dans les données réelles, interprété à l'échelle journalière comme une influence augmentée de EURGBP sur EURUSD à 2h de décalage environ. L'existence de ce *lag* permet une "bonne" prédiction de EURUSD due à la fréquence fondamentale, perturbée par le bruit ajouté, de façon proportionnelle à sa corrélation : plus les bruits sont corrélés, plus le modèle les prendra en compte et se trompera plus à cause du caractère markovien des browniens simulés⁸.

L'exemple présenté ici est un *modèle jouet* et n'a pas d'application pratique, mais démontre l'intérêt de l'utilisation des données synthétiques simulées. On peut imaginer simuler des données plus proches de la réalité (existence de motifs réalistes de *lagged correlation* par exemple, modèles plus réalistes que le Black-Scholes) et appliquer la méthode sur des modèles plus opérationnels.

⁸ En théorie le modèle utilisé n'a aucun pouvoir prédictif sur des browniens purs

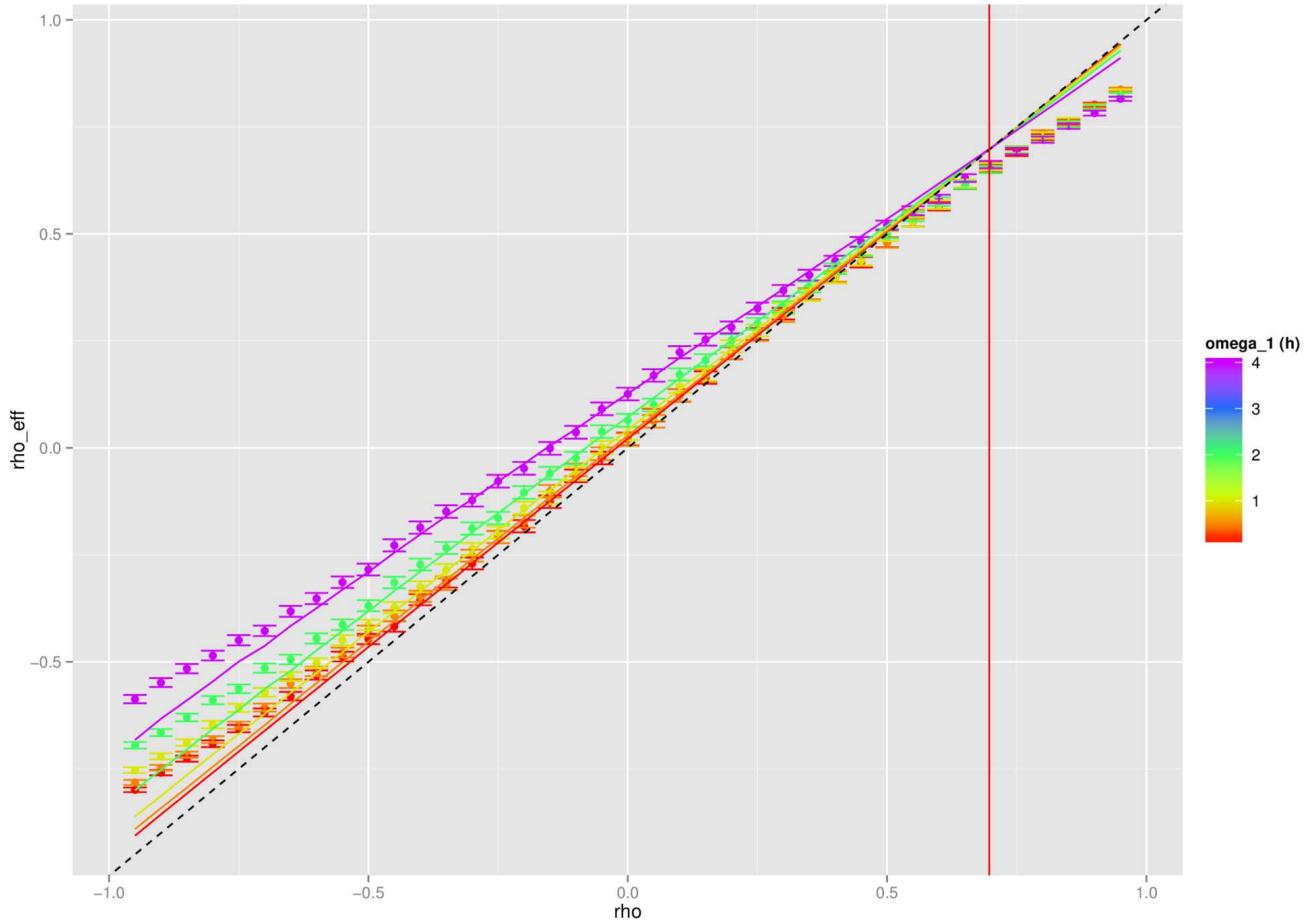


FIGURE 141: Corrélations Effectives obtenues sur données synthétiques. Les points donnent les corrélations estimées sur un jeu de données synthétiques basé sur 6 mois entre juin et novembre 2015 (les barres d'erreur donnent les intervalles de confiance à 95% obtenus par méthode de Fisher standard); l'échelle de couleur donne la fréquence de filtrage $\omega_1 = 10\text{min}, 30\text{min}, 1\text{h}, 2\text{h}, 4\text{h}$; les lignes pleines donnent les valeurs théoriques obtenues par l'équation 27 avec les volatilités estimées (la diagonale en pointillé donne la référence); la ligne verticale rouge est à la position de la valeur théorique telle que $\rho = \rho_e$ avec les valeurs moyennes de ε_i sur l'ensemble des points. Nous observons pour les fortes valeurs de corrélations absolues une déviation des valeurs corrigées, qui devrait être dues à la non-vérification des hypothèses d'indépendance et de centrage des retours. L'asymétrie est due à la forte valeur de $\rho_0 \simeq 0.71$.

