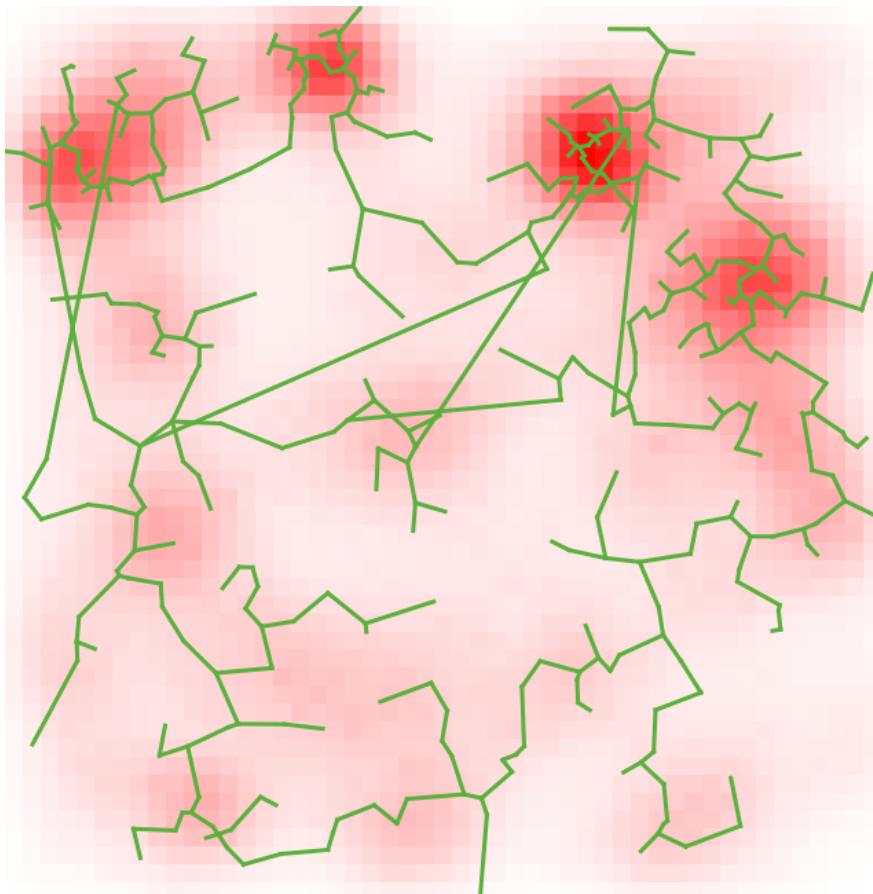


VERS DES MODÈLES COUPLANT DÉVELOPPEMENT URBAIN ET CROISSANCE DES RÉSEAUX DE TRANSPORT

JUSTE RAIMBAULT



Mémoire de Thèse de Doctorat

Sous la direction de ARNAUD BANOS et FLORENT LE NÉCHET

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

Université Paris Diderot - Paris 7

21 Octobre 2017 – version 3.3.2

Juste Raimbault : *Vers des Modèles Couplant Développement Urbain et Croissance des Réseaux de Transport*, Mémoire de Thèse de Doctorat, © 21 Octobre 2017

RÉSUMÉ

C : (Florent) trop de concepts dans l'abstract, peut pas apporter qqchse à tous

C : (Florent) commencer par expliquer ce que sont causalités circulaires et pourquoi difficiles à modéliser

C : (Arnaud) complexly : ?

C : (Arnaud) théorie des systèmes territoriaux en réseau co-évolutifs ?

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. 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 projet à <https://github.com/JusteRaimbault/Cit>. 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).

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). Identification of Causalities in Spatio-temporal Data, forthcoming in *Sageo 2017 Proceedings*. arXiv :1709.08684

Raimbault, J. (2017). An Applied Knowledge Framework to Study Complex Systems, forthcoming in *CSD&M 2017 Proceedings*. arXiv :1706.09244

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

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, 2017a]

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

Raimbault, J. (2016). Generation of Correlated Synthetic Data, forthcoming in *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

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.*

ACKNOWLEDGEMENTS

Un certain nombre de 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.

TABLE DES MATIÈRES

Introduction	3
I FOUNDATIONS	21
1 INTERACTIONS ENTRE RÉSEAUX ET TERRITOIRES	25
1.1 Territoires et Réseaux	27
1.2 De Paris à Zhuhai	42
1.3 Elements de terrain	53
2 MODÉLISER LES INTERACTIONS ENTRE RÉSEAUX ET TERRITOIRES	65
2.1 Modéliser les Interactions	68
2.2 Une Approche Epistémologique	82
2.3 Revue Systématique et Modélographie	99
3 POSITIONNEMENTS	111
3.1 Modélisation, données massives et calcul intensif	113
3.2 Reproductibilité	127
3.3 Positionnement Epistémologique	137
II BRIQUES ÉLÉMENTAIRES	149
4 THÉORIE EVOLUTIVE URBAINE	155
4.1 Corrélations Statiques	157
4.2 Causalités Spatio-temporelles	172
4.3 Modèle de croissance macroscopique	186
5 MORPHOGENÈSE URBAINE	207
5.1 Une Approche Interdisciplinaire de la Morphogenèse	209
5.2 Morphogenèse Urbaine par Agrégation-diffusion	223
5.3 Génération de configurations territoriales corrélées	239
III SYNTHÈSE : MODÈLES DE CO-ÉVOLUTION	251
6 CO-ÉVOLUTION À L'ÉCHELLE MACROSCOPIQUE	257
6.1 Modèles existants	259
6.2 Extension dynamique du modèle d'interaction	267
7 CO-ÉVOLUTION À L'ÉCHELLE MESOSCOPIQUE	281
7.1 Modèles de Croissance de Réseau	282
7.2 Co-évolution à l'échelle mesoscopique	290
7.3 Modélisation de la Gouvernance du Système de Transport	294
IV OUVERTURE	299
8 ÉCHELLES ET ONTOLOGIES	301
8.1 Equilibre Utilisateur Statique	303
8.2 Transport Routier et Déterminants des Coûts	317
9 CADRE THÉORIQUE	331

9.1 Contributions et Perspectives	333
9.2 Une Théorie Géographique	336
9.3 Un Cadre de Connaissances Appliqué	344
Conclusion	361
BIBLIOGRAPHIE	371
V APPENDICES	415
A INFORMATIONS SUPPLÉMENTAIRES	417
A.1 Elements de Terrain	418
A.2 Epistémologie Quantitative	420
A.3 Modélographie	424
A.4 Correlations Statiques	428
A.5 Régimes de causalité	436
A.6 Effets de réseau	439
A.7 Morphogenèse par agrégation-diffusion	440
A.8 Données Synthétiques Corrélées	448
A.9 Heuristiques de génération de réseau	449
B DÉVELOPPEMENTS MÉTHODOLOGIQUES	451
B.1 Un cadre unifié pour les modèles stochastiques de croissance urbaine	452
B.2 Sensibilité des Lois d’Echelle Urbaines à l’Etendue Spatiale	457
B.3 Correlations spatio-temporelles	462
B.4 Génération de Données Synthétiques Corrélées	465
B.5 Un Cadre basé sur la Discrépance	468
B.6 Exploration de l’Interdisciplinarité	484
B.7 Un Cadre pour les Systèmes Socio-techniques	485
C DÉVELOPPEMENTS THÉMATIQUES	497
C.1 Ponts entre Géographie et Economie	497
C.2 Generation of Correlated Synthetic Data	498
C.3 Classifying Patents Based on their Semantic Content .	505
D DONNÉES	531
D.1 Données de Traffic du Grand Paris	531
D.2 Prix de l’Essence aux Etats-Unis	531
D.3 Réseau Routier Européen	531
D.4 Interviews	531
E OUTILS	533
E.1 Softwares and Packages	534
E.2 Architecture and Sources for Algorithms and Models of Simulation	535
E.3 Tools and Workflow for an open Reproducible Research	540
F QUANTITATIVE ANALYSIS OF THESIS REFLEXIVITY	543

TABLE DES FIGURES

FIGURE 1	Impact du Grand Paris Express	46
FIGURE 2	Projets de transport successifs du Grand Paris	48
FIGURE 3	Corrélations retardées empiriques	49
FIGURE 4	Réseau à grande vitesse en Chine	56
FIGURE 5	TOD à Hong-Kong et Zhuhai	58
FIGURE 6	Algorithme de revue systématique	85
FIGURE 7	Réseau de citations	92
FIGURE 8	Motifs d'interdisciplinarité	96
FIGURE 9	Revue Systématique	102
FIGURE 10	Types de couplages	104
FIGURE 11	Usage naïf de la fouille de données	119
FIGURE 12	Distance des diagramme de phase à la référence	125
FIGURE 13	Exemples de diagrammes de phase	126
FIGURE 14	Reproductibilité et visualisation	130
FIGURE 15	Distribution spatiale des morphologies	161
FIGURE 16	Distribution spatiale des indicateur de réseau .	166
FIGURE 17	Corrélations Spatiales	167
FIGURE 18	Variation des corrélations avec l'échelle	169
FIGURE 19	Correlations dans le modèle RDB	177
FIGURE 20	Identification de régimes d'interactions	178
FIGURE 21	Evolution des mesures de réseau	182
FIGURE 22	Corrélations retardées	184
FIGURE 23	Corrélations temporelles	195
FIGURE 24	Sortie du modèle	196
FIGURE 25	Effets de réseau	197
FIGURE 26	Calibration du modèle de gravité	198
FIGURE 27	Valeurs des paramètres calibrés	199
FIGURE 28	Calibration du modèle complet	202
FIGURE 29	Exemple de formes urbaines générées	229
FIGURE 30	Comportement des indicateurs	232
FIGURE 31	Dépendance au chemin	234
FIGURE 32	Calibration du modèle	236
FIGURE 33	Exploration par PSE	237
FIGURE 34	Espace faisable des corrélations	244
FIGURE 35	Génération de configurations couplées	245
FIGURE 36	Comportement du modèle	265
FIGURE 37	Schématisation du modèle	268
FIGURE 38	Comportement du modèle de co-evolution .	271
FIGURE 39	Correlations dans le modèle abstrait	272
FIGURE 40	Corrélations empiriques pour le système de villes français	275

FIGURE 41	Fronts de Pareto	277
FIGURE 42	Evolution des paramètres optimaux	278
FIGURE 43	Example de réseau auto-renforçant	278
FIGURE 44	Exemple du réseau biologique	285
FIGURE 45	Exemples de réseaux	286
FIGURE 46	Espace topologique faisable	288
FIGURE 47	Comparaison aux réseaux réels	289
FIGURE 48	Morphogenèse mesoscopique	291
FIGURE 49	Calibration du modèle de morphogenèse . .	292
FIGURE 50	Régimes de causalité	292
FIGURE 51	Application web pour les données de trafic .	307
FIGURE 52	Variabilité spatiale des plus courts chemins .	308
FIGURE 53	Variabilité des temps de trajet	309
FIGURE 54	Stabilité temporelle de la centralité	311
FIGURE 55	Auto-corrélation spatiale	313
FIGURE 56	Prix moyen par Contés	322
FIGURE 57	Autocorrelation spatiale	323
FIGURE 58	Résultats des analyses GWR	326
FIGURE 59	Réseau de citations de la Théorie Evolutive Urbaine	350
FIGURE 60	Réseau complet des domaines de connaissance	354

LISTE DES TABLEAUX

TABLE 1	Proximités lexicales stationnaires	87
TABLE 2	Communautés sémantiques	94
TABLE 3	Type de modèles	107
TABLE 4	Espace des paramètres	192
TABLE 5	Résultats de l'AIC empirique.	204
TABLE 6	Résumé des paramètres	227
TABLE 7	254
TABLE 8	Sensibilité à l'espace	264
TABLE 9	Prix des carburants	320
TABLE 10	Régressions au niveau du comté C (FL) : quel rapport avec la these ?	328
TABLE 11	431
TABLE 12	432
TABLE 13	Résultats numériques des simulations synthétiques	477
TABLE 14	530

C : (Florent) cf receuil articles du Monde sur Grd Paris (numériser)

A1 : est ce que je te les ai données, sinon redemande moi.

C : (Florent) HDR Anne ? A1 : Florent : ?

C : (Florent) trop peu ancré concrètement dans le champ des interactions transport/ville - enchainement idée ok mais revoir granularité info. Catalogue de situations complexes d'interactions forme urbaine/transport à reproduire. A1 : Personnes intéressantes avec qui discuter de ces sujets au lvmt (liste non exhaustive) : Jean Laterasse, Caroline Gallez.

C (Florent) : Lu Aout 17 : Modelographie plus tard ; Faire 2 chapitres de 1,2,6; 3 à la fin avec chap 9 ; chap 4-2,4-3 : mal dit, on n'est pas vraiment dans la coévolution ; chap 5 hors sujet à supprimer ; Annexes : à vue de nez il y en a trop ce n'est pas évident de dire a priori ce qui est en trop mais des annexes = 50% de la thèse, cela fait bizarre ; Titre des figures : bien sûr à ajuster au fil de la lecture mais ces titres semblent non explicites, il faut pouvoir les comprendre hors du contexte, à minima.

INTRODUCTION

INTRODUCTION

C'est quand on donne un coup de pied dans la fourmilière qu'on se rend compte de toute sa complexité.

- ARNAUD BANOS

"En conséquence d'un problème technique, le trafic est interrompu sur la ligne B du RER pour une durée indéterminée. Plus d'information seront fournies dès que possible". Il y a des fortes chances pour que quiconque ayant vécu ou passé un peu de temps en région parisienne ait déjà entendu cette annonce glaçante et en ait subi les conséquences pour le reste de la journée. Mais il ne se doute sûrement pas des ramifications des cascades causales induites par cet évènement presque banal. Les systèmes territoriaux, quelles que soient les aspects considérés pour leur définition, ont une complexité croissante lorsqu'on augmente leur nombre, les interrelations à de nombreuses échelles spatiales et temporelles participant à la production des comportements émergents observés à tout niveau du système. Martin est un étudiant qui fait l'aller-retour journalier entre Paris et Palaiseau and manquera un examen crucial, ce qui aura un impact profond sur sa vie professionnelle : implications à une longue échelle de temps, une petite échelle spatiale et à la granularité de l'agent. Yuangxi était en train de relier les aéroports d'Orly et Roissy dans son voyage de Londres à Pékin et va manquer son avion ainsi que le mariage de sa soeur : grande échelle spatiale, courte échelle temporelle, granularité de l'agent. Une pétition collective émerge des voyageurs, conduisant à la création d'une organisation qui mettra la pression sur les autorités pour qu'elles augmentent le niveau de service : échelle temporelle et spatiales mesoscopique, granularité de l'aggrégation d'agents. La recherche de cause possible à l'incident conduira à des processus intriqués à diverses échelles, parmi lesquels aucun ne semble être une meilleure explication ; 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 également des éléments d'explication de la structure du réseau parisien de transports en commun qui conditionne la perturbation [GLEYZE, 2005] ; les motifs pendulaires dus à l'organisation territoriale induisent une surcharge de certaines ligne et ainsi nécessairement une augmentation des incidents d'exploitation. La liste pourrait être ainsi continuée un certain temps, chaque approche apportant sa

vision mature correspondant à un corpus de connaissances scientifiques dans des disciplines diverses comme la géographie, l'économie urbains, les transports. Cette anecdote quotidienne est suffisante pour faire ressentir la complexité des systèmes territoriaux. 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 position 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. **C (AB) : Déclaration discutable :** si j'ai bien compris l'argumentaire, ton absence d'ambition tient 1) à l'absence de quête d'universalité (dépendance au contexte) et 2) à la dimension politique inhérente à ton objet. Je te ferai dire que 1) l'absence d'universalité n'est pas une tare (c'est même la norme en SHS) et 2) la dissociation science/politique ne tient plus la route depuis bien longtemps et ne concerne qu'un micro domaine des sciences les plus abstraites et théoriques (maths fondamentales, physique théorique peut-être,... what else?) 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 vu comme une métaphore des processus étudiés. Enfin, sur un plan plus enclin à l'interprétation, cela suggérera la volonté d'une position délicate liant un positionnement 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 réductionnisme¹ strict qui a échoué dans la majorité de ses tentatives de synthèse [ANDERSON, 1972]. Arthur a rappelé récemment [ARTHUR, 2015] 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 des particules, la majorité des disciplines les prenant comme objet d'étude n'auraient jamais émergé puisque la thermodynamique aurait alors résolu la majorité des problèmes sociaux **C : (Florent)attention phrases asimoviennes**². Elles sont au contraire moins connues et acceptées en sciences "dures" comme la physique : LAUGHLIN développe dans [LAUGHLIN, 2006] une vision de la physique à la même position de "frontière des connaissances" que d'autres 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

¹ 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.

² 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 PRIGOGINE dans [PRIGOGINE et STENGERS, 1997].

correspond au premier cauchemar du Démon de Laplace développé dans [DEFFUANT et al., 2015].

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 ANDERSON [ANDERSON, 1972]. 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 [WIENER, 1948], et surtout en développant les notions de retroaction et de contrôle. Plus tard, la Synergétique [HAKEN, 1980] a posé les bases d'approches théoriques des phénomènes collectifs en physique. Les causes possibles de la croissance récente du nombre de travaux se 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 [VARENNE, 2010b]. Elles peuvent aussi être à chercher auprès de progrès en épistémologie : introduction de la notion de perspectivisme [GIERE, 2010c], reflexions plus fine autour de la nature des modèles [VARENNE et SILBERSTEIN, 2013]³. Les potentialités théoriques et empiriques de telles approches jouent nécessairement un rôle dans leur succès⁴, comme le confirme les domaines très variés d'application (voir [NEWMAN, 2011] pour une revue très générale), comme par exemple la Science de Réseaux [BARABASI, 2002] ; les Neurosciences [KOCH et LAURENT, 1999] ; les Sciences Humaines et Sociales, dont la Géographie [MANSON, 2001][PUMAIN, 1997] ; la Finance avec les approches écononophysiques [STANLEY et al., 1999] ; l'Eco-logie [GRIMM et al., 2005]. La Feuille de Route des Systèmes Complexes [BOURGINE, CHAVALARIAS et AL., 2009] 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.

C : (Florent) donner ici exemples dans champ transports/urba

C : (Florent) plus de détails sur les disciplines CS ?

³ dans ce cadre, les progrès scientifiques et épistémologiques ne peuvent pas être dissociés et peuvent être vus comme étant en co-évolution

⁴ même si l'adoption de nouvelles pratiques scientifiques est souvent largement biaisé par l'imitation et le manque d'originalité [DIRK, 1999], ou de façon plus ambiguë, par des stratégies de positionnement puisque le combat pour les fonds est un obstacle croissant à une recherche saine [BOLLEN et al., 2014].

INTERDISCIPLINARITÉ

Il est important d'insister sur le rôle de l'interdisciplinarité dans la position de recherche prise ici. Il s'agit moins d'un travail en Géographie ou en Modélisation de Systèmes Complexes Adaptatifs, pouvant difficilement être vraiment les deux à la fois, mais en *Science des Systèmes Complexes* que nous réclamons discipline propre comme le propose PAUL BOURGINE. **C (Florent) : pas vraiment fondateur de la discipline A1 : non mais du point de vue particulier que nous défendons - théories intégratives roadmap etc. - trouver une ref là dessus ?** 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 [DUPUY et BENGUIGUI, 2015]. Il faut se rappeler l'importance de la spirale vertueuse de BANOS entre disciplinarité et interdisciplinarité [BANOS, 2013]. Celle-ci doit nécessairement impliquer différents agents scientifiques, et il est compliqué pour un agent de se positionner dans les deux branches ; notre fond scientifique ne nous permet pas de nous positionner dans la *disciplinarité géographique* **C (AB) : je trouve que tu abandonnes bien vite la partie. Pourquoi renoncer dès le début alors que tu pourrais laisser la liberté à celui/celle qui te lit de décider si c'est AUSSI de la géo ou non ?** mais bien 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.

Le positionnement de BATTY lorsqu'il propose *Une Nouvelle Science des Villes* [BATTY, 2013b] (qu'il présente avec humour comme *La nouvelle science des villes*), se présente comme une intégration des disciplines et méthodes vers une science définie par son objet d'étude, les villes. Its theoretical and epistemological weaknesses (no theoretical constructions of studied geographical objects on the one hand, approximative contextualization of complexity) combined with an overall impression of *pot-pourri* of forgotten works (space syntax, land-use models), unfortunately avoid us to use it as we will use geographical theories (e.g. evolutive urban theory) in an appropriated epistemological complexity context. Yet our reading of this work may be the result of a misunderstanding due to different cultural backgrounds. **C (Arnaud) : j'espère que tu abuses ? :)!! Argument d'autorité A1 : yes, changer positionnement complètement malvenu C (Florent) : attention arguments autorité; insister sur difficulté à intégrer paradigmes plutot que juger précédents A1 : idem**

L'évolution scientifique des sciences de la complexité, qui est vue par certains comme une révolution [COLANDER, 2003], ou même comme *un nouveau type de science* [WOLFRAM, 2002], pourrait affronter des difficultés intrinsèques dues aux comportements et a-priori des chercheurs en tant qu'être humains. Plus précisément, le besoin d'interdis-

ciplinarité 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, competition, 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 [AGHION et al., 2015], et dont les conclusions ont été commentées comme s'opposant aux thèses de Piketty dans [PIKETTY, 2013], est typique de ce schéma. Alors que Piketty 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 sur-simplifié. D'autre part, Bergeaud *et al.* montrent par un modèle d'économie de l'innovation que *sous certaines hypothèses* les écarts de revenus peuvent être bénéfique à l'innovation et donc à une utilité globale. **C (JR) : revoir (et cité plus loin)** D'où des conclusions divergentes sur le rôles des capitaux personnels dans une économie. 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, [HOLMES et al., 2017] 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.

C (AB) : tout ceci remet en question ton « absence d'ambition » me semble-t-il

Comme déjà mentionné, DUPUY et BENGUIGUI soulignent dans [DUPUY et BENGUIGUI, 2015] 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, [BARTHELEMY et al., 2013] 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. [CHEN, 2009] 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 [O'SULLIVAN et MANSON, 2015] mais la route semble être bien longue encore.

Des conflits similaires se rencontrent en économie : comme décrit par [MARCHIONNI, 2004], la discipline de l'économie géographique, traditionnellement proche de la géographie, a fortement critiqué un nouveau courant de pensé nommé *économie géographisée*, C : (Arnaud) New economic geography? dont le but est la prise en compte de l'espace par les techniques économiques classiques. Elles n'ont pas les mêmes desseins et buts, et le conflit apparaît comme un malentendu complet vu d'un oeil extérieur.

C : (Florent) a développer ou ne pas en parler, un peu loin du cœur du sujet tel que abordé

Des conflits disciplinaires peuvent aussi se manifester sous la forme d'un rejet de méthodes nouvelles par les courants dominants. Suivant FARMER [FARMER et FOLEY, 2009], 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 semble rebu-

ter la plupart des économistes. Or, [BARTHELEMY, 2016] insiste sur la déconnexion exacerbée entre une grande partie des modèles et théories économiques et la réalité, du moins dans le domaine de l'économie urbaine. Celle-ci pourrait être un symptôme de la déconnexion disciplinaire évoquée ci-dessus.

C : (Arnaud) Difficile de se positionner de manière crédible sur ces sujets en 5 lignes et 1 référence !

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 [BARNDORFF-NIELSEN et al., 2011]). D'autre part, l'éconophysique a pour but d'étudier des faits stylisés empiriques et inférer les lois correspondantes pour tenter d'expliquer les phénomènes liés à la complexité des marchés financiers [STANLEY et al., 1999], comme par exemple 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 introductory, 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, géogra-

phie humaine, géographie de la santé, etc. pour en nommer quelques) 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 [PUMAIN, 2003] une histoire subjective de l'émergence des paradigmes de la complexité en géographie. La cybernétique a produit des théories des systèmes comme celle utilisée pour les premiers modèles systémiques visant à . **C : (Florent)**
pas dvlpé, difficile à lire 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 [SANDERS, 1992] 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. Par exemple, l'étude de la nature fractale de la forme urbaine a été introduite par [BATTY et LONGLEY, 1986], plus tard synthétisée par [BATTY et LONGLEY, 1994] et a eu de nombreuses applications jusqu'à des développements plus récents comme [KEERSMAECKER, FRANKHAUSER et THOMAS, 2003]. Les automates cellulaires, introduit en géographie par TOBLER [COUCLELIS, 1985], 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 [BATTY, 2007]. Relativement récemment, la Théorie du *Scaling* a été importée 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 [BETTENCOURT et al., 2007]. Une autre introduction de la complexité en géographie fut pour le cas des systèmes urbains à travers la théorie évolutive des villes de PUMAIN. En interaction intime avec la modélisation dès ses débuts (le premier modèle Simpop décrit par [SANDERS et al., 1997] rentre dans le cadre théorique de [PUMAIN, 1997]), 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 [PUMAIN, 2012a] 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 [BRETAGNOLLE et PUMAIN, 2010a]. A d'autres échelles de temps et dans d'autres contextes, le modèle SimpopLocal [SCHMITT, 2014] a pour but d'étudier les conditions pour l'émergence de systèmes urbains hiérarchiques à partir d'établissements disparates. Un modèle mi-

nimal (au sens de paramètres nécessaires et suffisants) a été isolé grâce à l'utilisation de calcul intensif via le logiciel d'exploration de modèles OpenMole [SCHMITT et al., 2014], ce qui était un résultat impossible à atteindre de manière analytique pour un tel type de modèle complexe. Les progrès techniques d'OpenMole [REUILLO, LECLAIRE et REY-COYREHOURCQ, 2013] 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 [REY-COYREHOURCQ, 2015], et de nouveaux concepts comme la modélisation incrémentale [COTTINEAU, CHAPRON et REUILLO, 2015] ont été découverts, avec de puissantes applications concrètes : [COTTINEAU, 2014] 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 [BANOS, 2013]. Il conclut par neuf principes⁵, 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 reproducibilité au centre de l'étude des systèmes complexes géographiques selon le point de vue décrit précédemment. Nous nous positionnons 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. 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. [ROBIC, 1982] 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 l'incubateur du changement social et de l'innovation [PUMAIN, 2010]. La définition concrète d'une ville en terme d'entités géographique et d'étendue spatiale est également problématique : des définitions morphologiques (c'est à 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

⁵ Je me rappelle RENÉ DOURSAT insister pour la recherche du dernier commandement de BANOS

par aire de déplacement domicile-travail dominant), administratives, etc., sont partiellement orthogonales et plus ou moins adaptées au problème étudié [GUÉROIS et PAULUS, 2002]. Récemment, un certain nombre d'études ont montré la forte sensibilité des lois d'échelles urbaines⁶ aux délimitations choisies pour l'estimation, pouvant entraîner une inversion des propriétés qualitatives attendues (voir par exemple [ARCAUTE et al., 2015]). Les variations des exposants estimés en fonction de paramètres de définition, comme effectué par [COTTINEAU et al., 2015], 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*⁷. Un système urbain peut être considéré comme un ensemble de villes en interaction, dont les dynamiques seront plus ou moins fortement couplées. [BERRY, 1964] 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 Evolutive des Villes détaillée précédemment (voir aussi 9.2). 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, permet dans une première définition simple, de capturer la distribution spatiale des activités urbaines, des agents les exerçant ou les développant, et des artefacts techniques, dont l'infrastructure, les supportant.

PROBLÉMATIQUE

Une caractéristique fondamentale des systèmes urbains et des territoires est leur nature spatio-temporelle, qui transparaît dans leur dynamiques spatio-temporelles. La notion de *processus* au sens de [*Hypergeo*], c'est à dire l'enchaînement dynamique de faits aux propriétés causales⁸, 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

⁶ 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 $Y_i = Y_0 \cdot (P_i/P_0)^\alpha$.

⁷ 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.

⁸ Nous prendrons la causalité au sens de causalité circulaire dans les systèmes complexes, qui considère des cycles d'entraînement 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

au choix d'*échelles*, qui doit être comprise ici au sens opérationnel (caractéristiques physiques), et d'une *ontologie* qui correspond aux objets réels étudiés⁹. Une ontologie particulière retiendra notre attention : au sein des territoires émergent des *Réseaux Physiques*, qui selon peuvent être compris selon [DUPUY, 1987] 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 [OFFNER, 1993] 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 *dans quelle mesure est-il possible de faire correspondre des dynamiques territoriales à une évolution de l'infrastructure de transport ?* On peut poser la question de manière réciproque, et plus généralement 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 couplés¹⁰ d'évolution des villes comme utilisé par [PAULUS, 2004], et appliqué aux relations entre réseaux et villes par [BRETAGNOLLE, 2009a]. Elle 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, qui elle correspond à des relations circulaires causales difficiles à séparer.

⁹ cet usage de la notion d'*ontologie* biaise naturellement la recherche vers des paradigmes de modélisation puisque qu'elle est proche de celle utilisée dans [LIVET et al., 2010a], 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.

¹⁰ 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 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.

Nous ferons ici le pari de challenger cette notion, pour éclaircir ses apports potentiels pour la compréhension des interactions entre réseaux et territoires. 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 développé ci-dessous, postulant 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. Ce positionnement fera l'objet d'un développement plus approfondi en [3.1](#).

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 clarifier le positionnement au regard de la modélisation de la co-évolution. 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 co-évolution 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 bâtissons dans la troisième partie sur les fondations et avec les briques élémentaires notre construction fondamen-

tale, qui consiste en différents modèles (ou famille de modèles) de co-évolution, que nous différencions selon les deux approches considérées :

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.

★ ★

★

Première partie

FOUNDATIONS

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 concrètes importantes.

INTRODUCTION DE LA PARTIE I

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 [DIDEROT, 1965] est éclairante : “*Si la question de la priorité de l'œuf sur la poule ou de la poule sur l'œuf vous embarrasse, 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 réciproquement, est-il possible et même pertinent de vouloir isoler les processus correspondants ? Cette question est bien connue des planificateurs des transports, comme le rappelle la notion des “effets structurants” qui fait débat depuis un certain temps au moins dans la communauté scientifique **C (FL) : non cela va trop vite** . Une vision simplifiée, selon laquelle on peut attribuer des rôles systématiques à une composante particulière, est souvent présente dans les approches réductionnistes qui ne postulent pas une complexité intrinsèque au sein des systèmes étudiés. **C (FL) : mots pas clairs; a la place : amener co-évolution** L'idée suggérée par DIDEROT est celle de *co-évolution* qui est un phénomène central dans les dynamiques évolutionnaires des Systèmes Complexes Adaptatifs. [HOLLAND, 2012] en propose une théorie. Il fait le lien entre l'émergence de structures à une échelle supérieure avec les interactions entre agents à une échelle donnée, en général concrétisée par un systèmes de limites **C (FL) : pas clair** , qui devient cruciale pour la co-évolution des agents à toutes les échelles : l'émergence d'une structure sera simultanée avec une autre, chacune exploitant leurs interrelations et environnements générés conditionnés par le système de limites. **C (FL) : rupture : le fil n'est pas clair** Nous explorerons ces idées pour le cas des systèmes territoriaux par la suite. Ceux-ci illustrent parfaitement ces problématiques, et sont typiques de systèmes dans lesquels cette complexité **C (FL) : laquelle** est cruciale pour une appréhension raisonnable des mécanismes impliqués dans leurs dynamiques. Un certain nombre d'illustrations concrètes **C (FL) : de quoi?** seront d'abord proposées pour formuler nos questionnements dans des contextes géographiques donnés. **C (FL) : phrase inutile**

Ce chapitre introductif est destiné à poser le cadre thématique, les contextes géographiques sur lesquels les développements suivants se baseront. **C (FL) : plus haut** Il n'est pas supposé être compris

comme une revue de littérature exhaustive ni comme les fondations théoriques fondamentales de notre travail, le premier point étant l'objet du chapitre 2 tandis que le second sera traité systématiquement dans le chapitre 9 lorsque le recul nécessaire aura été progressivement construit. Il doit plutôt être lu comme une construction narrative ayant pour but d'introduire nos objets et positions d'étude. **C (FL) : cela sera dans l'intro generale** La notion de co-évolution est particulièrement pertinente **C (FL) : oui : a remonter** pour comprendre les interactions entre territoires et réseaux. 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. Ces considérations abstraites seront illustrées par des cas d'étude concrets dans la deuxième section 1.2, choisis très différents pour comprendre les enjeux d'universalité sous-jacents. Enfin, dans la troisième section 1.3, des éléments d'observation de terrain effectués en Chine préciseront encore ces exemples aux échelles microscopique et mesoscopique. **C (FL) : reprendre**

* * *

*

Ce chapitre est entièrement inédit.

1.1 TERRITOIRES ET RÉSEAUX

1.1.1 Territoires et Réseaux : There and Back Again

TERRITOIRES 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 [TILMAN et KAREIVA, 1997], qu'on peut identifier comme notion de territoire. Les *Territoires Humains* impliquent des dimensions supplémentaires, par exemple par l'importance de leur représentations sémiotiques. Celles-ci jouent un rôle significatif dans l'émergence des constructions sociétales, dont la genèse est profondément liée à celle des systèmes urbains. Selon [RAFFESTIN, 1988], 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. Une illustration concrète est donnée par une vision simplifiée de la construction de la Métropole du Grand Paris, qui témoigne des ajustements successifs des territoires administratifs (émergence d'un nouveau niveau de gouvernance), des territoires fonctionnels (partiellement par l'évolution des possibilités d'accessibilité), des territoires perçus (dépassement de l'opposition Paris-banlieue), des territoires vécus (nouvelles pratiques de mobilité ou de mobilité résidentielle potentiellement induites par les dynamiques territoriales et du nouveau réseau de transport), l'ensemble de ces processus étant liés de manière complexe et ceux-ci étant loin d'être systématiques.

Bien que cette approche ne donne pas de conditions explicites pour l'émergence d'un système séminal d'établissements agrégés (c'est à dire l'émergence des villes, qui est une question centrale pour des théories crédibles des systèmes urbains), elle insiste sur leur rôle comme lieu de pouvoir et de création de richesse au travers des échanges. **C (FL) : tous ces termes doivent être définis** Mais la

¹ Nous utiliserons le terme *concept* pour des connaissances construites, plutôt que celui de *notion*, qui suivant [RAFFESTIN, 1978] est plus proche d'une information empirique. Cette distinction peut être mise en perspective avec les domaines de connaissance théorique et empirique de [LIVET et al., 2010a], que nous approfondissons en 9.3.

ville n'a pas d'existence sans son hinterland et le système territorial peut difficilement être résumé par ses villes, comme un système de villes. **C (FL) : question sous-jacente : quel est le territoire d'une ville ? merite un développement spécifique** En se restreignant à ce sous-système, il y a toutefois compatibilité entre la théorie des territoires de RAFFESTIN et la théorie évolutive des villes de PUMAIN [PUMAIN, 2010], **C (FL) : dans cette section on ne sait pas pourquoi tu parles de cela** qui interprète les villes comme des systèmes complexes dynamiques 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 **C (FL) : je ne vois pas ce que cela apporte : va moins vite** (voir ?? pour une entrée sur la notion d'innovation). Les villes sont ainsi des agents compétitifs qui co-évoluent (au sens donné précédemment). **C (FL) : trop affirmatif c'est une grille de lecture possible** Le système territorial peut ainsi être compris comme une structure sociale organisée dans l'espace, qui comprend ses artefacts concrets et abstraits. Une étendue spatiale imaginaire avec des ressources potentielles qui n'aurait jamais connu de contact avec l'humain ne pourra pas être un territoire si elle n'est pas habitée, imaginée, vécue, exploitée, même si ces ressources pourraient être potentiellement exploitée le cas échéant. **C (FL) : tu as déjà défini territoire → ne pas y revenir** En effet, ce qui est considéré comme une ressource (naturelle ou artificielle) dépendra de la société (par exemple de ses pratiques et de ses capacités technologiques).

Ces visions complémentaires du territoire peuvent enrichies par une perspective historique. [DI MEO, 1998] procède à une analyse historique des différentes conceptions de l'espace (qui aboutissent entre autre 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. [GIRAUT, 2008] 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'avantage d'un objet interdisciplinaire capturant une certaine complexité des systèmes étudiés, ce qui confirme la pertinence de la notion dans notre cas.

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 inévitable.

RÉSEAUX Un réseau doit être compris au sens très large de motifs de connectivité entre entités d'un système, qui peuvent être vus comme relations, liens, interactions. [HAGGETT et CHORLEY, 1970] pos-

tule que l'existence d'un réseau est nécessairement liée à celle de flux, 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. L'analyse topologique révèle déjà un certain nombre de propriétés du système, mais [HAGGETT et CHORLEY, 1970] 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 [BARTHÉLEMY, 2011] 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 de diffusion).

Pour approfondir le concept de réseau en appuyant sur sa forte interdépendance avec celui de Territoire, nous reprenons [DUPUY, 1987] 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* (auxquels appartiennent les réseaux concrets, incluant les réseaux matériels et donc les réseaux de transport, les réseaux sociaux étant également des réseaux réels sur lesquels nous ne nous attarderons pas) et les *réseaux virtuels*, eux-même 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'interactions² potentielles entre les éléments du système territorial, notamment des agents et des ressources. [DUPUY, 1987] désigne ces interactions potentielles comme *projets transactionnels*. Cela justifie de manière thématique l'utilisation de modèles de potentiels **C (FL) : de quoi parles tu ? c'est enigmatique** pour capturer les interactions entre agents **C (FL) : c'est la premiere fois que tu parles d'interaction, c'est dommage que cela arrive comme cela**, que ce soit à l'échelle des villes ou au sein de celles-ci. Par exemple, de nos jours les actifs se doivent d'accéder à la ressource qu'est l'emploi, et des échanges économiques s'effectuent entre les différents territoires spécialisés dans les productions de différents types. Une distribution spatiale d'agents suffit à introduire des interactions potentielles.

Le réseau d'interaction potentiel est concrétisé quand l'offre s'adapte à la demande **C (AB) : me semble une generalite trop grande**, et résulte en la combinaison de contraintes économiques et géographiques avec les motifs de demande, de manière non-linéaire via des agents qu'on peut désigner comme *opérateurs* **C (FL) : c'est theorique a ce stade → pourquoi ce vocabulaire?**. Un tel processus est

²

loin d'être immédiat, et conduit à de forts effets de non-stationnarité et de dépendance au chemin **C (AB) : definir les 2 concepts** : l'extension d'un réseau existant **C (FL) : la tu es déjà dans le rai-sonnement de ta these mais le lecteur ne le sait pas** dépendra de la configuration précédente, et selon les échelles de temps impliquées, la logique et même la nature des opérateurs peut avoir évolué. Les exemples de trajectoires concrètes peuvent être très variées : [KASRAIAN, MAAT et WEE, 2015] montre par exemple dans le cas de la Randstad sur le temps long, un premier régime d'adaptation **C (FL) : sens ?** du réseau ferré au développement urbain suivi par des effets inverses plus récemment. A une échelle urbaine sur le temps long, la dépendance au chemin **C (FL) : sens ?** est montrée pour Boston par [BLOCK-SCHACHTER, 2012] puisque l'environnement bâti et la distribution de la population sont montrés fortement dépendants des lignes de tramway passé même lorsqu'elles n'existent plus. RAFFESTIN souligne dans sa préface de [OFFNER et PUMAIN, 1996] qu'une théorie géographique articulant espaces, réseaux et territoires n'a jamais été formulée de manière cohérente **C (FL) : plutot dans la conclusion generale**, chaque approche ayant une vision réduite à certaines composantes seulement et ne visant pas à construire une théorie globalement cohérente. Il semble que c'est toujours le cas aujourd'hui, même si la théorie évoquée ci-dessus semble être un bon candidat bien qu'elle reste à un niveau conceptuel. La présence d'un territoire humain implique nécessairement **C (FL) : non pas d'accord mais surtout : pourquoi est-ce important ?** la présence de réseaux d'interactions abstraites et de réseaux concrets utilisés pour transporter les individus et les ressources (incluant les réseaux de communication puisque l'information est une ressource essentielle).

Le statut du réseau par rapport au territoire est fortement conditionné par le contexte socio-économique et technologique. Selon DURANTON [DURANTON, 1999], un facteur influençant la forme des villes pré-industrielles était la performance des réseaux de transport. Les progrès technologiques 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. On s'attendra ainsi à l'existence de multiples processus d'interaction, potentiellement superposables de manière complexe.