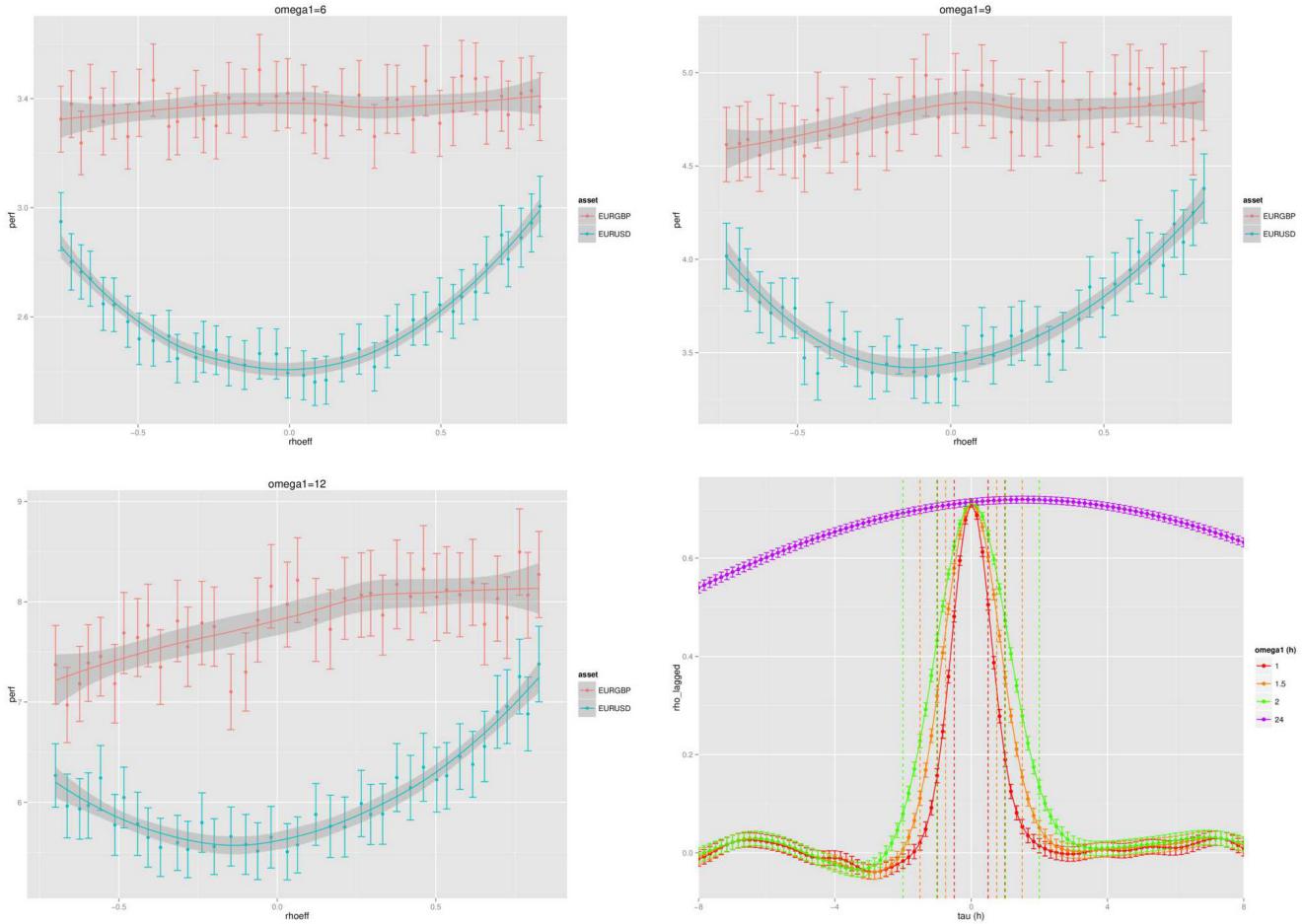


FIGURE 142: Performance d'un modèle prédictif en fonction des corrélations simulées. De gauche à droite et de haut en bas, les trois premiers graphes donnent pour chaque actif la performance normalisée d'un modèle ARMA ($p = 2, q = 0$), définie par $\pi = \left(\frac{1}{T} \sum_t (\tilde{X}_i(t) - M_{\omega_1} [\tilde{X}_i](t))^2 \right) / \sigma [\tilde{X}_i]^2$ (intervalles de confiance à 95% calculés par $\pi = \bar{\pi} \pm (1.96 \cdot \sigma[\pi])/\sqrt{T}$, le lissage polynomial local est pour l'aide à la lecture). Il est intéressant de noter la forme de U pour EUR/USD, due aux interférences entre composantes à différentes échelles. La corrélation entre les bruits simulés détériore le pouvoir de prédiction. L'étude des corrélations retardées (ici données par $\rho[\Delta X_{EURUSD}(t), \Delta X_{EURGBP}(t - \tau)]$) sur les données réelles clarifie ce phénomène : le quatrième graphe présente une asymétrie des courbes à toutes les échelles par rapport au décalage nul ($\tau = 0$) ce qui conduit les composantes fondamentales à accroître le pouvoir prédictif pour le dollar, amélioration qui est ensuite perturbée par les corrélations entre les composantes simulées. Les lignes pointillées donnent les pas de temps (en équivalent en unités de τ) utilisés par les ARMA à chaque échelle, ce qui permet de lire la corrélation retardée correspondante sur les composantes fondamentales.

C.6 MODÉLISATION MULTI-SCALAIRE DES DYNAMIQUES RÉSIDENTIELLES

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 [losavio2017modeling].

★ ★

★

Cette section introduit un modèle basé agent des dynamique de migration intra-régionales, appliqué à la Méga-city Region du Delta de la Rivière des Perles. Il se concentre sur les dynamiques résidentielles de travailleurs migrants et vise à prendre en compte la variété des profils des migrants, en se basant sur des observations qualitatives de terrain. L'exploration intensive du modèle, à la fois sur des configurations synthétiques et réelles, fournit des faits stylisés sur les processus de migration, et des effets spécifiques de la géographie régionale, qui pourraient être appliqués aux régulations des migrations. Nous supposons que de telles approches de modélisation intégrées seront par l'avenir de l'étude des villes chinoises.

c.6.1 *Introduction*

CONTEXTE Ces dernières décades, les travailleurs migrants du rural vers l'urbain ont été une force motrice pour l'économie chinoise, attirant l'attention sur les questions socio-économiques associées. Cependant, l'importance de leur diversité économique et la mobilité sociale ont été peu considérés dans l'analyse des stratégies de développement urbaines.

Nous utilisons un modèle basé agent pour simuler les dynamiques résidentielles des migrants dans la méga-région urbaine du Delta de la Rivière des Perles (PRD), prenant en compte l'ensemble des statuts socio-économique des migrants et leur évolution. Les méga-régions urbaines sont devenues une nouvelles échelle de régulation pour l'Etat Chinois, et le PRD représente le plus prospère et dynamique en termes de vagues de migrations. Cela en fait ainsi une parfaite unité d'analyse.

MÉGA-RÉGIONS URBAINES Les *Mega-city Regions* (MCR) sont définies par [florida2008rise] comme “*des ensembles de villes intégrés et leur territoires suburbains environnants, au sein desquels le travail et le capital peuvent être réalloués à moindre coût*”. Cette configuration urbaine correspond à ce que [gottman1961megalopolis] définit comme *megapolis* en référence à la côte nord-est des Etats-unis. Malgré cette similarité dans leur configuration spatiale et fonctionnelle, les MCRs fonctionnent à une échelle différentes de celle de la megapolis : elles opèrent à la fois à une échelle régionale et à une échelle globale. En effet, l'une des caractéristiques principales des MCRs est leur connectivité : spatiallement elles s'étendent aux régions urbaines et métropolitaines proches, et économiquement elles ont un impact international, bien au delà de leur frontière physique. Ces régions densément peuplées n'ont pas un unique barycentre, mais consistent en de multiples centres fortement connectés. La forte densité de connections et le polycentrisme caractérisant ces nouvelles unités économiques facilitent les flux de migration et encouragent l'intégration régionale.

En Chine, le développement des méga-régions urbaines a commencé juste après l'implémentation de la politiques des portes ouvertes en 1978. Mais c'est la décentralisation progressive du pouvoir de l'Etat, qui a eu lieu au début des années 1990, qui promeut les villes et plus récemment les méga-régions urbaines comme une nouvelle échelle de régulation de l'Etat Chinois [IJUR:IJUR12437]. Le processus de croissance économique rapide et le développement urbain conduisent à de nouvelles méga-régions urbaines densément peuplées et industriellement dynamiques, parmi lesquelles le Delta de la Rivière des Perles (PRD)⁹ est l'exemple le plus représentatif. La

⁹ La *Mega City Region* du PRD est composée de neuf villes : les villes cœur sont Guangzhou et Shenzhen, entourées de Dongguan, Foshan, Zhongshan, Zhuhai,

zone a été choisie en 1988 comme une "zone de réforme économique complète", et il lui a été accordé de nombreuses politiques de "pas en avant" pour attirer le capital étranger. Evoluant vers le centre d'exportation le plus important depuis les réformes économiques, le Delta de la Rivière des Perles représente la MCR la plus dynamique en termes de vagues de migration [IJUR:IJUR820].

TRAVAILLEURS MIGRANTS Considérant le PRD comme l'unité spatiale d'étude, nous visons à reproduire les motifs résidentiels des travailleurs migrants, en prenant en compte l'ensemble des possibilités de leur statut socio-économique. Les motifs de migration et les questions essentielles qui y sont rattachées ont été largement étudiés selon diverses perspectives, s'étendant par exemple de questions ethniques [dong2010policing] aux analyses par données massives de leur comportement spatio-temporel [2017arXiv170600682Y]. Cependant, les travailleurs migrants sont généralement considérés et traités comme une catégorie uniforme, qui est placée au bas de la société urbaine, portant les stigmates du système d'enregistrement rural. La structure duale rurale-urbain a depuis des années été l'unique approche pour définir et comprendre les travailleurs migrants, mais le processus de croissance économique rapide que la Chine a expérimenté a accéléré la transformation sociale. Nous postulons que l'étude des travailleurs migrants en considérant seulement leur statut de *Hukou* et le lieu d'enregistrement n'est plus suffisant pour appréhender une telle catégorie sociale complexe et diversifiée. D'autres aspects comme le capital économique, culturel et humain des travailleurs migrants doivent être pris en compte.

En particulier, trois dimensions peuvent être utiles pour différencier un certain nombre de sous-catégories de travailleurs migrants : (i) la dimension professionnelle, qui détermine non-seulement la situation économique des migrants mais influence également leur trajectoires et la durée de leur séjour dans la ville ainsi que leurs choix résidentiels ; (ii) la dimension résidentielle qui influe sur l'ensemble des aspects des vies urbaines des migrants : établissements urbains, choix de logement, conditions résidentielles, relations à la ville, activités de voisinage, etc. ; (iii) la dimension de la génération¹⁰.