Cette approche territoriale des réseaux semble naturelle en géographie, puisque les réseaux sont étudiés conjointement avec des objets géographiques qu'ils connectent, en opposition à la science des réseaux qui étudie les réseaux spatiaux de manière relativement décon-

nectée de leur fond thématique [DUCRUET et BEAUGUITTE, 2014]. C (AB) : pas d'accord → vrai pour sa composante *statistique* mais pas dans ses applications en sociologie, archi, histoire ou geo par exemple
⇒ pas une question de "science des réseaux" donc

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. C (FL) : B - en parler avant pour fixer les idées Dans le cas des *réseaux techniques*, une autre désignation des réseaux réels C (AB) : bof donnée dans [OFFNER et PUMAIN, 1996], 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 des réseaux de transport permet des motifs de mobilité multi-échelles formant ainsi le territoire vécu. A une plus petite échelle, des changements de l'accessibilité peuvent induire l'adaptation d'un espace fonctionnel urbain. 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. C (FL) : c'est le cœur de ta these : mieux mettre en valeur Revenir à la citation de Diderot C (FL) : laquelle ? 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 mutuellement par des relations causales C (FL) : sens ?, mais comme des composantes fortement couplées C (FL) : sens ? d'un système plus large. La confusion autour de possibles relations causales simples a nourri un débat scientifique encore actif aujourd'hui. Les méthodologies pour identifier ce qui est nommé *effets structurants* des réseaux de transport ont été proposées par les planificateurs dans les années 1970 [BONNAFOUS et PLASSARD, 1974; BONNAFOUS, PLASSARD et SOUM, 1974]. Il aura fallu un certain temps pour un positionnement critique sur l'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. C (FL) : c'est elliptique. je crois voir ce que tu veux dire mais il faut ancrer davantage ton propos Cela a été fait en premier par OFFNER dans [OFFNER, 1993]. Une édition spéciale relativement récente de l'Espace Géographique sur ce débat [L'ESPACE GÉOGRAPHIQUE, 2014] a rappelé d'une part que les mauvaises interprétations et les mauvais usages C (FL) : la encore : il faut du concret étaient encore largement présents aujourd'hui dans les milieux opérationnels de la planification comme confirmé par [CROZET et DUMONT, 2011a], et d'autre part qu'il faudrait encore une certaine quantité de progrès C (FL) : formulation curieuse scientifiques pour comprendre en pro-

fondeur les relations entre réseaux et territoires. Les débats récents 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 : spéculation dans les quartiers de gare, questionnements des pratiques de mobilité quotidienne mais aussi sociale³. La complexité et la portée des sujets montrent bien la difficulté d'une compréhension systématique d'effets du transport sur les territoires. PUMAIN souligne que des travaux récents ont révélé des effets systématiques **C (FL) : def** sur de très longues échelles temporelles, comme par exemple le travail de BRETAGNOLLE sur l'évolution des chemins de fer, qui montre une sorte d'effet structurel **C (FL) : est ce different des effets structurants** sur la nécessité de connexion au réseau des villes, afin de rester actives, mais qui n'est ni suffisant ni totalement causal. Certaines trajectoires de villes correspondent à un renforcement de la hiérarchie par une accessibilité accrue, tandis que des villes non connectées seront a priori hors-jeu pour une période considérable **C (FL) : c'est caricatural : etre pedagogique ce n'est pas etre emphatique**, avec de grandes fluctuations pouvant conduire à une relative indépendance du taux de croissance **C (FL) : de quoi?** et de l'augmentation d'accessibilité pour certains cas.

A un niveau macroscopique **C (AB) : pas clair // petite echelle** des motifs typiques d'interaction émergent, mais les trajectoires microscopiques du systèmes sont essentiellement chaotiques **C (AB) : meilleur terme?** : la compréhension des dynamiques couplées dépend fortement de l'échelle considérée. A une petite échelle il est difficile de s'attendre à identifier facilement des comportements systématiques **C (FL) : sens?**, comme le rappelle OFFNER. Par exemple, sur des territoires de montagne français comparables, [BERNE, 2008] 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 par la nouvelle connectivité **C (FL) : sens**, d'autres au contraire devenant plus fermés. Ces potentiels processus antagonistes sont examinés plus en détails par [BERNIER, 2007] toujours dans le cadre des territoires de montagne **C (FL) : tu en parles deux fois : organiser differemment**, 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 avoir un fort potentiel touristique, et donc d'ouverture, mais avoir une faible accessibilité. Au contraire, des contraintes douanières **C (FL) : tu prends des exemples trop variés, au risque d'etre imprecis sur la plupart** par exemple peuvent freiner le potentiel d'ouverture d'une

³ voir http://www.libération.fr/futurs/2017/07/02/immobilier-plus-de-parisiens-comment-les-bordelais-voient-l-arrivée-de-la-lgv_1580776

infrastructure performante. Ces retroactions potentielles des réseaux sur les territoires n'agissent pas nécessairement sur des composantes concrètes : CLAVAL montre dans [CLAVAL, 1987] 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. A l'échelle mesoscopique, on peut montrer des effets forts de la présence d'infrastructures pour des types particuliers d'usage du sol : [NILSSON et SMIRNOV, 2016] l'illustre par exemple pour les fast food dans deux villes aux Etats-Unis. **C (FL) :** en montrant quoi ?

C (AB) : je pense que la lecture Peter Hagget et de Peter Gould te fera pas de mal → j'essaye de les retrouver :)

SYSTÈMES TERRITORIAUX Cet aperçu introductif, des territoires aux réseaux, nous permet de clarifier notre approche des systèmes territoriaux qui sera sous-jacente dans l'ensemble de la suite. 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 auquel peuvent être associés **C (FL) : ok mais il faut etre plus constructif** à la fois des réseaux d'interactions et des réseaux réels **C (AB) : mieux definir. reeels → concret**. Les réseaux réels, et plus particulièrement les réseaux concrets, 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. Cette lecture des systèmes territoriaux est conditionnée à l'existence des réseaux et pourrait écarter certains territoires humains, mais il s'agit d'un choix délibéré justifié par les considérations précédentes, et qui précise notre sujet vers l'étude des interactions entre réseaux et territoires. Le réseau n'est pas une composante en tant que telle du territoire, mais bien du système territorial en notre sens. **C (FL) : pourquoi opposer? on peut dire que nw ∈ territoires et aussi ∈ systeme territorial. plus largement, je ne vois pas encore en quoi cette distinction va t'etre utile.** Il faut aussi garder à l'esprit que le transport en lui-même est différent des réseaux de transport **C (FL) : en quoi estce un argument?**, puisqu'il correspond à l'utilisation de ceux-ci par les agents territoriaux. Dans une grande partie des approches que nous décrirons par la suite, et typiquement les approches appliquée en planification urbaine, la modélisation du transport s'axe sur des question de demande, d'offre, de congestion, c'est à dire à des échelles relatives à la mobilité, et est liée au réseau mais ne se concentre pas directement sur celui-ci **C (FL) : ce n'est pas clair → la croissane du reseau?, l'usage du reseau? la croissance**

de l'usage du réseau ? comme notre positionnement propose C
(AB) : ≈

1.1.2 Réseaux de Transport

LA PARTICULARITÉ DES RÉSEAUX DE TRANSPORT Déjà évoqués dans le cas des effets structurants des réseaux, les réseaux de transports jouent un rôle central dans l'évolution des territoires, mais il n'est évidemment pas question de leur attribuer des effets causaux déterministes. Même si d'autres types de réseaux sont également fortement impliqués dans l'évolution des systèmes territoriaux (voir e.g. 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 semblent dominer C (FL) : question très controversée dans les motifs d'évolution territoriale, en particulier dans nos sociétés contemporaines pour lesquelles les réseaux de transport jouent un rôle privilégié [BAVOUX et al., 2005]. Le développement du réseau français à grande vitesse est une illustration de l'impact des réseaux de transport sur les politiques de développement territorial. Présenté comme une nouvelle ère de transport sur rail, une planification par le haut C (FL) : sens ? de lignes totalement nouvelles et indépendantes C (FL) : non elles ne le sont pas vraiment de par leur vitesse deux fois plus élevée, a été défendu par les acteurs politiques entre autres comme central pour le développement [ZEMBRI, 1997]. Le manque d'intégration C (FL) : sens de ces nouveaux réseaux avec l'existant et avec les territoires locaux est à présent observé comme une faiblesse structurelle C (FL) : sens et des impacts négatifs sur certains territoires C (FL) : de quel type ? être plus explicite ont été prouvés [ZEMBRI, 2008]. Une revue faite dans [BAZIN et al., 2011] 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 élus. Ces exemples illustrent la manière dont les réseaux de transport peuvent avoir des effets à la fois directs et indirects sur les dynamiques territoriales. Le développement des différentes Lignes à Grande Vitesse s'inscrit dans des contextes territoriaux très différents, et il est dans tous les cas délicat de penser pouvoir interpréter des processus hors de ceux-ci C (FL) : mal tourne mais oui c'est une remarque importante : par exemple, les lignes LGV Nord et LGV Est s'inscrivent dans des échelles européennes plus vastes que la LGV Bretagne ouverte en juillet 2017. Les effets de l'ouverture d'une ligne peuvent s'étendre au delà des seuls territoires directement concernés : [L'HOSTIS, LEYSENS et LIU, 2014] montre par l'utilisation d'indicateurs issus de la *Time Geography* (mesurant une quantité de temps de tra-

vail 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. La planification intégrée, au sens d'une planification coordonnée entre les infrastructures de transport et le développement urbain, considère le réseau comme une composante déterminante du système territorial. Les Villes Nouvelles parisiennes sont un tel cas qui témoigne de la complexité de ces actions de planification qui le plus souvent ne mènent pas au effets initialement désirés [OSTROWETSKY, 2004]. Des projets récents comme [L'HOSTIS, SOULAS et WULFHORST, 2012] ont tenté d'implémenter des idées similaires, mais le recul manque encore pour juger de leur succès à produire un territoire effectivement intégré. On sait que sur des échelles de temps relativement courtes allant de l'année à la dizaine d'année, les effets observés sur les mobilités quotidiennes et mobilité résidentielles peuvent être significatifs. **C (FL) : c'est ce genre de choses qu'il faut creuser** Les réseaux de transports sont dans tous les cas au centre de ces approches des territoires urbains. Nous nous concentrerons par la suite sur les réseaux de transport de manière générale pour toutes ces raisons évoquées ici.

TRANSPORTS ET ACCESSIBILITÉ La notion d'accessibilité émerge naturellement **C (AB) : rapidement?** lorsqu'on s'intéresse aux réseaux de transport. **C (FL) : phrase toute faite** 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 spatiale⁴ [BAVOUX et al., 2005]. Une approche axiomatique **C (FL) : sens** a été proposée par [MILLER, 1999], en mettant en valeur trois façons de comprendre l'accessibilité, basées sur la *Time Geography* et les contraintes, les mesures d'utilité basée sur l'utilisateur, et le temps de transport **C (AB) : (...)**, les mesures correspondantes étant dérivées dans un cadre mathématique uniifié **C (FL) : et alors?**. 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 impactées par un projet de transport. Il faut cependant rester prudent sur son usage inconditionnel. Plus précisément, il peut s'agir d'une construction qui ignore une partie conséquente des dynamiques territoriales. La co-construction de la notion de *mobilité* **C (FL) : def** et des solutions techniques impliquant les solutions de modélisation mais aussi la production de l'infrastructure, a été montrée par COMMENGES dans [COMMENGES, 2013b] pour le contexte français **C (FL) : trop rapide**. Il révèle qu'une partie des débats sur la modélisation

⁴ et souvent généralisée comme une *accessibilité fonctionnelle*, par exemple les emplois accessibles aux actifs d'un lieu. Les potentiels d'interaction spatiaux **C (AB) : spatial/spatiaux** s'exprimant dans les lois de gravité **C (AB) : gravitaires?** peuvent aussi être compris de cette façon.

de la mobilité et les notions correspondantes étaient majoritairement construites de manière ad-hoc par les administrateurs de transports issus du *Corps des Ponts* qui importaient les outils et méthodes des Etats-Unis sans mener de réflexion approfondie pour leur adaptation au contexte français. L'accessibilité pourrait **C (AB) : conditionnel ambigu : préciser** de même être une construction sociale et n'avoir que peu de fondement théorique, puisqu'il s'agit en grande partie d'un outil de modélisation et de planning. Les débats récents sur la planification du *Grand Paris Express* [MANGIN, 2014], cette nouvelle infrastructure de transport métropolitaine planifiée pour les vingt prochaines années, a révélé l'opposition entre une vision de l'accessibilité comme un droit pour les territoires désavantagés, contre l'accessibilité comme un moteur du développement économique pour des zones déjà dynamiques, les deux étant difficilement compatibles car correspondent à des corridors de transport très différents : l'un initialement porté par l'Etat dans la perspective des pôles de compétitivité, l'autre par la région dans une perspective d'équité territoriale. De tels problèmes opérationnels confirment la complexité du rôle des réseaux de transports dans les dynamiques des systèmes territoriaux, et nous devrons donner dans notre travail des éléments de réponse pour une définition de l'accessibilité qui intégrerait les dynamiques territoriales intrinsèques.

ECHELLES ET HIERARCHIES Un aspect des réseaux de transport qu'il est important de considérer est la notion de hiérarchie. Les réseaux de transport sont par essence hiérarchiques, cette propriété dépendant des échelles dans lesquelles ils sont intégrés. [LOUF, ROTH et BARTHELEMY, 2014] montre empiriquement des propriétés de loi d'échelle pour un nombre conséquent d'aires métropolitaines à travers la planète, et 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 [NITSCH, 2005] ou d'autres lois d'échelle urbaines [ARCAUTE et al., 2013; BETTENCOURT et LOBO, 2015]. La topologie du réseau de transport suit de telles lois pour la distribution de ses mesures locales comme la centralité [SAMANIEGO et MOSES, 2008], celles-ci étant directement liées au motifs d'accessibilité à différentes échelles : cette notion est donc nécessaire pour les comprendre, mais induit aussi des choix d'échelles pour préciser la définition de ceux-ci. De plus, la topologie du réseau fait partie des facteurs induisant la hiérarchie d'usage, se retrouvant dans les externalités négative de congestion, en relation avec la distribution spatiale de l'usage du sol [TSEKERIS et GEROLIMINIS, 2013]. La hiérarchie joue un rôle particulier dans les processus d'interaction. [BRETAGNOLLE, 2009a] souligne ainsi 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 i.e. 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 [BRETAGNOLLE, PUMAIN et ROZENBLAT, 1998]. 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 sans effet systématique toujours

C (AB) : reformuler : ce contexte particulier de conquête d'un espace vierge d'infrastructures implique un régime particulier au système territorial. Cela met l'emphase sur la présence de dépendances au chemin pour les trajectoires des systèmes urbains, que nous retrouverons régulièrement par la suite : la présence en France d'un système préalable de villes et de réseau (routes postales) a fortement influencé le développement du réseau ferré, tandis que son absence aux Etats-unis a conduit à une histoire complètement différente. La question reste ouverte si **C (FL) : reformuler** des processus génériques sont implicites aux deux évolutions, chacun correspondant à des réalisations différentes avec des conditions initiales et des métaparamètres différentes, c'est à des *régimes* différents au sens des transitions des systèmes de peuplement, puisqu'une transition entre deux régimes peut être comprise comme un changement de stationnarité des métaparamètres d'une dynamique plus générale. En termes de systèmes dynamiques, cela revient à se demander si les dynamiques des ensembles de catastrophes (composantes à plus grandes échelles temporelles) obéissent à des équations similaires à la position et nature des attracteurs pour un système dynamique stochastique qui donne son régime courant, en particulier si le système est dans un état local divergent (exposant de Liapounov local positif) ou en train de converger vers des mécanismes stables [SANDERS, 1992]. Pour répondre à cette question en même temps que l'isolation des processus de coévolution pour ce régime, [BRETAGNOLLE, 2009a] propose la modélisation comme élément de réponse constructif. Nous verrons dans le chapitre suivant comme la modélisation peut être source de connaissance sur les processus territoriaux.

C (AB) : AERER

TRANSPORTS ET MOBILITÉ La notion de mobilité et l'ensemble des approches associées, peuvent capturer nos questionnements à échelle fine : 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. [FUSCO, 2004] révèle par exemple une relation causale de la mobilité sur la structure urbaine, l'offre d'infrastructure et ses propriétés ayant cependant des effets joints à la fois sur la mobilité et sur la structure urbaine. Dans le cas des réseaux autoroutiers, [FAIVRE, 2003] 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.

1.1.3 *Interactions entre Réseaux et Territoires*

A ce stade, nous avons identifié que les processus d'interaction entre réseaux de transport et territoires jouent un rôle significatif dans la complexité des systèmes territoriaux. Dans le cadre de l'approche d'un système territorial par la définition donnée ci-dessus, 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 spatiale et temporelle 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. Donnons à présent des exemples concrets clarifiant la complexité des interactions et la nécessité de considérer une coévolution.

DIVERSITÉ DES INTERACTIONS [HEDDEBAUT et ERNECQ, 2016] montre pour l'impact des infrastructures sur le long terme, dans le cas du tunnel sous la Manche, 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 les retombées économiques directes locales se sont rapidement estompées : on rejouit l'idée défendue par BRETAGNOLLE dans [L'ESPACE GÉOGRAPHIQUE, 2014] **C (AB) : cit?** selon laquelle des "effets de structure" effectivement existent mais que ceux-ci se manifestent sur le temps long en terme de dynamiques systémiques pour lesquelles une vision locale courte n'a que peu de sens. 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. A l'échelle intra-urbaine, [FRITSCH, 2007] prend l'exemple du Tramway de Nantes pour montrer que les dynamiques de densification urbaines sont bien en dessous des anticipations des élus et planificateurs. [KASRAIAN et al., 2016] procède à une revue systématiques des études empiriques des impacts à moyen terme des infrastructures de transport, et montre qu'une densification urbaine à proximité des nouvelles infrastructures est relativement systématique, résidentielle dans le cas d'une infrastructure ferroviaire et pour les emplois et l'activité industrielle et commerciale pour le réseau routier. Pour reprendre l'exemple bien particulier des villes de montagne, [TORRICELLI, 2002] 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 re-dynamiser, par le tourisme et l'industrie, et enfin l'autoroute a encore plus récemment induit une déstructuration par des effets de périurbanisation par exemple. Les effets sont naturellement différents selon l'échelle d'observation, comme [RIETVELD, 1994] le montre dans une revue des approches économiques des interactions, notamment l'importance de différencier l'intra-urbain et l'intra-régional. Les effets des territoires sur les infrastructures sont plus complexes. 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 absolument pas systématique. [OTAMENDI, PASTOR et GARCIA, 2008] 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 difficile 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 de la gare sur la ligne à grande vitesse, ou des facteurs purement économiques. Il s'agit très probablement d'une combinaison complexe de multiples facteurs, difficiles à séparer.

C (AB) : Aerer

GOUVERNANCE ET PLANIFICATION Certains aspects de la gouvernance territoriale peuvent avoir un impact déterminant sur le développement des infrastructures de transport : [DENG et LIU, 2007] 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 en terme de *Transit Oriented Development* (TOD) doivent être prises pour anticiper ces externalités négatives. 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 coévolution volontaire, dans laquelle l'articulation est pensée et planifiée. Ces concepts ne sont pas nouveaux, puisqu'ils étaient implicites par exemple dans l'aménagement des villes nouvelles, sous une forme différentes puisque celles-ci étaient également fortement zonées et dépendantes de l'automobile pour certains quartiers. [L'HOSTIS, SOULAS et WULFHORST, 2012] 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, [L'HOSTIS, SOULAS et VULTURESCU, 2016] **C (AB) : L'Hostis ⊂ Jury?** utilise une analyse multicritères pour comprendre les facteurs déterminants dans la sélection des stations de la ville planifiée, incluant densité urbaine et temps d'accès aux stations. [LIU et ALAIN, 2014] 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 être opérée de façon fortement couplée, d'autant plus que le projet est spécialisé : [] rappelle l'anecdote du métro SK de Noisy-le-Grand montre un cas 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.

CO-ÉVOLUTION DES RÉSEAUX ET DES TERRITOIRES La complexité des interactions entre réseaux et territoires nécessite de se placer dans une ontologie particulière, celle d'une *co-évolution*. [LEVINSON, 2011] souligne la difficulté de la compréhension de la coévolution 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. [OFFNER, 1993] parle de congruence, qu'on peut comprendre en terme de dynamique systémique impliquant des corrélations fortuites ou non, à lier avec la vision systémique de l'époque, ce qui serait une vision préliminaire de la coévolution. La nécessité de dépasser les approches réductrices des effets structurants, tout en capturant la complexité des interactions entre réseaux et territoires par leur coévolution, est confirmée par le cas des effets économiques des trains à grande vitesse (HSR) : [BLANQUART et KONING, 2017] 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, à des effets en apparence aléatoires si les sujets sont considérés hors contexte, témoignant de situation locales bien plus 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 : il y a bien co-évolution entre les différentes composantes du système. Cette revue confirme le décalage entre les discours politiques et techniques prévalant aux projets de transports et les analyses effectives *a posteriori* révélée par [BAZIN et al., 2010]. [BAZIN, BECKERICH et DELAPLACE, 2007] avait d'autre part procédé à une étude ciblée du marché immobilier à Reims en anticipation de l'arrivée du TGV Est, et avait conclu que seul des opérations très localisées pouvaient être directement reliées au TGV, l'ensemble du marché répondant à une dynamique globale indépendante. Ainsi, cette notion de co-évolution que nous préciserons par la suite est une bonne candidate pour capturer la complexité de relations circulaires causales.

★ ★

*

1.2 DE PARIS À ZHUHAI

Nous développons dans cette section des cas d'étude géographique à l'échelle métropolitaine, que nous choisissons très différents pour montrer la diversité des situations possibles mais aussi les motifs récurrents généraux qui pourraient se dégager. 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.

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, au cours du temps et à l'échelle intermédiaire d'une région métropolitaine globalement monocentrique. [GILLI, 2005] rappelle l'importance de l'hinterland du Bassin Parisien et l'importance de ne pas considérer l'hypercentre de manière isolée. Si la moyenne couronne possède un certain niveau de polycentricité, notamment grâce à l'effet des villes nouvelles qui sont d'importants pôles d'emplois locaux [BERROIR et al., 2005] qui même s'il a rapidement divergé des intentions planificatrices [OSTROWETSKY, 2004], est bien réel, le Bassin Parisien étendu peut être également vu ayant un certains nombre de centres importants à une heure de Paris : Chartres, Orléans, Rouen, Reims et Lille grâce à la grande vitesse. Le rôle des différentes infrastructures de transport dans les différentes dynamiques économiques en Ile-de-France n'est pas trivial, comme le montre [PADEIRO, 2013] 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, généralement par ruptures, que nous détaillerons par la suite, sont liés à des dynamiques de planification et des processus de gouvernance qu'il convient de comprendre de manière intégrée. La métropole parisienne, qu'on désignera par *Grand Paris*, et sa région d'influence élargie, témoignent ainsi de relations complexes entre territoires et réseaux.

GOUVERNANCE [GILLI et OFFNER, 2009] propose en 2009 un diagnostic de la situation institutionnelle de la région parisienne, et des pistes pour une approche couplée entre gouvernance et aménagement. La préfiguration de "l'instauration d'un acteur collectif métropolitain" correspond à la métropole du Grand Paris qui sera inaugurée 7 ans plus tard, puisque le conseil métropolitain est mis en place fin 2016. Son effectivité concrète reste quasi-nulle au moment de l'écriture, confirmant une certaines inertie des structures de gouvernance,

qui a nécessairement un impact sur celle des réseaux de transport. La mise en place de ce nouveau niveau de gouvernance a été disséquée plus récemment toujours par GILLI dans [GILLI, 2014], où il la situe dans un contexte plus large socio-économique et urbain, en quelque sorte un diagnostic territorial qui explique certains aspects de ce besoin de mutation. En perte de vitesse sur le plan de l'aménagement, mais aussi sur le plan social au vu d'inégalités socio-économiques locales très fortes, la métropole a besoin de se réinventer, et se nouveau souffle se cristallise naturellement dans le Grand Paris, c'est à dire "l'avenir de Paris est sa banlieue". Cette initiative se concrétise par la convergence d'une auto-organisation des élus locaux, et d'une redéfinition du rôle de l'état, voulue centralisatrice jusqu'en 2012 puis laissant la place libre à la gouvernance métropolitaine avec le nouveau gouvernement, 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'état et celle poussée par la région. Suivant [DESJARDINS, 2016], 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.

RÉSEAU DE TRANSPORT DU GRAND PARIS L'histoire du développement du réseau de transport de la métropole francilienne est rappelée dans [BEAUGUITTE, 2014]. La particularité centralisatrice française a conduit à une structure particulière du réseau ferré à l'échelle nationale, mais aussi à celle 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 subit des évolutions conséquentes. Avant 1975, la distribution de l'accessibilité est clairement centralisée et le centre de Paris fortement congestionné. La mise en place du réseau RER entre 1975 et 1990, même si elle ne suit pas la volonté initiale de Delouvrier qui comprenait un plus grand nombre de branches transversale, permet la concrétisation d'un certain niveau de polycentrisme grâce à la construction conjointe des Villes Nouvelles. La période qui suivra jusqu'à aujourd'hui visera surtout à des désenclavements, mais aucun changement majeur d'accessibilité n'est observé. Les schémas directeurs successifs, qui aboutiront au SDRIF de 2013 [SDRIF, 2013], préfigurent le futur réseau du Grand Paris Express, qui aura un impact crucial en terme de cohésion territoriale en favorisant les liaisons de banlieue à banlieue qui sont les plus problématique dans le réseau actuel. De plus, le schéma se veut pensé de manière intégrée, par densification autour des gares et articulation des opérations d'aménagement et les 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)⁵

IMPACT DU GRAND PARIS EXPRESS 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 par exemple. La construction d’un réseau de métro en rocade permettant des liaisons de banlieue à banlieue est un besoin ancien, et a mené à plusieurs propositions sur lesquelles se sont opposés l’Etat et la Région au tournant des années 2010 [DESJARDINS, 2010]. Le projet Arc Express [STIF, 2010], 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 [SDRIF, 2013]) est un compromis et permet un rééquilibrage est-ouest de l’accessibilité [BEUCIRE et DREVILLE, 2013].

Les impacts immédiats d’une nouvelle de transport en terme d’accessibilité 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. La Fig. 1 cartographie les gains de temps moyens permis par ce projet sur les départements métropolitains (75, 92, 93 et 94)⁶, avec le tracé le plus récent pour l’ensemble des lignes. On observe, conformément à l’analyse de [BEUCIRE et DREVILLE, 2013], un rééquilibrage des différentiels d’accessibilité entre Est et Ouest. A distance égale du centre, l’accessibilité temporelle moyenne est plus élevée pour la Seine-Saint-Denis et le Val-de-Marne que pour les Hauts-de-Seine. 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 est en effet précisément dédiée au désenclavement du nord-est de la Seine-Saint-Denis [DESJARDINS, 2016]. 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

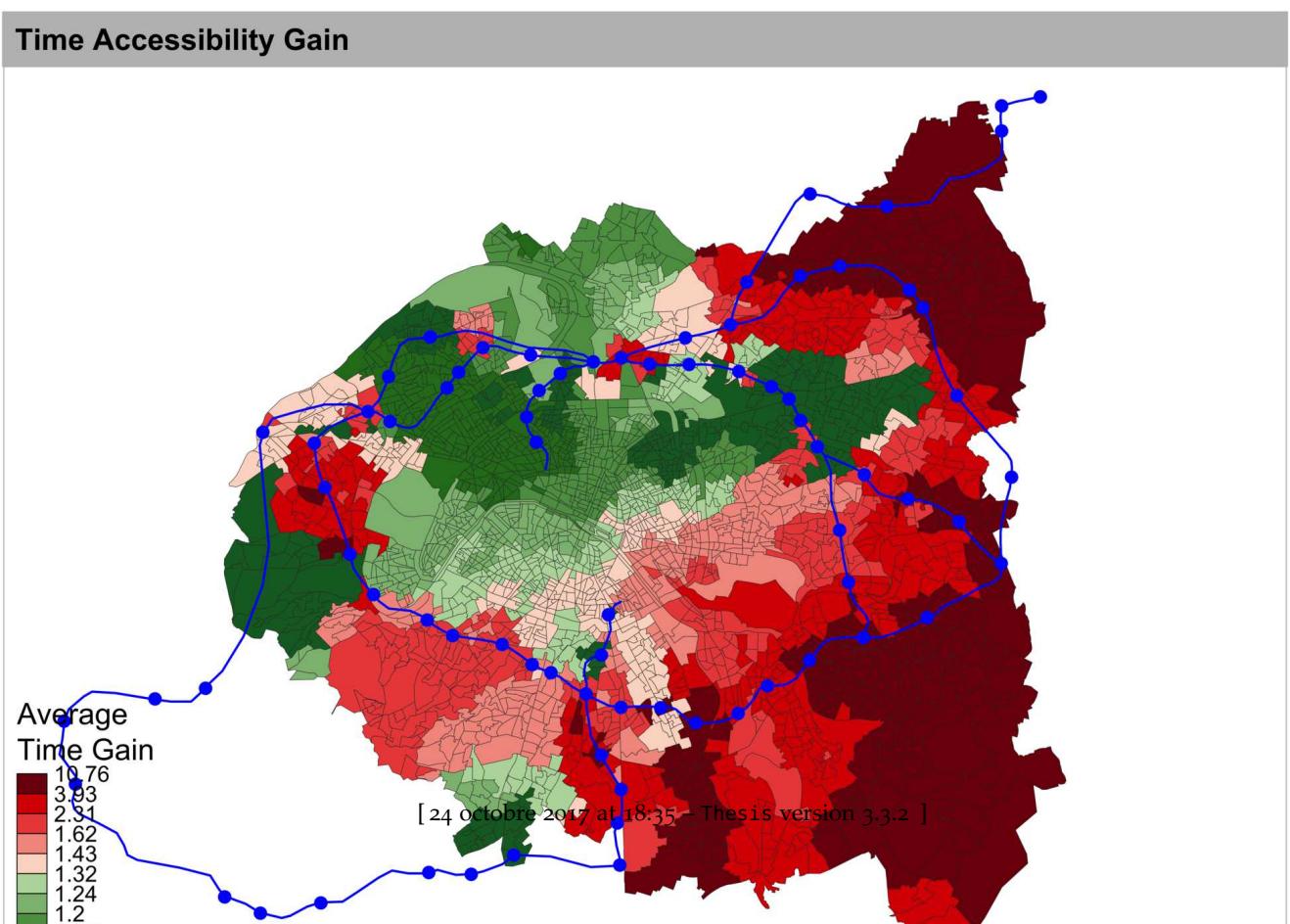
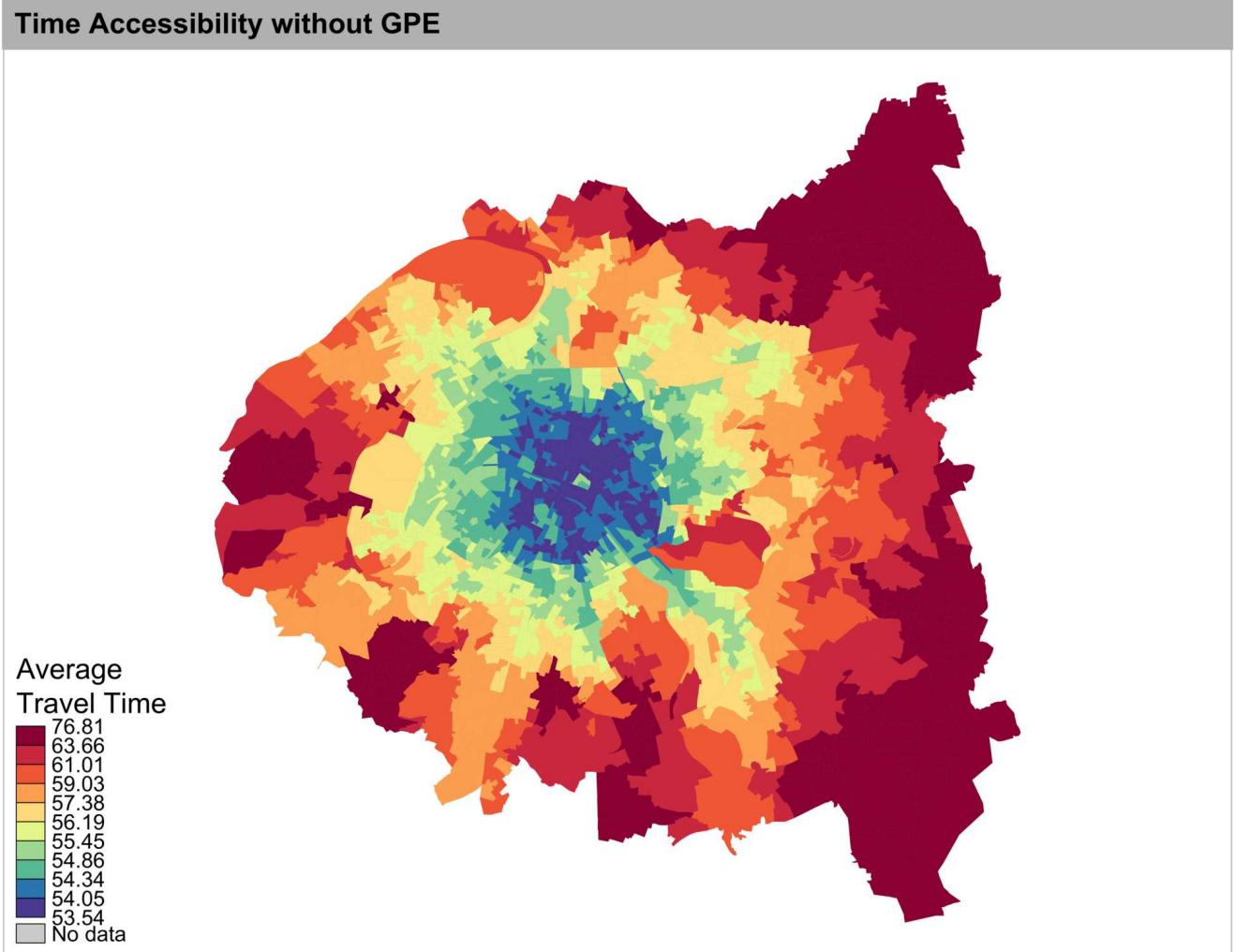
⁵ voir l’actualité du 4 octobre 2017 sur <https://www.iledefrance-mobilites.fr/actualites/un-reseau-de-tr> qui souligne que “Avec 29 km de réseau supplémentaires et l’ouverture de 28 points de desserte, les territoires se rapprochent”, témoignant de l’importance de l’accessibilité pour les territoires, notion par ailleurs floue. Les mêmes orientations de discours se retrouvent pour les différents projets d’extension ou de construction de nouvelles lignes.

⁶ voir la section ?? pour le détail de la méthodologie de calcul.

plus parler de proche banlieue mais de quartiers faisant partie intégrante du Grand Paris. Les effets peuvent cependant être mitigés selon les zones : [L'HORTY et SARI, 2013] montre 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 accroître. 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 sûre comme nous l'avons déjà développé. 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.

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, 2009] étudie les choix résidentiels des ménages pour comprendre les interactions entre usage du sol et transport. 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. [GUÉROIS et LE GOIX, 2009] 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. En effet, les liens entre nouvelles lignes et évolution du foncier sont parfois remarquables [DAMM et al., 1980].

C : sur les anticipations des acteurs : [CARROUET et al., 2013]

[GUÉROIS et LE GOIX, 2009] : bulles immobilières locales ?

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. ?? 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 (RER 60km.h⁻¹, Transilien 100km.h⁻¹, Metro 30km.h⁻¹, Tramway 20km.h⁻¹). 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 [R CORE TEAM, 2015a] et l'ensemble des données, du code source et des résultats sont disponibles sur un dépôt git ouvert⁷.

Nous calculons pour chaque projet, le différentiel ΔT_i d'accessibilité en temps moyen de trajet à partir de chaque IRIS en comparaison à celui dans le réseau sans le projet, défini par $T_i = \sum_k \exp -t_{ik}/t_0$ avec k communes, t_{ik} temps de trajet, et t_0 paramètre d'atténuation. A chaque projet est associée une date⁸, 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). Nous étudions les corrélations décalées de cette variable avec les variations ΔY_{ij} des variables socio-économiques suivantes : population, revenu médian, indice de Gini des revenus, prix moyen des transactions immobilières et montant moyen des crédits immobiliers. 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 a également été menée mais moins in-

⁷ A l'adresse

<https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/SpatioTempCausality/GrandParis>.
Les données de la base BIENS ne sont fournies que de manière agrégée à l'IRIS et pour les variables de prix et de crédit, pour des raisons de fermeture contractuelle de la base brute.

⁸ 2006 pour Arc Express, 2008 pour le Réseau du Grand Paris, 2010 pour le Grand Paris Express

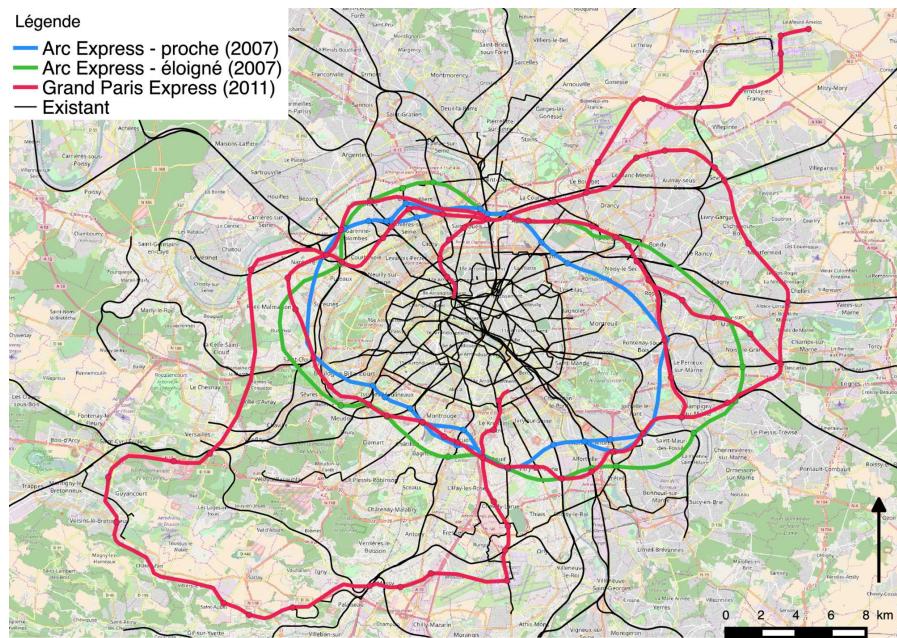


FIGURE 2 : Projets de transport successifs de la métropole du Grand Paris. Nous montrons les deux alternatives du projet Arc Express porté par la région, et le Grand Paris Express (GPE) porté par l'état. Le Réseau du Grand Paris, précurseur du GPE, n'est pas montré ici pour des raisons de visibilité à cause de sa proximité avec celui-ci.

téressante car très peu sensible à la composante mobilité (réseau et atténuation) par rapport aux variables elle-même, informe uniquement sur des relations entre celles-ci et n'est donc pas présentée ici. Nous présentons en Fig. ?? les résultats pour l'ensemble des réseaux et variables. Il est remarquable tout d'abord de noter l'existence d'effets significatifs 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, l'effet des autres projets serait alors fallacieux de par leur proximité dans les grands tronçons : cela impliquerait que les zones où ils diffèrent fondamentalement comme le Plateau de Saclay ne soient que très peu sensibles au projet de transport, ce qui confirmerait 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. Enfin, comme attendu les prix immobiliers sont tirés par l'arrivée potentielle des nouveaux réseaux, effet qui disparaît à deux ans pour le Grand Paris Express, suggérant une bulle immobilière passagère. 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 article, car supposerait des études de terrain qualitatives, des études de cas ciblées, etc. Cet exemple illustre cependant le caractère opérationnel de notre méthode sur un cas d'étude réel.

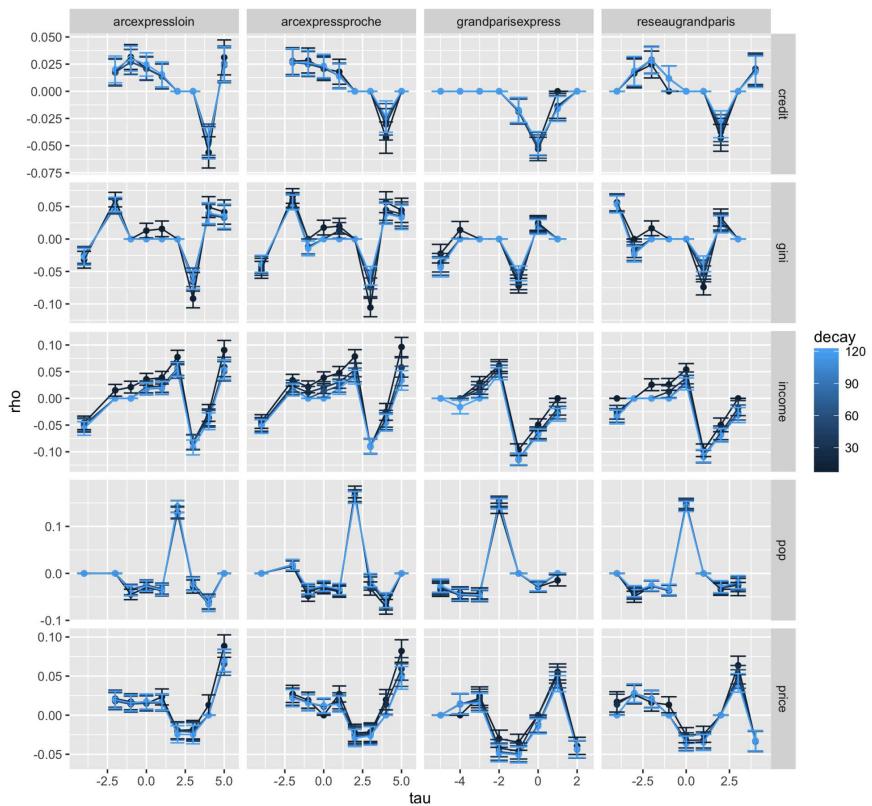


FIGURE 3 : Corrélations retardées empiriques. Les graphiques donnent la valeur de la corrélation entre le différentiel d'accessibilité en temps de trajet moyen ΔT pour chaque projet (en colonnes) et le différentiel des différentes variables socio-économiques et de transactions immobilières (en lignes), pour différentes valeurs du paramètre d'atténuation (decay). Les barres d'erreur donnent l'intervalle de confiance à 95%.

1.2.2 Le Delta de la Rivière des Perles

NOUVEAUX RÉGIMES URBAINS ET MEGA-RÉGION URBAINE Si la notion de megalopolis peut être tracée jusqu'à GOTTMANN [GOTTMANN, 1964], et qu'elle est à l'origine de celle de *Mega-city Region* (MCR) consacrée par HALL [HALL et PAIN, 2006], il est clair que cette dernière est toujours plus d'actualité avec l'apparition récente de nouveaux régimes, notamment par l'urbanisation accélérée dans des pays à forte croissance économique et en mutation très rapide comme la Chine [SWERTS et DENIS, 2015]. Le second cas que nous développons ici rentre dans cette catégorie : le Delta de la Rivière des Perles 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 Zones Economiques Spéciales (经济特区) 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.

[YE, 2014] 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'Etat 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 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 d'un niveau de gouvernance qui expliquerait tel ou tel processus urbain est épingleuse : [LIAO et GAUDIN, 2017] interprète les transferts progressifs des initiatives économiques du pouvoir central vers les autorités locales comme une forme de gouvernance multi-niveau. Dans le cadre des transports pour la MCR, il n'existe pas d'autorité d'organisation des transports, 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 retrouveront 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. Cette situation serait cependant transitoire, étant donné les infrastructures déjà en construction et celles planifiées sur un plus long terme : le métro de Shenzhen, qui couvre aujourd’hui 285km, est planifié jusqu’à 30 lignes et 1142km en 2030 [SHENZHEN-SHI, 2016]. 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. Différents terminus seront connectés au métro 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.

IMPACT DU PONT ZHUHAI-HONG-KONG-MACAO Un projet iconique d’infrastructure de transport dans la région est le pont fermant l’embouchure du Delta, reliant Zhuhai et Macao à Hong-Kong. En réalité un Pont-tunnel. La longueur de la traversée est de 36.5km, ce qui en fait un ouvrage d’art exceptionnel. L’ouverture au trafic a été retardée de plusieurs années et est prévue finalement pour fin 2017⁹. [ZHOU, 2016] montre que les changements de motifs d’accessibilité attendus pour l’Ouest du Delta sont relativement forts, et ceux-ci peuvent potentiellement induire de fortes bifurcations dans les trajectoires des villes. La nécessité du projet est supportée par une narration 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 s’opère cependant pour le mode routier uniquement, ce qui conduit à questionner ses impacts potentiels : d’une part l’accès systématique à l’automobile reste réservé à une partie de la population seulement, d’autre part les externalités négatives de congestion peuvent rapidement modérer les gains d’accessibilité. Les impacts à moyen et long terme du pont sont ainsi difficile à estimer, une piste étant de poser le problème différemment et de chercher comprendre la dynamique du système métropolitain de manière intégrée, comme nous l’évoquerons en 7.3.

1.2.3 Comparabilité des études de cas

La possibilité de transfert des modèles urbains est délicate, et 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

⁹ voir <http://www.hzmb.org/cn/default.asp>

le faire, comme la vision de la Banque Mondiale [AMSDEN, 1994]. La comparabilité de systèmes urbains est une question ouverte au centre des enjeux de la Théorie Evolutive Urbaine, et est par exemple liée au caractère ergodique de ces systèmes : si la trajectoire d'une ville dans le temps capture l'ensemble des états urbains possibles, alors les différentes villes sont différentes manifestations du même processus stochastique à différentes périodes, et un ensemble de villes permettrait d'avoir une idée des trajectoires temporelles. Intuitivement ce n'est pas le cas, et la Théorie Evolutive postule en effet la non-ergodicité [PUMAIN, 2012b], que nous étudierons plus en détail en 4.1.

★ ★

★

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 subjective à l'échelle microscopique, par des exemples concrets de terrain. Le terrain géographique majoritaire est le Delta de la Rivière des Perles en Chine (珠江三角洲), dans la province du Guangdong (广东), que nous avons décrit ci-dessus, et plus particulièrement en grande partie la ville de Zhuhai (珠海). Des observations ont également été effectuées en métropole Parisienne. Nous assumons le terme de *Terrain géographique*, sachant les débats épistémologiques que peuvent poser l'utilisation de celui-ci. En effet,

Dans le cadre du projet européen Medium, visant à une approche interdisciplinaire de la soutenabilité pour les villes Chinoises en se concentrant sur les villes moyenne, 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 comptes-rendus narratifs sont disponibles en Annexe A.1. Le format de ceux-ci est "à-la-volée" suivant les recommandations de [GOFFMAN, 1989] pour la prise de notes en terrain d'immersion notamment, tandis que la position volontairement subjective rejoint [BALL, 1990] qui souligne l'importance de la réflexivité pour tirer des conclusions rigoureuses à partir d'observations qualitatives de terrain duquel le chercheur est partie intégrante.

1.3.1 Développement d'un réseau de transport

L'un des axes principaux de la recherche de terrain en Chine a été de construire une vue subjective des multiples facettes et couches d'un système de transport public complexe et en mutation permanente. 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'accèsibilité 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 microscopiques qui peuvent devenir cruciaux au regard de l'utilisation effective du réseau. Par exemple, la multi-modalité 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" [LIU, JIA et CHENG, 2012], 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 a l'embarquement peuvent influencer considérablement les motifs d'usage.

Différents voyages sur le territoire Chinois ont été effectués pour percevoir les mécanismes d'usage du réseau à grande vitesse (HSR). Depuis 2008, la Chine a établi le plus grand réseau de HSR du monde a partir de zero, qui a reçu un grand succès et dont les lignes sont actuellement saturées. Celui-ci répond à des motifs de demande primaires en terme de taille de ville, mais d'autres dimensions ont été prises en compte dans leur planification, comme le développement du tourisme. Ainsi, la ligne entre Guangzhou et Guiyang a vu la construction de stations spécifiques au développement du tourisme, comme Yangshuo dans le Guangxi, dont la fréquentation a alors fortement augmenté. Un an après l'ouverture de la gare, le lien routier majeur avec la ville est toujours en construction, mais un grand nombre de trains s'y arrêtent 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. On observe qu'une stratégie similaire concernant l'exploitation du potentiel touristique a pu être employée pour la ligne Chengdu-Emeishan, dont le principal objectif est pour l'instant de desservir les destinations touristiques très fréquentées d'Emeishan et de Leshan, même si le lien manquant entre Leshan et Guiyang est déjà bien avancé dans sa construction et complétera le lien direct entre Guangzhou et Chengdu. Cette ligne fait partie du squelette structurant des "8+8" reformulées récemment par le gouvernement central¹⁰, et les territoires traversés en attendent beaucoup comme le montre [LU et al., 2012] pour la ville de Yibin à mi-chemin entre Chengdu et Guiyang. La complexité du réseau est accrue par la diversité des missions et la congestion, qui semble impliquer que les vitesses maximales potentielles ne sont pas systématiquement exploitées. Nous illustrons en Fig. 4 l'insertion du HSR dans ses territoires et les attentes socio-économiques envers celui-ci et les agents locaux qui doivent contribuer à son succès : les publicités vantant les mérites de la grande vitesse, et la vente d'appartement dans des opérations immobilières semble contribuer à la construction

¹⁰ 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.

d'une "classe moyenne" et au rôle qu'elle doit jouer dans le dynamisme des territoires [ROCCA, 2008]¹¹. L'insertion des lignes dans les territoires semble dans certains cas forcée, comme le montre la gare de Yangshuo qui exploite le potentiel touristique du passage de la ligne dans une zone très peu peuplée, ou les nouvelles opérations immobilières peu accessibles à Zhuhai.

1.3.2 Variations locales du TOD

Des branches régionales 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. A cela s'ajoute le réseau de train classique, et certaines connexions requièrent l'utilisation des deux réseaux et des transports urbains, comme la liaison entre Zhuhai et Hong-Kong expérimentée par voie terrestre seulement¹². Le réseau urbain local et les opérations de développement immobilières sont planifiées 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) 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 anti-urbaine et une organisation peu pratique peut mener au questionnement de l'efficacité de l'approche et à des interrogations sur la nature auto-prophétique du projet, comme suggéré par les publicités pour du nouvel immobilier à vendre, appuyant sur l'importance de la présence de la ligne ferroviaire. Toute une narration autour du TOD, parfois sans réel fondement empirique dans le cas donné, semble être utilisée par différents acteurs du développement. D'autres observations de terrain, comme dans les *new territories* à Hong-Kong, témoignent d'un TOD efficace et réalisant son objectif, avec une combinaison intelligente 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. 5 la comparaison des deux cas de TOD évoqués ci-dessus, sous forme de schéma syn-

¹¹ Construction, comme le souligne 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.

¹² À la suite du Typhon Hato le 23/08/2017, les liaisons maritimes ont été interrompues pour une grande partie du delta.



FIGURE 4 : Manifestations locales des mutations du 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; (Bas Gauche) La gare de Yangshuo, au milieu de nulle part sur la ligne à grande vitesse, rappelant les gares type "betterave-TGV" typiques en France; (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.

théтиque 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, dont l'infrastructure 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, prend avantage du tracé de la nouvelle ligne ferroviaire, pour essayer de reorganiser le 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 porte toujours les stigmates de cette mise en place déphasée, et l'accessibilité en transport en commun est relativement faible, les lignes de bus étant sujettes à une congestion croissante due à la forte augmentation du nombre d'automobiles. 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 à présent d'ébaucher une entrée qualitative d'un certain type, pour suggérer une façon de compléter nos connaissances et mieux cerner les processus de manière concrète.

L'entrée prise suit la méthode *d'observation flottante*, introduite à l'interface de l'anthropologie et la sociologie par [PÉTONNET, 1982], 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 [CHOAY, 2009] qui explore la direction inverse, c'est à dire le propre des sociétés humaines de façonne l'espace, et la capacité de construire un environnement bâti à différentes échelles par l'architecture et l'urbanisme. 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". Sans s'y méprendre et considérer la méthode comme une négligence métho-

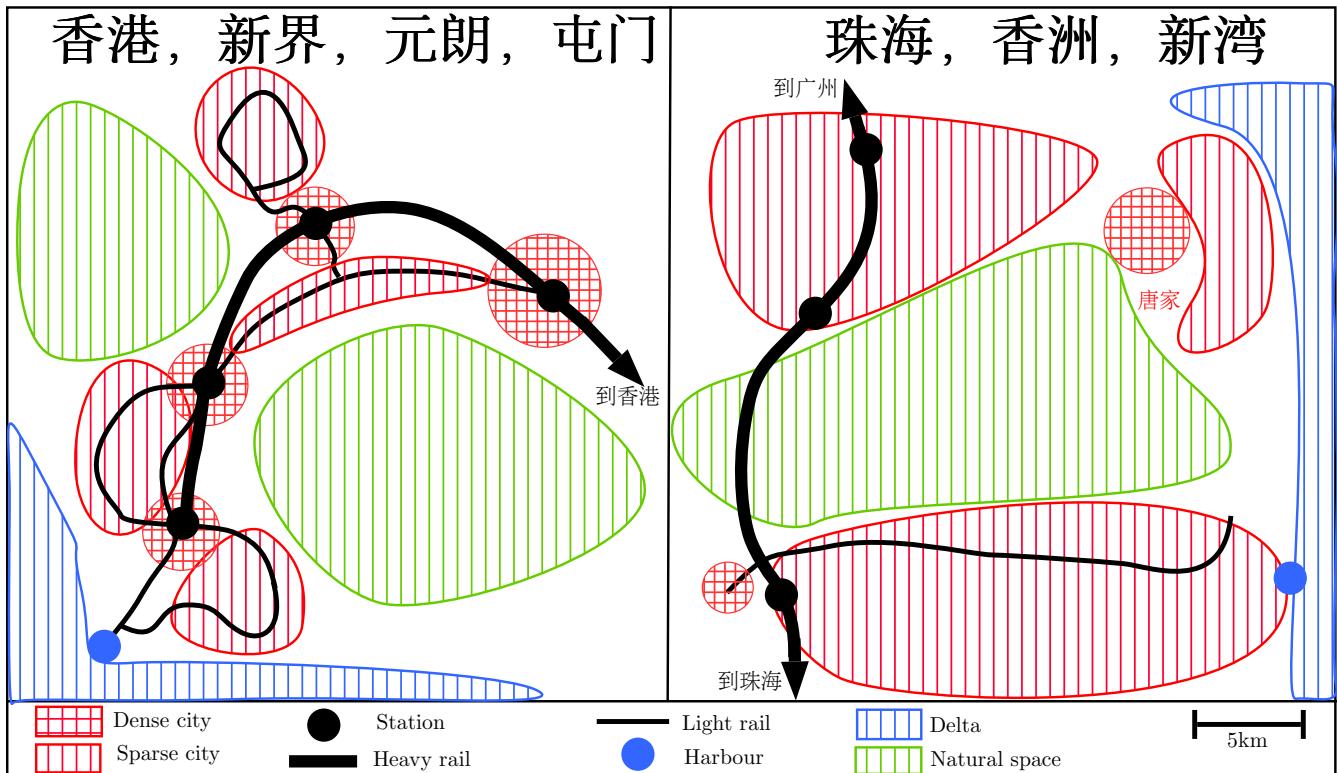


FIGURE 5 : Analyse comparative de deux implémentations du TOD en PRD. A une échelle comparable, nous synthétisons la configuration urbaine de Yuenlong (元朗) et Tuenmun (屯門), Hong-Kong New Territories (香港, 新界), et de Xinwan, Xiangzhou, Zhuhai (珠海, 香洲, 新湾), qui contient la High-Tech zone de Zhuhai dans 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.

dologique, nous y voyons une opportunité d'un accès rapide et à faible coût dans le monde du qualitatif, tout en restant conscient de sa portée très limitée. La 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 [ALBA et AGUILAR, 2012].

MÉTHODE Les mouvements pendulaires à échelle moyenne 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) 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 : se laisser errer 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 routine¹³. 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 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 [BALL, 1990] pour la place de la subjectivité dans la recherche ethnographique de terrain.

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 majoritaire est le bus de ville, suivi par le train régional, puis le train à grande vitesse et le ferry ; la portée des déplacements cor-

¹³ Cette contrainte sera respectée dans notre cas pour le Guangdong, mais pas pour l'Ile-de-France.

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

respondent 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.1. Les observations pour la région parisienne sont quasi-quotidiennes et non consignées ; celles-ci ont eu lieu en plus grande partie sur la ligne 4 du métro et sur la ligne A du RER entre février 2016 et octobre 2016, sur la ligne R du Transilien et la ligne A du RER entre novembre 2016 et septembre 2017 puis entre février 2017 et mai 2017, puis sur la ligne 9 et la ligne 4 entre septembre 2017 et octobre 2017.

Les deux synthèses d'observation flottante pour chacune des régions, matériaux produit à partir des notes brutes, sont présentées dans les encadrés ci-dessus. Celles-ci illustrent entre autre par des exemples subjectifs certaines instances d'interactions entre réseaux et

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 ” : la perte de son passeport 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 : *Une expérience en observation flottante, Guangdong, Zhuhai*

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 meta :

1. La complexité du système de transport et de son intégration 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 l'efficacité émergente, tandis qu'en France la complexité semble être source de freins et finalement d'externalités négatives.

2. La résilience, qui peut être caractérisée par exemple par 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.

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

★ ★

★

CONCLUSION DU CHAPITRE

Les territoires, que nous avons défini comme territoires humains, interagissent de manière complexe avec les réseau, en particulier ceux de transport, comme montré par les nombreux cas d'étude empiriques ou les constructions théoriques passés en revue. A différentes échelles temporelles typiques, l'année, la décennie et le siècle, correspondent plus ou moins 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 déclinés 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, et 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, pour limiter les effets d'intermédiaire qui augmenteraient la distance au terrain géographique réel qui est selon [LEFORT, 2012] déjà très présente dans les études le prenant comme matériau empirique principal, 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 restreindre la portée des entrées prises sur le sujet, et s'intéresser plus particulièrement aux approches impliquant une modélisation, faisant le choix d'un rôle fondamental du *modèle* (que nous définirons 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 des relations circulaires causales entre dynamiques territoriales et dynamiques des réseaux de transports que l'on se proposera de désigner par *co-évolution* C (FL) : faire une tournure de phrase plus simple pour définir le mot *coevolution* , ceux-ci semblent diverger sur toute explication potentiellement simple C (FL) : mal dit ou systématique, comme le rappelle par exemple les débats autour des effets structurants des infrastructures [OFFNER, 1993] C (FL) : qu'il convient ici d'expliquer en 2.3 - φ : qui dit quoi dans ce débat? . Au contraire, les multiples situations géographiques poussent à privilégier des études ciblées très fortement dépendantes du contexte et du travail de terrain C (FL) : phrase un peu rapide tu donnes l'impression que tu tranches le débat . Or l'explication géographique et la compréhension des processus est très vite limitée dans cette approche, et intervient un besoin d'un certain niveau d'abstraction et de généralisation C (FL) : or justement il convient d'expliquer ce débat théorie vs empirisme (ou autre)→dans la phrase en pointilles tu parles d'un besoin d'abstraction, ce n'est pas le terme scientifique . C'est sur un tel point que la Théorie Evolutive des Villes se concentre particulièrement C (FL) : il faut absolument se départir d'un ton enthousiaste non scientifique. tu te proposes d'appliquer une théorie/un cadre analytique à une question - ça marche plus ou moins, c'est tout, il ne doit pas y avoir d'a priori ou de préférence de ton côté , puisqu'elle permet de combiner des schémas et modèles généraux aux particularités géographiques, et en tire même parti, tandis C (FL) : faire deux phrases que certaines théories issues de la physique comme la Théorie du Scaling de WEST [WEST, 2017] peuvent être plus difficiles à accepter pour les géographes C (FL) : tu pars déjà dans de l'interprétation épistémique : il faut séparer les choses, d'abord de quoi s'agit-il? 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* C (FL) : l'italique ne joue pas le même rôle pour les 2 mots cela peut créer confusion (voir 9.3 pour un développement des domaines de connaissance et du rôle du modèle). Comme le rappelle J.P. MARCHAND [RAIMBAULT, 2017e],

“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 qui pourraient être aujourd’hui à leur balbutiements **C (FL) : ok avec le sens de cette phrase mais le conditionnel sonne mou. si tu as des arguments avance les, sinon dis clairement quelle est ta position**. Sans développer les innombrables 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, objet (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). Ce chapitre peut être vu comme un “état de l’art” des démarches 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. Les modèles de changement d'usage du sol très appliqués en planification sont tout autant concernés que des modèles totalement abstraits issus de la biologie ou de la physique, que des approches intégrées en géographie ou spécifiques en économie

C (FL) : pas utile ici tu detailleras en 2.1 : ici affirmer les objectifs (quelle information en retirer) et les moyens (par ez meme automatique tu as bien un (des) points d'entrees → lesquels et pourquoi?). Cet aperçu suggère des structures de connaissances assez indépendantes et des disciplines ne communiquant que rarement. Nous procédons à une revue systématique algorithmique

C (FL) : est-ce une façon standard de nommer cela ? A1 : (JR) l'exploration itérative de la façon dont elle est faite n'a jamais été faite à ma connaissance, j'introduis donc une “nouvelle” façon de faire.

C (AB) : définir dans 2.2 pour reconstruire leur paysage scientifique, dont les résultats tendent à confirmer ce cloisonnement

C (FL) : c'est intéressant bien sûr mais il faut dire pourquoi tu vises cela. à mon avis il vaut mieux commencer par des exemples de manières dont la littérature sci. prend en compte ces interactions, dans différentes disciplines, avant d'atteindre l'épistémologie quantitative.

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 du corpus utilisé pour la modélographie et la meta-analyse

C (FL) : mots qui n'ont pas encore été introduits effectuée en dernière section 2.3,

qui dissèque la nature d'un certain nombre de modèles et la relie au contexte disciplinaire, ce qui pose les bases et le cadre précis des efforts de modélisation qui seront développés par la suite.

★ ★

★

Ce chapitre est inédit pour sa première section ; reprend dans sa deuxième section le texte traduit de [RAIMBAULT, 2015a], puis pour sa deuxième partie la méthodologie de [RAIMBAULT, 2016b], les outils de [BERGEAUD, POTIRON et RAIMBAULT, 2017] ; et est enfin inédit pour sa dernière partie.

2.1 MODÉLISER LES INTERACTIONS

2.1.1 Modélisation en Géographie Quantitative

La modélisation joue en Géographie Théorique et Quantitative (TQG) un rôle fondamental. **C (FL)** : manque peut etre quelques phrases avec des modeles "landmark" anciens (Lowry 1964 par exemple) mais encore utilises, de plus ne pas perdre le fil du champ applicatif auquel tu te destines CUYALA procède dans [CUYALA, 2014] 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 [PUMAIN et ROBIC, 2002]. 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. Une histoire étendue de la genèse des modèles de simulation en géographie est faite par REY dans [REY-COYREHOURCQ, 2015] avec une attention particulière pour la notion de validation de modèles **C (FL)** : cest tres utile, a chaque fois fois profites en pour dire la facon dont tu vas t'approprier, plus ou moins fortement, ces manieres de faire . L'utilisation de ressources de calcul pour la simulation de modèles est antérieure à l'introduction des paradigmes de la complexité, remontant par exemple à FORRESTER **C (FL)** : est ce qu'on ne peut pas lier le courant systemique et le courant "sciences de la complexité" , informaticien qui a été pionnier des modèles d'économie spatiale inspirés par la cybernétique. Avec l'augmentation des potentialités de calcul, des transformations 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 Performance (calcul sur grille à l'utilisation transparente, c'est à dire sans besoin de compétences techniques pointues liées au mécanismes de la parallélisation **C (FL)** : quest ce que cela signifie? par ailleurs il faut clairement hierarchiser les choses. ce point est crucial, il justifie pleinement ta demarche, mais il doit venir apres plusierus choses, dont la definition du mot systeme. ; voir [SCHMITT et al., 2014] pour 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), qui est également accompagnée par une évolu-

tion des pratiques [BANOS, 2013] et techniques [CHÉREL, COTTINEAU et REUILLOON, 2015] de modélisation. La modélisation, et en particulier les modèles de simulation, est vue par beaucoup comme une brique fondamentale de la connaissance : [LIVET et al., 2010a] 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 **C (FL) : entre autres choses(ne pas donner l'impression que tu clos le debat) - OK**. Ce sont des exemples de fonctions parmi d'autres : Varenne donne dans [VARENNE, 2010b] **C (FL) : notation ?** une classification raffinée **C (FL) : sens ?** des diverses fonctions d'un modèle. Nous considérons la modélisation comme un instrument fondamental de connaissance des processus au sein de systèmes complexes adaptatifs **C (FL) : est ce que les modeles se limitent a cela? dire surtour qu'en l'occurrence, tu t'attaques a ce type d'objets**, et précisons encore notre question de recherche, qui s'intéressera aux *modèles dont l'ontologie contient une part non négligeable d'interactions réseaux et territoires.*

MODÉLISATION ET EQUILIBRE Lorsqu'on se détache des approches proposées par l'Economie Géographique, la plupart des approches en Géographie Théoriques et Quantitatives sont généralement basées sur des hypothèses de systèmes hors-équilibre [PUMAIN, 2017]. Les premières contributions de la théorie de PRIGOGINE **C (FL) : pourquoi en majuscules ?** à l'étude des systèmes urbains, comme par exemple les modèles d'entropie de ALLEN comme celui étudié par [PUMAIN, SAINT-JULIEN et SANDERS, 1984], ont permis de consacrer les ontologies de l'auto-organisation dans des modèles formels, puis plus tard de simulation, pour les dynamiques urbaines. Les modèles que nous considérerons par la suite rentreront a priori dans cette catégorie **C (FL) : quelle categorie? ce n'est pas clair** pour leur grande majorité. Si un équilibre **C (FL) : est il besoin d'evoquer cela a ce stade?** est supposé à certaines échelles d'espace ou de temps, ce sera souvent dans un contexte de déséquilibre au niveau supérieur, et donc de non-stationnarité du premier niveau **C (FL) : sens ?** . **C (AB) : il faudra que ces termes (équilibre, stationnarité) soient définis qque part**

2.1.2 Modéliser les territoires et réseaux

Au sujet de notre question précise **C (FL) : non adapte : il faut toujours garder ce fil** des interactions entre réseaux de transport et

territoires, nous proposons un aperçu des différentes approches. Selon [BRETAGNOLLE, PAULUS et PUMAIN, 2002], "les idées des spécialistes de la planification cherchant à donner des définitions des systèmes de ville C (FL) : tu n'as pas encore défini les villes ni les systèmes de villes , depuis 1830, sont étroitement liées aux transformations des réseaux de communication". C'est en quelque sorte la prophétie auto-réalisatrice inversée, au sens où elle est déjà réalisée avant d'être formulée C (FL) : sens ? . Cela implique que les ontologies et les modèles correspondants proposés par les géographes et les planificateurs sont fortement liés aux préoccupations historiques courantes, ainsi forcément limités en portée et raisons C (FL) : sens ? . Au delà de la question de la définition du système sur laquelle nous reviendrons maintes fois C (FL) : mal dit , on comprend bien l'impact que peut avoir cette influence sur la portée des modèles développés. Dans une vision perspectiviste de la science [GIERE, 2010c] 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 C (FL) : pas par essence mais cela peut l'être effectivement A1 : (JR) pas d'accord : opérer un couplage de modèle implique un couplage des ontologies et nécessairement un accroissement des connaissances (si celui-ci est utile c'est un autre problème) une source de connaissance.

Modèles LUTI

C (FL) : tu arrives trop tot la dessus - les processus clés ne sont pas ammenes suffisamment clairement : travaille par échelle de temps, d'espace, d'acteurs

Un partie importante de la littérature proposant des modélisations des interactions entre réseaux et territoires se trouve dans le domaine C (FL) : d'application ? de la planification urbaine, avec les *modèles d'interaction entre usage du sol et transport (LUTI, pour Land-Use Transport Interaction)*. Ces travaux peuvent être difficiles à cerner car liés à différentes disciplines scientifiques. Par exemple, du point de vue de l'Economie Urbaine, les propositions de modèles intégrés C (FL) : qu'est ce que sont des modèles intégrés existent depuis un certain temps [PUTMAN, 1975] C (AB) : préciser . La variété des modèles existants a conduit à des comparaisons opérationnelles [PAULLEY et WEBSTER, 1991] C (FL) : ce type de phrase n'est pas informative : quelles conclusions ? qu'apportent-elles ? . Plus récemment, les avantages respectifs des approches statiques et dynamiques C (FL) : que sont les approches statiques et dynamiques ? a été étudié par [KRYVOBOKOV et al., 2013], dans un cadre métropolitain sur des échelles de temps moyennes C (FL) : sens ? . Dans tous les cas, ce type de modèle opère généralement à des échelles temporelles et spatiales relativement faibles. [WEGENER et FÜRST, 2004] C (FL) : 1999 donne un état de l'art des études empiriques et

de modélisation sur ce type d'approche des interactions entre usage du sol et transport **C (FL) : il convient de definir et de discuter de ces deux termes**. Le positionnement théorique est plutôt proche des disciplines de la socio-économie des transports et de la planification **C (FL) : il serait bon de dresser un panorama de ces disciplines** (voir les paysages dressés en 2.2), et pas forcément proche de nos raisonnements géographiques **C (FL) : sens ?** qui cherchent à comprendre également des processus sur le temps long. Pas moins de dix-sept modèles sont comparés **C (FL) : par qui ?** et classifiés, parmi lesquels aucun n'inclut une évolution endogène du réseau de transport sur les échelles de temps relativement petites **C (AB) : préciser** des simulations **C (FL) : c'est un point cle de la these, cela doit etre mieux ammené**. Une revue complémentaire est faite par [CHANG, 2006], 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 **C (FL) : sens ? est ce de la localisation des menages ? des entreprises ? des investissements etc ?**), les modèles microscopiques d'utilité aléatoire, et les modèles de marché foncier. Différents aspects du même système peuvent être traduits par divers modèles (comme e.g. [WEGENER, MACKETT et SIMMONDS, 1991]), 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. Toutes ces techniques opèrent également à une petite échelle **C (FL) : qu'est ce qu'une petite echelle ?** et considèrent au plus **C (FL) : cela suggere que c'est un probleme, ok ameliorable** l'évolution de l'usage du sol. [IACONO, LEVINSON et EL-GENEIDY, 2008] couvre un horizon similaire avec une emphase **C (FL) : sens ?** supplémentaire sur les modèles à automates cellulaires d'évolution d'usage du sol et les modèles basés agent **C (FL) : a base d'agents**. Les modèles LUTI sont toujours largement étudiés et appliqués, comme par exemple [DELONS, COULOMBEL et LEURENT, 2008] qui est utilisé pour la région métropolitaine parisienne. La courte portée **C (FL) : ne pas abuser de ces termes relatifs, donc flous (a moins d'avoir explicite cela)** temporelle d'application de ces modèles 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. **C (FL) : ce point est tres important et merite (un dvlpmnt?)** [TIMMERMANS, 2003] 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'arte-

facts 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 plus être centrées sur une meilleure intégration des nouvelles technologies et une meilleur interaction avec les politique et la planification, par exemple via des interfaces de visualisation [WEE, 2015], **C (FL) : par rapport a quoi ?** mais en rien des problématiques de dynamiques territoriales incluant le réseau sur de plus longues échelles par exemple, ce qui confirme la portée et la logique autour de ce type de modèles. **C (FL) : cette phrase est alambiquee et difficile a comprendre** Une généralisation de ce type d'approche à une plus grande échelle, comme celle proposée par [RUSSO et MUSOLINO, 2012], consiste au couplage du LUTI à l'échelle mesoscopique à des modèles macroéconomiques à l'échelle macroskopique. Ceux-ci ne considèrent pas l'évolution du réseau de transport de manière explicite mais s'intéresse seulement aux motifs abstraits d'offre et demande. L'économie urbaine a développé des approches spécifiques similaires dans leur démarche : [MASSON, 2000] décrit par exemple un modèle intégré couplant développement urbain, relocalisation et équilibre des flux de transports.

Croissance du Réseau

La croissance de réseaux est pratiquée **C (FL) : mal dit** dans des démarches de modélisation qui cherchent à expliquer de manière endogène, au sens de modèles génératifs **C (FL) : pas clair**, la croissance des réseaux de transport. Ils prennent généralement un point de vue *bottom-up*, i.e. en mettant en évidence des règles locales qui permettraient de reproduire la croissance du réseau sur de longues échelles de temps (souvent le réseau de rues). Les économistes ont proposé des modèles de ce type : [ZHANG et LEVINSON, 2007] passe en revue la littérature en Economie des Transports sur la croissance des réseaux dans le contexte d'une théorie endogène de la croissance [AGHION et al., 1998] **C (FL) : c'est beaucoup trop elliptique, on ne peut pas comprendre les apports de cet article**, 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. [XIE et LEVINSON, 2009c] propose une revue étendue de la modélisation de croissance des réseaux, en prenant en compte d'autres champs : 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 selon [XIE et LEVINSON, 2009c] **C (FL) : point tres important faire une section a part**; 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 jouet **C (FL) : quest ce qu'un modele jouet ?** 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. Nous donnons pour commencer des exemples d'études utilisant des concepts économiques ou géométriques pour modéliser la croissance de réseau. Les mécanismes induisant la croissance du réseau, sur le plan de la gouvernance ou économique, peuvent être très détaillés **C (FL) : avant d'arriver a ce niveau de detail, il faut expliciter ces mecanismes**, comme [LEVINSON, XIE et OCA, 2012] qui se base sur des enquêtes qualitatives et des modèles statistiques calibrés sur des données réelles **C (FL) : idem, tu es la dans un aspect methodologique qui doit intervenir plus tard** pour paramétrier un modèle de croissance de réseau. [XIE et LEVINSON, 2009b] compare l'influence relative des processus de croissance centralisés et décentralisés **C (FL) : on ne sait pas de quoi tu parles**. [LEVINSON et KARAMALAPUTI, 2003] procède à une étude empirique des déterminants de la croissance du réseau routier pour les Twin Cities **C (FL) : sens ?**, 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 dans l'accessibilité minimale **C (FL) : qu'est ce que l'accessibilité ?**. Ces données sont utilisées par [ZHANG et LEVINSON, 2016] pour calibrer un modèle de croissance de réseau qui superpose les décisions d'investissement aux motifs d'utilisation du réseau. [YERRA et LEVINSON, 2005] montre avec un modèle économique basé sur des processus auto-renforçants **C (FL) : sens ?** 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 [XIE et LEVINSON, 2011]. Les physiciens se sont largement inspiré de cette littérature économique : un modèle très similaire au dernier cité est donné par [LOUF, JENSEN et BARTHELEMY, 2013] avec des fonctions coûts-bénéfices plus simples mais obtenant une conclusion similaire. Etant donné une distribution de noeuds (villes) **C (FL) : ce tyoe d'hypothese est tres interessante et doit faire l'objet d'une discussion** dont la population suit une loi puissance, deux villes seront connecté par un lien routier si une fonction d'utilité coût-bénéfice combinant linéairement flux gravitaire potentiel (loi puissance de la distance **C (FL) : pas clair**) et coût de construction (linéaire de la distance) a une valeur positive. Ces hy-

pothèses 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. Alors que ces modèles basés sur des processus cherchent à reproduire des motifs macroscopiques des réseaux (typiquement les lois d'échelle) **C (FL) : la aussi on est dans les indicateurs de sortie, cest une discussion a part**, les modèles d'optimisation géométrique cherchent à ressembler à des réseaux réels dans leur topologie. La simplicité des hypothèses dans ce genre de modèle permet dans certains cas d'inclure des processus qui serait par ailleurs difficile à intégrer : [BOTTINELLI, LOUF et GHERARDI, 2016] étudie ainsi un modèle de croissance d'arbre appliquée aux pistes de fourmis, dans lequel coût de maintenance et coût de construction influencent tous les deux les choix de nouveau lien. [BARTHÉLEMY et FLAMMINI, 2008] décrit un modèle basé sur une optimisation locale de l'énergie, mais ce modèle reste abstrait et non validé **C (FL) : alors, ne pas en parler ?** quantitativement ou sur des faits stylisés. Le modèle de morphogenèse de [COURTAT, GLOAGUEN et DOUADY, 2011] qui utilise des potentiels locaux et des règles de connectivité, même s'il n'est pas calibré, semble **C (FL) : selon quel point de vue ?** reproduire de manière plus raisonnable des motifs réels des réseaux de rues. Un modèle très proche est décrit dans [RUI et al., 2013] **C (FL) : qu'apporte cette phrase ?**. D'autres tentatives comme [DE LEON, FELSEN et WILENSKY, 2007; YAMINS, RASMUSSEN et FOGEL, 2003] sont plus proches de la modélisation procédurale [LECHNER et al., 2004; WATSON et al., 2008] et pour cette raison n'ont pas d'intérêt pour notre cas puisqu'ils peuvent difficilement être utilisés comme modèles explicatifs. La modélisation procédurale génère des structures à la manière des grammaires de forme **C (FL) : sens ?**, mais celle-ci se concentre généralement sur la reproduction fidèle de forme locale, sans tenir compte des propriétés macroscopiques émergentes. Les classifier comme modèles de morphogenèse n'est pas correct et correspond à une incompréhension des mécanismes du *Pattern Oriented Modeling* **C (FL) : tu n'as pas du tout explique ce que c'est** d'une part et de l'épistémologie de la Morphogenèse d'autre part (voir 5.1). Nous utiliserons ce type de modèle (mixtures d'exponentielles ou réseau par connexification) **C (AB) : mixture ==> mélange; connexification : faut que tu déposes un brevet** pour générer des données synthétiques initiales uniquement pour faire tourner d'autres modèles complexes (voir 3.1 et 5.3). Enfin, une approche originale et intéressante pour la croissance des réseaux est le réseau biologique. Cette approche appartient au champ de l'ingénierie morphogénétique dont DOURSAT est un pionnier, qui vise à concevoir des systèmes complexes artificiels inspirés de systèmes complexes naturels et sur lesquels un contrôle des propriétés émergentes est possible [DOURSAT, SAYAMA et MICHEL, 2012]. Les *Machines Physarum*, qui sont des modèles d'une moisis-

sure auto-organisée (*slime mould*) ont été prouvés comme résolvant de manière efficiente et par le bas des problèmes computationnellement lourds comme des problème de routage [TERO, KOBAYASHI et NAKAGAKI, 2006] ou des problèmes de navigation NP-complets comme le Problème du Voyageur de Commerce [ZHU et al., 2013a], ce qui est porteur de sens au regard des liens entre différents types de complexité développés en 3.3. Ils produisent des réseaux ayant des propriétés de coût-robustesse Pareto-efficientes [TERO et al., 2010] qui sont typiques des propriétés empiriques des réseaux réels, et de plus relativement proches en forme de ceux-ci (sous certaines conditions, voir [ADAMATZKY et JONES, 2010]). 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, 2016], faisant un pont intéressant avec les modèles d'investissement de LEVINSON. Les critères de validation appliqués restent cependant 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.

C (FL) : la encore c'est confus : il faut hierarchiser, structurer, expliquer

2.1.3 Modéliser la co-évolution

Modélisation Hybride

Les modèles de simulation qui incluent un couplage des dynamiques de la croissance urbaine et du réseau de transport sont relativement rares, et pour la plupart au stade de modèles stylisés. **C (FL) : TB** Les efforts étant assez disparates et dans des domaines très variés, il est difficile de percevoir une unité dans ce type de modèle **C (FL) : pourquoi y en aurait-il une ?**, si ce n'est l'abstraction de l'hypothèse d'interdépendance entre réseaux et caractéristiques du territoire dans le temps. Une généralisation du modèle d'optimisation locale géométrique décrit précédemment a été développé dans [BARTHÉLEMY et FLAMMINI, 2009]. Comme pour le modèle de croissance de réseau routier dont il est l'extension, les mécanismes locaux n'ont pas de justification théorique ou thématique, et le modèle n'est de plus pas exploré et aucune connaissance géographique ne peut en

être tirée. **C (FL) : pourquoi en parler alors?** [DING et al., 2017] 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 d'un modèle de co-évolution entre réseau de transport et territoire. [LEVINSON, XIE et ZHU, 2007] prend une approche économique plus intéressante **C (FL) : de quel point de vue?** du point de vue des processus de développement de réseau impliqués, similaire à un modèle à quatre étapes **C (FL) : il faut expliquer un peu plus en détail** (génération de flux origine-destination basés sur la gravité, attribution du traffic par Equilibre Utilisateur Stochastique) 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 met à jour les actifs et emplois par modélisation de choix discrets **C (FL) : c'est une méthode : pas a mettre en avant a ce stade A1 : (JR) pas forcement d'accord avec cette hierarchisation des domaines de connaissance.(cf chap 9)**. Les expériences **C (FL) : lesquelles?** montrent que l'usage du sol et le réseau en co-évolution conduisent à des retroactions positives renforçant les hiérarchies, **C (FL) : faire deux phrases** mais sont loin d'être satisfaisantes pour deux raisons : d'une part 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 **C (FL) : par exemple? construction?** sur de plus longues échelles de temps ne sont pas pris en compte, et d'autre part les conclusions sont assez limitées puisque le comportement du modèle n'est pas connu, les analyses de sensibilité étant faites sur un petit nombre d'espaces unidimensionnels **C (FL) : la aussi c'est une autre discussion** : les mécanismes exhaustifs restent ainsi inconnus comme seuls des cas particuliers sont donnés dans l'analyse de sensibilité. [LI et al., 2016] 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, [LEVINSON et CHEN, 2005] est aussi présenté comme un modèle de co-évolution mais correspond plus à une analyse statistique couplée **C (FL) : de quoi? quest ce qui est couple?** puisqu'elle repose sur un modèle prédictif à chaîne de Markov. [RUI et BAN, 2011] 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 **C (FL) : en parlant comme cela, tu considere comme aquis que dans ce type de modele, on travaille par sous-blocs ayant des liens avec les autres : cest discutable donc a discuter A1 : (JR) ontologies separees (cf chap 1), donc necessairement decomposition modeulaire avec on-**

tologie de couplage (cf chap 9) , le modèle d'usage du sol étant conditionné à la croissance du réseau autonome. Ce modèle est mis en perspective avec d'autres modèles d'usage du sol et de croissance de réseau dans [RUI, 2013] C (FL) : et alors ? . [ACHIBET et al., 2014] décrit un modèle de co-évolution à une très petite échelle (échelle du bâtiment), dans lequel l'évolution du réseau et des bâtiments sont tous les deux régis par un agent commun (qui est influencé différemment par la topologie du réseau et la densité de population) ce qui implique une simplification trop grande des processus sous-jacents C (FL) : voir aussi les travaux d'Anne Ruas . Enfin, un modèle hybride simple exploré et appliqué à un exemple jouet de planification dans [RAIMBAULT, BANOS et DOURSAT, 2014], 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 les processus qui nous intéressent C (FL) : lesquels ? , mais le modèle produit à une petite échelle une large gamme de formes urbaines qui reproduisent les motifs typiques des établissements humains. Ce modèle est s'inspire de [MORENO, BADARIOTTI et BANOS, 2012] pour ses mécanismes de base mais permet une génération de formes bien plus larges par la prise en compte des fonctions urbaines. A cette échelle, i.e. urbaine ou 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, [WU et al., 2017] 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. De manière conceptuelle, une certaine forme de couplage fort est utilisé dans [BIGOTTE et al., 2010] 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. Enfin, le modèle proposé par [BLUMENFELD-LIEBERTHAL et PORTUGALI, 2010] peut être vu comme un pont vers 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. A une échelle macroscopique et également plus proche de la modélisation de systèmes urbains que nous développerons dans la section suivante, [BAPTISTE, 1999] propose de coupler le modèle de croissance urbaine basé sur les migrations (introduit par l'application de la synergétique au système de ville par [SANDERS, 1992]) avec un mécanisme d'auto-renforcement pour le réseau routier sans modification topologique (retroaction positive par seuils du différentiel flux-capacité sur la capacité C (FL) : difficilement comprehensible). Sa dernière version est présentée par [BAPTISTE, 2010]. Peu de conclusions générales ne peuvent cependant être tirées de ce travail, outre

que ce couplage permet de faire émerger une configuration hiérarchique (mais on sait par ailleurs que des modèles plus simples, un attachement préférentiel uniquement par exemple, permettent de reproduire ce fait stylisé) 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.