Toutes ces sous-catégories ont différents motifs de mobilité, que nous simulons dans le modèle. Considérant cette diversité et la traduisant en faits stylisés qualitatifs qui correspondent à des motifs précis de données synthétiques, ce modèle vise à établir une nouvelle perspective pour comprendre la mobilité résidentielle urbaine et régionale

Huizhou, Jiangmen, et Zhaoqing. Le modèle n'inclut pas Hong-Kong et Macao, qui font partie de la méga-région urbaine du PRD mais ne sont pas en Chine continentale.

¹⁰ La dimension génération n'est pas prise en compte dans le modèle, puisque les dynamiques simulées correspondent à des échelles temporelles plutôt courtes, entre 10 et 20 ans.

en Chine, en utilisant une approche plus qualitative par la spécification des mécanismes par lesquels l'Etat-Parti contrôle les paramètres des choix des migrants.

c.6.2 Modèle

MODÉLISATION DES MIGRATIONS RURAL-URBAIN EN CHINE La plupart des travaux existant sur la modélisation de la migration rural-urbain en Chine sont principalement des études économétriques, qui se basent sur des données des sondages ou d'études ciblées. [zhang2013measuring] estime des modèles de choix discrets pour étudier le trade-off entre distance de migration et différence de salaire. [fan2005modeling] montre que des modèles gravitaires peuvent bien expliquer les motifs de migration inter-provinciaux, impliquant de forts processus dominants d'agrégation sous-jacents. L'association positive entre les écarts salariaux et les taux de migration a été obtenue à partir d'analyse de séries temporelles dans [zhang2003rural]. Une étude empirique des dynamiques résidentielles intra-urbaines des migrants est faite par [wu2006migrant]. [xie2007simulating] utilise un modèle basé agent pour simuler l'émergence des villages urbains. Au meilleur de notre connaissance, il n'existe pas de tentative précédente dans la littérature pour modéliser les migrations régionales en Chine à partir d'une perspective basée agent.

MODÈLE Le modèle est conçu pour inclure des faits stylisés précis et des expériences associées, en particulier le rôle de la structure socio-économique de la population de migrants. Plus précisément, un changement récent dans la structure socio-économique de la population migrante a été observée, incluant une augmentation du nombre de migrants aux salaires médians et une relativisation du rôle du *Hukou* dans les dynamiques migratoires. Le cœur du modèle est pour cette raison centré sur l'exploration de l'impact d'une variation de la structure économique de la population des migrants sur les dynamiques du système, et l'influence des politiques migratoires gouvernementales.

La région est représentée dans le modèle par N patches, caractérisé par leur population $P_i(t)$ et une structure économique $E_i^{(c)}(t)$ qui représente un nombre potentiel d'emplois pour une classe socio-économique c . Le nombre effectif de travailleurs associés est noté $W_i^{(c)}(t)$. Dans un souci de simplicité, nous supposons un ensemble discret de classes. A l'instant initial, les variables sont initialisées soit selon un processus de génération de données synthétiques (voir ci-dessous), ou à partir de données géographiques réelles (abstraites et simplifiées pour répondre à notre contexte).

Les centres urbains sont caractérisés par une population agrégée $\tilde{P}_k(t)$ et les variables économiques correspondantes $\tilde{E}_k^c(t)$. Un agent

est un foyer de migrants, avec une localisation résidentielle et pour le travail. La structure socio-économique de la population est capturée par la distribution de richesse $g(w)$, qui est ensuite stratifiée en catégories. A un instant donné, la différence d'utilité entre l'action de relocalisation de la cellule j vers la cellule i et rester sur place, pour une catégorie c , est donnée par

$$\Delta U_{i,j}^{(c)}(t) = \frac{Z_j^{(c)} - Z_i^{(c)}}{Z_0} + \gamma \cdot \frac{C_i^{(c)} - C_j^{(c)}}{C_0} - u_i^{(c)} - h_j^{(c)}$$

avec $Z_i^{(c)}$ une mesure d'accessibilité généralisée définie comme

$$Z_i^{(c)} = P_i \cdot \sum_k [E_k^{(c)} - W_k^{(c)}] \cdot \exp\left(\frac{-d_{ij}}{d_0}\right)$$

avec d_{ij} une distance effective de trajet¹¹ et d_0 distance de migration pendulaire typique; le paramètre γ est un rapport donnant l'importance relative du coût de la vie en comparaison à l'accessibilité dans les décisions migratoires; $C_i^{(c)}$ est le coût de la vie qui est une fonction à la fois de la cellule et des variables de la ville, que nous prenons comme $C_i^{(c)} \propto P_i^{\alpha_0} \cdot \tilde{P}_i^{\alpha_1}$; $u_i^{(c)}$ est une aversion au mouvement de référence et $h_j^{(c)}$ une variable exogène correspondant aux politiques de régulation; Z_0 et C_0 sont des paramètres de dimensionnement.

A chaque pas de temps, le système évolue séquentiellement selon les règles suivantes :

1. les variables au niveau de la ville sont mises à jour et distribuées aux variables de patch (dans les premières expériences, nous supposerons une courte échelle temporelle et sauterons cette étape);
2. de nouveaux migrants entrent dans la région et se fient à leur réseau social pour s'établir;
3. les migrations ont lieu dans la région, tirées aléatoirement à partir des probabilités de choix discrets obtenues avec les différences d'utilités entre patches données ci-dessus;
4. les migrants mettent à jour leur richesse et possiblement leur catégorie économique, selon une "qualité du lieu" abstraite que nous associons au GPD par tête qui suit une loi d'échelle de la population.

¹¹ Comme le modèle ne se concentre pas sur le rôle des transports, nous prenons la distance euclidienne, et d_0 capture une distance typique de migration pendulaire à la fois par transport public et voiture. Un modèle plus compliqué pourrait inclure un réseau de transport explicite ainsi que des choix modaux dépendants des catégories socio-économiques.

c.6.3 Résultats

Le modèle est implémenté en NetLogo, l'implémentation ouverte étant disponible avec les résultats à [https://www.github.com/JusteRaimbault/](https://www.github.com/JusteRaimbault/MigrationDynamics)
MigrationDynamics. Le modèle est d'abord exploré sur des systèmes de villes synthétiques, afin d'isoler les résultats dus aux processus en eux-même des résultats dus à la configuration géographique. Avec un tel modèle stochastique pour lequel de nombreux paramètres ne peuvent être fixé par des valeurs observées, il est nécessaire d'explorer de manière extensive l'espace des paramètres pour obtenir des conclusions robustes. Grâce au logiciel OpenMole [reuillon2013openmole], nous procédons à 1,599,495 simulations du modèle sur grille de calcul, réalisant autour de 15 ans de calcul en équivalent CPU en environ 2 jours. Nous validons le modèle de manière interne en vérifiant la convergence statistique des indicateurs.

Les expériences de contrôle (valeur de référence des paramètres) fournissent les faits stylisés suivants sur les dynamiques intrinsèques au cœur du modèle :

1. Lorsque les migrants ont une forte potentialité de mobilité, la répartition spatiale des emplois devient sous-optimale dans des régimes intermédiaires de stochasticité (au sens des valeur prises par le paramètre de choix discrets), ce qui correspond à un régime où la congestion domine.
2. Ce régime de congestion implique une décroissance linéaire de la distance à l'emploi avec la diminution du caractère aléatoire, ce qui signifie que le déterminisme social crée des inégalités spatiales.
3. L'importance relative de l'accessibilité influe très peu les dynamiques agrégées. Ainsi, un gain croissant en mobilité (c'est à dire une importance accrue pour l'accessibilité), qui peut être encouragé par des politiques locales telles des subventions, n'aura que très peu d'effet sur les motifs de migration.
4. Les configurations avec des valeurs intermédiaires de l'aversion au mouvement (dans lesquelles la situation réelle se trouve) induisent un effet de feedback négatif du temps au cours des trajectoires, témoignant d'une saturation progressive. Dans un comportement "en-U", les configurations très mobiles et celles très stables donnent un effet positif du temps (augmentation du nombre de migrations).

Nous étudions ensuite des expériences ciblées.

L'ajout des catégories socio-économiques ne change pas fondamentalement le comportement qualitatif du modèle. La catégorie la plus basse semble cependant plus vulnérable aux inégalités spatiales induites par le déterminisme spatial. Concernant l'influence des pa-

ramètres économiques, en particulier les inégalités de revenu et la croissance des revenus, nous trouvons que : (i) de plus fortes inégalités de revenus induisent de plus fortes inégalités spatiales dans l'accès à l'emploi ; (ii) une croissance des revenus plus forte (un enrichissement plus grand) lors d'une migration conduit à un régime sous-optimal pour la catégorie supérieure.

L'application du modèle sur la configuration observée pour la population et les emplois dans le delta de la rivière des Perles change légèrement les conclusions : celle-ci témoigne par exemple de plages optimales pour les paramètres comportementaux au regard des indicateurs de distance de mobilité pendulaire. Cela signifie que les incitations aux migrations doivent être spécifiquement conçues selon la configuration de la région. La plupart des conclusions tirées précédemment tiennent toujours, et sont ainsi spécifiques au processus considérés.

DISCUSSION Une dernière application en cours de développement est l'exploration de l'impact de politiques de régulation localisées, i.e. en ayant le terme $h_j^{(c)}$ qui varie selon les villes et selon les catégories socio-économiques, ce qui correspondrait à des politiques effectivement observées en pratique. Les divers faits stylisés décrits précédemment peuvent de plus être porteurs de sens pour l'élaboration de politiques plus générales, comme l'impact d'une mobilité accrue, ou bien l'existence de régimes optimaux pour des valeurs intermédiaires du caractère aléatoire. Un développement futur pourra consister en une calibration du modèle sur des trajectoires de migration avec les jeux de données appropriés, mais également en un retour des résultats de simulation sur le travail de terrain qualitatif, en les comparant aux situations concrètement observées.

Cette entreprise de modélisation vise à être intégrée, puisque le modèle est initialement construit en prenant en considération des observations qualitatives du travail de terrain¹², et ses sorties devraient en retour être utiles pour la recherche qualitative. Nous sommes convaincus que de telles approches de modélisation intégrée seront des outils importants pour le futur de la recherche Urbaine en Chine, en particulier en lien avec l'émergence de nouveaux régimes urbains dans les villes chinoises qui n'ont jamais été observés ailleurs auparavant, rendant difficile l'utilisation de certaines connaissances empiriques précédentes sur les villes.

¹² Pour rappeler le contexte dans le cadre plus général de la thèse, il ne s'agit pas du travail de terrain décrit en 1.3, mais de celui de CINZIA LOSAVIO réalisé dans le cadre de sa thèse en cours (voir contributions ci-dessus).