Modélisation de Systèmes Urbains

Une approche relativement proche des précédentes, mais ayant des caractéristiques propres, est celle de la modélisation intégrée des systèmes de villes. Dans la continuité des modèles Simpop pour modéliser les systèmes de villes, [SCHMITT, 2014] décrit le modèle Simpop-Net 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. COTTINEAU propose une croissance endogène des réseaux de transport comme la dernière brique de construction du cadre de modélisation MARIUS [COTTINEAU, 2014] mais cela reste à un niveau conceptuel puisque cette brique n'a pas 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. Nous nous positionnerons particulièrement dans cette lignée de recherche dans cette thèse, vu l'importance que prendra la Théorie Evolutive dans notre démarche théorique et de modélisation comme nous le détaillerons par la suite. L'ensemble des briques est nécessaire pour comprendre les implications de ce positionnement, mais le lecteur pressé pourra directement consulter le chapitre 9 **C (FL) : a eviter** pour une synthèse des implications théoriques à différents niveaux d'abstraction. Typiquement, les hypothèses épistémologiques **C (FL) : peut on vraiment parler d'hyp. epistemo?** fondamentales telles le rôle des relations et de la configuration spatiales, ou la présence d'un équilibre **C (FL) : termes a definir** - nous considérons les systèmes urbains comme des systèmes complexes adaptatifs **C (FL) : idem A1 : deja fait en intro?**, auto-organisés loin de l'équilibre, sont représentatives de cette approche si on les considère conjointement. On voit bien l'opposition aux principes épistémologiques de l'économie géographique : [Fujita, Krugman et Mori, 1999] introduisent par exemple un modèle évolutionnaire capable de reproduire une hiérarchie urbaine et une organisation typique de la Théorie des Places Centrales **C (FL) : au moins cite Christaller**, 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 homogène **C (FL) : sens ? C**
(AB) : attention, ça n'est pas forcément une limite fondamentale, tout dépend des objectifs. Peut être suffisant en particulier pour étudier l'apparition de ces « particularités » dont tu parles ensuite **A1 :** (JR) oui en effet, on rejoint la question de representation des territoires, qu'est ce qui est nécessaire a mettre ou non, selon l'objectif poursuivi (// la même discussion plus loin) // idée de com pour le cist . Cette approche peut être instructive sur les processus économiques en eux-mêmes mais aucunement sur les processus géographiques **C (FL) : pourquoi ?** , qui incluent le déroulement des processus économiques dans l'espace géographique dans lequel les particularités 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).

Co-évolution

Après cet aperçu de la littérature, incluant différents degrés de couplage entre les composantes des réseaux et territoires, nous sommes en mesure de préciser ce que nous entendrons par *modéliser la co-évolution*. La vision donnée ici a un but opérationnel, puisqu'il ne nous paraissait pas pertinent **C (FL) : c'est dommage car en l'état tu as surtout juxtapose des lectures sans en tirer tout ce qui pouvait être tiré** de donner d'emblée une vision trop théorique et abstraite (qui sera développée en 9.2). En Géographie Economique, la notion de co-évolution a également été mobilisée, notamment dans sa branche évolutionnaire **C (FL) : sens ?** . Ainsi, [WAL et BOSCHMA, 2011] 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. L'idée d'entités évolutionnaires en économie est difficilement compatible avec le courant néoclassique *mainstream* **C (FL) : on ne peut pas dire les choses comme cela c'est un jugement de valeur** , mais trouve un écho de plus en plus pertinent [NELSON et WINTER, 2009]. Pour la géographie, 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 s'est cristallisée **C (FL) : sens ?** , mais il est évident qu'elle était présente dès les fondements de la théorie comme le rappelle DENISE PUMAIN (voir D.4) : le système complexe adaptatif est composé d'entités en dépendance causales **C (FL) : sens ?** fortes **C (AB) : pas d'ac-**

cord ! les dépendances faibles (interactions indirectes) font également partie du dispositif. Défini ce que tu appelles « entités ». 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. [PAULUS, 2004] 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 [SCHMITT, 2014] repose sur une entrée par la Théorie Evolutive, mais n'approfondi pas au delà d'une lecture des systèmes de villes comme entités fortement interdépendantes. Or l'interdépendance est une notion aussi lâche que le fameux "tout interagit avec tout" C (FL) : mal dit , c'est à dire qu'elle est particulièrement creuse si elle n'est pas quantifiée C (FL) : c'est un postulat fort. je ne suis pas certain d'être d'accord A1 : (JR) ok c'est mal dit, malentendu sur la notion de quanti - je veux dire une comprehension fine et integree . Elle permet comme prémisses épistémologique de considérer certaines ontologies et certaines démarches de modélisation, mais ne permet pas de comprendre finement la structure et les processus d'un système. Par exemple, étant donné un réseau topologique d'interaction entre entités et des motifs temporels de propagation correspondants, on peut se demander quels sont les motifs de corrélations statiques et dynamiques correspondants, s'il existe des causalités et à quelles échelles C (FL) : a supprimer cet exemple n'éclaire pas la proposition . Il existe en pratique une infinité de "régimes" de co-évolution possibles C (FL) : discutable , liés à la structure du réseau écologique de la niche correspondante si on interprète celle-ci de cette façon [HOLLAND, 2012]. L'idée de diffusion hiérarchique de l'innovation dans la théorie évolutive capture par exemple qualitativement certains de ces aspects, mais la quantification des régimes correspondants et donc de la co-évolution reste une question ouverte. L'une de nos contributions principales C (FL) : c'est un point important qui se trouve un peu dilué ici , qui aboutira comme produit des efforts empiriques et de modélisation C (AB) : sens de cette incise ? , sous forme théorique en 9.2, sera de clarifier cette notion et d'en donner une définition précise. A ce stade, l'état de l'art fait 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 faible épaisseur C (FL) : sens ? 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 ci-dessus sont précurseurs.

* *

*

2.2 UNE APPROCHE EPISTÉMOLOGIQUE

Un corolaire de la matière thématique introduite **C (FL) : phrase a reprendre** en chapitre 1 est le besoin d'une compréhension des disciplines impliquées elles-même pour être en mesure de construire des modèles hétérogènes intégrés. Les possibilités de couplage et d'intégration **C (FL) : definir, differencier, clarifier A1 : (JR) cf intro** sont hautement déterminées par les approches existantes et les lacunes correspondantes qui ont été exposées dans la section précédente 2.1. Cela implique une étude épistémologique avancée dans chaque champ, que nous proposons de mener de manière quantitative et systématique. Ce choix délibéré **C (FL) : choix qui n'est pas justifie de facon satisfaisante. pourquoi = pourquoi faire? quattends tu de cela et pourquoi l'approche classique ne convient pas? A1 : (JR) importance de la reflexivite et de l'epistmo quanti pour les approches integratives** pourrait occulter des considérations épistémologiques élaborées mais suit notre objectif d'investigations préliminaires pour la construction de modèles, en révélant potentiellement des directions de recherche.

Nous décrivons et explorons d'abord un algorithme de revue systématique algorithmique, qui reconstruit des corpus de références par une extraction sémantique itérative **C (FL) : ajouter precisions**. 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.

2.2.1 Revue Systématique Algorithmique

Une étude bibliographique étendue suggère une rareté **C (FL) : par rapport a quoi?** des modèles quantitatifs de simulation qui intègrent à la fois la croissance urbaine et la croissance des réseaux. Cette absence pourrait être due aux intérêts divergents des acteurs des disciplines **C (FL) : peut on parler d'interets de disciplines de facon univoque? A1 : (JR) acteurs des disciplines?** concernées qui induiraient un manque de communication. Nous proposons de procéder à une revue de la littérature systématique et algorithmique pour donner des éléments de réponse quantitatifs à cette question. 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 procédon 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.

En recherche de modèles de co-évolution

Comme développé en 1.1, les réseaux de transport et l'usage du sol urbain sont connus pour être des composantes au couplage complexe au sein des systèmes urbains **C (FL) : trop rapide, cest une lecture possible mais pas une realite bien connue**, à différentes échelles [BRETAGNOLLE, PUMAIN et VACCHIANI-MARCUZZO, 2009]. Une approche commune est de les considérer comme étant en co-évolution, tout en évitant les interprétations trompeuses **C (FL) : la aussi tu simplifies trop** comme le mythe des effets structurants des infrastructures de transport [OFFNER, 1993]. Une question qui se présente rapidement **C (FL) : mal dit** est l'existence de modèles endogénéisant cette co-évolution, i.e. prenant en compte simultanément la croissance urbaine et celle du réseau. Nous essayons d'y répondre par une revue systématique algorithmique **C (FL) : 1) tu l'as déjà dit, 2) tu ne justifies pas cette approche**. Nous proposons dans cette section de développer cette approche en formalisant l'algorithme, dont les résultats sont ensuite présentés et discutés.

Modéliser les interactions entre croissance urbaine et croissance des réseaux

Nous avons revu selon divers point de vue les efforts de modélisation des interactions entre territoires et réseaux dans la section précédente 2.1. Cet état de l'art nous suggère fortement des domaines relativement cloisonnés et s'intéressant à des problématiques différentes.

Analyse Bibliométrique

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 [RUCKER, 2012]. Passer sous silence certaines références peut même être considéré comme une erreur scientifique dans le contexte de l'émergence des systèmes d'information qui par l'accès plus aisément à l'information rend difficilement justifiable l'omission de références clés [LISSACK, 2013]. **C (FL) : phrase risquée : est-tu certain de n'être passé à côté de rien ? par exemple as tu lu Brotchie 1984 ? A1 : (JR) justement la est la contradiction et la justification de l'approche : maximiser couverture, tout en sachant qu'on ne peut pas tout couvrir (perspectivisme) - à développer** Nous proposons de tirer parti de telles techniques pour traiter notre problème. En effet, l'observation de la bibliographie obtenue dans la section précédente

soulève une hypothèse. On peut postuler sans risques **C (FL) : tournure a reprendre** à partir de la revue précédente **2.1 C (AB) : ?** Il semble clair que toutes les briques sont présentes pour l'existence de modèles co-évolutifs mais des questionnements et objectifs différents semblent la stopper. Comme montré par [COMMENGES, 2013b] pour le concept de mobilité, pour lequel un "petit monde d'acteurs" **C (FL) : sens ?** relativement fermé, en l'occurrence les Corpsards de l'Ecole des Ponts et Chaussées **C (FL) : tu generalises trop** **C (AB) : phrase à reformuler**, a inventé une notion *ad hoc*, utilisant des modèles sans connaissance préalable d'un contexte scientifique plus général. On pourrait se trouver dans un cas similaire pour le type de modèles auxquels on s'intéresse. Des interactions restreintes entre des champs scientifiques travaillant sur les mêmes objets mais avec des objectifs et contextes divergents, et à des échelles différentes, pourrait être à l'origine de l'absence de modèles co-évolutifs. Tandis que la majorité des études en bibliométrie se reposent sur les réseaux de citation [NEWMAN, 2013] ou les réseaux de co-auteurs [SARIGÖL et al., 2014], nous proposons d'utiliser un paradigme moins exploré, basé sur l'analyse textuelle, introduit par [CHAVALARIAS et COINTET, 2013], qui produit une cartographie dynamique des disciplines scientifiques en se basant sur leur contenu sémantique. Nous postulons que cette couche supplémentaire d'information apporte une information complémentaire **C (AB) : reformuler** **C (FL) : c'est une posture originale, dommage de ne pas avoir applique de methode plus classique pour commencer A1 : (JR) c'est l'objet du 2.1**, nécessaire pour appréhender la diversité des domaines. La méthode est particulièrement adaptée pour notre étude puisque nous voulons comprendre la structure du contenu des recherches sur le sujet. Nous appliquons une approche algorithmique décrite par la suite. 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.

DESCRIPTION DE L'ALGORITHME Soit A un alphabet (un ensemble arbitraire de symboles), A^* les mots correspondants **C (FL) : sens ?** et $T = \cup_{k \in \mathbb{N}} A^{*^k}$ **C (FL) : que signifie cette notation ? A1 : (JR) INF101** les textes de longueur finie sur celui-ci. 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 **C (FL) : on ne voit pas le lien avec Astar** à l'itération n sera noté $\mathcal{C} \subset T^3$ **C (FL) : pourquoi ny a til pas n ? A1 : (JR) on parle d'un element** : il s'agit d'un sous-ensemble de triplets de textes. Nous supposons l'existence d'un ensemble de mots-clés \mathcal{K}_n ,

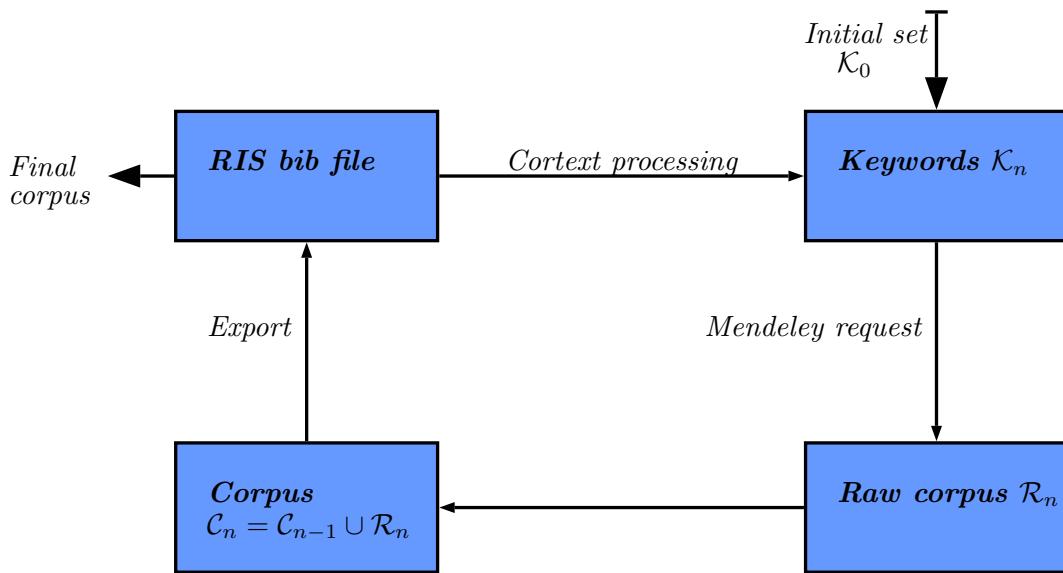


FIGURE 6 : Architecture globale de l'algorithme, incluant des détails d'implémentation : la requête au catalogue est faite via l'API Mendeley **C (FL) : tu ne l'as pas dit auparavant** ; les corpus finaux sont sous forme de fichiers RIS. **C (FL) : traduire ; pourquoi mendeley ?**

les mots-clés initiaux étant \mathcal{K}_0 , spécifiés par l'utilisateur¹. Une itération procède de la manière suivante :

C (FL) : Commencer par expliquer les objets qui seront utilisés

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 mots-clés \mathcal{K}_n sont extraits du corpus par Traitement du Langage Naturel (NLP), étant donné un paramètre fixé N_k **C (FL) : pourquoi cette notation ?** donnant le nombre de mots-clés. **C (FL) : expliquer ce dont il s'agit**

L'algorithme s'arrête quand la taille du corpus devient stable **C (FL) : quel sens ? quel critère ?** ou quand un nombre maximal d'itérations défini par l'utilisateur est atteint. La figure 6 synthétise le processus général.

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. **C (FL) : oui mais justement l'explication du choix de ce corpus est cruciale**

² 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.

RÉSULTATS Les détails précis concernant l'implémentation de l'algorithme ainsi qu'une analyse de sensibilité pour vérifier la convergence sur un échantillon de requêtes initiales (typiques des champs étudiés **C (FL) : ?**) sont donnés en Appendice A.2. Lorsque l'algorithme a été partiellement validé par cette analyse, nous l'appliquons à notre question. **C (AB) : a revoir** Nous partons de cinq différentes requêtes initiales qui ont été manuellement extraites des divers domaines identifiés dans la bibliographie (qui sont "city system network", "land use transport interaction", "network urban modeling", "population density transport", "transportation network urban growth") **C (FL) : ces mots sont nombreux, quel impact sur l'algo?** **C (FL) : du coup tu exclues totalement la literature francophone. pourquoi ce choix?**³. Nous prenons l'hypothèse la plus faible pour le paramètre $N_k = 100$ (plus N_k est grand, plus les domaines atteints devraient être moins restreints et donc plus des résultats de distance seront significatifs) **C (FL) : phrase non compréhensible**. Après avoir construit les corpus, nous étudions leur cohérence lexicale 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 1 montre les valeurs **C (FL) : tu n'as pas explique comment tu quantifies cela** de la proximité lexicale relative, qui est significativement faibles **C (FL) : sens?** sachant que les chiffres peuvent directement s'interpréter comme une proportion de mots en co-occurrence, ce qui tend à confirmer notre hypothèse. Pour situer ces résultats de manière relative, il faudrait un modèle nul **C (FL) : ?** avec des corpus aléatoires par exemple, ce qui pourrait faire l'objet de développements futurs.

Les développements possibles incluent la construction de réseaux de citation via un accès automatique à Google Scholar **C (FL) : pourquoi cette plateforme et pas une autre?** **C (AB) : c'est ton unique source?** Si je fais une recherche à la main dans google scholar avec les termes « urban growth networks » j'obtiens des références, par exemple Streets tree networks and urban growth : optimal geometry for quickest access between a finite-size volume and one point A Bejan, GA Ledezma - Physica A : Statistical Mechanics and its ..., 1998 qui fournit les citations entrantes. La confrontation des coefficients inter-clusters pour le réseau de citations entre les différents corpus avec la cohérence lexicale est un aspect clé d'une validation approfondie des résultats.

L'absence peu explicable **C (FL) : pourquoi cette hypothese?** **C (AB) : qu'entends tu par « absence » de modèles?** a priori de

³ Ce choix est arbitraire, **C (FL) : on espere que non** 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.

TABLE 1 : Matrice symétrique des proximités lexicales entre les corpus finaux, définies comme la somme des co-occurrences totale de mots-clés finaux entre corpus, normalisé par le nombre de mots-clés finaux (100). La taille des corpus finaux est donnée par W . Les valeurs obtenues pour les proximités sont considérablement faibles, ce qui confirme que les corpus sont éloignés de manière significative (voir texte). **C (FL)** : ce tableau est difficilement comprehensible; traduire; trop de chiffres significatifs. **C (AB)** : C'est vrai que sans valeurs de références c'est difficile à interpréter. Quelle serait la valeur max pour un corpus de taille X et un nombre de mots clés Y ?

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

modèles qui simulent la co-évolution des réseaux de transport et de l'usage du sol urbain, qui se confirme à première vue par un état de l'art couvrant des domaines disparates, pourrait être due à l'absence de communication entre les disciplines 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. Les premiers résultats numériques semblent confirmer l'hypothèse **C (AB) : laquelle ?**. Cependant, une telle analyse quantitative ne doit pas être considérée seule, mais devrait plutôt venir comme soutien à des études qualitatives qui peuvent être l'objet de développements futurs, comme celle menée dans [COMMENGES, 2013b], dans laquelle des questionnaires avec des acteurs historiques fournissent des informations extrêmement pertinentes. **C (FL) : ET A UNE REVUE DE BIBLIO CLASSIQUE A1 : (JR) faite en 2.1**

2.2.2 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 scienti-

fique 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 **C (FL) : sens?** 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 de l'interdisciplinarité **C (FL) : est ce que si les disciplines sont (re) construites par l'algo, la notion d'interdisciplinarite a le meme sens?** et du paysage scientifique sur notre sujet.

Contexte

La majorité des disciplines scientifiques présentent un besoin **C (FL) : n'y a t'il pas la deja un postulat fort?** fort en interdisciplinarité et approches transversales, comme illustré par exemple par l'édition spéciale récente de Nature sur le sujet ([“[Interdisciplinarity, Nature Special Issue](#)”]), pour diverses raisons qui peuvent inclure le développement de champs intégrés verticalement **C (FL) : sens?** conjointement aux questions horizontales comme détaillé dans la feuille de route des Systèmes Complexes **C (FL) : expression grandiloquente A1 : (JR) pas d'autre facon de nommer la roadmap** ([BOURGINE, CHAVALARIAS et AL., [2009](#)]). Les débats courants sur la nature exacte de l'interdisciplinarité sont bien sûr nombreux (d'autres termes existent comme transdisciplinarité ou cross-disciplinarité), et celle-ci dépend en fait des domaines impliqués : des disciplines hybrides apparues récemment (voir par exemples celles soulignées par [BAIS, [2010](#)] comme l'astro-biologie **C (FL) : la geomatique**) sont une bonne illustration du cas où les intrications sont très fortes, tandis que des champs plus mou **C (FL) : mauvaise strategie je pense : avant de designer une discipline en particulier, il faut preciser le sens de "mou"** comme “l’urbanisme” qui n’ont pas de définition précise **C (FL) : mal dit** montrent comment l’intégration horizontale est nécessaire et comment de la connaissance transversale peut être produite (menant à des possibles malentendus **C (FL) : et alors? est ce que les malentendus sont necessairement contre-productifs?** lorsque récemment introduite trop brutalement à des physiciens comme montré par [DUPUY et BENGUIGUI, [2015](#)]). Cette question se transfère naturellement au champ de la communication scientifique : quelles sont les alternatives correspondantes pour une dissémination efficace de la connaissance ? Des éléments de réponse à une question si générale impliquent, dans une perspective *evidence-based* **C (FL) : traduire A1 : (JR)non traduisible - appuyer quelque part sur le sens de l'evidence-based** , des mesures quantitatives de l'in-

terdisciplinarité, qui font partie d'une approche multi-dimensionnelle de l'étude de la science, en quelque sorte "au-delà de la bibliométrie" [CRONIN et SUGIMOTO, 2014] **C (FL) : que tu n'as pas introduite**

Les méthodes potentielles pour des entrées quantitatives en épistémologie sont nombreuses. En utilisant les caractéristiques des réseaux de citation, un bon pouvoir prédictif pour les motifs de citation est par exemple obtenu par [NEWMAN, 2013]. Les réseaux de co-auteurs peuvent également être utilisés pour des modèles prédictifs [SARIGÖL et al., 2014]. Une approche multi-couches a récemment été proposée par [OMODEI, DE DOMENICO et ARENAS, 2016], utilisant des réseaux bipartites des papiers et des chercheurs, dans le but de produire des mesures d'interdisciplinarité. Les disciplines peuvent être stratifiées en couches pour révéler des communautés entre elles et ainsi des motifs de collaboration [BATTISTON et al., 2015]. Les réseaux de mots-clés sont utilisés dans d'autres champs comme l'économie de l'innovation : par exemple, [CHOI et HWANG, 2014] introduit une méthode pour identifier les opportunités technologiques en détectant des mots-clés importants au sens des mesures topologiques. [SHIBATA et al., 2008] utilise l'analyse topologique du réseau de citations pour déterminer des fronts de recherche émergents.

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 [PALCHYKOV et al., 2016] pour les articles en physique mais sans extraction des mots-clés, ou par [GURCIULLO et al., 2015] pour un analyse des réseaux sémantiques de débats politiques) avec celle du réseau de citations pour extraire par exemple des mesures d'interdisciplinarité. Notre 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. [NICHOLS, 2014] introduit une approche similaire, utilisant le modèle d'extraction de thématiques *Latent Dirichlet Allocation*⁴ pour caractériser l'interdisciplinarité de récompenses dans des sciences précises. [LARIVIÈRE et

⁴ Le modèle LDA, introduit par [BLEI, NG et JORDAN, 2003], suppose les documents comme produits par des thèmes sous-jacents, avec une distribution de Dirichlet pour

GINGRAS, 2014] quantifie l'interdisciplinarité sur une longue période temporelle en étudiant l'étendue de la bibliographie des publications.

C (FL) : vague

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. C (FL) : B Pour cela, nous rassemblons et compilons 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 standard⁵ qui peut être citée par une autre (articles de journaux, livre, chapitre de livre, article d'actes, communication, etc.) et contient des informations de base (titre, résumé, auteurs, année de publication). Nous travaillons par la suite sur le réseau des références. Il est important de noter qu'une contribution fondamentale C (FL) : le dire avant de cette partie 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.

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.2. Celui-ci est pris de taille raisonnable C (FL) : sens ?, mais les méthodes utilisées ici ont été développées sur des données massives, pour les brevets par exemple [BERGEAUD, POTIRON et RAIMBAULT, 2017].

DONNÉES DE CITATION Le réseau de citations est reconstruit à partir de Google Scholar qui est souvent l'unique source des citations entrantes [NORUZI, 2005] puisqu'en science humaines les ouvrages ne sont pas systématiquement référencés par les bases fournit des services (payants) comme le réseau de citation.⁶ Nous sommes conscient des biais possibles de l'utilisation de cette source unique (voir par exemple [BOHANNON, 2014])⁷, 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-

leur composition ainsi que pour la distribution des mots par thèmes. Son estimation donne la composition des thèmes en terme de mots-clés.

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

6 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.

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

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). Le corpus n'est pas très international (contrairement par exemple au thème de la croissance urbaine, étudié pour le développement thématique sur les liens entre économie et géographie développé en C.1).

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 $\gamma = 0.0013$ C (AB) : comparer avec des valeurs de référence ? . Le degré entrant moyen (qui peut être interprété comme un facteur d'impact stationnaire) est de 1.26, ce qui est relativement élevé pour des sciences humaines. Il est important de noter sa connexité faible, ce qui signifie que les domaines initiaux ne sont pas en isolation totale : les références initiales sont partagées à un degré minimal par les différents domaines. Nous travaillons sur la suite sur le sous-réseau des noeuds comprenant au moins deux liens, pour extraire le cœur de la structure du réseau et se débarrasser de l'effet "grappe". De plus, le réseau est nécessairement complet entre ces noeuds puisqu'on est remonté au deuxième niveau. Nous procédons à une détection de communautés par l'algorithme de Louvain, sur le réseau non-dirigé correspondant. On obtient 13 communautés C (FL) : comment ? c'est elliptique A1 : (JR) dit juste avant, algo de Louvain , de modularité dirigée 0.66, 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 suivants : LUTI (18% du réseau), Géographie Urbaine et des Transports (16%), Planification des infrastructures (12%), Planification intégrée - TOD (6%), Réseaux Spatiaux (17%), Etudes d'accessibilité (18%). C (FL) : quelle est leur cohérence interne ? comment as tu donné les noms ? A1 : (JR) cf remarque de Caruso a ECTQG La Fig. 7 permet de visualiser les relations de 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 est un résultat en lui-même témoignant d'une certaine séparation. **C (FL) : phrase alambiquee** Ceux-ci font le pont entre les approches Réseaux spatiaux et les approches géographiques, qui comportent une partie importante de sciences politiques par exemple. Les liens entre physique et géographie restent très faibles. Ce panorama dépend bien sûr du corpus initial, mais nous permet de mieux comprendre le contexte de celui-ci dans son environnement disciplinaire.



FIGURE 7 : 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 [CHAVALARIAS et COINTET, 2013]. 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)

C (FL) : statut de ces mots ? . 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.2. 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.2.

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 [YANG et al., 2000] 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 2. 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

TABLE 2 : Description des communautés sémantiques. On donne leur taille, leur proportion en quantité de mots-clés cumulés sur l'ensemble du corpus, et des mots-clés représentatifs sélectionnés par degré maximal.
C (FL) : pourquoi ces mots sont ils tronqués ? **A1 :** (JR) c'est explique dans le texte (surtout en annexe) il s'agit de stem

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

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 constante l'importance des réseaux, des problématiques de sciences politiques et socio-économiques. Nous 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 7.3.

MESURES D'INTERDISCIPLINARITÉ La distribution des mots clés dans les communautés permettent de définir une mesure d'interdis-

ciplinarité 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'interdisciplinarité [BERGEAUD, POTIRON et RAIMBAULT, 2017] est alors $I_i = 1 - \sum_j p_{ij}^2$. Soit \mathbf{A} la matrice d'adjacence du réseau de citation, et soit \mathbf{I}_k les matrices de sélection des lignes correspondantes à la classe k de la classification de citation : $\text{Id} \cdot \mathbb{1}_{c(i)=k}$, telle que $\mathbf{I}_k \cdot \mathbf{A} \cdot \mathbf{I}_{k'}$ donne exactement les citations de k vers k' . La proximité de citation entre les communautés de citation est alors définie par $c_{kk'} = \sum \mathbf{I}_k \cdot \mathbf{A} \cdot \mathbf{I}_{k'} / \sum \mathbf{I}_k \cdot \mathbf{A}$. On définit la proximité sémantique en définissant une matrice de distance entre références par $\mathbf{D} = d_{ii'} = \sqrt{\frac{1}{2} \sum (p_{ij} - p_{i'j})^2}$ puis la proximité sémantique par $s_{kk'} = \mathbf{I}_k \cdot \mathbf{D} \cdot \mathbf{I}_{k'} / \sum \mathbf{I}_k \sum \mathbf{I}_{k'}$. Nous montrons en Fig. 8 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 exemples, 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 de la sous-section précédente en terme 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

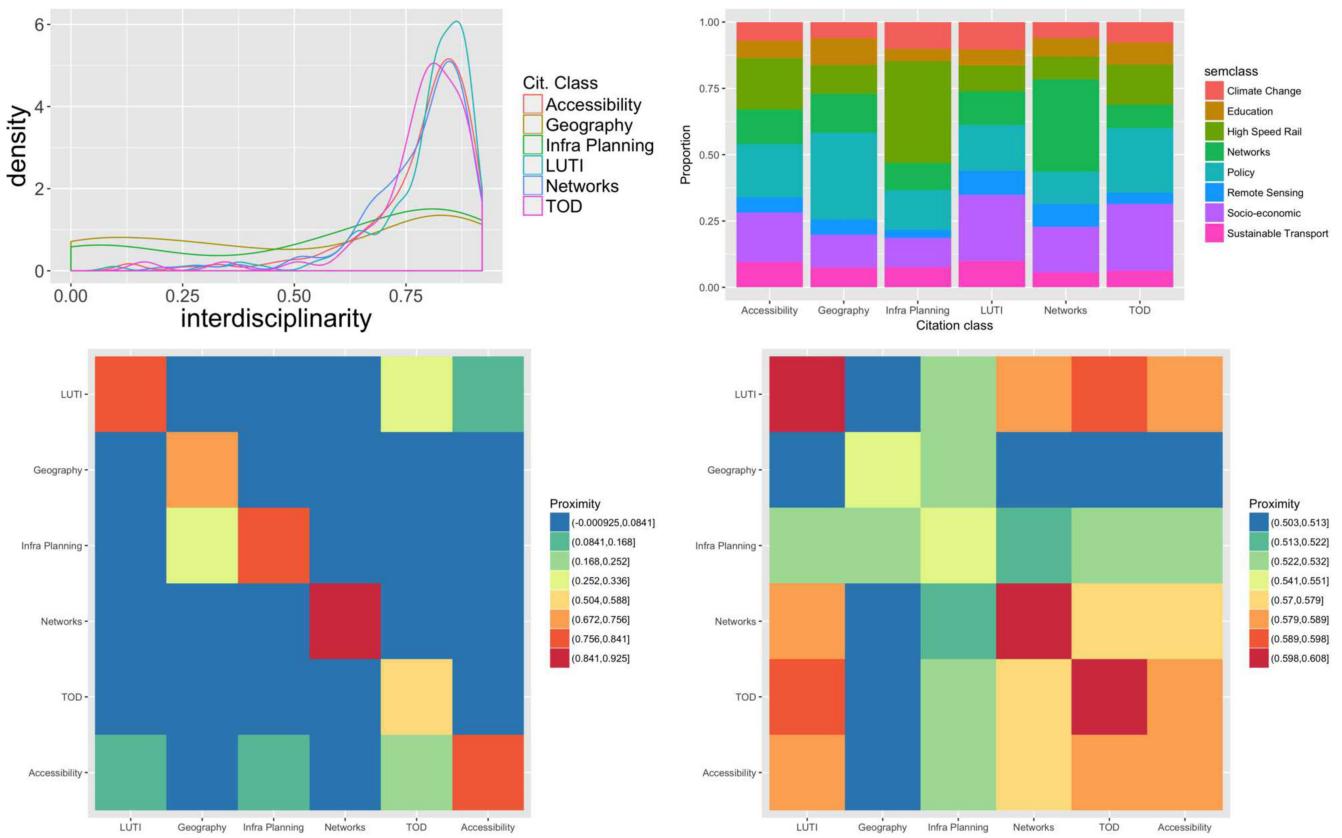


FIGURE 8 : **Motifs d'interdisciplinarité.** (Haut Gauche) Distribution des I_i par classes de citations; (Haut Droite) Composition sémantiques des classes de citation; (Bas Gauche) Matrice de proximité de citation $c_{kk'}$ entre classes de citations; (Bas Droite) Matrice de proximité sémantique $s_{kk'}$ entre classes de citations. **C (FL) :** tu compares les disciplines entre elles mais pas la façon dont elles attaquent les questions au coeur de ta thèse. c'est dommage. **A1 :** (JR)c'est l'objet de la section suivante

méthodes des ces champs liés au planing 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. On ne se parle pas alors qu'on parle des langages pas si lointains, d'où la pertinence accrue de les faire parler d'une commune voie dans nos travaux : nos modèles devront mobiliser des éléments, ontologies et échelles de ces différents champs.

Nous concluons cette analyse par une approche plus robuste pour quantifier les proximités entre couches de l'hyperréseau. Il est aisément de construire une matrice de corrélation entre deux classifications, par les corrélations de leur colonnes. Nous définissons les probabilités P_C toutes égales à 1 pour la classification de citation. La matrice de correlation de celle-ci avec P s'étend de -0.17 à 0.54 et a une moyenne de valeur absolue de 0.08, ce qui est significatif par rapport à des

classifications aléatoire puisque un bootstrap à $b = 100$ répétitions avec les matrices mélangées donne un minimum à -0.08 ± 0.012 , un maximum à 0.11 ± 0.02 et une moyenne absolue à 0.03 ± 0.002 . Cela montre que les classifications sont complémentaires et que cette complémentarité est significative statistiquement par rapport à des classifications aléatoires. L'adéquation de la classification sémantique par rapport au réseau de citation peut également être quantifiée par la modularité multi-classes [NICOSIA et al., 2009] (voir ?? pour une définition mathématique), qui traduit la probabilité qu'un lien soit dû à la classification étudiée, en prenant en compte l'appartenance simultanée à de multiples classes. Ainsi, la modularité multi-classes des probabilités sémantiques pour le réseau de citation est de 0.10, ce qui d'une part est significativement signe d'adéquation, un bootstrap toujours à $b = 100$ donnant une valeur de 0.073 ± 0.003 , qui reste limitée vu la valeur maximale fixée par les probabilités de citations dans leur propre réseau qui donnent une valeur de 0.81, ce qui confirme d'autre part la complémentarité des classifications.

2.2.3 Discussion

Vers une modélisation des thèmes et une extraction automatique du contexte

Une direction possible pour renforcer cette analyse en épistémologie quantitative serait de travailler sur les textes complets des références contenant des efforts de modélisation des interactions entre réseaux et territoires, avec le but d'extraire automatiquement les thématiques des articles. Des méthodes plus adaptées pour les long texte que celle utilisée ici incluent par exemple l'Allocation Latente de Dirichlet [BLEI, NG et JORDAN, 2003]. L'idée serait de procéder à une sorte de modélographie automatique **C (AB) : référence thèse Clara ?**

A1 : oui ! , pour extraire des caractéristiques 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 du langage **C (FL) :?** , 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 potentialités 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 [RUCKER, 2012]). Dans notre cas, l'exercice est bien plus aléatoire pour les raisons exposées dans les deux sections précédentes : les objets sont hybrides, les problématiques diverses, et les disciplines variées. Les différents points soulevés par la suite auront souvent autant de valeur thématique que de valeur méthodologique, suggérant des points cruciaux lors de la réalisation d'une telle revue systématique hybride.

Nous proposons une méthodologie hybride couplant les deux méthodologies développées précédemment avec une procédure plus classique de revue systématique. Nous souhaitons à la fois une représentativité de l'ensemble des disciplines que l'on a découvertes, mais aussi un bruit limité dans les références prises en compte pour la modélographie. Nous adoptons pour cela le protocole suivant :

1. Partant du corpus de citation isolé en 2.2.2, 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. 9, 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-hub 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 paraît 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 de requête. La littérature grise ainsi que les pre-prints sont pris en compte différemment selon les champs.

- Le passage en revue manuel des grands corpus permet de pas louper des “poids lourds” qui auraient pu être omis en amont [LIS-SACK, 2013]. La question de la mesure dans laquelle on peut s’attendre d’être au courant de la manière la plus exhaustive des découvertes récentes liées au sujet étudié évolue très probablement vu l’augmentation de la quantité totale de littérature produite et la fragmentation des domaines pour certains toujours plus pointus [BASTIAN, GLASZIÖU et CHALMERS, 2010]. Rejoignant les points précédents, on peut supposer que des outils d’aide à l’analyse systématique permettront de garder cet objectif raisonnable.
- Les résultats de la revue automatique sont sensiblement différents des domaines dessinés dans la revue classique : certaines associations conceptuelles, notamment l’inclusion des modèles de croissance de réseaux, ne sont pas naturelles et existent peu dans le paysage scientifique comme nous l’avons montré précédemment.

D’autre part, l’opération de construction du corpus permet déjà 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 sélection 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 8.1), 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, surtout en France où les Corps des Ponts ou des Mines ont une main mise relative sur les deux aspects simultanément.