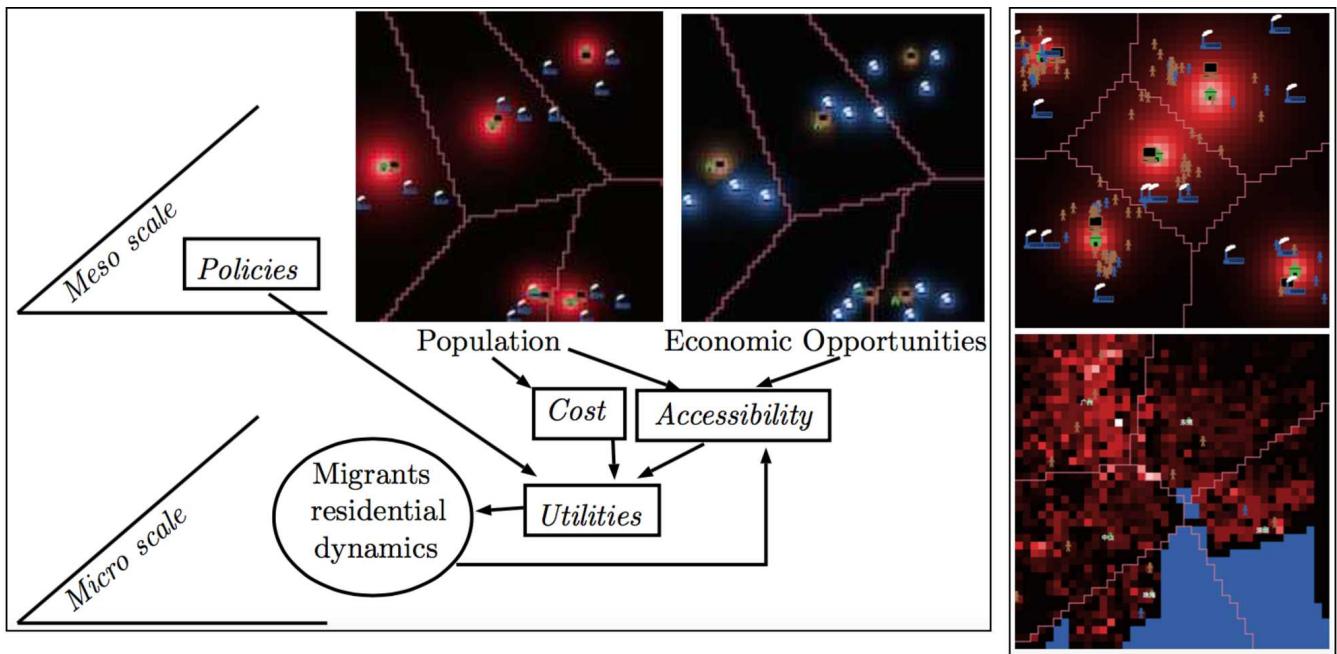


FIGURE 143: Structure du modèle de migrations intra-régionales. (Gauche) Schéma des processus pris en compte et des agents, dans une perspective multi-scalaire