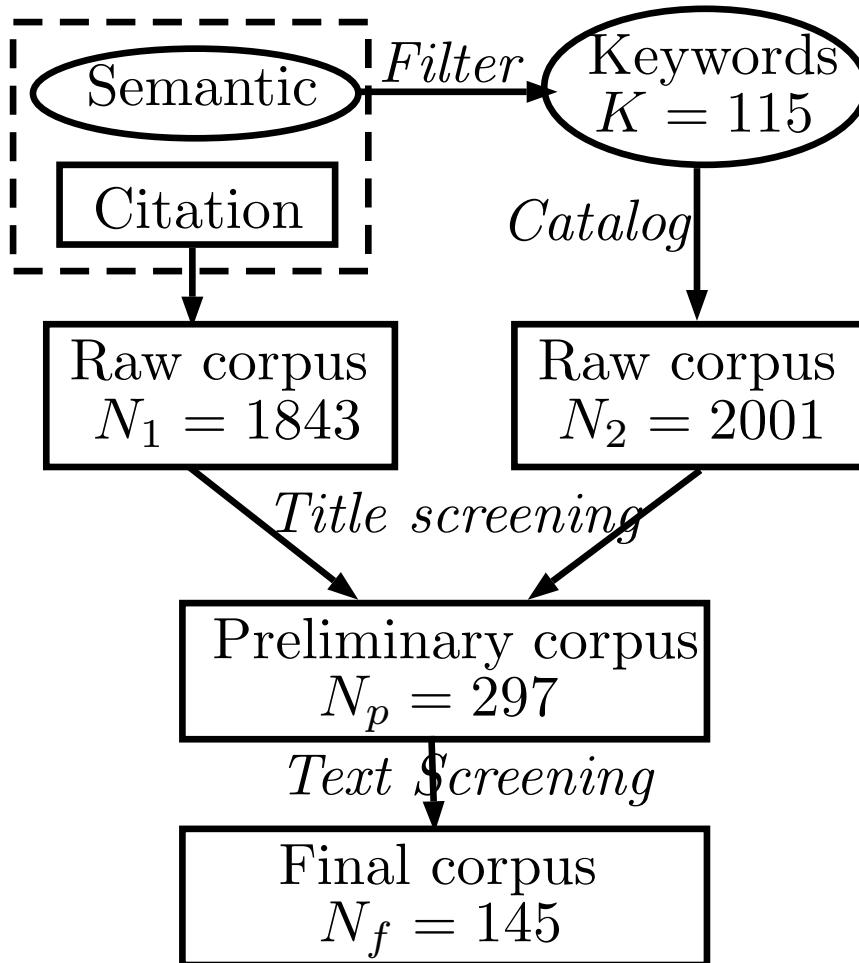


FIGURE 9 : Méthodologie de la revue systématique.

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'étudier 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 [COTTINEAU, 2016] 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 10 : {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.

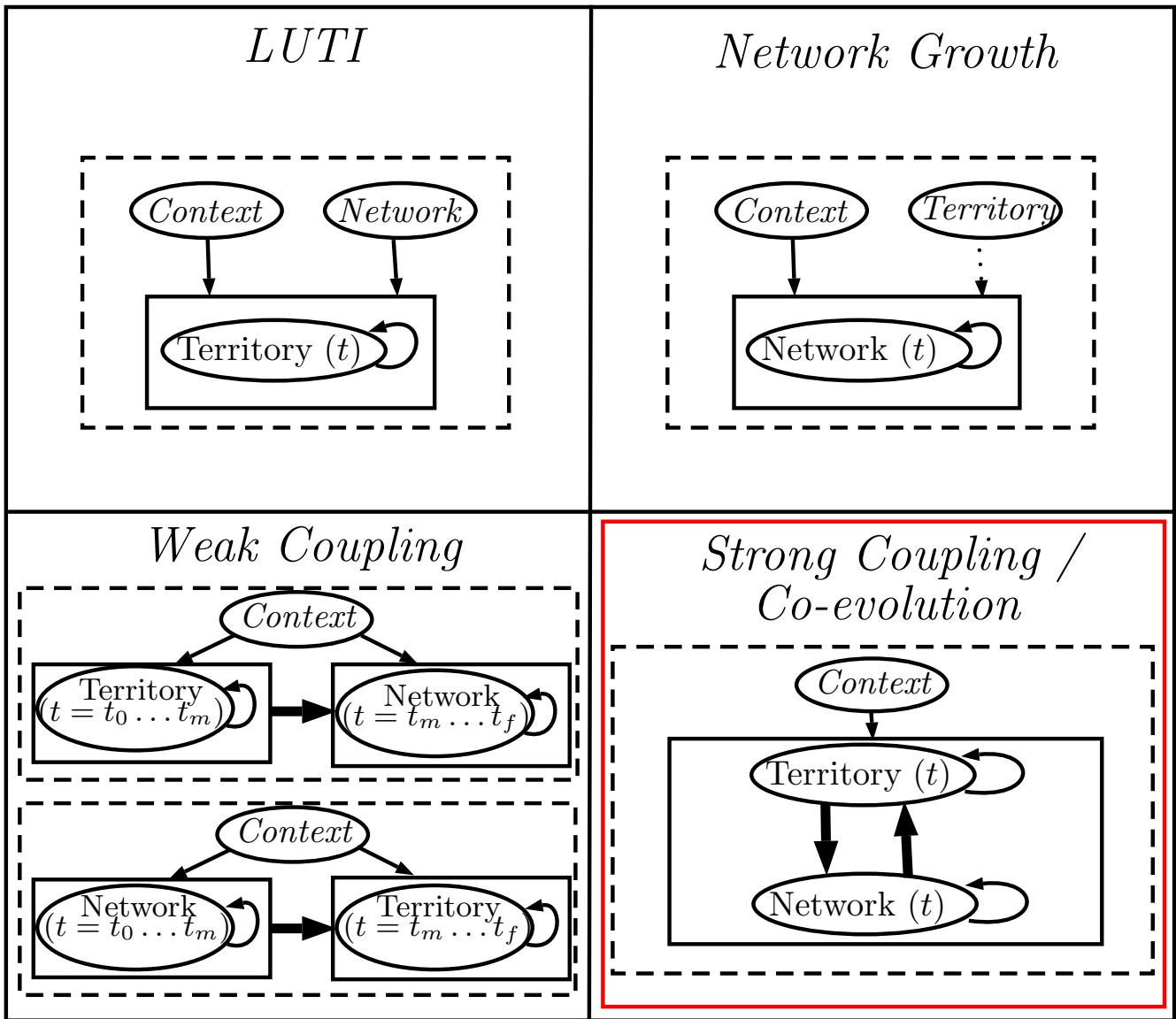


FIGURE 10 : **Représentation schématique de la distinction entre différents types de modèles couplant territoires et réseaux.** Les ontologies sont représentés par des ovales, les sous-modèles par les boîtes pleines, les modèles par les boîtes pointillées, les couplages par les flèches. Nous soulignons 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 **C (FL) : B** : 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 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.

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 ?? 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.

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

TABLE 3 : Type de modèles étudiés selon les différentes classifications. Tables de contingence de la variable discrete 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	Plan-	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	

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 l'influence de divers facteurs sur les caractéristiques des modèles par des régressions linéaires simples. Dans une démarche de multi-modélisation, nous testons l'ensemble des modèles possible pour expliquer la variable à partir des autres, et sélectionnons le meilleur en terme de Critère d'Information d'Akaike. **C (AB) : expliquer rapidement et référencer en note de bas de page. ; Version simplifiée du tableau de résultats à placer ici ?** Les résultats complets des régressions sont donnés en Appendice A.3. Les échelles temporelle et d'espace sont les mieux expliquées par les modèles prenant en compte l'ensemble des autres variables. Pour l'échelle de temps, les variables les plus significative sont le fait d'utiliser des méthodes de simulation et le fait d'être en physique, qui tous deux influent négativement. L'échelle spatiale et le fait d'être en planification influent positivement. Au contraire pour

l'échelle d'espace, le fait d'être en planning influence négativement alors que le domaine sémantique du TOD est positif, ce qui veut dire que les journaux de planning privilégient des études localisées alors que des problématiques proches ont tendance à étendre l'aire d'étude. Le niveau d'interdisciplinarité est le mieux expliqué par une unique variable, l'année, qui l'influence de manière négative, ce qui confirme l'augmentation des spécialisations scientifiques dans le temps.

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

2.3.3 *Discussion*

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 [COTTI-

NEAU, CHAPRON et REUILLOU, 2015]. 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 mété-analyse 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.

SYNTHÈSE Nous proposons, pour les modèles identifiés comme incluant la co-évolution, de faire une synthèse des ontologies, des processus, des contextes et des échelles pris en compte.

C : tableau de type intro partie III pour résumer le contenu des modèles

C (JR) : here fix scales and process, and why morphogenesis and evolutive urban theory

★ ★

★

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 [PUMAIN, 2005] comme "une nouvelle mesure de complexité", qui serait liée aux nombre de point de vue nécessaires pour appréhender un système à un niveau donné d'exhaustivité. Cette idée rejoint la position de *perspectivisme appliqué* que la section B.7 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, RIPOLL alerte contre l'illusion d'une recherche objective et les dangers de la technocratie [RIPOLL, 2017]. Nous ne rentrerons pas dans ces débats bien trop vastes pour être traités même en un chapitre, puisqu'il rejoignent des thèmes de sciences politiques, d'éthique, de philosophie, liés par exemple à la gouvernance scientifique, à l'insertion de la science dans la société, à la responsabilité scientifique. Il est clair que même des sujets *a priori* intrinsèquement objectifs, comme la physique des particules et des hautes énergies, ont des implications regardant d'une part les choix de leur financements et les externalités associées (par exemple, l'existence du CERN a largement contribué au développement du calcul distribué), mais d'autre part aussi les applications potentielles des découvertes qui peuvent avoir des répercussions sociales considérables. En biologie, l'éthique est au cœur des principes fondateurs des disciplines, comme en témoignent les débats soulevés par l'émergence de la biologie synthétique [GUTMANN, 2011]. Les tenants d'approche prudentes dans celle-ci se recoupent avec la biologie intégrative, or les Sciences Intégratives défendues par PAUL BOURGINE, mises en oeuvre par l'intermédiaire du campus digital Unesco CS-DC¹, ont typiquement la responsabilité sociale et l'implication citoyenne au cœur de leur cercle vertueux. En Sciences Humaines et Sociales, comme les recherches interagissent avec les objets étudiés (en quelque sorte l'idée des *interactive kind* de HACKING [HACKING, 1999]), les implications politiques et sociales de la recherche sont bien évidemment indiscutables. 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 certains critères de scientificité,

¹ <https://www.cs-dc.org/>

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

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, qui consiste en le développement précis de certains positionnements qui ont une influence particulière dans notre démarche de recherche. Par exemple, le travail en données quasi-intégralement ouverte et en architecture modulaire résulte de notre exigence de reproductibilité. L'utilisation des modèles et la manière de les explorer de notre vision du calcul intensif. 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 [BANOS, 2013] pourra éventuellement sauter les deux premières sections à part s'il est intéressé par des illustrations pratiques originales, notre positionnement étant très similaire et ne divergeant que sur des subtilités mineures pour les sujets évoqués dans ces sections.

* * *

*

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 [RAIMBAULT, 2016a], et pour sa deuxième partie des idées présentées dans [COTTINEAU et al., 2017]. La troisième section reprend dans sa première partie les bases épistémologiques de [RAIMBAULT, 2017h] approfondies par [RAIMBAULT, 2017b], est inédite pour sa deuxième partie et rend compte de [RAIMBAULT, 2017d] 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.

3.1.1 *Pourquoi modéliser ?*

Nous développons dans un premier temps le rôle de la modélisation dans notre démarche de production scientifique. Chaque discipline aura sa ou ses propres définitions et utilisation des modèles : un modèle en physique sera partie d'une théorie 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.

Le modèle comme outil de connaissance indirecte

L'utilisation de modélisation générative peut être mise en correspondance forte avec la notion d'émergence faible introduite par [BEDAU, 2002]. 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 incompatible avec l'existence de *downward causation* et son autonomie. Certains systèmes comme la conscience 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. Connaitre les dynamiques de systèmes faiblement émergents nécessite par définition de les simu-

ler, et donc de les modéliser³, cette approche est ainsi naturelle pour connaître la structure ou les processus dans un système complexe.

Tous nos modèles sont à visée de comprehension, même quand calibration fine ! développer ça en lien avec fonction des modèles !

Epstein def generative social science

varenne : lien entre modélisation et simulations [PHAN et VARENNE, 2010]

differentes fonctions d'un modèle (Varenne aussi)

definition du modèle selon le perspectivisme.

partie sur la parcimonie

lien avec les laboratoires virtuels, test d'hypothèse. cf simplicial, possibilité de trouver des conditions nécessaires et suffisantes.

pattern oriented modeling [GRIMM et al., 2005]

l'équifinalité : le rôle des analyses de sensibilité pour être sûr que les patterns reproduits sont bien dus aux hypothèses (structure causale micro) on détaille alors cuisine à la suite :

Comment explorer un modèle de simulation

Lien entre modélisation et Science Ouverte

Enfin, il est important de souligner brièvement les liens entre pratiques de modélisation et science ouverte, comme le lien entre reproductibilité et Science Ouverte souligné 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 [OPENSHAW, 1983], 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 : [COTTINEAU, REY et REUILLOU, 2016] 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 plate-forme d'exploration de modèle OpenMole [REUILLOU, LECLAIRE et REY-COYREHOURCQ, 2013] 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

³ sur la différence entre simulation de modèle et modèle de simulation, Varenne explique dans quelle mesure ces deux notions sont relativement équivalentes : Ces remarques se vérifient particulièrement dans le cas de la modélisation générative. Nous utiliserons les deux de manière interchangeable par la suite.

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 tant que le modèle peut prendre la configuration spatiale **C (AB) :?** comme entrée ou dans un fichier d'entrée. D'autre part, une idée des workflow est de favoriser des constructions ouvertes et collaboratives, puisque le "marketplace" d'OpenMole, directement intégré au logiciel, 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 couplage de modèle hétérogènes et multi-échelles, ne peuvent guère être viables sans des pratiques d'ouverture, de partage et de construction collaborative des modèles, comme le rappelle [BANOS, 2013].

C : Sur la pédagogie : [CHEN et LEVINSON, 2006] : la simulation comme outil pour apprendre aux élèves ingénieurs. Intéressant à utiliser pour l'aspect performatif, feedback des modèles sur les situations réelles / illustration des différents objectifs de chaque domaine : pourquoi et comment c'est intéressant de prendre en compte certains aspects selon les objectifs / perspectivisme appliqué : faire ce projet , l'évoquer ici.

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* ([ARTHUR, 2015]) est central pour une science consciente de la complexité et est sans doute la base des pratiques de modélisation futures en géographie comme [BANOS, 2013] 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, [GLEYZE, 2005] 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 : [LAGESSE, 2015] 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 [SANDERS et al., 1997] étaient "calibrés" à la main, tandis que [COTTINEAU et al., 2015a] calibre le modèle Marius en multi-modélisation et [SCHMITT et al., 2014] 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 [HILLIER et HANSON, 1989] jusqu'à ses récentes applications à grande échelle [HILLIER, 2016].

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 [O'BRIEN, CHESHIRE et BATTY, 2014], des réseaux sociaux [FRANK et al., 2014], ou des données plus exotiques comme des données de téléphonie mobile [DE NADAI et al., 2016]. 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 méta-analyses. 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, [COTTINEAU, 2016] décrit une application web ayant pour but de présenter une méta-analyse 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 méta-analyse. D'autres applications permettent l'ex-

ploration interactive de la littérature scientifique pour une meilleure connaissance d'un horizon scientifique complexe, comme [CHASSET et al., 2016] 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 sont flous et on peut facilement s'y perdre. Des illustrations parleront d'elles-même. [BARTHELEMY et al., 2013] 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 [AXTELL, 2016], 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 web⁴ qui utilise des ressources de calcul financées par l'argent public 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 [GABAIX, 1999]. 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 principe⁵ : 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 collection de dérives. Récemment, sur la liste de diffusion de géographie franco-phone *Geotamtam*, un soudain engouement autour des données issues

⁴ voir <http://shiny.parisgeo.cnrs.fr/gibratsim/>

⁵ 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.

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 historiques 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 : [LOUAIL et al., 2016] 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 11 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 massives, de par les algorithmes spécialement conçus pour être capable de faire les traitements, est une condition nécessaire au résultat obtenu, mais à la fois l'échelle est les objets (c'est à dire les indica-

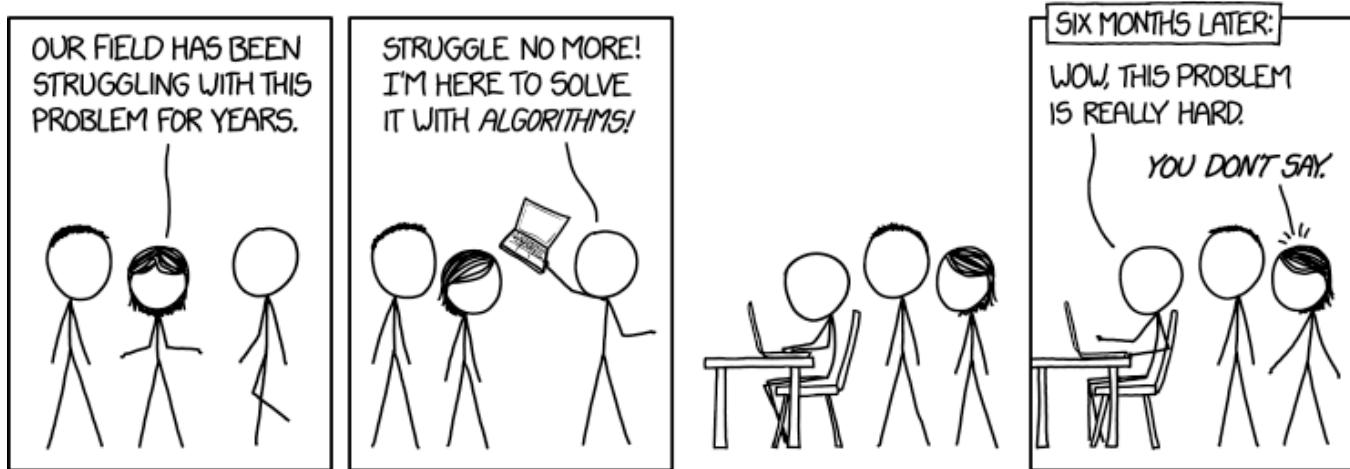


FIGURE 11 : De l'usage naïf de la fouille de données et du calcul intensif. Source : xkcd

teurs 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 [BATTY, 2016], 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.

FORMALISATION Commençons par donner une formulation abstraite de l'idée, d'un point de vue du couplage de modèle. Le générateur est considéré comme un modèle amont, couplé simplement (les sorties devenant les entrées) avec le modèle aval étudié. Si M_u est le modèle amont, M_d le modèle aval et α les meta-paramètres, on a la composition de la dérivée le long des meta-paramètres

$$\partial_\alpha [M_u \circ M_d] = (\partial_\alpha M_u \circ M_d) \cdot \partial_\alpha M_d$$

Cela implique que la sensibilité du modèle aval aux meta-paramètres peut être déterminée en étudiant le couplage séquentiel et le modèle amont. Nous gagnons de la connaissance thématique, dans la sensibilité à un meta-paramètre implicite, mais il y a aussi un gain computationnel : la génération de différentielles contrôlées dans l'espace initial (c'est à dire tester directement la comparaison entre deux grilles proches) serait compliquer à atteindre directement. La question de la stochasticité dans de tels modèles couplés simplement ne pose pas

de problème supplémentaire puisque $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$. Cela multiplie naturellement le nombre de répétitions pour converger bien évidemment. Nous resterons dans l'application pratique ici à une étude de l'espace faisable de sortie et non à une étude différentielle, cette considération théorique n'influe pas à cet ordre, mais doit être gardée à l'esprit pour d'éventuelles applications plus fines.

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 [PUMAIN, 2012b] (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 [GELL-MANN, 1995], 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 [WILSON, 1981] ou l'échelle intra-urbaine [ALLEN et SANGLIER, 1979]. 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 [ZEMBRI, 2010]. 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'accès 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 [GUENÉE, 1968]. 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) [OPENSHAW, 1984]. Plus récemment, [KWAN, 2012] 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 [THOMAS et al., 2017], ou des processus de diffusion étudiés par modèles basé-agents [LE TEXIER et CARUSO, 2017].

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 introduisons 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 . The 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 vision par ordinateur pour comparer des distributions de probabilité [RUBNER, TOMASI et GUIBAS, 2000], 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 [VISSER et DE NIJS, 2006]. 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 performance similaire à la comparaison visuelle directe lorsqu'on prend la moyenne sur un ensemble de contributions crowdsourcées [Koch et STISEN, 2017]. Pour rester simple et car de telles considérations méthodologiques sont auxiliaires 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 [EPSTEIN et AXTELL, 1996]. 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 [LI et WILENSKY, 2009].

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. 12 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 moyen 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. 12 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 [LE NÉCHET, 2015] ainsi que la

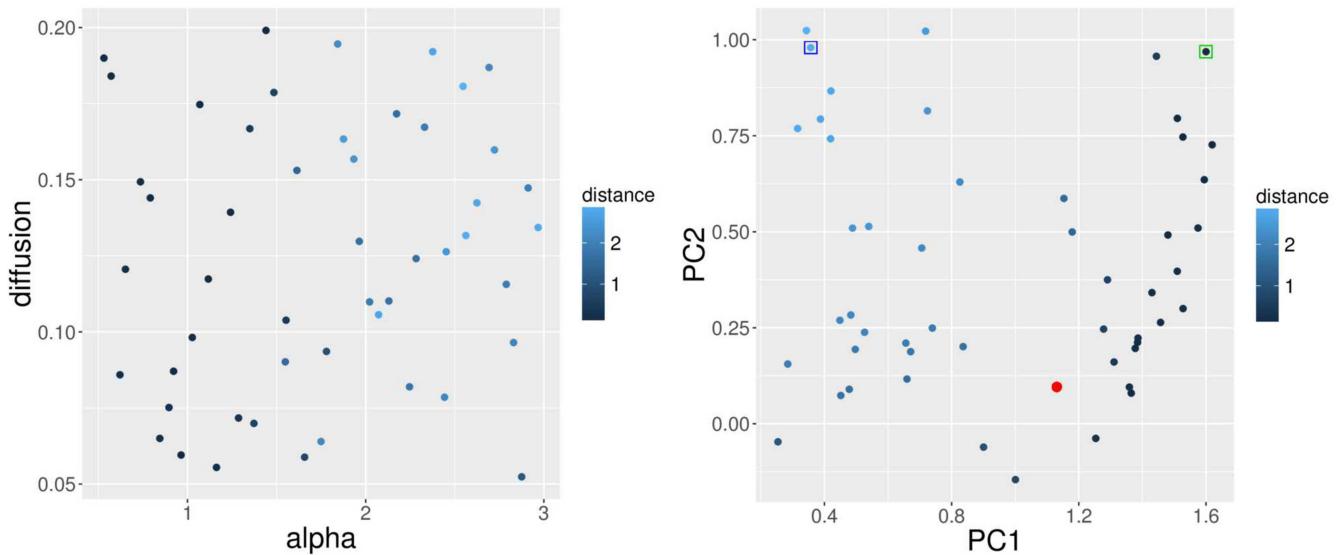


FIGURE 12 : Distance relative des diagrammes de phase à la référence pour l’ensemble des grilles. (Gauche) Distance relative comme fonction des meta-paramètres α (force de l’attachement préférentiel) et la diffusion (β , force du processus de diffusion). (Droite) Distance relative comme fonction des 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. 13.

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.⁶ 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. 13 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. 12. Les comportements sont relativement stables pour s_+ variant, ce qui signifie que les agents les plus pauvres ont un rôle déterminant dans les trajectoires. Les deux exemples ont non seulement une inégalité de base très distance (le plafond du premier 0.35 est environ le plancher

⁶ nous avons $PC1 = 0.76 \cdot \text{distance} + 0.60 \cdot \text{entropy} + 0.03 \cdot \text{moran} + 0.24 \cdot \text{slope}$ et $PC2 = -0.26 \cdot \text{distance} + 0.18 \cdot \text{entropy} + 0.91 \cdot \text{moran} + 0.26 \cdot \text{slope}$.

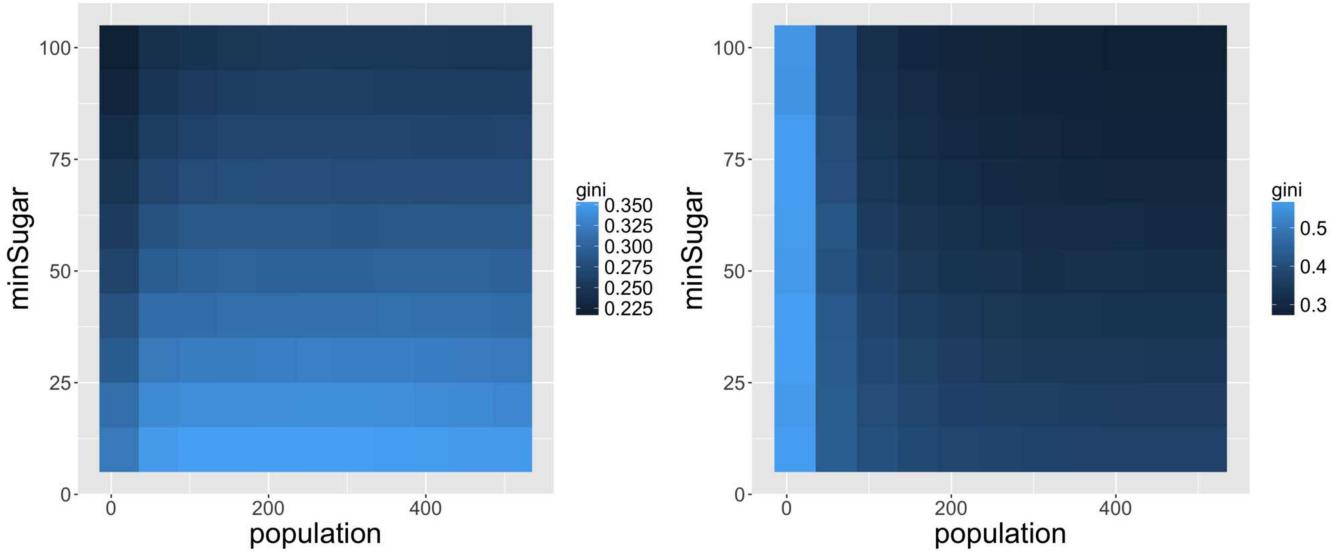


FIGURE 13 : Exemples de diagrammes de phase. Nous montrons deux diagrammes bi-dimensionnels sur (P, s_-) , obtenus à $s_+ = 110$ fixé. (Gauche) Cadre vert, obtenu avec $\alpha = 0.79$, $n = 2$, $\beta = 0.14$, $N = 157$; (Droite) Cadre bleu, obtenu avec $\alpha = 2.56$, $n = 3$, $\beta = 0.13$, $N = 128$.

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 force **C (FL) :?** de la Science vient de 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 [BAIS, 2010].

La reproductibilité semble être de plus en plus pratiquée de manière effective [STODDEN, 2010] 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 [RAM, 2013], ou pour intégrer la création de document et l'analyse de données comme knitr [XIE, 2013]), 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.

C : [CRICK, HALL et ISHTIAQ, 2015]

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 [HATTON et WARR, 2016]. Pour résoudre ce problème, [HUNG et al., 2016] 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 bibliothèques et des modèles en binaire (cf. R. REUILLOU dans [RAIMBAULT, 2017e]). Dans tous les cas, la 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, doivent être autant que possible indépendants de l'implémentation (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 [DRUMMOND, 2009].

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 fourni avec la plate-forme de modélisation agent NetLogo [WILENSKY, 1999]. Le modèle est également disponible en ligne [DE LEON, FELSEN et WILENSKY, 2007] 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 [LECHNER et al., 2006], 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 projet⁷.

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

⁷ 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 [BANOS, 2013], 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 fournir 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 "sauter" **C (FL) : sens ?** 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 subtils 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 C (AB) : ces variables sont-elles définies quelque part ? Si oui, le préciser. Sinon, revoir ton argumentation...** 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



FIGURE 14 : 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.

transport C (AB) : idem dans le modèle (même si aucune erreur ne survient 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 [GIERE, 2010c], 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'exécution et l'exploration du modèle, comme le montre la figure 14.

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 [BARTHÉLEMY et FLAMMINI, 2008] 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 [ZILSEL, 2015], 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 [WILENSKY, 1999] ou Gamma [Cit. gamma]. C'était d'ailleurs un objectif couplé qu'avait initialement [REY-COYREHOURCQ, 2015], c'est à dire une intégration complète de l'exploration fine des modèles et de la production des graphes de sortie ainsi que leur exploration interactive. Comme le rappelle R. Reuillon (Entretien du 11/04/2017, voir ??), la plateforme OpenMole qui devait accueillir cette couche supplémentaire était à ses débuts à l'époque et ne l'est toujours pas aujourd'hui, puisque l'état de l'art de telles pratiques est en pleine construction et bouleversements réguliers [HOLZINGER, DEHMER et JURISICA, 2014]. Des difficultés au regard de la reproductibilité, qui nous concernent particulièrement ici, sont récurrentes et loin d'être résolues. En effet, il faut bien situer la position de ces outils et méthodes comme une aide cognitive préliminaire⁸, 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'extraction de caractéristiques est ainsi

⁸ que nous ne jugeons pas superficielle puisque nous les mobilisons au moins par deux fois par la suite, voir 8.1 et 8.2

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. De manière très visionnaire, BANOS avait déjà mis en garde contre "les dangers de la jungle" des données dans [BANOS, 2001], quand il souligne très justement que l'exploration interactive doit nécessairement se doubler d'indicateurs locaux adaptés, mais surtout d'outils d'exploration automatisés et de critère d'évaluation des choix faits et des motifs découverts par l'utilisateur. On revient encore à l'idée d'une plate-forme 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 [RADDICK et al., 2010]. Les résultats d'un crowdsourcing de la cognition humaine peuvent rivaliser avec les techniques automatiques les plus avancées comme le montre [KOCHE et STISEN, 2017] pour l'exemple de la comparaison de cartes spatiales. Ces possibilités ne doivent cependant pas être sur-estimées ou utilisées à mauvais escient, et les questions d'intégration efficiente homme-machine sont d'ailleurs totalement ouvertes. Dans le domaine de la visualisation de l'information géographique, [PFAENDER, 2009] introduit une sémiologie spécifique visant à favoriser l'exploration de grands jeux de données hétérogènes, et l'expérimente sur une application 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

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 [CAMERER et al., 2016]) 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 investissement au détriment de sa propre recherche [CHAVALARIAS et al., 2005]. 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 : [TEPLITSKIY, LU et DUEDE, 2015] 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 [PEREZ-RIVEROL et al., 2016]. Plus généralement, les sciences computationnelles nécessitent l'adoption de certains standards et pratiques pour assurer une bonne reproductibilité, et ceux-ci restent majoritairement à développer : [WILSON et al., 2017] donne des premières pistes. Concernant la qualité des données, de nombreux efforts sont faits pour introduire des cadres de standardisation des données : par exemple [VEIGA et al., 2017] décrit un cadre conceptuel visant à guider la résolution de problème 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).

C : citer Romain sur le blockchain, en lien avec ce papier ? [FURLANELLO et al., 2017]

L'accès aux données est également un point crucial pour la reproductibilité, et sans nous y attarder car cela impliquerait des développements sur la définition, la philosophie, le droit des données etc. qui sont des sujets de recherche en eux-même, nous donnons des perspectives sur les potentiels d'une ouverture systématique des données en recherche. En géographie, les *data paper* sont une pratique inexistante, et la règle est plutôt de garder la main jalousement sur un jeu produit, capitalisant sur le fait d'être le seul à y avoir accès. 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 richesse. La fermeture induira plutôt des effets négatifs, comme par exemple du temps perdu à recoder un 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 ? Une expérience de revue induit à réellement s'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. Mais en suivant l'adage de Framasoft, "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* 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 considérable pour la gagner ; les données immobilières proviennent d'une base propriétaire achetée avec de l'argent public, et nous pouvons profiter d'un flou du contrat pour les rendre disponibles de manière agrégées avec les résultats ; les données des stations essence 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 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 CybergeoNetworks⁹, 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

⁹ <http://shiny.parisgeo.cnrs.fr/CybergeoNetworks>

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.

★ ★

★

3.3 POSITIONNEMENT ÉPISTÉMOLOGIQUE

3.3.1 Approche cognitive et Perspectivisme

Notre positionnement épistémologique se fonde sur une approche cognitive de la science, introduite par GIERE dans [GIERE, 2010b]. L'approche se concentre sur le rôle des agents cognitifs comme porteurs et producteurs de la connaissance. Son opérationnalité a été montrée par [GIERE, 2010a] qui étudie un modèle basé-agent de la science. Ces idées convergent avec le jeu Nobel de CHAVALARIAS [CHAVALARIAS, 2016] qui teste de manière stylisée l'équilibre entre exploration et falsification **C (AB) : est-ce vraiment le bon couple? A vérifier** dans l'entreprise scientifique collective. Ce positionnement épistémologique a été présenté par GIERE comme *perspectivisme scientifique* [GIERE, 2010c], 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 concrétiser, nous pouvons le positionner sur la *check-list* du constructivisme de HACKING [HACKING, 1999], un outil pratique pour positionner une position épistémologique dans un espace simplifié à trois dimensions dans lequel les dimensions sont différents aspects sur lesquels les approches réalistes et constructivistes généralement divergent : d'abord la contingence (dépendance au chemin du processus de construction de connaissances) est nécessaire l'approche perspectiviste qui est pluraliste, deuxièrement le "degré de constructivisme" est assez haut car les agents produisent la connaissance, et enfin la stabilité des théories dépend des interactions complexes entre les agents et leur perspectives **C (AB) : reprendre plus clairement**. Cela a pour ces raisons été présenté comme un chemin intermédiaire et alternatif entre le réalisme absolu et le constructivisme sceptique [BROWN, 2009]. La notion de *perspective* jouera 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 FEYERABEND [FEYERABEND, 1993]. 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. L'auto-organisation et l'émergence des connaissances nécessitent un certain anarchisme pour échapper aux préconceptions cadrant par le haut **C (AB) : a revoir**. 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 [DUDA, 2013]. 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 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*

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 [BETTENCOURT et al., 2007]. 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 ? BANOS file dans [BANOS, 2013] 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 [MONOD, 1970], 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 les différencier d'autres systèmes : (i) la téléonomie, 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 produisent¹⁰; (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. Un aspect divergent est cette notion de multi-objectif qui est typique des systèmes complexes socio-techniques. 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 résumons 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 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. [GABORA et STEEL, 2017] propose un modèle de réseau autocatalytique 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 au molécules de s'auto-entretenir et s'auto-reproduire, les représentations mentales faisant office de molécules. Cet exemple montre bien que le parallèle n'est pas toujours absurde. Mais si les processus à l'origine sont analogues, la nature de l'évolution est bien différente par la suite, comme le montre [LEEUW, LANE et READ, 2009], les critères darwiniens d'évolution n'étant pas suffisant pour expliquer l'évolution de nos sociétés organisées. Il s'agit d'un degré de complexité supérieur **C (AB) : vraiment? Eléments factuels à avancer** et le rôle des flux d'information est crucial (voir le rôle de la complexité informationnelle dans la sous-section suivante). Enfin, l'un des points sur lequel il s'agit d'être attentif, est la plus grande difficulté de définir les niveaux d'émergence pour les systèmes sociaux : [ROTH, 2009] 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

¹⁰ à ne pas confondre avec la téléologie, propres aux animismes, qui consiste à prêter un projet ou un sens à l'univers

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.

C : [MESOUDI, 2017]

3.3.3 *Nature de la Complexité et Production de Connaissances*

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'élèver 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.7, 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 réflexive¹¹. 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 d'une part 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 [DEFFUANT et al., 2015]. 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, la nature des systèmes ne jouant pas de rôle dans notre réflexion. **C (AB) : je vois l'idée mais c'est loin d'être clair ==> à clarifier** D'autre part, nous distinguons deux autres "types" de complexité, la

¹¹ Nous remercions D. Pumain d'avoir pointé cette vue alternative du problème que nous allons développer par la suite

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 degrés 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 [FEBRES, JAFFÉ et GERSHENSON, 2013]. 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. D'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 [CHU, 2008]¹². Nous proposons de nous concentrer sur ces trois concepts en particulier, pour lesquels les relations ne sont déjà pas évidentes. 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.

C (JR) : Gell-mann : effective complexity : AIC of CAS observing an other CAS : prelude to perspectivism? check link between information and Kolmogorov complexity - add Kolmogorov complexity in computational complexity.

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.

Le "paradoxe" du chat de Schrödinger n'en est pas une sous une perspective réductionniste, c'est à dire si l'on 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. **C (AB) :** trop elliptique là aussi : on ne voit pas clairement le lien avec el gato. Sois plus précis Cette vision intuitive a récemment été démontrée rigoureusement par [BOLOTIN,

¹² Dans une approche en un sens réflexive, [CHU, 2008] 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).

[\[2014\]](#) qui prouve que l'acceptation de $P \neq NP$ implique une séparation qualitative entre le niveau quantique microscopique et le niveau d'observation macroscopique. En d'autres termes, la complexité computationnelle est suffisante pour la présence d'émergence. A priori, cette séparation effective des échelles n'implique pas que le niveau inférieur ne joue pas un rôle crucial, puisque [VATTAY, SALAHUB et CSABAI, [2015](#)] 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 : [VERLINDE, [2016](#)] 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".

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 [MOORE et MERTENS, [2011](#)]. Des automates cellulaires, qui sont par ailleurs cruciaux pour la compréhension de divers systèmes complexes, ont été montrés Turing-complets¹³ **C (AB) : definition A1 : a finir** (comme le Jeu de la Vie). Des organismes sans système nerveux central sont capables de résoudre des problèmes décisionnels difficiles [REID et al., [2016](#)]. Un algorithme à base de fourmis est montré par [PINTEA, POP et CHIRI, [2017](#)] comme résolvant un Problème du Voyageur de Commerce Généralisé (GTSP), problème NP-difficile. Ce lien fondamental avait 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 [TURING, [1952](#)] (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).

C : [Tošić et ORDONEZ, [2017](#)] lower bound for computational complexity of simple ABM when adding interactions with the environment.

C : [ELLIOTT et GU, [2017](#)] quantum computation reduces drastically memory needed

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

¹³ un système est Turing-complet s'il est capable de calculer les mêmes fonctions qu'une machine de Turing

ce qui a permis de résoudre le paradoxe apparent du 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'information¹⁴. 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 [MAINZER et CHUA, 2013]. 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 : [CROSATO et al., 2017] 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'interpréter comme un lien entre information et sens, puisque la sémantique d'un objet cogni-

¹⁴ Le démon de Maxwell est plus qu'une construction intellectuelle : [COTTET et al., 2017] implémente un démon expérimentalement au niveau quantique.

tif ne peut se passer d'une fonction. HOFSTADER rappelle dans [HOFSTADER, 1980] 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. Comme ces systèmes sont généralement multi-niveaux, ou présentent au moins un certain niveau de complexité computationnelle, leur connaissance se doit de la capturer, puisque même des modèles *simples* devront capturer leur complexité de manière conceptuelle et impliquer une structure conceptuelle sous-jacente complexe, même si celle-ci n'est pas explicitement explorée. **C**
(AB) : ton raisonnement pourrait être clarifié un peu Ainsi, toute *connaissance du complexe* embrasse non seulement toutes les complexités mais aussi leur relations, dans son contenu et dans sa nature : elle doit nécessairement avoir un certain degré de réflexivité pour alors être cohérente. On peut tenter d'étendre à la réflexivité en tant que réflexion sur le positionnement disciplinaire : suivant PUMAIN dans [PUMAIN, 2005], 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 complexité, qui serait une dimension supplémentaire liée à la connaissance des systèmes complexes, la réflexivité doit être au cœur de la démarche. [READ, LANE et LEEUW, 2009] 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 [GERSHENSON, 2015] comme extension de la *requisite variety* d'ASHBY¹⁵, la *connaissance du complexe* devra nécessairement être *connaissance complexe*. Cet autre point de vue renforce la nécessité de la réflexivité, puisque suivant MORIN (voir par exemple voir le Tome 4 de la méthode centré sur la production de connaissance [MORIN, 1991]), la *Connaissance de la Connaissance* est centrale dans l'établissement d'une pensée complexe.

¹⁵ 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 [ALLEN, STACEY et BAR-YAM, 2017] qui introduit la *requisite variety* multi-échelle, démontrant la compatibilité avec une théorie de la complexité basée sur la théorie de l'information.

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 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 [LIVET et al., 2010a]) 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 ([MORIN, 1991]) est ni inductive ni deductive, 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 extrapolier 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 [PAULUS, 2004]), mais la relative fermeture de ce type de données ne les rend pas utilisables dans notre démarche. Ensuite, notre positionnement sur le rôle du calcul intensif et les besoins d'exploration des modèles 3.1 est source de l'ensemble des expériences numériques et des méthodologies utilisées ou développées. Enfin, notre positionnement épistémologique 3.3 percole dans l'ensemble de notre travail, et permet de poser les premières briques pour des formalisations théoriques plus systématiques qui seront développées en Chapitre 9.

Deuxième partie

BRIQUES ÉLÉMENTAIRES

This part provides building blocks for the final objective of constructing models of co-evolution. These contain both stylized facts from empirical analyses and toy and hybrid modeling. They correspond to three distinct components of our overall construction : first analyses at the micro-scale confirming the chaotic and non-stationary nature of interactions between networks and territories, secondly a morphogenetic vision of these that corresponds roughly to a meso-scale, and finally an application of the evolutive urban theory at the macro-scale.

INTRODUCTION DE LA PARTIE II

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 par la suite. Sauf indication, les spécifications données ici feront référence lors de l'utilisation des termes correspondants.