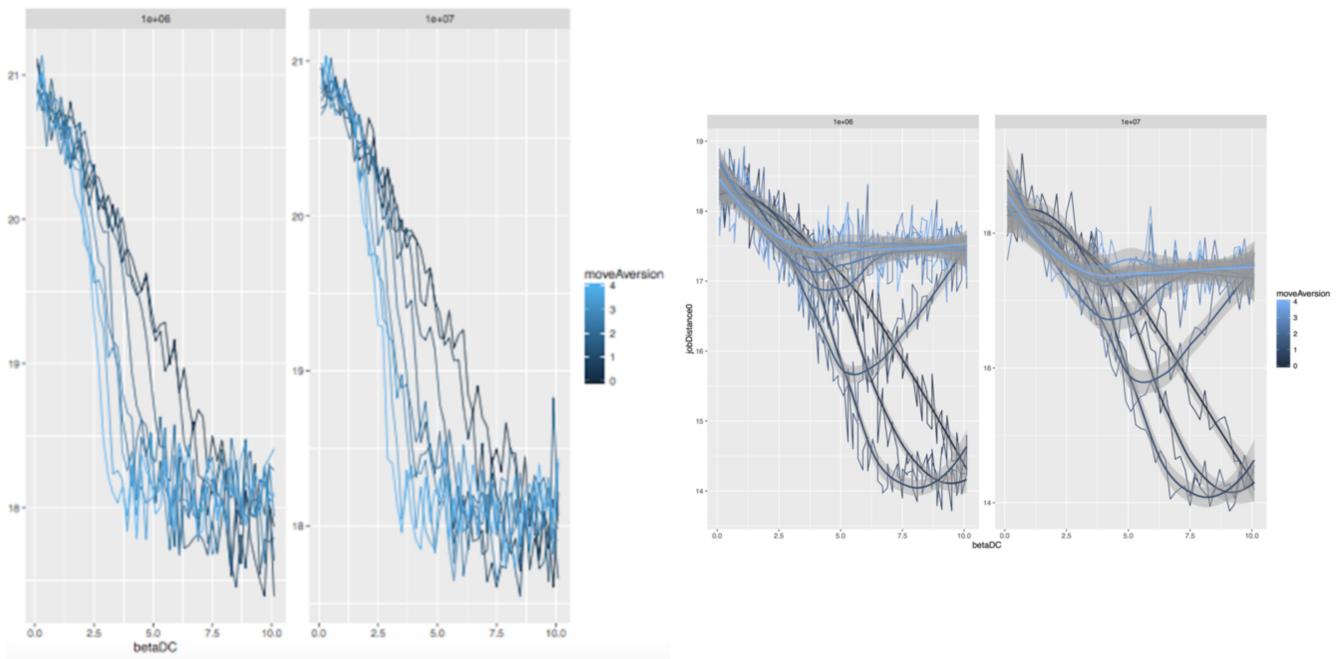


FIGURE 144: Exemple de faits stylisés obtenus par le modèle. Nous comparons la distance moyenne aux emplois pour la catégorie économique la plus basse, en fonction du paramètre d'aléatoire β , entre des systèmes de villes synthétiques (Deux graphes de gauche) et la configuration réelle (Deux graphes de droite). La couleur donne l'aversion au mouvement constante $u_i^{(c)}$ et les graphes sont donnés pour deux valeurs du ratio entre coût et accessibilité γ . Nous voyons émerger des valeurs optimales pour β dans la situation réelle, probablement causées par la géographie.

* *

*

D

DONNÉES

Cette annexe liste et décrit les différents jeux de données ouvertes que nous avons été amenés à créer et à utiliser dans la thèse. Les données sont en effet bien un domaine de connaissance propre, et les opérations de collecte et de consolidation sont une étape scientifique à part entière.

D.1 DONNÉES DE TRAFFIC DU GRAND PARIS

D.1.1 *Spécification*

CITATION

TYPE ET FORMAT

LICENCE

DISPONIBILITÉ

D.1.2 *Description*

D.2 GRAPHES TOPOLOGIQUES DES RÉSEAUX ROUTIERS

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.

D.2.1 *Description*

D.2.2 *Spécification*

CITATION Raimbault, Juste, 2018, "Simplified road networks, Europe and China", doi :10.7910/DVN/RKDZMV, Harvard Dataverse, V1

TYPE ET FORMAT Edge lists of graphs. Format : postgis dumps

LICENCE CCo

DISPONIBILITÉ 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 retranscriptions est essentielle pour la reproductibilité puisqu'il s'agit du dernier (et du premier) stade avant la traduction non reproductive 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_kapiszewski_2018].

D.3.1 Description

Entretien avec Denise Pumain, 2017/03/31

Cet entretien est intervenu dans le contexte d'une collecte de matériau empirique pour la rédaction de [raimbault2017applied], qui a permis entre autre la construction du cadre de connaissances développé en 9.3. L'entretien est principalement centré sur la genèse de la Théorie Evolutive des Villes.

Entretien avec Romain Reuillon, 2017/04/11

Cet entretien intervient dans le même contexte, en cherchant à apporter un éclairage du point de vue des méthodes et outils. Il retrace en particulier la genèse d'OpenMole.

Entretien avec Clémentine Cottineau, 2017/05/05

Géographe à l'interface interdisciplinaire

Entretien avec Denise Pumain, 2017/12/15

Effets structurants des infrastructures de transport et co-évolution, du point de vue de la géographie.

Entretien avec Alain Bonnafous, 2018/01/09

Effets structurants des infrastructures de transport, du point de vue de l'économie des transports.

¹ <https://qdr.syr.edu/>

D.3.2 Spécification

CITATION

TYPE ET FORMAT Données textuelles : transcription des entretiens au format texte.

LICENCE

DISPONIBILITÉ Dépôt git : <https://github.com/JusteRaimbault/Entretiens>

D.4 DONNÉES SYNTHÉTIQUES ET RÉSULTATS DE SIMULATIONS

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>
- Simulation et calibration du modèle macroscopique d'interactions
- Simulation et calibration du modèle de morphogenèse pour la densité <http://dx.doi.org/10.7910/DVN/WSUSBA>
- 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](http://dx.doi.org/10.7910/DVN/V3KI2N)

E

OUTILS

E.1 SOFTWARES AND PACKAGES

Cette annexe recense les contributions logicielles significatives, qui ont fait l'objet d'un *packaging* dans l'esprit d'une science ouverte.

E.1.1 *largeNetwoRk : Import de réseau et simplification pour R*

E.1.2 *Fouille de Corpus scientifique*

E.1.3 *Réseaux de transports et accessibilité en R*

E.1.4 *morphology : extension NetLogo pour mesurer la forme urbaine*

Disponible à <https://github.com/JusteRaimbault/nl-spatialmorphology>

E.2 ARCHITECTURE AND SOURCES FOR ALGORITHMS AND MODELS OF SIMULATION

You must not be afraid of putting code in your thesis, code is not dirty
 - ALEXIS DROGOUL PhD defense
 of [rey2015plateforme]

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. The following reports are partially generated automatically using experimental tools aimed at workflow improvement.

E.2.1 Revue Systématique Algorithmique

OBJECTIFS Implement systematic literature review algorithm.

LOCALISATION <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Biblio/AlgoSR/AlgoSRJavaApp>

CARACTÉRISTIQUES

- Language : Java
- Size : 7116

PARTICULARITÉS

- HashConsing used for unique bibliography object, specific hashCode 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.

SCRIPTS ADDITIONNELS R for result exploration and visualization.

E.2.2 Bibliométrie Indirecte

OBJECTIFS Multi-layer network analysis of scientific corpuses : cybergeo journal, corpus in [2.2](#)

LOCALISATION <https://github.com/Geographie-cites/cybergeo20/tree/master/HyperNetwork>
<https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Biblio/AlgoSR/AlgoSRJavaApp> for common Java part.

CARACTÉRISTIQUES

- Language : Python, R and Java.
- Size : -

PARTICULARITÉS Polyglot

ARCHITECTURE See schema chapter 3.