Processus stochastiques

STATIONNARITÉ Nous utiliserons une notion faible de la stationnarité des processus stochastiques, puisqu'on la mobilisera pour des analyses empiriques sur lesquelles la taille des données ne permettra pas des tests sur les distributions complètes. Soit X_i un processus stochastique

ERGODICITÉ La notion d'ergodicité sera utilisée dans le cadre de l'universalité des processus spatio-temporels. Soit $X(t, \vec{x})$ un processus stochastique spatio-temporel

Systèmes dynamiques

CHAOS Un système dynamique $\dot{X} = \Phi(X, t)$ sera dit chaotique si ses exposants de Liapounov sont strictement positifs dans une région non-négligeable de l'espace des conditions initiales.

Statistiques

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

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

CAUSALITÉ DE GRANGER Une série temporelle multi-dimensionnelle $X(t)$ présente une causalité de Granger si avec $X(t) = A \cdot (X(t-\tau))_{\tau>0} + \varepsilon$, il existe τ, i tels que $a_{i\tau} > 0$ significativement. On peut montrer que dans ce cas $\rho_\tau = a_{i\tau}^2$. Nous utiliserons cette version faible de la causalité de Granger, c'est à dire un test sur les corrélations retardées.

REGRESSION 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

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.

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.

Exploration de modèles

DISCRÉPANCE

ECHANTILLONNAGE LHS

NSGA

4

CO-EVOLUTION : UNE ENTRÉE PAR LA THÉORIE EVOLUTIVE URBAINE

On a illustré en Chapitre 1 les particularités de chaque cas à l'échelle microscopique pour les interactions entre territoires et réseaux de transport. De quelle manière ces processus font-ils émerger des caractéristiques de ces relations à d'autres échelles comme l'échelle mesoscopique ou macroscopique ? Dans quelle mesure est-il possible d'extraire certaines régularités, qu'on qualifierait alors de structurales, et qui permettraient une certaine connaissance générale des processus impliqués dans les systèmes urbains. Il s'agit de la question fondamentale de *l'universalité des processus* dans les systèmes urbains. Une autre caractéristique fondamentale des interactions est la question de l'existence d'équilibres : on peut également se demander s'il existe des échelles spatiales et temporelles sur lesquelles des équilibres peuvent exister, ce qui revient à se poser la question des *échelles de stationnarité des processus*. Ces deux questions ouvertes sont au centre des préoccupations de la *Théorie Evolutive Urbaine*, qui vise à identifier les faits stylisés réguliers dans les systèmes de villes tout en mettant l'emphase sur la particularité de leur éléments et les bifurcations associées (voir 9.3). Il est alors légitime d'explorer cette première entrée et les solutions qu'elle propose, par des analyses empiriques et de modélisation, ce qui est l'objet de ce chapitre. Nous étudions d'abord à l'échelle mesoscopique les propriétés de non-stationnarité spatiale entre des manifestations simples des caractéristiques des territoires et des réseaux, capturées dans des indicateurs morphologiques pour chacun, par l'étude des correlations spatiales entre ces indicateurs. Nous introduisons ensuite la dimension temporelle en étudiant la notion de causalité spatio-temporelle dans la section 4.2, qui 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 régimes de causalité mis en évidence pour un modèle simple de morphogenèse couplant fortement croissance du réseau et densité témoignent de causalités circulaires qui sont bien des marques d'une co-évolution. Dans le cas d'une non-stationnarité, qui a été mise en valeur par 4.1, ces régimes peuvent alors évoluer dans le temps et l'espace, impliquant alors une co-évolution sur le temps long. L'application au cas du réseau ferroviaire en Afrique du Sud montre que cette multiplicité des régimes existe bien pour des données réelles. Nous explorons en-

fin dans une dernière section 4.3 les possibilité offerte par les modèles d’interaction issus de la théorie évolutive, à une grande échelle spatiale et temporelle, ce qui permet de démontrer l’existence d’effets de réseau de manière indirecte, sans même introduire d’aspects de coévolution dans un premier temps. Ainsi, nous posons les premières fondations, sur différents aspects des relations 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 [] pour l’analyse morphologique, puis les résultats présentés par [RAIMBAULT, 2016a] pour l’analyse des correlations ; la deuxième section correspond à la majorité de [] pour la formulation théorique et l’illustration sur données synthétiques, puis présente les résultats de [] pour l’application. Enfin la dernière section est une traduction de [].

4.1 CORRÉLATIONS STATIQUES ENTRE FORME URBAINE ET FORME DE RÉSEAU

Une première entrée en matière empirique, et qui se voudra simple sur les objets étudiés, est de s'intéresser à des caractéristiques directement mesurables des territoires et réseaux. De manière phénoménologique, les agrégats urbains se qualifient au dessus d'une certaine échelle par une forme urbaine, de même que les réseaux de transport présentent des propriétés topologiques synthétiques. **C (FL) : le fait de quantifier ne signifie pas que ces propriétés existent A1 : (JR)** pas d'accord, a partir du moment où on a un modèle on a nécessairement une ontologie sous-jacente - je sais pas ce que tu veux dire exactement par existe ? On peut alors s'interroger sur des liens directement mesurables entre ceux-ci, c'est à dire quelle information contiennent les corrélations statiques entre forme urbaine et topologie du réseau routier, au sens de corrélations estimées sur un échantillon local dans l'espace sur des données fixes. Dans une perspective de la Théorie Evolutive, on devine bien les implications de cette démarche : les liens entre corrélations dynamiques et statiques sont liées aux propriétés d'ergodicité du système, et la variation des estimateurs dans l'espace et selon les échelles informera sur le degré de stationnarité des interactions. Il s'agit d'une manière indirecte de lier statique et dynamique. **C (JR) : les termes sont encore mal définis là, ils le seront dans le préliminaire de la partie II qui introduira toutes les définitions mathématiques.**

Les processus spatio-temporels impliquant une diffusion ou une propagation (ce qui est a priori le cas de la forme urbaine comme suggéré par 5.2) peuvent généralement être compris partiellement par leur structure de correlation dans le temps et l'espace. Dans certains cas, on peut espérer que l'étude d'une correlation statique entre différentes instances d'un système peuvent sous certaines conditions informer sur les correlations dynamiques sous-jacentes, ce que nous ferons de manière empirique ici.

A l'échelle macroscopique du système de ville, le caractère spatial du système urbain est capturé de manière raisonnable par les positions des villes, associées aux variables agrégées au niveau de la ville qui représentent entièrement le système, comme la plupart des modèles liés à la Théorie Evolutive postulent. A l'échelle mesoscopique, à laquelle nous nous attendons à capturer des manifestations morphologiques des interactions entre ville et transport, la structure du système territorial peut être spécifiée par des indicateurs plus raffinés pour l'aspect morphologique. Le choix des indicateurs de forme urbaine pertinents pour répondre à un type de question donnée n'est pas évident, et dépendra de l'échelle et du contexte : on

peut par exemple s'intéresser au caractère polycentrique¹ pour lequel les indicateurs seront différents si on s'intéresse à des phénomènes de concentration. Notre but est de capturer le maximum de dimensions de variation de la forme urbaine, nous calculerons pour cela un certain nombre d'indicateurs arbitraire satisfaisant une certaine convergence de la variance cumulée des composantes principales.

Nous étudions de manière systématique les indicateurs morphologiques pour des zones de taille constante couvrant une région donnée. Le choix de zones de taille fixe peut être interrogé au regard de la définition d'un système territorial, qui peut par ailleurs être compris comme une entité spatiale consistante à une échelle donnée et selon certains critères : les *Territoires Humains* comme nous avons déjà défini en 1.1 ou plus généralement des espaces fonctionnels autonomes². Le choix de limites "pertinentes" pour le territoire ou la ville est un problème relativement ouvert qui dépendra souvent de la question à laquelle on cherche à répondre : [GUÉROIS et PAULUS, 2002] montre 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 ($\simeq 50\text{km}$) 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.

Le but n'étant pas de comparer les territoires sur lesquels ces indicateurs sont calculés entre eux, mais de calculer une valeur "locale" et d'établir un champ discret régulier dans l'espace, la taille fixe de la fenêtre est nécessaire. **C (FL) : pourquoi ce choix ? que font les autres auteurs ? A1 : (JR) rien a priori, car jamais fait de cette façon** Cette taille est arbitraire, mais l'analyse a été menée pour des tailles voisines également (voir A.4). Les "territoires" qu'une approche plus classique voudra comparer, comme des aires urbaines fonctionnelles par exemple, pourront émerger de manière endogène si ceux-ci font sens pour les variations des indicateurs, comme des unités régionales cohérentes au regard des valeurs prises.

¹ la polycentricité peut être vue selon [SERVAIS et al., 2004] comme l'émergence de centre secondaire d'intérêts. Pour une distribution de population, cela correspondra à des noyaux de peuplement distincts pouvant chacun faire office de centre. Ces auteurs montrent les difficultés de quantifier finement cette notion, vu ses multiples caractérisations numériques possibles.

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.

4.1.1 Mesures morphologiques

MORPHOLOGIE URBAINE [GUÉROIS et PUMAIN, 2008] étudie 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 de robustesse et d'invariance. Par exemple, deux villes polycentriques devraient être classifiées comme morphologiquement proches tandis qu'une comparaison directe des distributions (avec une distance de Monge par exemple) pourra donner une distance très élevée entre les configurations selon la position des centres. L'utilisation d'indices issus de l'analyse fractale est une possibilité par exemple appliquée systématiquement par [CHEN, 2016]. Le lien entre morphologie urbaine et topologie du graphe de relations correspondant a été suggéré par[BADARIOTTI, BANOS et MORENO, 2007]. Nous choisissons de nous référer à la littérature en morphologie urbaine qui propose des jeux d'indicateurs variés pour décrire la forme urbaine [TSAI, 2005]. Le nombre de dimensions peut être réduit pour obtenir une description robuste avec un petit nombre d'indicateurs indépendants [SCHWARZ, 2010]. Il faut noter que nous ne considérons ici des indicateurs sur la densité de population seule, et que des considérations plus élaborées sur la forme urbaine incluent 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 dans [LE NÉCHET, 2015] où une typologie morphologique des grandes villes européennes est obtenue.

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 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 $\ln(P_i/P_0) \sim k + \gamma \cdot \ln(\tilde{i}/i_0)$ **C (FL) : k : ? A1 : la constante de la regression, dummy dans l'indicateur** où \tilde{i} sont les indices de la distribution triée de manière décroissante. Elle est toujours négative ou nulle, et des valeurs proches de zéro signifient une distribution plate.
2. Entropie de la distribution, qui exprime l'uniformité de la distribution :

$$\mathcal{E} = \sum_{i=1}^M \frac{P_i}{P} \cdot \ln \frac{P_i}{P} \quad (2)$$

C (FL) : si $P_i = 0$ A1 : (le terme dans la somme est nul, car x log(x) eq 0 en 0) $\mathcal{E} = 0$ 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, avec des poids spatiaux simples donnés par $w_{ij} = 1/d_{ij}$

$$I = M \cdot \frac{\sum_{i \neq j} w_{ij} (P_i - \bar{P}) \cdot (P_j - \bar{P})}{\sum_{i \neq j} w_{ij} \sum_i (P_i - \bar{P})^2}$$

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, qui témoigne de la dispersion spatiale de la population

$$\bar{d} = \frac{1}{d_M} \cdot \sum_{i < j} \frac{P_i P_j}{P^2} \cdot d_{ij}$$

où d_M est une constante de normalisation que nous prenons comme la diagonale de l'étendue considérée dans notre cas.

Les deux premiers index sont pas spatiaux, mais nécessaires pour une bonne qualification des distributions de population, 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, 2014]. Cette base a certains défauts de précision qui ont été reconnus [BRETAGNOLLE et al., 2016] mais nous agrégerons les données à un niveau suffisant pour les éviter. 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, ce qui permet de plus d'être en accord avec l'ontologie du modèle de morphogenèse que l'on développera en 5.2. Comme cela ne fait pas sens d'avoir une résolution trop détaillée à cause de la qualité des données, nous agrégeons les données raster initiales à une résolution de 500m pour avoir des fenêtres de taille $N = 100$. 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 d'une certaine façon résout le problème du biais de la forme de la fenêtre par la "continuité" des valeurs. Nous avons 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 assez similaires. L'implémentation

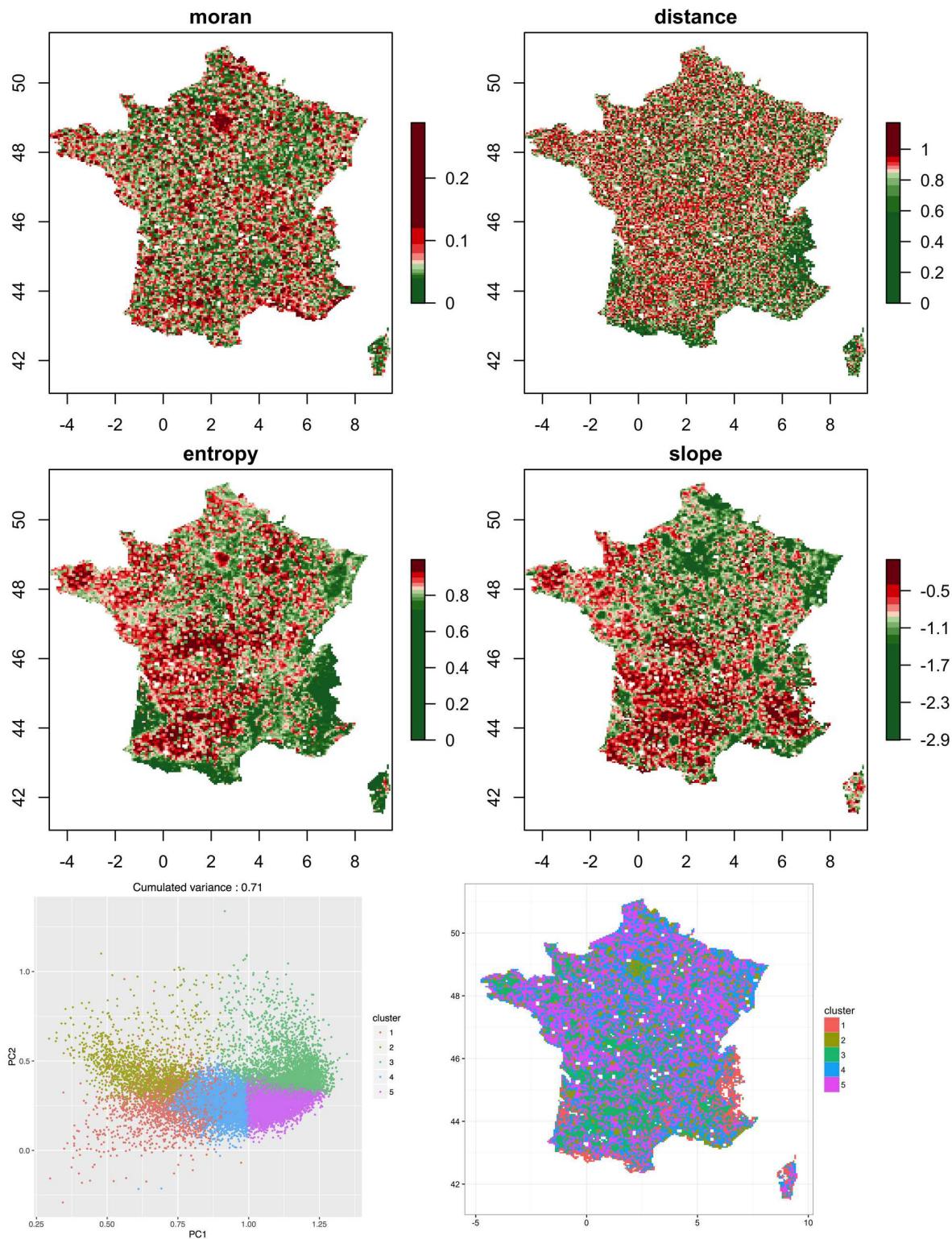


FIGURE 15 : Valeurs empiriques des indicateurs morphologiques. (*Quatre cartes du haut*) Distribution spatiale des indicateurs morphologiques pour la France. La détermination de l'échelle de couleur est faite par quantiles pour faciliter la lecture des cartes. (*Bas gauche*) Projection des valeurs morphologiques 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.

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 $O(\log^2 N \cdot N^2)$. Nous montrons en Fig. 15 des cartes donnant les valeurs des indicateurs, pour la France seulement afin de permettre une lisibilité. Pour avoir une idée des valeurs typiques de chacun des indicateurs, on pourra se référer aux distributions empiriques données en Appendice A.4. 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 métropolitaines, 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 dans les zones montagneuses. Des régions à très forte entropie sont observées dans le centre et le Sud-ouest. Pour avoir une meilleure compréhension des classes morphologiques existantes, nous utilisons une classification non-supervisée³ avec un algorithme des k-means simple. 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 15, 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 [FU, JIANG et HUANG, 0001]. Les cartes sont disponibles en Appendice A.4.

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 la morphologie, pour être en mesure d'explorer les relations entre ces mesures statiques. L'analyse statique de réseau a été intensément do-

³ qui consiste à partitionner l'espace des données selon leur structure endogène

cumentée dans la littérature, voir par exemple [LOUF et BARTHELEMY, 2014a] pour une étude comparative des villes ou [LAGESSE, 2015] pour l'exploration de nouvelles mesures pour le réseau de rues. [MOOSAVI, 2017] 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.

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, 2012]. Sa qualité a été étudiée pour différents pays comme l'Angleterre [HAKLAY, 2010] et la France [GIRRES et TOUYA, 2010]. 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 [ZHENG et ZHENG, 2014] soulève une récente accélération de la complétude 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. [ZHANG et al., 2015] souligne quatre régimes de qualité des données, fournissant 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.

Pour les segments de rue primaires, nous calculons le réseau topologique pour l'ensemble des zones étudiées, à une granularité de 100m pour pouvoir être utilisé de manière cohérente avec les grilles de population et pour être robuste aux imperfections locales de codage ou données très locales manquantes. Les données OSM sont importées dans `pgsql` en utilisant `osmosis` [TEAM, 2016]. Le réseau est ensuite agrégé à la granularité fixe pour créer un graphe topologique, qui est finalement 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 est détaillé en Appendice A.4.

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 résumant 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 ayant les positions spatiales $\vec{x}(v)$ et des populations $p(v)$ obtenus par agrégation de la population dans la partition de Dirichlet correspondante, les liens des *distances effectives* $l(E)$ qui prennent en compte les impédances et les distance réelle (pour inclure la hiérarchie primaire du réseau), nous utilisons :

- Caractéristiques basiques : nombre de noeuds $|V|$, nombre de lien $|E|$, densité d , degré moyen $\bar{\delta}$, nombre cyclotomique μ , connectivité α , longueur moyenne des liens \bar{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 (définie par [BANOS et GENRE-GRANDPIERRE, 2012]).
- Mesures de centralité agrégées (le niveau de hiérarchie étant calculé par un ajustement OLS d'une loi rang taille) :
 - *Betweenness Centrality*, 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*, moyenne \bar{cl} et hiérarchie α_{cl}
 - Temps de trajet moyen vers les autres noeuds, moyenne \bar{t} et hiérarchie α_t
 - Accessibilité, qui correspond au temps de trajet pondéré par les populations, moyenne \bar{a} et hiérarchie α_a
- Modularité

L'accessibilité est bien considérée comme un indicateur de réseau, puisque son calcul implique d'attribuer des poids aux noeuds par un population correspondante, et revient ensuite à un temps de trajet moyen pondéré. 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 ($\rho = 0.86$ pour l'ensemble de la Chine par exemple).

Forme de Réseau et Résilience

L'idée fondamentale motivant le calcul d'indicateurs de réseau est d'obtenir une réduction de dimension drastique, s'il est possible d'associer certains "types" de réseau à des valeurs typiques d'indicateurs. On est très loin d'une connaissance fine de typologies qui associeraient propriétés topologiques, dynamiques et processus de génération du réseau, le tout dans des typologies. De même que de pouvoir relier systématiquement ces propriétés à des caractéristiques dérivées, comme la résilience qui est une propriété aux définitions diverses pour laquelle [GAO, BARZEL et BARABÁSI, 2016] introduit une approche par la sensibilité des processus dynamiques. Afin d'illustrer d'une part la difficulté de caractériser les réseaux et d'autre part les potentialités offertes par notre base de données, nous développons en Appendice A.4 une courte analyse des propriétés de résilience au sens de [ASH et NEWTH, 2007] pour des réseaux typiques.

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 correlations par la suite. Nous montrons en Figure 16 un échantillon pour la France. Le comportement spatial des indicateurs est très instructif, et révèle comme pour la forme urbaine des régimes locaux (urbain, rural, métropolitain), mais aussi des régimes régionaux très marqués. Ceux-ci peuvent être dus au 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 d'un foncier très fragmenté. 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 Betweenness moyenne, en accord avec une forte hiérarchisation. Pour la Chine, pour laquelle une sélection d'indicateurs est également donnée en A.4, on observe des variations locales et régionales encore plus marquées, ainsi que par exemple les mega-régions urbaines qui 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 ty-

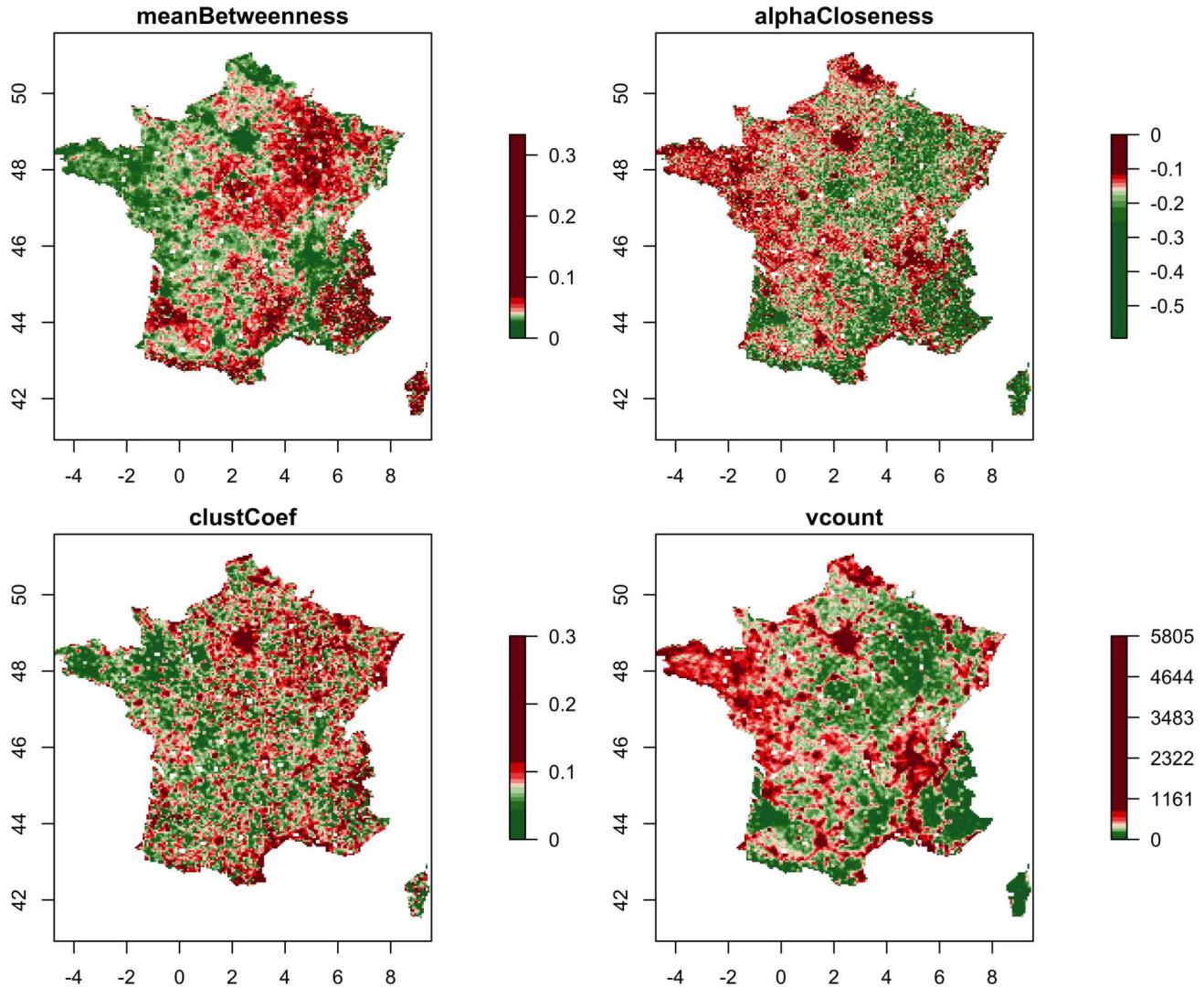


FIGURE 16 : **Distribution spatiale des indicateur de réseau.** Nous donnons les indicateurs pour la France, en correspondance avec les indicateurs morphologiques décrits précédemment.

piquement 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 Figure 17 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 avoisine les 800. 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 inform sur différents régimes urbains. Les métropoles semblent exhiber une corrélation positive dans ce cas, et des espaces ruraux négatifs. 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 Appendice A.4 d'autres exemples, pour l'ensemble de l'Europe et la Chine. Ces propriétés de non-stationnarité sont régulières pour l'ensemble de ces cas d'étude.

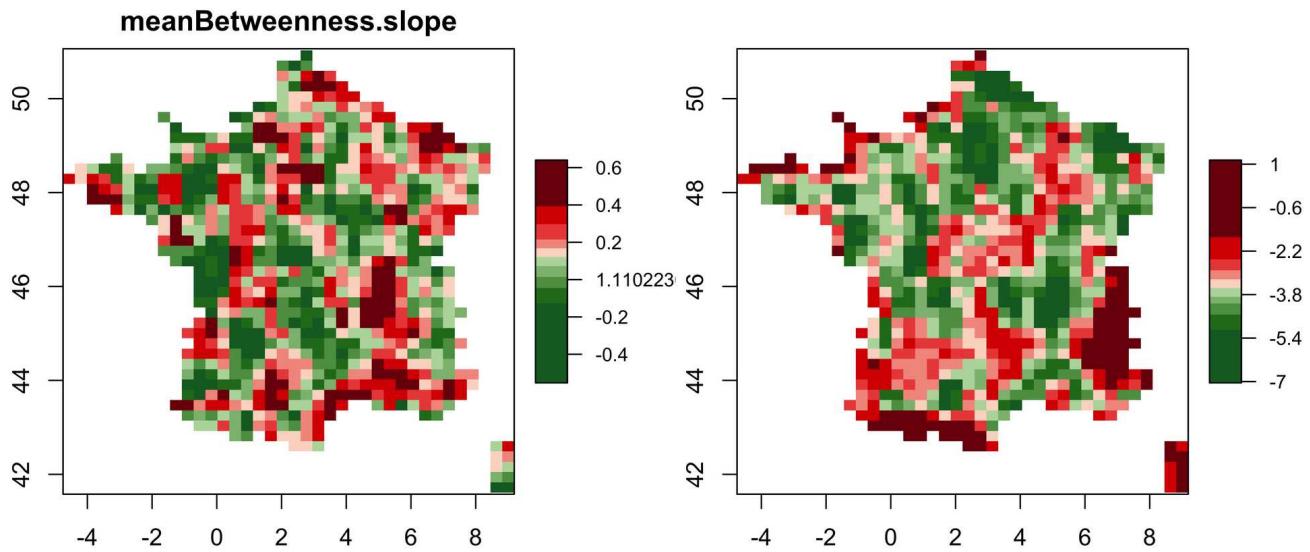


FIGURE 17 : Exemples de corrélations Spatiales. Pour la France, les cartes donnent $\rho [\bar{bw}, \gamma]$ (gauche) et la première composante de la matrice réduite (droite).

NATURE MULTI-SCALAIRE DES PROCESSUS On observe une variation significative des corrélations fonction de δ , qui se reflète dans la valeur moyenne de la matrice. D'abord, la distribution statistique des corrélations suit une loi similaire à une log-normale pour la morphologie seule, et plutôt normale 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. On montre en Figure 18 ces distributions et les résultats des expériences de variation de δ pour l'Europe. On constate sur les nuages de points que les configurations où les corrélations croisées sont les plus fortes correspondent à celles où morphologique et réseau sont également fortes, confirmant l'imbrication des processus dans ce cas. L'augmentation de δ cause pour l'ensemble un décalage dans le positif, mais également un rétrécissement de la distribution, ces deux effets se traduisant par une décroissance des corrélations absolues moyennes, qui se stabilisent approximativement pour les grandes valeurs de δ . Cette variation est révélateur 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. Or la variation de la taille de l'intervalle de confiance normalisée, qui en théorie sous hypothèse de normalité devrait conduire $\delta \cdot |\rho_+ - \rho_-|$ à être constant dans ce cas (puisque les bornes varient comme $\sqrt{N} \sim \sqrt{\delta^2}$), confirme bien cette première hypothèse. Ainsi, les processus sont à la fois non-stationnaires et multi-scalaires.

Non-stationnarité spatiale et non-ergodicité

FORMALISATION Formalisons les conclusions empiriques obtenues. Soit $Y_i[\vec{x}, t]$ un processus stochastique spatio-temporel. Nous avons alors les hypothèses suivantes :

1. L'autocorrelation spatiale locale existe en dessous d'une échelle minimale l_0 (en d'autres termes le processus est continu dans l'espace) : pour tout \vec{x} et t , on a $|\rho_{\|\Delta\vec{x}\| < l_0}[Y_i(\vec{x} + \Delta\vec{x}, t), Y_i(\vec{x}, t)]| > 0$.
2. Les processus sont localement paramétrisés : $Y_i = Y_i[\alpha_i]$, où $\alpha_i(\vec{x})$ varie à l'échelle l_α , avec $l_\alpha \gg l_0$ et est localement stationnaire dans l'espace.
3. Les processus sont multi-scalaires : comme $\rho(\delta = \infty) > \rho(\delta = 0)$, une nécessaire correction non-linéaire sur les moyennes spatiales des processus est présente dans le calcul des corrélations.

SUR LA NON-ERGODICITÉ GLOBALE Nous proposons d'esquisser un lien entre les propriétés de non-stationnarité et la non-ergodicité

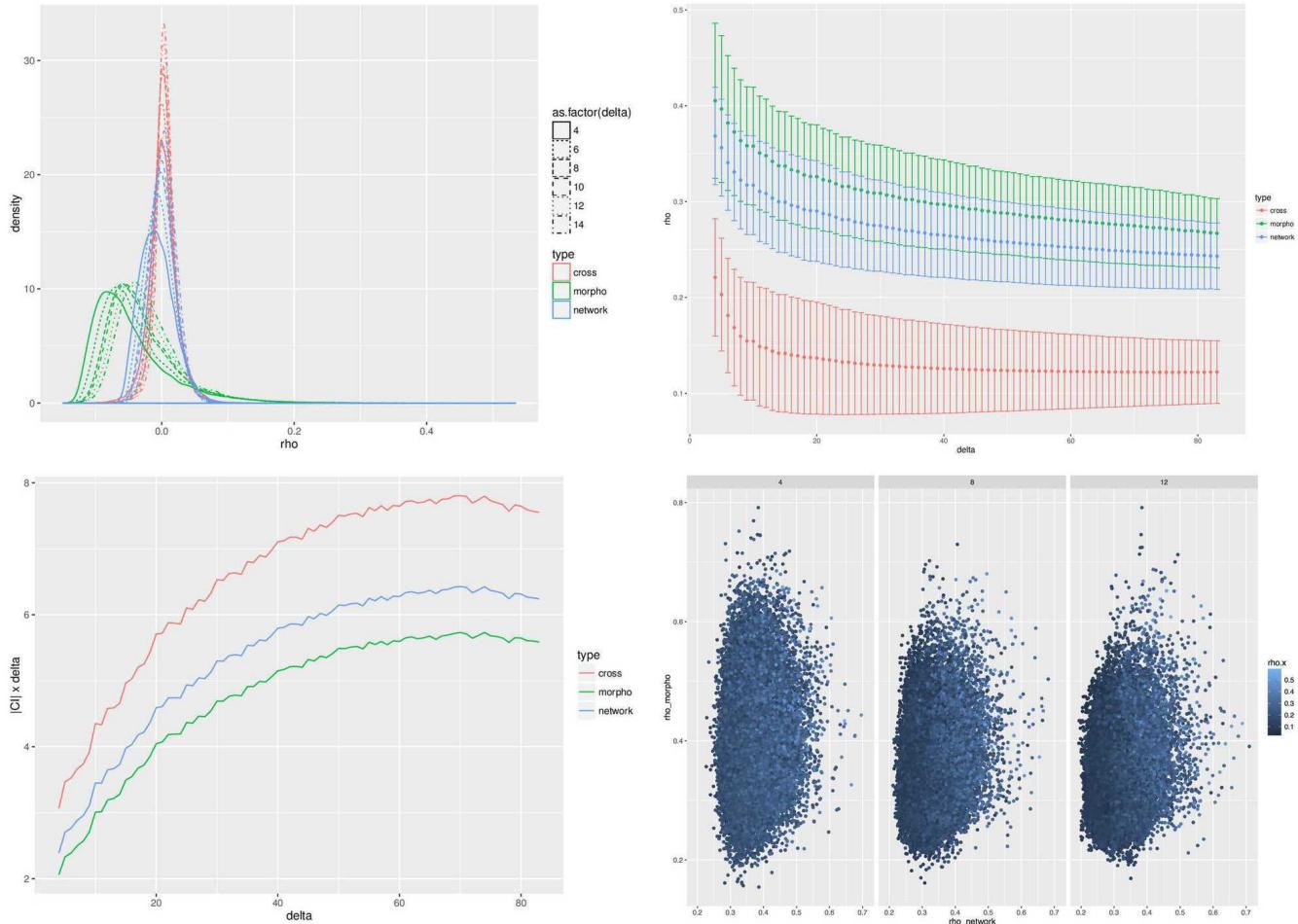


FIGURE 18 : Variation des corrélations avec l'échelle, pour les corrélations calculées sur l'Europe. (Haut Gauche)
 Distribution statistique des corrélations, pour les différents blocs morphologique, réseau et corrélations croisées (couleur), pour différentes valeurs de δ (type de ligne); (**Haut Droite**) Correlations absolues moyennes et leur déviation standard, pour les différents blocs, en fonction de δ ; (**Bas Gauche**) Taille de l'intervalle de confiance normalisé $\delta \cdot |\rho_+ - \rho_-|$ (IC estimé par méthode de Fisher) en fonction de δ ; (**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 .

globale des systèmes, qui est un aspect essentiel postulé par la Théorie Evolutive [PUMAIN, 2012b], conduisant à discuter les interprétation universelles des systèmes urbains proposées par les théories du Scaling [BETTENCOURT et al., 2007]. Nous suggérons que la non-stationnarité spatiale est reliée d'une part à différentes échelles de temps impliquées, et d'autre part à une non-ergodicité globale, sous l'hypothèse de stationnarité et d'ergodicité locale. Cette dernière paraît raisonnable, au sens où un régime local se manifestera de manière aléatoire sur ses différentes instances locales dans le cas d'indicateurs effectivement stochastiques à cette échelle (on pourra considérer les résultats de simulation de 5.2 pour se donner une idée). Empiriquement, la croissance urbaine et du réseau assez récente, rapide et étendue, laisse penser qu'on devrait être dans un cas analogue. Supposons ergodicité locale en \vec{x}_0 à l'échelle $\delta \cdot l_0$ à laquelle nous estimons les corrélations. Alors le théorème ergodique fournit un échantillonnage temporel \mathcal{T} tel que

$$\langle Y_i(t) \rangle_{\|\vec{x} - \vec{x}_0\| < \delta \cdot l_0} = \langle Y_i(\vec{x}_0) \rangle_{t \in \mathcal{T}}$$

En se plaçant en un autre point \vec{x}_1 assez loin, la stationnarité spatiale devrait impliquer $\langle Y_i \rangle_{\vec{x}_0} = \langle Y_i \rangle_{\vec{x}_1}$ et \mathcal{T} sera similaire pour garder invariance par translation. Par contraposition comme on a montré la non-stationnarité, les processus ont ainsi nécessairement des caractéristiques dynamiques différentes. Concernant la non-ergodicité globale, soit X_k une partition de l'espace en zones locales. On a $\langle \cdot \rangle_x = \sum_k w_k \langle \cdot \rangle_{x_k} = \sum_k w_k \langle \cdot \rangle_{\mathcal{T}_k}$. Mais d'autre part, l'ergodicité globale impliquerait que $\langle \cdot \rangle_t = \langle \cdot \rangle_{\mathcal{T}} = \sum_k w_k \langle \cdot \rangle_{\mathcal{T}}$ et donc $\sum_k w_k (\langle \cdot \rangle_{\mathcal{T}} - \langle \cdot \rangle_{\mathcal{T}_k}) = 0$. Pour que cette relation soit vraie sur la totalité des sous-ensembles, il est nécessaire que $\mathcal{T} = \mathcal{T}_k$, ce qui contredit la propriété montrée précédemment, et le système global est nécessairement non-ergodique. Ces résultats dépendent des hypothèses théoriques, mais nous postulons qu'ils devraient rester vrais de manière empirique vu les suggestions de la Théorie Evolutive.

Discussion

UNIVERSALITÉ Des grilles de densité de population existent pour l'ensemble des régions du monde, comme par exemple celles fournies par [STEVENS et al., 2015]⁴. 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 [BRETAGNOLLE et PUMAIN, 2010a], mais la différence se doit d'être étudiée quantitativement.

⁴ disponibles à <http://www.worldpop.org.uk/>

DÉVELOPPEMENTS Nous avons montré empiriquement la non-stationnarité des interactions entre forme urbaine et forme de réseau, qui suggère la non-ergodicité du système urbain concernant l'interaction entre ces composantes. Nous n'extrayons pas de résultats directs sur les dynamiques par ces analyses statiques, mais pouvons postuler des résultats indirects : les processus spatio-temporels n'ont pas les mêmes vitesses et réagissent et diffusent différemment. Certains développements de cette étude seraient potentiellement intéressants. La recherche d'échelle 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 validation de la taille resterait à déterminer : il peut s'agir comme ci-dessus de portée optimale pour des modèles locaux. La question de l'ergodicité doit 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 $\rho[\Delta_t Y]$ et celles dans l'espace $\rho[\Delta_x Y]$, mais la question de l'existence de base assez fines dans le temps paraît problématique. L'étude d'un lien entre $\Delta_\delta \rho(\delta)$ est les dérivées des processus est également une piste pour obtenir des informations indirectes sur la dynamique à partir des données statiques. D'autre part, 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. On suggère par exemple en Appendice ?? des cas idéaux pour lesquels un lien peut être directement obtenu, comme le cas de processus ondulatoires par exemple.

★ ★

★

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. Celle-ci 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 liens évolutifs, témoins des événements historiques entre les dynamiques démographiques territoriales et la croissance du réseau. L'utilisation de statistiques spatiales sur les relations dynamiques entre réseaux et territoires, c'est à dire cherchant à exhiber des relations de causalité, sont relativement rares. Par exemple, [LEVINSON, 2008] explique pour Londres les variables de population et de connectivité au réseau par ces mêmes variables décalées dans le temps, démontrant des effets causaux réciproques. [GARGI CHAUDHURI AND KEITH C CLARKE, 2015] 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. [CUTHBERT, ANDERSON et HALL, 2005] procède à des estimations économétriques des influences réciproques, et conclut que dans le cas d'étude (au Canada à une échelle sous-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, BLANQUART et DELAPLACE, 2013] 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 propre de la desserte négatif, après contrôle de l'endogénéité de la desserte par un modèle de selection, et un effet significatif des caractéristiques propres des unités urbaines. Cette étude reste limitée car non spatialisée et ne prenant en compte un décalage d'une unité de temps seulement. [CANNING et PEDRONI, 2008] 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.

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 excellente allégorie de ce phénomène : des méthodes isolant les “effets structurants” d’une infrastructure développées dans les années 70 [BONNAFOUS et PLASSARD, 1974] se sont révélées par la suite de l’instrumentation politique et sans fondement empirique [OFFNER, 1993]. Le débat est toujours d’actualité puisque la question se pose toujours par exemple pour la construction de lignes à grande vitesse [CROZET et DUMONT, 2011b]. La réalité des processus territoriaux est en fait bien plus compliqué qu’une simple relation causale entre la mise en place d’une infrastructure et les retombées sur le développement local, mais correspond bien d’une *co-évolution* complexe [BRETAGNOLLE, 2009b]. Sur le temps long et à grande 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 [L’ESPACE GÉOGRAPHIQUE, 2014], montrant que le démêlage 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 [CERQUEIRA, 2017] qui s’avèrent tout autant complexes. Ce type de problématique est bien sûr présent dans d’autres domaines : en Economie Géographique, l’exemple des liens entre innovation, impacts locaux de la connaissance et aggregation des agents économiques est une illustration typiques de processus économiques spatio-temporels présentant des causalités circulaires difficiles à démêler [AUDRETSCH et FELDMAN, 1996]. Des méthodes spécifiques sont introduites, comme l’utilisation d’instruments statistiques comme par [ACHION et al., 2015] 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 haut salaires, et permet de montrer que la correlation significative entre les deux est en fait une causalité de l’innovation sur les inégalités.