SCRIPTS ADDITIONNELS -

E.2.3 Croissance Urbaine

OBJECTIF Density-based urban morphogenesis model

LOCALISATION <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Synthetic/Density>

CARACTÉRISTIQUES

- Language : NetLogo, scala.
- Size : 4355

PARTICULARITÉS Morphological indicators in scala implemented with Fast Fourier transform; with R communication in NetLogo.

ARCHITECTURE Nothing particular.

SCRIPTS ADDITIONNELS R for result exploration and morphological analysis.
 oms for model exploration.

E.2.4 Génération des Données Synthétiques Corrélates

OBJECTIFS Weak coupling of density generation and network generation.

LOCALISATION https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Synthetic/Network_20151229

CARACTÉRISTIQUES

- Language : NetLogo (network) and scala.
- Size : 3188

PARTICULARITÉS Network heuristic easier to implement and explore in netlogo

ARCHITECTURE OpenMole allows coupling between modules through exploration script.

SCRIPTS ADDITIONNELS R for result exploration.
oms for model exploration.

E.2.5 Modèle Lutecia

OBJECTIF Implementation of Lutecia model, chapter ??.

LOCALISATION <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Governance/MetropolSim/Lutecia>

CARACTÉRISTIQUES

- Language : NetLogo
- Size : 4791

PARTICULARITÉS Shortest path dynamical programming using matrices.

ARCHITECTURE Pseudo object architecture in agent environment.

SCRIPTS ADDITIONNELS R for result exploration.
oms for model exploration.

E.2.6 Analyse des Réseaux

Package LargeNetwoRk

OBJECTIF Simplification of european road network

LOCALISATION <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/StaticCorrelations>

CARACTÉRISTIQUES

- Language : R, Shell, PostgreSQL
- Size : 505

PARTICULARITÉS Handling of large size databases imposes sequential processing; use of external program osmosis for conversion from osm data to postgresql.

ARCHITECTURE Shell script lead maneuvers.

SCRIPTS ADDITIONNELS -**E.2.7 Co-évolution par morphogenèse**

OBJECTIF Implémentation du modèle de co-évolution à l'échelle mesoscopique

LOCALISATION <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/MesoCoevol>

CARACTÉRISTIQUES

- Language : NetLogo
- Size :

PARTICULARITÉS -**ARCHITECTURE** -**SCRIPTS ADDITIONNELS** -**E.2.8 Co-évolution à l'échelle macroscopique**

OBJECTIF Implémentation du modèle de co-évolution à l'échelle macroscopique

LOCALISATION <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/MacroCoevol>

CARACTÉRISTIQUES

- Language : NetLogo
- Size :

PARTICULARITÉS -

ARCHITECTURE -

DONNÉES UTILISÉES

SCRIPTS ADDITIONNELS -

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 *Générateur de Documentation Netlogo*

La génération de documentation est centrale pour la reproductibilité, permettant d'automatiser la description de l'implémentation d'un modèle. NetLogo ne fournit pas de générateur de documentation. Nous avons implémenté un wrapper du logiciel Doxygen (génération de documentation pour divers langages dont Java) pour son application au langage NetLogo. Il repose sur le principe basique de génération d'un code Java intermédiaire, miroir du code NetLogo dans ses structures objet et reprenant les blocs de commentaires dans le code NetLogo. Une version expérimentale est disponible à <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Doc>.

E.3.2 *git comme outil de reproductibilité*

L'utilisation de git comme outil de reproductibilité et de transparence a été mis en valeur par [ram2013git], qui soulignent de nombreux avantages tels le suivi exact de l'historique du processus de production de connaissance, un clonage immédiat (en combinaison avec des dépôts publics, pour lesquels existent des sites collaboratifs comme github ou gitlab), une possibilité de branchage à partir de commits passés.

Cet outil permet d'autre part de faciliter le flux de travail individuel, permettant par exemple le backup automatique, l'organisation, le suivi des expériences.

Revue Ouverte

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.

E.3.3 *Vers un gestionnaire de métadonnées compatible avec git*

La question de la conservation des métadonnées pour les figures est cruciale pour la reproductibilité, puisqu'il est souvent difficile de garder une trace de l'ensemble de la configuration ayant généré une figure, ainsi que le code correspondant, celui-ci pouvant être modifié

par des versions antérieures. L'utilisation d'environnements scriptés comme R ou python peuvent également être piégeurs puisque les variables peuvent être modifiées sans modification du code, et il faut garder alors l'ensemble de l'historique des commandes exécutées.

Le stockage exhaustif des données, de l'environnement, du code et de l'historique qui a conduit à la génération d'une figure précise sont une condition nécessaire pour une reproductibilité exacte. Une piste pour répondre à ce problème est l'élaboration d'un outil compatible avec git qui généreraient automatiquement ces métadonnées, par exemple en créant une branche propre et en conservant le hash du commit associé à la figure. L'idée finale serait d'avoir pour chaque figure un identifiant unique la reliant à l'environnement exact l'ayant produite, impliquant également une automatisation du système d'indexation au sein des documents les utilisant.

E.3.4 *TorPool*

TorPool est un wrapper java du logiciel tor, qui permet de maintenir une équipe d'instances en parallèle, et de renouveler ces instances sur demande. Une interface avec TorPool est disponible avec java par une bibliothèque dédiée. Cet utilitaire permet entre autres de faciliter la collection automatique de données.

Il est disponible sous forme exécutable à <https://github.com/JusteRaimbault/TorPool>.



QUANTITATIVE ANALYSIS OF THESIS REFLEXIVITY

Analyse de la réflexivité

concept maps : [novak2008theory]

C : faire un graphe des concepts; compare to semantic network of concepts in Gödel Escher Bach.