Le couplage fort spatio-temporel implique généralement l’introduction de la notion de causalité, à laquelle la géographie s’est toujours intéressée : [LOI, 1985] montre que les questions fondamentales que se pose la géographie théorique récente (isolation des objets, lien entre espace et structures causales, etc.) étaient déjà présentes dans la géographie classique de Vidal. [CLAVAL, 1985] critique d’ailleurs les nouveaux déterminismes ayant émergé, notamment celui proposé par certains tenants de l’analyse systémique : 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 devraient permettre de capturer la complexité qui fait la particularité des décisions humaines : l’école de Prigogine et la Théorie des

Catastrophes de Thom. Ce point de vue est remarquablement visionnaire, puisque comme le rappelle Pumain dans [PUMAIN, 2003], 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 [DURAND-DASTES, 2003], 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 *systémogenèse*. 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.

Les régimes sous lesquels des identifications de causalité sont cohérentes ne sont pas identifiés de manière évidente. Ceux-ci dépendront des définitions utilisées, de la même manière que les méthodes à disposition pour lesquelles nous pouvons donner quelques illustrations. [LIU et al., 2011] propose la détection de relations spatio-temporelles entre perturbations des flots 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 trafic [MIN et WYNTER, 2011]. Également dans le domaine des transports et de l'usage du sol, [XIE et LEVINSON, 2009a] 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. [LUO et al., 2013] 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 [KE, SUKTHANKAR et HEBERT, 2007] 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 concrète comme la compression de fichiers vidéos par extrapolation des vecteurs de mouvement [CHALIDABHONGSE et KUO, 1997]. Dans l'ensemble de ces cas, l'étude des corrélations spatio-temporelles rejoint les notions faibles de causalité vues précédemment. Cette contribution cherche 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'équilibrisme de concilier un certain niveau de sim-

plicité 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 Granger⁵, 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 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, i.e. $\hat{\rho}[X, Y] = \hat{E}_{i,t,k}[XY] - \hat{E}_{i,t,k}[X]\hat{E}_{i,t,k}[Y]$. Il est important de noter ici l'hypothèse de stationnarité spatiale et temporelle, qui peut toutefois aisément se relâcher 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 GWR⁶ [BRUNSDON, FOTHERIN-

⁵ Se référer au chapitre préliminaire à la deuxième partie pour les définitions mathématiques des notions non triviales utilisées.

⁶ voir également le chapitre de définition

GHAM et CHARLTON, 1998] par exemple). Cela nous permet de définir la corrélation retardée par

$$\rho_\tau [X_{j_1}, X_{j_2}] = \hat{\rho} [x_{i,j_1,t-\tau}^{(k)}, x_{i,j_2,t}^{(k)}] \quad (3)$$

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. Les critères de significativité dépendront du cas d'application et de l'estimateur utilisé, mais peuvent par exemple inclure la significativité du test statistique (test de Fisher dans le cas d'un estimateur de Pearson), la position des bornes d'un intervalle de confiance à un niveau donné, ou même un seuil exogène θ sur $|\rho_\tau|$ pour forcer un certain degré de corrélation.

4.2.2 Données Synthétiques

Cette méthode doit dans un premier temps être testée et partiellement validée, ce que nous proposons de faire sur des données synthétiques, méthode qui permet une connaissance plus fine des comportements des modèles [RAIMBAULT, 2016]. 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. [RAIMBAULT, BANOS et DOURSAT, 2014] décrit et explore un modèle simple de morphogenèse urbaine (modèle RBD) répondant parfaitement à ces contraintes. En effet, 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 n'exhibent pas de causalités évidentes : c'est donc un parfait cas pour tester si la méthode est capable d'en détecter. Nous utilisons une implémentation adaptée⁷ 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

⁷ disponible sur le dépôt ouvert du projet à

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

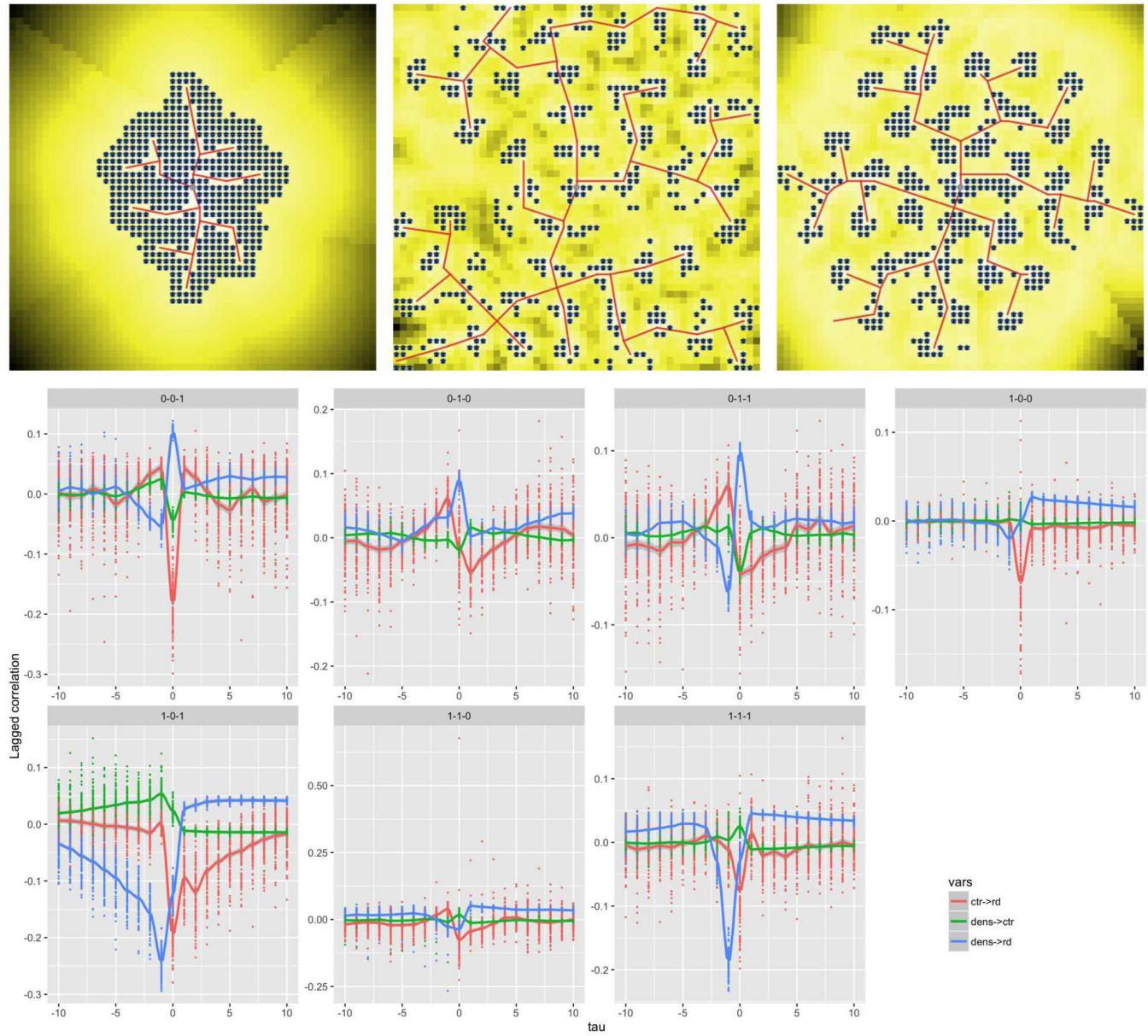


FIGURE 19 : Correlations dans le modèle RDB (Première ligne) Exemples de configurations finales, obtenues avec (w_d, w_c, w_r) valant respectivement $(0, 1, 1), (1, 0, 1)$, et $(1, 1, 1)$. (Deuxième ligne) Corrélations retardées, pour chaque combinaison des paramètres, en fonction du retard τ . Les différentes 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.

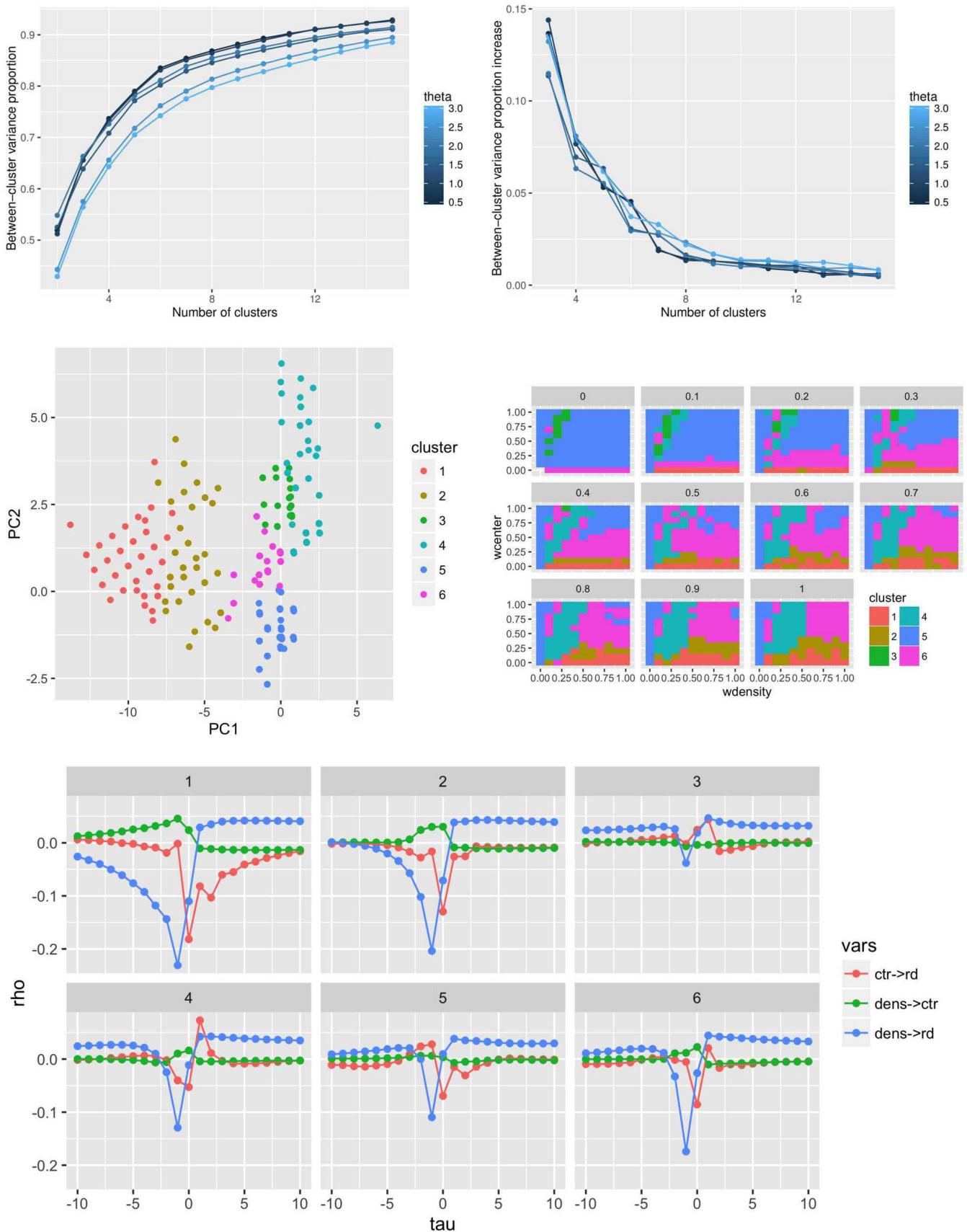


FIGURE 20 : Identification de régimes d'interactions endogènes par classification non-supervisée (Haut Gauche) Variance inter-cluster comme fonction du nombre de clusters. (Haut Droite) Dérivée de la variance inter-cluster. (Milieu Gauche) Features dans un plan principal (81% de variance expliquée par les deux premières composantes) (Milieu Droite) Diagramme de phase des régimes dans l'espace (w_d , w_c , w_r), w_r variant entre les différents sous-diagrammes. Théorie urbaine (Bas) Trajectoires correspondantes des centroïdes.

modèle RBD, faisant varier les paramètres de poids de la densité, de la distance au centre et de la distance au réseau⁸, que l'on note respectivement (w_d, w_c, w_r), dans $[0; 1]$ avec un pas de 0.1. Les autres paramètres sont fixés à leur valeurs par défaut données par [RAIMBAULT, BANOS et DOURSAT, 2014]. Pour chaque valeur des paramètres, nous procémons à $N = 100$ répétitions ce qui est suffisant pour une bonne convergence des indicateurs. Les explorations sont effectuées via le logiciel OpenMole [REUILLO, LECLAIRE et REY-COYREHOURCQ, 2013], le grand nombre de simulations (1,330,000) nécessitant l'utilisation d'une grille de calcul. Nous calculons sur l'ensemble des patches les corrélations retardées par estimateur de Pearson non biaisé entre les variations des variables suivantes⁹ : densité locale, distance au centre et distance au réseau. La Fig. 19 montre le comportement de ρ_τ pour chaque couple de variable (non dirigé, τ prenant des valeurs négatives et positives), pour les combinaisons des valeurs extrêmes des paramètres. On peut voir déjà différents régimes émerger : par exemple, $(1, 0, 1)$ conduit à une causalité de la densité sur la distance au centre avec un retard 1, et une causalité négative de la densité sur la distance au réseau avec le même retard, tandis que distance au centre et au réseau sont corrélés de manière synchrone.

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 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* 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 [HAMERLY et EL-KAN, 2003], dans notre cas le système possède une structure agréable : les courbes de la proportion de variance inter-cluster et de sa dérivée en Fig. 20, 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

8 Le modèle fonctionne de la façon suivante : une valeur des patches est déterminée par la moyenne pondérée de ces différentes variables explicatives, valeur qui détermine la croissance de nouveaux patches à l'instant suivant.

9 Calculer les corrélations sur les variables directement n'a pas de sens puisque leur valeur n'en a pas en absolu.

modèle, ainsi que les trajectoires des centres des clusters (calculées comme barycentre dans l'espace complet initial) en Fig. ??.

INTERPRÉTATION Nous proposons finalement d'interpréter les régimes obtenus. Le comportement obtenu est particulièrement intéressant : les régions du diagramme correspondant aux régimes sont clairement délimitées et connexes. Par exemple, on observe l'émergence du régime 6 où la distance au réseau cause fortement la densité de manière négative, mais la distance au centre cause la distance au réseau, régime dont l'étendue maximale sur (w_d, w_r) 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, 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 5, où la distance au réseau influence la densité de la même manière, mais la relation entre distance au centre et route est inversée, est tout aussi intéressant, et est prédominant dans les faibles w_r . Le régime 1, extrême, correspond à une situation isolée dans laquelle la distance au centre n'importe pas : cet aspect domine alors totalement les autres processus d'interaction entre densité et réseau. Cette application sur données synthétique 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 socio-économique, 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 montre dans [BAFFI, 2016] que lors de l'apartheid la planification du réseau ferré était utilisée comme un outil de ségrégation raciale par l'établissement de motifs de mobilité et d'accessibilité fortement contraints. En particulier, il est montré qualitativement que les dyna-

miques 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 minimisée 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 ségrégatives ont effectivement modifié la trajectoire du système couplé, ce qui correspondrait à des impacts plus larges et profonds que leurs effets immédiats.

Résultats

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, 2016], mais ne sont pas ouvertes, nous mettons ainsi à disposition uniquement les données agrégées utilisées dans l'analyse. **C (FL) : a mettre dans une partie type ch3, ultérieurement A1 : (JR) la discussion sur les données tu veux dire ?**

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 l'existence d'une planification les ayant influencées. Nous montrons en Figure 21 l'évolution des mesures de réseau dans le temps, 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, la planification après cette date a dans le meilleur des cas eu aucun effet sur cette propriété, dans le pire des cas 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 égalitaire.

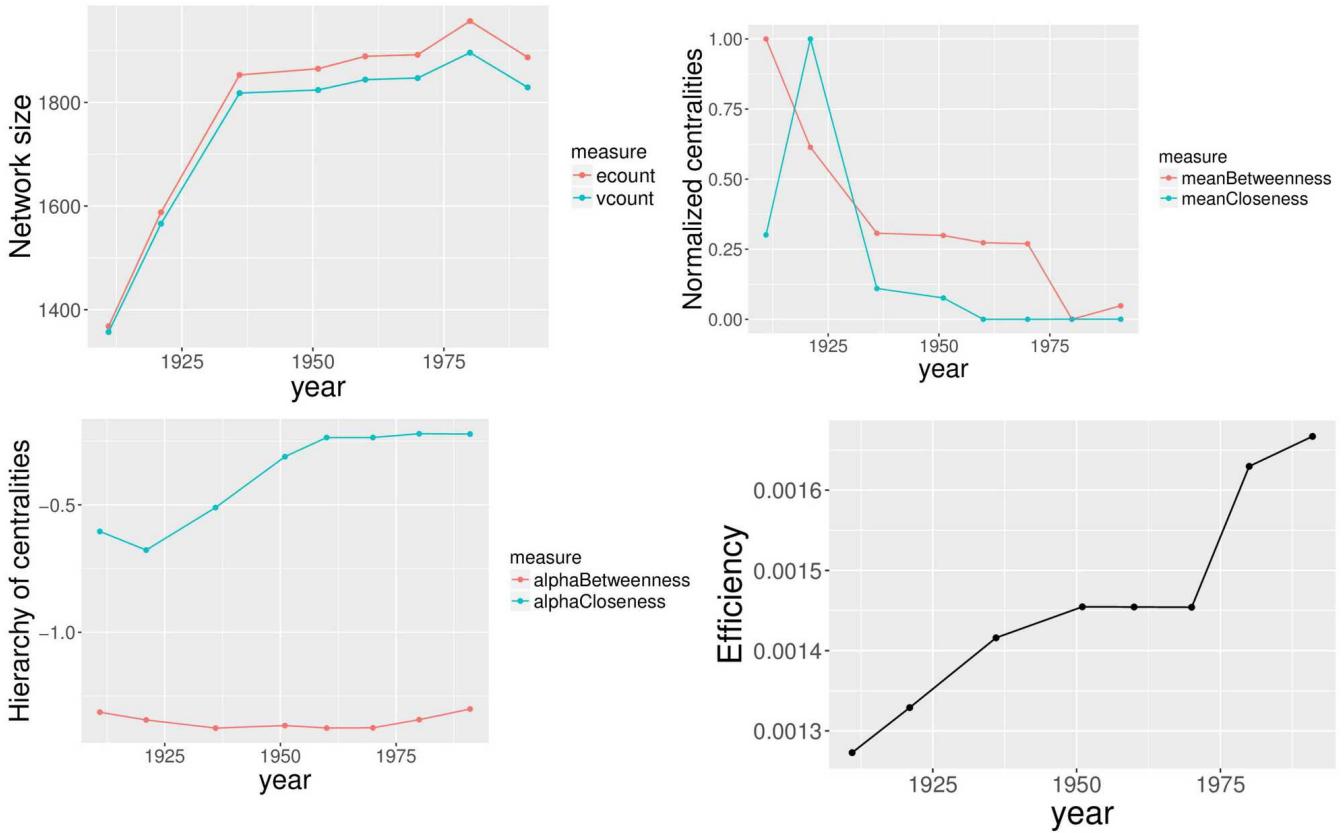


FIGURE 21 : Evolution des mesures de réseau. On calcule pour l'ensemble des dates les mesures basiques de réseau : taille, centralités résumées par leur hiérarchie et leur moyenne, efficience. Les centralités sont normalisées pour comparaison de leur variation respective ($\max \bar{bw} = 0.07$, $\max \bar{cl} = 1.5e-4$).

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 22. Nous obtenons des résultats significatifs avec

l'accessibilité non-pondérée seulement, l'auto-corrélation devant dominer 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. 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 le niveau moyen 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 exhibe clairement une transition de phase dans ses valeurs intermédiaires, ce qui devrait correspondre au passage entre l'échelle spatiale des aires urbaines et celle du pays, ce qui donne l'échelle locale de stationnarité spatiale. Quand on examine le comportement des corrélations retardées pour la distance, on observe des motifs de causalité assez évident, 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 après l'apartheid (avec un délai de 20 ans). 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 motif rejoignent les suggestions de la littérature empirique sur le sujet de l'Apartheid, comme le suggère [BAFFI, 2016], qui montrent par exemple un fort effet des mesures sur les déplacement 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 précise, pour quantifier par 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 [ANGRIST, IMBENS et RUBIN, 1996] 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 approches.

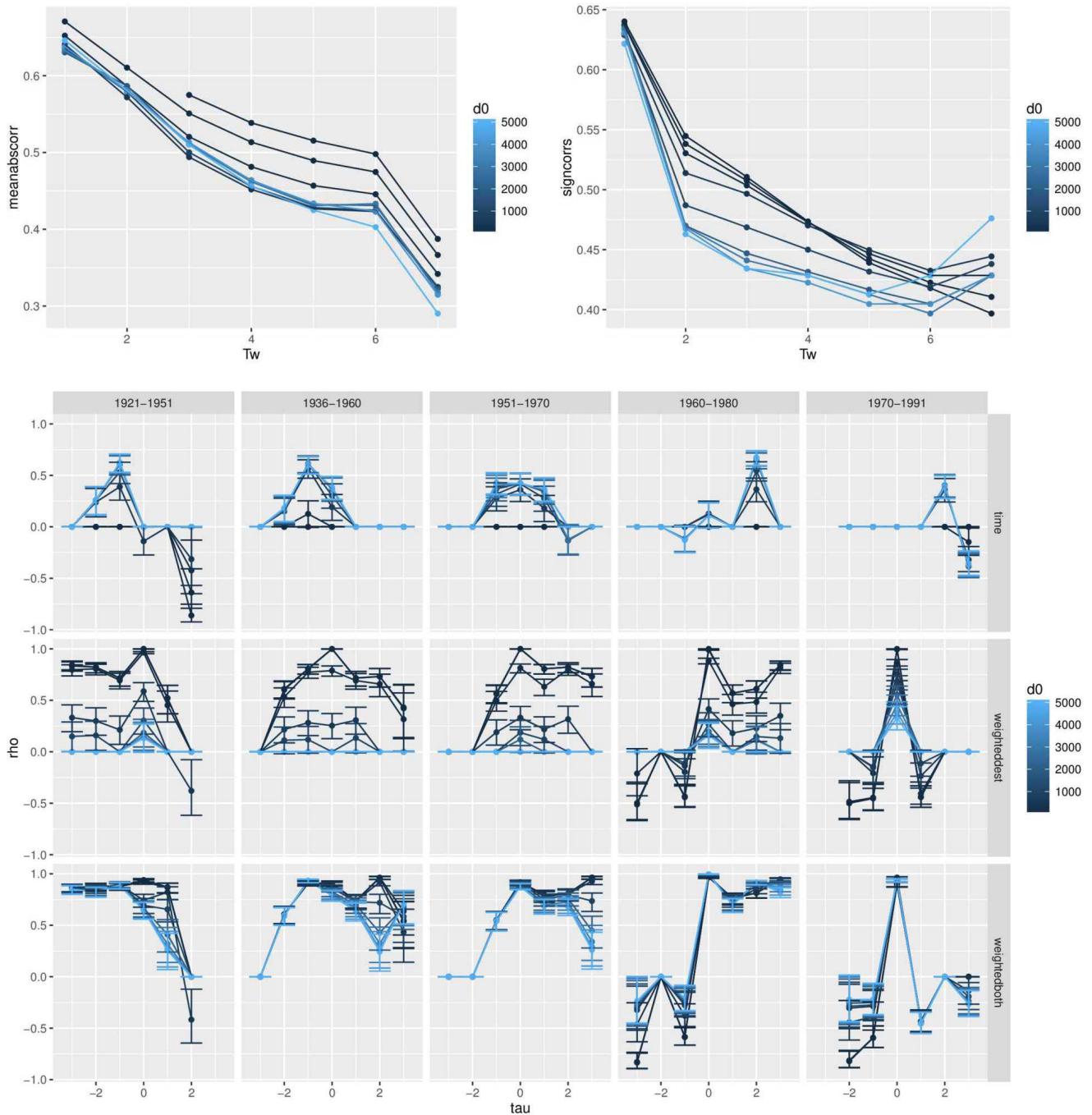


FIGURE 22 : Corrélations retardées. (Haut Gauche) Corrélations absolues moyennées sur l'ensemble des retards, en fonction de la taille de la fenêtre temporelle T_W (en nombre d'observations temporelles), pour différentes valeurs du paramètre de décroissance d_0 ; (Haut Droite) Proportion de corrélations 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). Pour interpréter, on observe pour l'accessibilité temporelle (les autres étant basées par l'autocorrélation) 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.

★ ★

★

4.3 MODÈLE DE CROISSANCE MACROSCOPIQUE

C (FL) : pourquoi a la suite de 4.1 et 4.2 ? la démarche est fort différente A1 : (JR) entree thematique par theorie evolutive, démarche de modelisation logique dans ce cadre

Nous décrivons 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 du fit 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.

4.3.1 Contexte

Replaçons la présente démarche dans un contexte plus global de modélisation de la croissance urbaine, notion plus générale que notre problématique précise mais qui nous permet ici 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. [PUMAIN, PAULUS et VACCHIANI-MARCUZZO, 2009] suggère que les villes sont l'incubateur du changement social, leur destin étant étroitement lié à celui des sociétés. 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 [KRUGMAN, 1998] mais ceux-ci sont extrêmement détaillés en termes de processus de marché, même pour des modèles en économie géographique, tandis que la géographie se concentre plus sur les spécificités territoriales et les interactions dans l'espace mais produira des conclusions générales avec plus de difficulté. L'exemple de ces deux disciplines montre comment il est difficile de créer des ponts, comme il a fallu des effort exceptionnel pour effectuer des traductions de l'une à l'autre (comme P. HALL le fit avec le travail de VON THUNEN [TAYLOR, 2016]), 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 expliquer la croissance urbaine, le modèle de Gibrat, qui assume des taux de croissance aléatoires, a été montré par [GABAIX, 1999] 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 [NITSCH, 2005]. Expliquer les loi d'échelles urbaines est étroitement lié à la compréhension de la croissance urbaine, comme [BETTENCOURT, LOBO et WEST, 2008] suggère que celles-ci reflètent des processus universels sous-jacents et que toutes les villes sont des versions à l'échelle l'une de l'autre. Cette approche reflète cependant peu les relations complexes entre agents économiques pour lesquelles [STORPER et SCOTT, 2009] se positionnent. Par l'utilisation d'une reconstruction par le bas des aires urbaines via des données microscopiques dynamiques de population, [ROZENFELD et al., 2008] montre en effet que des déviations positives à la loi rang-taille existent systématiquement, et qu'elles doivent être un effet des interactions spatiales entre les aires urbaines. Les approches par la complexité sont de bon candidates pour intégrer celles-ci dans les modèles. [ANDERSSON, FRENKEN et HELLERVIK, 2006] introduit par exemple un modèle d'économie urbaine comme un réseau complexe de relations en croissance. La Théorie Evolutive Urbaine, introduite par [PUMAIN, 1997], se concentre sur les villes comme des entités en co-évolution et produit des explications pour la croissance au niveau du système de villes. [PUMAIN et al., 2006] montre que les lois d'échelles pourraient être dues à la différenciation 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, puisque [PUMAIN, 2012b] 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, généralement les villes, qui produisent les motifs émergents à l'échelle du système. [PUMAIN et SANDERS, 2013] 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.

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 interactions avec les flux portés par les réseaux physiques, principalement les réseaux de transport. Nous étendons les modèles existants de manière correspondante. Notre contribution

consiste 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, 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 revoyons d'abord les approches de modélisation de la croissance urbaine basées sur les interactions spatiales.

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 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. [BRETAGNOLLE et al., 2000] a proposé une extension spatiale du modèle de Gibrat. Le modèle d'interaction basé sur la gravité que [SANDERS, 1992] utilise pour appliquer les concepts de la Synergétique aux villes est également proche de cette idée de croissance 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 [FARVO et PUMAIN, 2011], fournissant une version du cœur des modèles Simpop [PUMAIN, 2012a] en termes de systèmes dynamiques. Cette famille de modèles a commencé avec un modèle jouet basé sur les interactions économiques entre les villes comme agents, qui produit des motifs de hiérarchie à l'échelle du système [SANDERS et al., 1997]. 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 [BRETAGNOLLE et PUMAIN, 2010a]. Le modèle SimpopLocal [PUMAIN et REUILLOU, 2017a] est utilisé pour montrer l'émergence des motifs initiaux d'établissement humains. Le modèle Marius [COTTINEAU, 2014] couple la croissance de la population et économique avec les interactions entre les ville, permettant de reproduire assez

fidèlement les trajectoires réelles sur l'ancien Union Soviétique après calibration avec multi-modélisation des processus.

CROISSANCE URBAINE ET RÉSEAUX DE TRANSPORTS 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 [CHANG, 2006]. Les modèles de croissance de réseau [XIE et LEVINSON, 2009c], prolifiques en économie et physique, ne peuvent pas être utilisés pour expliquer la croissance urbaine. [BIGOTTE et al., 2010] 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. [BAPTISTE, 1999] 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 [SCHMITT, 2014] 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 soient 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. **C (FL) : cela devrait etre dans le chapitre 2 A1 : (JR) j'en ai déjà parlé; ca ne marche pas là en mini etat de l'art local pour bien situer le modèle ?**

La suite de cette section est organisée de la manière suivante : le modèle est d'abord introduit et décrit de manière formelle ; puis nous décrivons 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.

4.3.2 Modèle et Résultats

Description du modèle

CONTEXTE 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, au sens de l'émergence faible [BEDAU, 2002]. En particulier, la distribution complète des modèles de croissance aléatoire peut être déterminée ana-

lytiquement à tout instant [GABAIX, 1999], 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 [COTINEAU, 2014]. 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. 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 dire que la force d'interaction décroît avec la distance. D'autres approches introduites plus récemment ont des performances similaires à cette échelle [MASUCCI et al., 2013].

DESCRIPTION DU MODÈLE Nous considérons une extension déterministes du modèle de Gibrat, ce qui est équivalent à considérer seulement les espérances dans le temps. Soit $\vec{P}(t) = (P_i(t))_{1 \leq i \leq n}$ la population des villes dans le temps. Sous les hypothèses d'indépendance de Gibrat, nous avons $\text{Cov}[P_i(t), P_j(t)] = 0$. Une version étendue linéaire 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 modèles et plus flexible. Notant $\vec{\mu}(t) = \mathbb{E}[\vec{P}(t)]$, nous prenons $\vec{\mu}(t+1) = \Delta t \cdot f(\vec{\mu}(t))$. 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 $\mathbf{G} = G_{ij} = w_G \cdot \frac{V_{ij}}{\langle V_{ij} \rangle}$ 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),

$$V_{ij} = \left(\frac{\mu_i \mu_j}{(\sum_k \mu_k)^2} \right)^{\gamma_G} \cdot \exp(-d_{ij}/d_G) \quad (5)$$

Le terme d'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 sept paramètres du modèle sont détaillés ci-dessous.

La première composante 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 [SANDERS, 1992]. La rationnelle pour le 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 assumer 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 ?? 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 la séparabilité entre l'effet de la distance et 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. Nous proposons d'interpréter le paramètre de décroissance de la distance de la façon suivante. 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 cette requête intuitive. 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 4 : Espace des paramètres

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 assumé hybride car il repose sur une semi-paramétrisation sur des vraies données. 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 [PUMAIN et RIANDEY, 1986], 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 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, 2009a], qui construit une définition fonctionnelle des villes comme entités dont les limites évoluent dans le temps. Nous travaillons avec les 50 plus grandes villes en 1999. Nous isolons de plus des périodes de longueur similaires excluant les guerres, obtenant 9 périodes de 20 ans sur lesquelles le fit 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 construisons 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 plus petites.

Evaluation du modèle

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 $\bar{\mu}(t_0) = \bar{p}(t_0)$ 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 carré 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 carré moyenne des logarithmes

$$\varepsilon_L = \frac{1}{T} \sum_t \frac{1}{n} \sum_i (\ln p_i(t) - \ln \mu_i(t))^2$$

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 STYLISÉS Des faits stylisés basiques peuvent être extraits d'une telle base de données, comme il a été déjà largement été exploré dans la littérature [GUÉRIN-PACE et PUMAIN, 1990]. Nous retrouvons les meilleurs fits des distributions log-normales des taux de croissance à toutes les dates comparé à des distributions normales, et aussi le fait que les taux de croissance sont essentiellement positifs, sur les villes que nous considérons et en 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 $\hat{\rho}_{ij} = \rho [\Delta X_i, \Delta X_j]$ avec un estimateur de Pearson classique, où $\Delta X_i = X_i(t) - X_i(t-1)$ et $X_i(t) = \ln \left(\frac{P_i(t)}{P_i(t_0)} \right)$. Cette méthode, utilisée principalement en éconophysique [MANTEGNA et STANLEY, 1999], révèle des interactions dynamiques sans être biaisée par les tailles. Nous montrons en Figure 23 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 différenciations 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 [THÉVENIN, SCHWARTZ et SAPET, 2013], 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

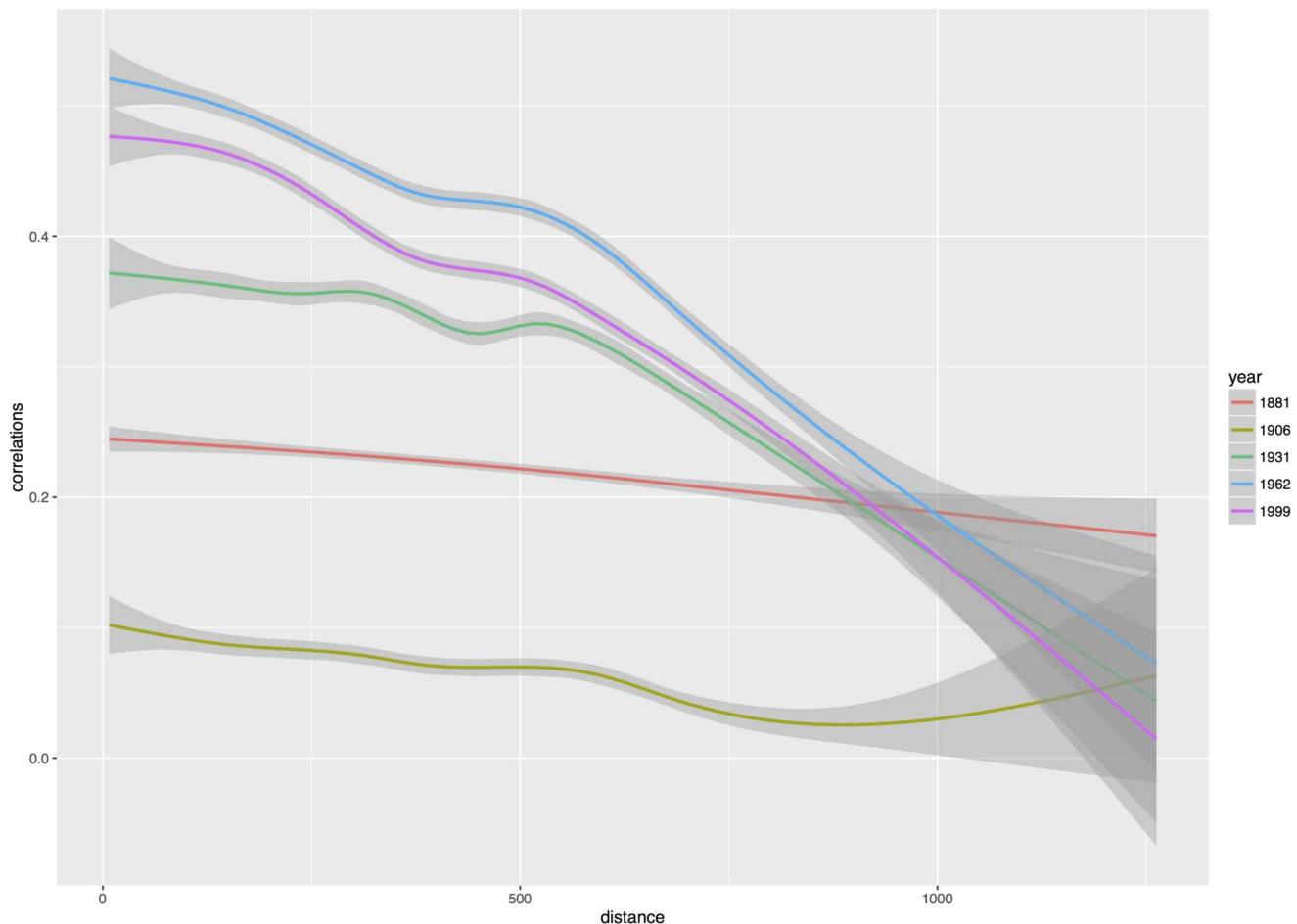


FIGURE 23 : Corrélations entre séries temporelles en fonction de la distance. Les lignes pleines correspondent aux corrélations lissées, calculées entre chaque paire des log-returns normalisés des séries temporelles de population, sur des périodes successives données par la couleur de la courbe.

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 [REUILLOU, LECLAIRE et REY-COYREHOURCQ, 2013], 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

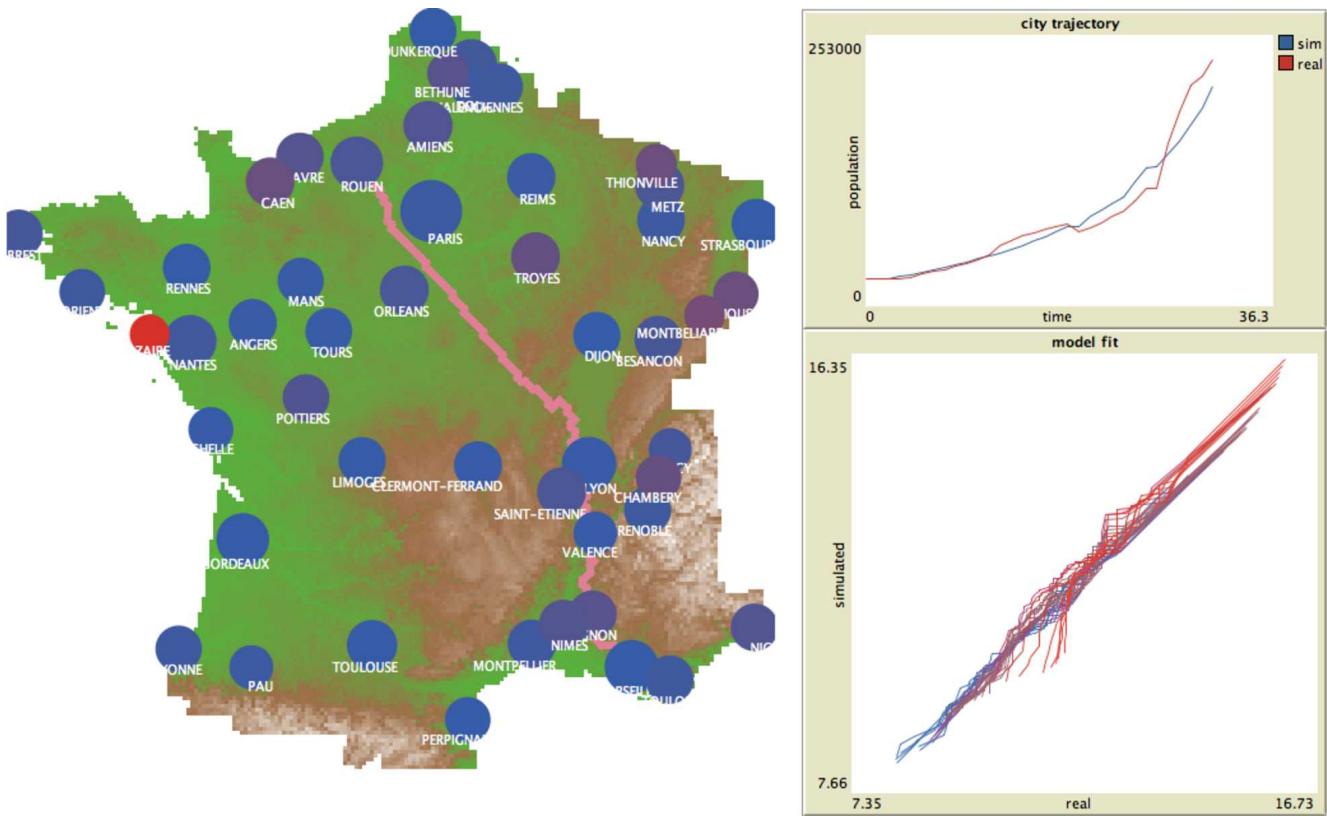


FIGURE 24 : Exemple de sortie du modèle. L’interface graphique permet d’explorer de manière interactive sur quelles villes les changements s’opèrent après un changement de paramètres, ce qui est nécessaire pour interpréter les résultats bruts de calibration.

des modèles, les données brutes nettoyées, les données de simulation, et les résultats utilisés ici sont disponibles sur le dépôt ouvert du projet à <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/NetworkNe>. Nous montrons en Fig. 24 un exemple de sortie du modèle. Les couleurs des villes donnent l’erreur de fit 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 simplement parcourant 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 les taux de croissance. Nous montrons en

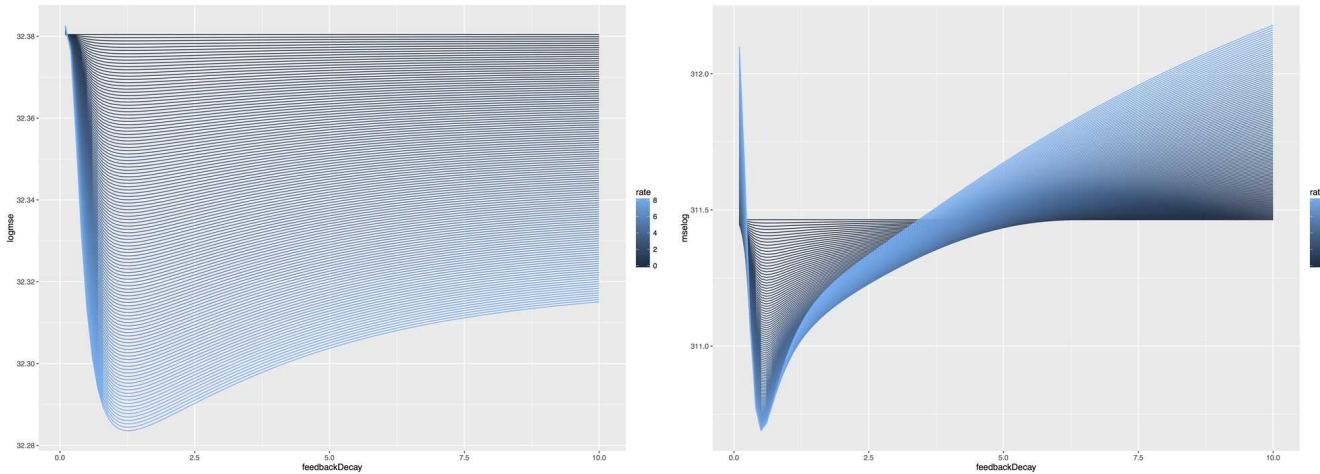


FIGURE 25 : Effets de réseau révélés par l’exploration du modèle. Le graphe de gauche donne ϵ_G comme fonction de d_N pour r_0/w_N variant, à effet de gravité fixe et $\gamma_N = 3$. Le graphe de droite est similaire pour ϵ_L .

Fig. 25 une configuration dans laquelle c'est le cas. A 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 calibrons 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 (BATTY et MACKIE, 1972) comme une descente du gradient échouent à cause de la forme assez compliquée du paysage 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

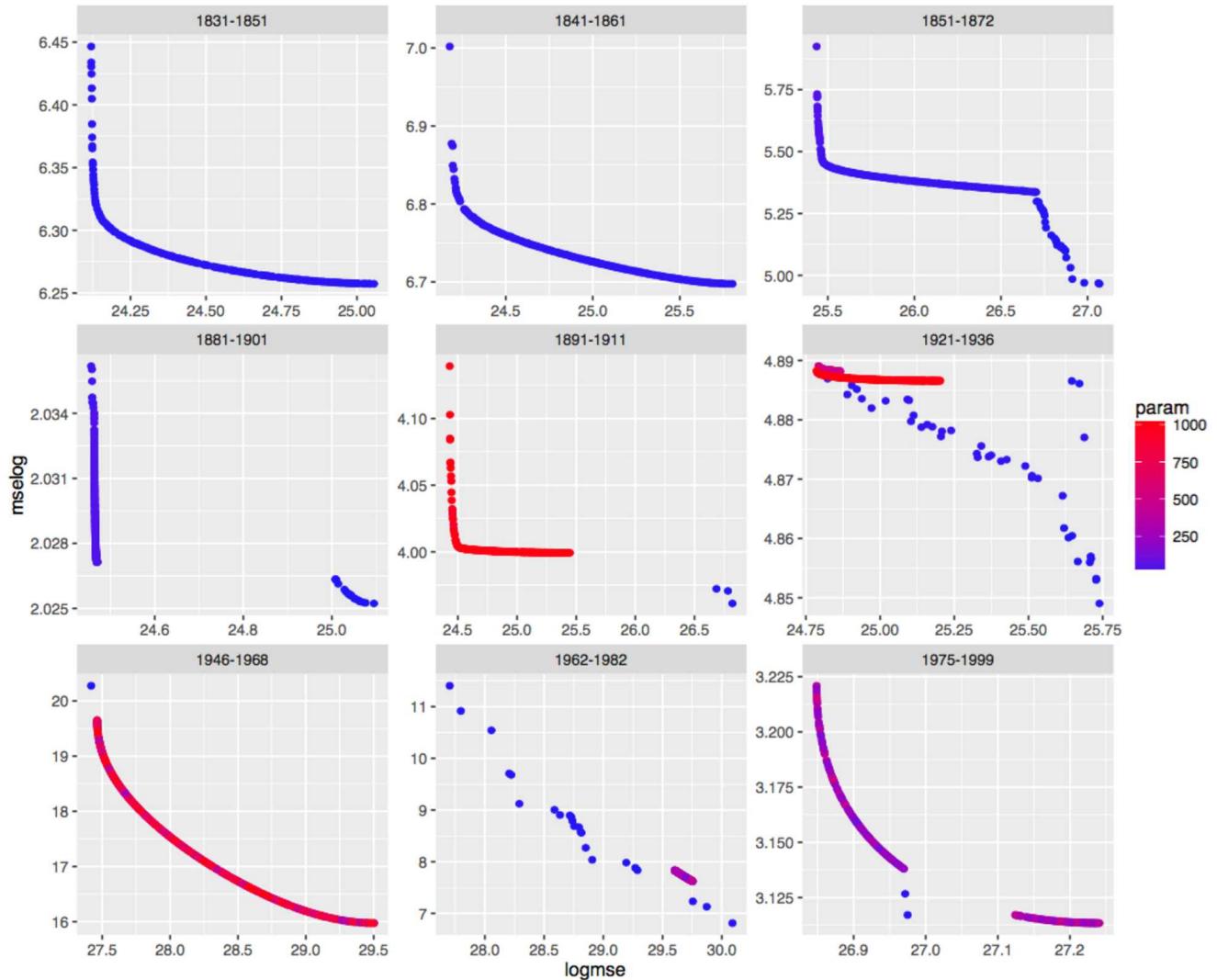


FIGURE 26 : Calibration du modèle de gravité. Fronts de Pareto sur les périodes successives. La couleur donne la valeur du paramètre de décroissance de la distance.

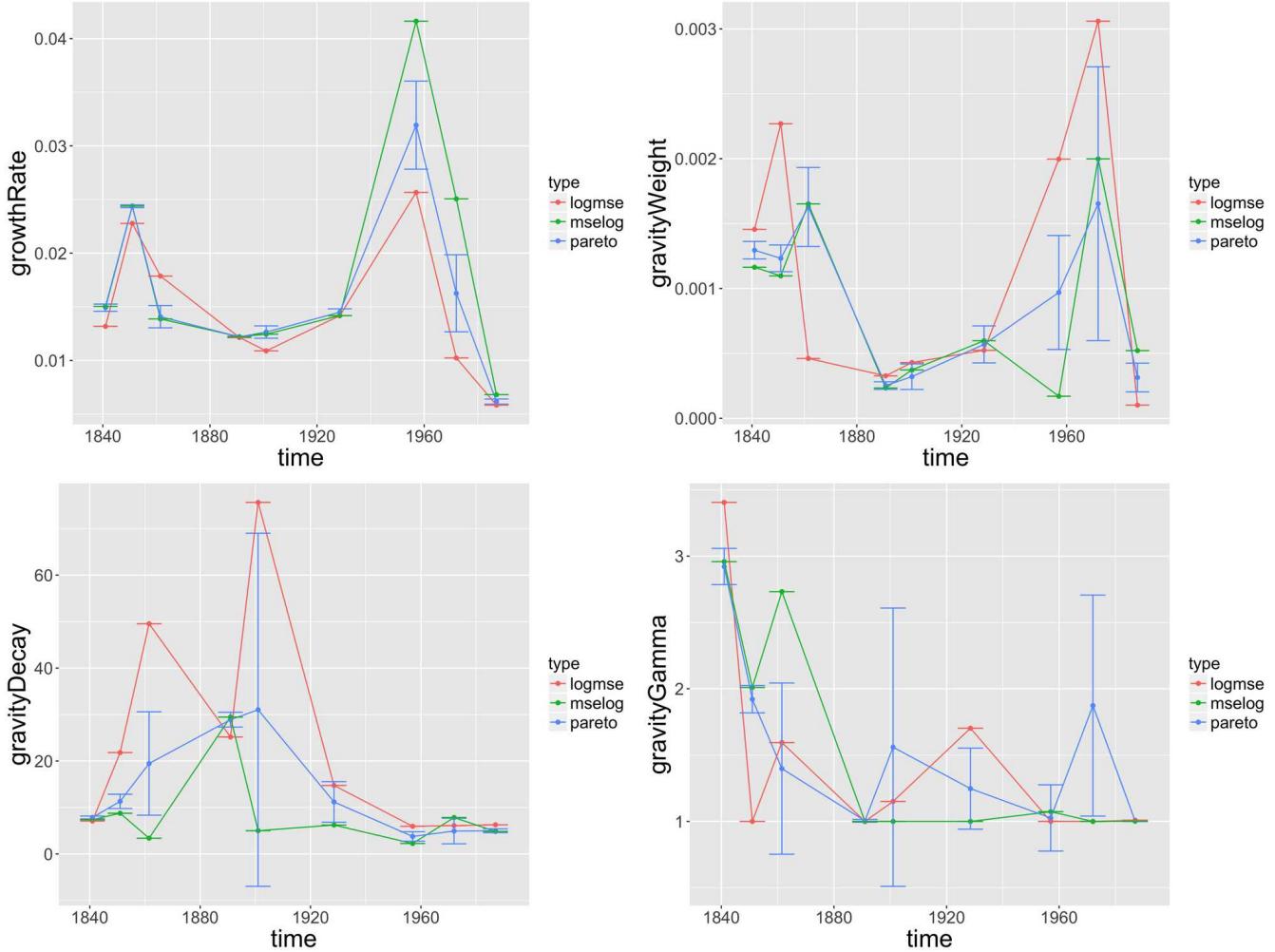


FIGURE 27 : Valeurs des paramètres calibrés pour le modèle de gravité seul. Chaque graphique donne les valeurs ajustées dans le temps pour chaque paramètre. Les courbes rouge et verte correspondent aux points optimaux pour ε_G (respectivement ε_L), tandis que la courbe bleue donne la valeur moyenne sur l'ensemble du front de Pareto avec la déviation standard.

méta-heuristiques pour différents buts : [SCHMITT et al., 2014] 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 fourni par OpenMole, distribué sur 25 îles, avec une population de 200 et 100 générations. Nous montrons en Fig. 26 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 éparpillés. 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. 27 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 hiérarchie γ_G a été régulièrement décroissant, ce qui correspond à 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'administration.

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. 28 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, quand les transports à grande vitesse ne s'arrêtent peu. 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.

PERFORMANCE DU MODÈLE 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 l'overfitting, qui a été reconnue depuis un certain temps en Apprentissage Statistique par exemple [DIETERICH, 1995], mais pour lequel il manque des méthodes applicables aux modèles de simulation. Nous avons besoin d'introduire un outil qui confirme que l'amélioration de l'ajustement n'est pas uniquement artificiellement due aux paramètres supplémentaires. 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 [AKAIKE, 1998], corrigeant l'amélioration du fit par le nombre de paramètres. Des méthodes similaires incluent le critère d'information Bayésien (BIC), qui repose sur des hypothèses légèrement différentes et corrige différemment. [BIERNACKI, CELEUX et GOVAERT, 2000] propose une likelihood intégrée comme une généralisation de ces critères pour la classification non-supervisée. [POHLE et al., 2017] montre que dans le cas de la sélection du nombre d'états pour des Modèles de Markov Cachés, les cas réels induisent trop

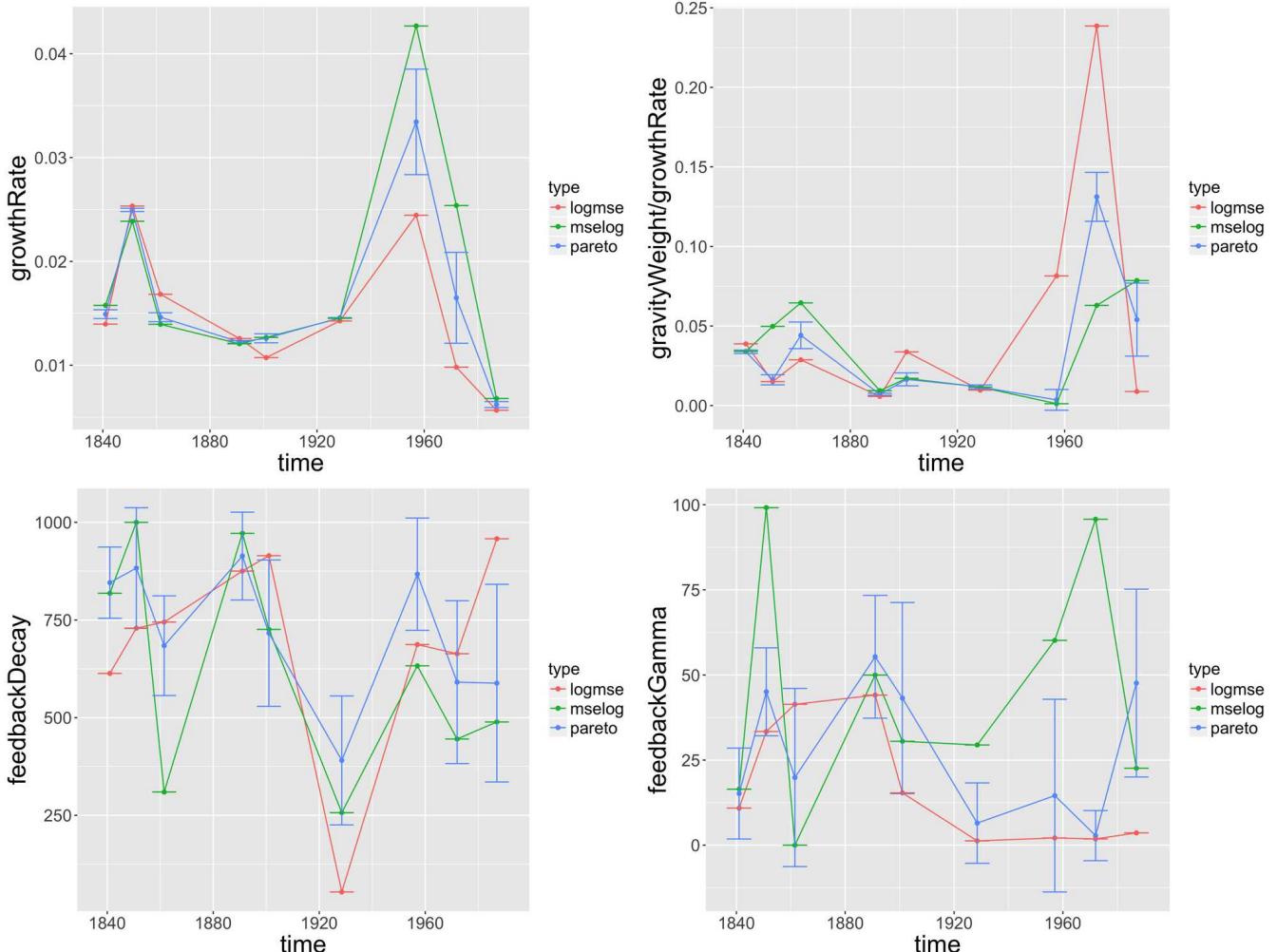


FIGURE 28 : Paramètres ajustés pour le modèle complet. Nous donnons les valeurs de r_0 , w_G/r_0 , d_N et γ_N dans le temps, pour les points optimaux pour les objectifs simples (courbes rouge et verte) et moyen sur le front de Pareto (bleu).

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. [BASTANI, KIM et BASTANI, 2017] 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 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 calibrions 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 le 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 5. 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 5 : Résultats 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 [BRETAGNOLLE et al., 2000] 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 de train utilisé par [THÉVENIN, SCHWARTZ et SAPET, 2013]. 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, 2009a]. Ce type de modèle est très rare, et [SCHMITT, 2014] fournit avec SimpopNet l'un des exemples. Il est montré par [RAIMBAULT,

2016] et 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 [DUPUY et STRANSKY, 1996] 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

Il est bien établi en géographie l'importance des relations spatiales et de la mise en réseau, comme le formule TOBLER par sa "première loi de la géographie" [TOBLER, 2004] **C (FL) : la citer** . Nous l'avons mis en évidence pour les relations entre réseaux et territoires par exemple en section 4.3. Toutefois, les travaux sur la non-stationnarité et la non-ergodicité, ainsi que la mise en valeur d'échelles locales endogènes, suggèrent une certaine pertinence à l'idée de sous-système relativement indépendant, au sens où il serait possible d'isoler certaines règles locales régissant celui-ci étant fixés certains paramètres exogène 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 agents considérés. Le manque de coordination générale entre transports et urbanisme laisse supposer que les dynamiques urbaines en terme de populations et d'emplois y sont relativement insensibles à court terme **C (FL) : sens ?** . Le Bureau des Transports de la Municipalité ainsi que le bureau technique Européen ont pu subir des répercussions politiques et économiques bien plus graves **C (FL) : ~** . D'autre part, que ce soit à Zhongshan, Macao ou Hong-Kong le problème a une répercussion quasi-nulle. Généralisant au système de transport local, celui-ci peut être relativement bien isolé des systèmes voisins ou à plus grande échelle, et donc ses relations avec la ville considérées dans un contexte locale. On supposera à la fois une certaine forme de stationnarité locale ("régime urbain local") mais aussi une certaine indépendance avec l'extérieur. 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 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, très mal comprise, contient un potentiel d'application immense aux problèmes qui nous concernent, comme l'a effleuré DURAND-DASTÈS [DURAND-DASTÈS, 2003] 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 certains aspects. 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 faisable des corrélations statiques entre indicateurs de forme urbaine et indicateurs de réseau, qui sont comme on l'a vu précédemment un témoin des relations locales entre réseaux et territoires. Celui-ci s'avère relativement large, ce qui confirmera l'utilisation de ce type de modèle de manière fortement couplée par la suite.

* * *

*

Ce chapitre est composé de divers travaux. La première section est adaptée d'un travail en anglais en collaboration avec C. ANTELOPE, L. HUBATSCH et J.M. SERNA à la suite de l'école d'été 2016 du Santa Fe Institute [ANTELOPE et al., 2016]; la deuxième section est traduite de []; et enfin la troisième section a été écrite pour les Actes des Journées de Rochebrune 2016 [RAIMBAULT, 2016].

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. 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 [ANTELOPE et al., 2016] 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 appliquée comme introduit en section 3.3, pour obtenir des concepts aussi génériques et larges que possible. **C (FL) : TB**

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. 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'emphase sur le rôle du changement et de la croissance, plutôt qu'un état statique. Comme [THOMPSON, 1942] mentionnait déjà, "l'histoire naturelle est concernée par 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 [BOURGINE et LESNE, 2010] 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 réfère 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 [TURING, 1952], 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. [GIERER et MEINHARDT, 1972] 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 d'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 [KONDO et MIURA, 2010]. 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 [GIERER et MEINHARDT, 1972]. Le système à l'origine d'une coloration le plus étudié sont les réactions responsables des rayures jaunes et noires du poisson zèbre [NAKAMASU et al., 2009]. L'émergence de la polarité cellulaire est expliquée chez certaines levures par un mécanisme similaire [GORYACHEV et POKHILKO, 2008]. Des exemples impliquant des fonctions comme la segmentation du corps de *Drosophila melanogaster* impliquent des réseaux d'espèces chimiques bien plus complexe 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 [WOLPERT, 1969]. 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 [ROGERS et SCHIER, 2011]). 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 [WOLPERT, 2011], aucun de ces systèmes n'est parfaitement bien compris, mais les évidences 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 [HEISENBERG et BELLAÏCHE, 2013], 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 [GUILLOT et LENCUIT, 2013]. 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 [LECUIT et LENNE, 2007].

Artificial Life

La notion de *Programmable Self-Assembly* semble être en *Artificial Life*¹ très proche du concept biologique de morphogenèse : [CROSATO, 2014] 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 [DOURSAT, SAYAMA et MICHEL, 2013]. 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 [MONOD, 1970] pour garder un certain niveau de généralité). Un exemple d'une nuée hétérogène de particules, produisant des architectures complexes, est décrit dans [DOURSAT, 2008]. 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. [SAYAMA, 2009] 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, STEFANEC et CRAILSHEIM, 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 "imi-

¹ au sens du domaine scientifique propre, animé par exemple par la *International Society for Artificial Life* (voir <http://alife.org/>).

tant la vie”². L’ajout d’un environnement avec ses propriétés propres influence fortement les dynamiques morphogénétiques, comme montré par [CUSSAT-BLANC et al., 2012] 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 et al., 2017] 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 bio-mimétisme 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 petites³. La théorie du Métabolisme Urbain de OLSEN [OLSEN, 1982] 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 [MOUDON, 1997], qui s’intéresse aux processus morphogénétiques, est présenté comme un champ émergent en lui-même, à l’interface de la géographie, l’architecture et la planification urbaine : cette vision appuie sur le rôle crucial de la forme dans ce genre de processus. [BURKE, 1972] 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 développeurs. A une autre échelle, en architecture, un bâtiment peut être vu comme le résultat de processus microscopiques faisant sens et un style architectural particulier peut être interprété par l’utilisation d’une grammaire générative de formes [CECCARINI, 2001]. Cette méthodologie se rapproche du travail de C. ALEXANDER, un architecte ayant produit une théorie des processus de design [MEHAFFY, 2007], 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,

² Nous développerons plus en détails plus loin les concepts nécessaires pour creuser cette affirmation.

³ nous n’incluons pas l’aménagement du territoire, mais considérons bien le contexte de projets urbains qui ne dépassent jamais l’échelle métropolitaine

puisque elle réfère au processus de la génération de forme en général, de la même façon que [WHITEHAND, MORTON et CARR, 1999] 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 [DOLLENS, 2014], 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. [DESMARAIS, 1992] soulève que les structures humaines sont porteuses d'une morphologie abstraite, et que celle-ci est générée par des processus porteurs de sens. 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. [LÉVY, 2005] 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 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-Muth, est proposé dans [BONIN et HUBERT, 2012]. [MAKSE et al., 1998] 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. [ACHIBET et al., 2014] décrit un modèle de co-évolution entre réseau de rues et la structure des blocs urbains. A une plus grande échelle et impliquant des fonctions plus abstraites, [RAIMBAULT, BANOS et DOURSAT, 2014] 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 [RENFREW, 1978]. 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, 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, MULLER et BRENT, 2007]. D'autres part, en neurosciences, la structure du cerveau en elle-même et la mise en place des réseau de neurones est typiquement l'issue de processus morphogénétiques [*Issues in Neuroscience Research and Application* 2013]. En psychologie sociale, la co-évolution de l'individu et de la société peut également être vu par ce prisme [ARCHER et ARCHER, 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 PICIONE et FRANCESCA FREDA, 2016]. Toutefois, au delà d'une unité systématique au travers de ces différents champs, les usages sont plutôt discontinus, et on pourrait supposer 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.

AUTOPOIÈSE La notion d'autopoïèse provenant de la biologie, que nous détaillerons plus loin, 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, 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, 2004]. Ces approches rejoignent également les visions du sujet comme dynamique et récursif [PICHON RIVIÈRE, 2004]. L'interpénétration du social et du psychologique trouvent écho chez l'anthropologie psychanalytique de FREUD qui appuie les relations entre les symptômes neurotiques et les phénomènes socio-culturels [FREUD, STRACHEY et FREUD, 1989]. **C**
(FL) : cf abercrombie : ?

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 mouvement cellulaires de bactéries ont été découverts [ABERCROMBIE, 1977]. 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.

Autres

EPISTÉMOLOGIE La morphogenèse peut aussi être utilisée pour étudier la science elle-même : par exemple [GILBERT, 2003] 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* [THOM, 1974] 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 différentiable, 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 usuellement : 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 occurrent 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 ciblé 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 [VILLANI, 2017], "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 La notion d'*autopoïèse*, déjà mentionnée ci-dessus, exprime la capacité d'un système à s'auto-reproduire. Une caractérisation basique est une frontière semi-perméable produite par le système et la capacité à reproduire ses composants. Une définition plus générale est proposée par [BOURGINE et STEWART, 2004] : "*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*". La notion de processus dynamique est clé, et pourrait être lié à la théorie 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 [BITT-BOL et LUISI, 2004]. Dans ce cadre par exemple, l'arbotron [JUN et HÜBLER, 2005] est cognitif mais pas autopoïétique. Un exemple de lien entre autopoïèse et morphogenèse est montré dans [NIIZATO, SHIRAKAWA et GUNJI, 2010], 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, au 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 [CHAPMAN et MARCULIS, 1998]. 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 **C (JR)** : en lien avec autopoièse et co-evol (signals and boundaries), et remarque de Bonin a TheoQuant : importance de systeme ouvert/ferme - def du sys-teme

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 **C (FL) : pour quoi penserait-on cela?**, 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 occurront, 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* [HOLLAND, 2012]). 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 [COTTINEAU et al., 2015]).
- **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 intrications 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 :

$$\text{Self-organization} \supsetneq \text{Morphogenesis} \supsetneq \text{Autopoiesis} \supsetneq \text{Life} \quad (8)$$

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 [BEDAU, 2002].

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*⁴.

Définition : Autopoïèse et Vie. Nous prenons la définition de BOURGINE pour l'autopoïèse [BOURGINE et STEWART, 2004], qui étend celle de BITBOL [BITBOL et LUISI, 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* [DOURSAT, SAYAMA et MICHEL, 2013], généralement émergente de manière *bottom-up*. 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 [BEDAU, 2002] diminue avec la profondeur, au sens que le nombre d'échelles autonomes, ainsi que le nombre de propriétés aux pouvoir causaux irréductibles, augmentent.

⁴ Au sens donné par MONOD dans [MONOD, 1970], 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 BEDAU 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 [MORIN, 1976].

- Le nombre de bifurcations augmente [THOM, 1974], 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 [LIVET et al., 2010b], 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.

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

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. [CHAVALARIAS et COINTET, 2013] par exemple reconstruit 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 [RAIMBAULT, 2015a].

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'autres modèles qui auraient été difficiles à formuler autrement.

★ ★

★

5.2 MORPHOGENÈSE URBAINE PAR AGRÉGATION-DIFFUSION

Nous étudions un modèle stochastique de croissance urbain générant des distribution spatiales de densité de population à une échelle intermédiaire mesoscopique. Le modèle se base sur le jeu antagoniste entre les deux processus abstrait opposé 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

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 [SETO, GÜNERALP et HUTYRA, 2012] 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 [BATTY, 2013b], 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. [GABAIX, 1999] 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 cependant été démontré empiriquement que des déviations systématiques à cette loi existent [ROZENFELD et al., 2008], et que les interactions spatiales pourraient en être responsables. Les mo-

dèles intégrant les interactions spatiales incluent par exemple [BRETAGNOLLE et al., 2000] 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, [FAVARO et PUMAIN, 2011] 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 [COTTINEAU, 2014] pour le cas de l'ex-Union Soviétique, pour lequel la performance du modèle est démontrée améliorée par rapport aux modèles sans interactions.

A de plus petites 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. Par exemple, [ANDERSSON et al., 2002] 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 [MAKSE et al., 1998], qui développe le modèle introduit dans [MAKSE, HAVLIN et STANLEY, 1995], pour moduler les motifs de croissance pour qu'ils ressemblent à des configurations réelles. Le monde des modèles de croissance urbaine à automates cellulaires (CA) [BATTY et XIE, 1994] offre aussi de nombreux exemples. [XIE, 1996] 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 [WARD, MURRAY et PHINN, 2000]. Le modèle Sleuth, introduit initialement par [CLARKE et GAYDOS, 1998] pour la zone de la Baie de San Francisco, et pour lequel un aperçu des diverses applications est donné dans [CLARKE et al., 2007], a été calibré sur des régions tout autour du monde, fournissant des mesures comparatives au travers des paramètres calibrés.

Assez proches des modèles CA, mais pas exactement similaires, sont les modèles de Morphogenèse Urbaine, qui visent à simuler la croissance de la forme urbaine à partir de règles autonomes. [FRANKHAUSER, 1998] 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. [COURTAT, GLOAGUEN et DOUADY, 2011] 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, [RAIMBAULT, BANOS et DOURSAT, 2014] couple un CA avec un réseau évolutif pour repro-

duire 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 [BATTY, 1991], 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 [BOSCH, METZ et DIEKMANN, 1990].

Nous étudions dans cette section un modèle de morphogenèse, à l'échelle mesoscopique, dont le but est d'être simple dans ses règles et variables, mais visant à être performant pour la reproduction de motifs existants. La question sous-jacente est l'exploration de la performance de mécanismes simples pour reproduire des formes urbaines complexes. Nous considérons des processus abstrait, précisément l'agrégation et la diffusion, comme candidats 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. **C (FL) : revue de litt pas assez hierarchisée**

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 [BATTY, 2006], comme une variation

cellulaire du modèle DLA [BATTY, 1991]. En effet, la tension entre les mécanismes antagonistes d'agrégation et d'étalement peut être un processus important pour la morphogenèse urbaine. Par exemple, [FUJITA et THISSE, 1996] 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. 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, puisqu'il a été montré que le modèle de Simon qui se base dessus génère des *power-law* qui sont typiques des systèmes urbains (lois d'échelles par exemple) [SHERIDAN DODDS et al., 2016]. 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 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 que 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 6, 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.

TABLE 6 : Résumé des paramètres

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}

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.

Données réelles

Nous travaillons sur les valeur 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 [WILENSKY, 1999] pour des raisons d'exploration et de visualisation, et en Scala pour des raisons de performance et d'intégration plus aisée dans OpenMole [REUILLO, LECLAIRE et REY-COYREHOURCQ, 2013], 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 [HJMANS, 2015]. Le code source et les résultats sont disponibles sur le dépôt ouvert du projet⁵. Les données des valeurs réelles des indicateurs et des résultats de simulation sont disponibles sur Dataverse⁶. Nous avons dans le cadre de l'implémentation Scala implémenté la convolution de distribution en deux dimension par Transformée de Fourier rapide, permettant de transformer une complexité $O(N^4)$ en $O(N^2 \log^2 N)$, puis implémenté les indicateurs qui ont pu être intégrés à une extension NetLogo dédiée E.1.4.

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. 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 6, pour une taille de monde $N = 100$. Fig. 29 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.

⁵ à <https://github.com/JusteRaimbault/Density>

⁶ à <http://dx.doi.org/10.7910/DVN/WSUSBA>

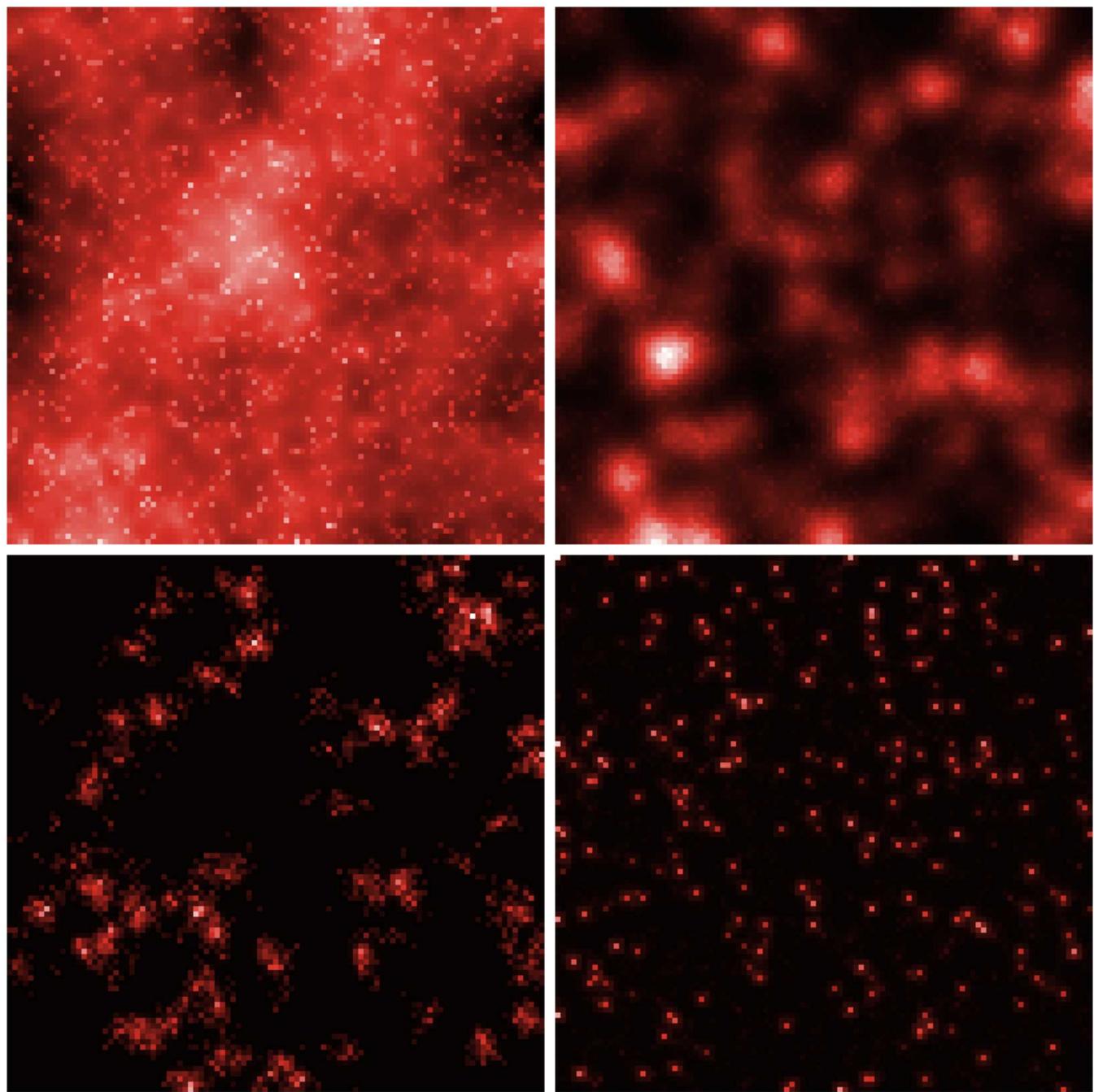


FIGURE 29 : Exemple de la variété de *formes territoriales* générées. (Haut Gauche) Configuration urbaine très diffuse, $\alpha = 0.4, \beta = 0.05, n_d = 2, N_G = 76, P_m = 75620$; (Haut Droite) Configuration polycentrique urbaine semi-stationnaire, $\alpha = 1.4, \beta = 0.047, n_d = 2, N_G = 274, P_m = 53977$; (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, $\alpha = 1.6, \beta = 0.006, n_d = 1, N_G = 268, P_m = 76376$.

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 [BANOS, 2013]. Ce type d'approche est typique de ce qu'ARTHUR nomme le *tournant computationnel dans la science moderne* [ARTHUR, 2015] : 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.7. Les indicateurs présentent de bonnes propriétés de convergence : la plupart des indicateurs sont aisément discernable de manière statistique entre les points de paramètres, et ceux-ci sont distinguables sans ambiguïté quand tous les indicateurs sont pris en compte. Nous utilisons cette expérience pour établir un nombre raisonnable de répétitions nécessaire 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'index 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 étant statistiquement significantes 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 variant 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.

Des graphes complets du comportement du modèle en fonction des paramètres sont donnés en A.7. Nous montrons en Fig. 30 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 configuration plus étalée 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 : des processus similaires ont par exemple été utilisés par TURING dans son article séminal sur la morphogenèse [TURING, 1952]. Une autre façon de formuler le modèle typique à ces approches est d'utiliser des Equations aux Dérivées Partielles (PDE). 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 $P(x, t)$. **C (FL) : notations ne sont pas claires** Nous travaillons sur les espérances $p(x, t) = \mathbb{E}[P(x, t)]$, et supposons que $n_d = 1$. Comme développé en

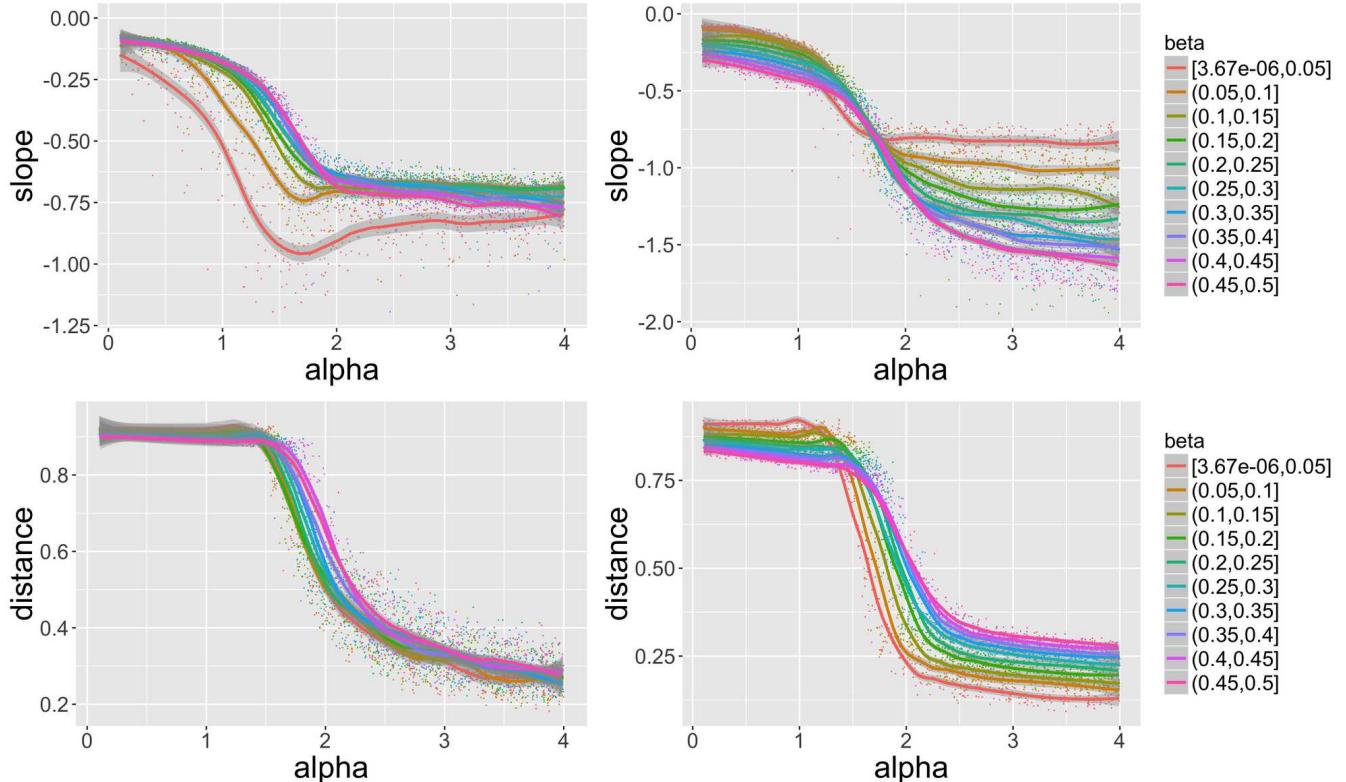


FIGURE 30 : **Comportement des indicateurs.** Pente γ (ligne du haut) et distance moyenne \bar{d} (ligne du bas) comme fonction de α , pour différentes valeurs de β données par la couleur des courbes, pour des valeurs particulières $n_d = 1, P_m/N_G \in [13, 26]$ (colonne de gauche) et $n_d = 4, P_m/N_G \in [41, 78]$ (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.

Information Supplémentaire A.7, on peut montrer que ce processus simplifié obéit à la PDE 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 [TADMOR, 2012]. Il est important de noter que le modèle simplifié peut être exprimé comme une PDE analogue aux équations de réaction-diffusion, comme celle partiellement résolue pour un modèle plus simple dans [BOSCH, METZ et DIEKMANN, 1990]. Nous montrons en A.7 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 obtenue 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. 31, 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, et confirme l'importance d'indicateurs morphologiques robustes décrits précédemment.

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

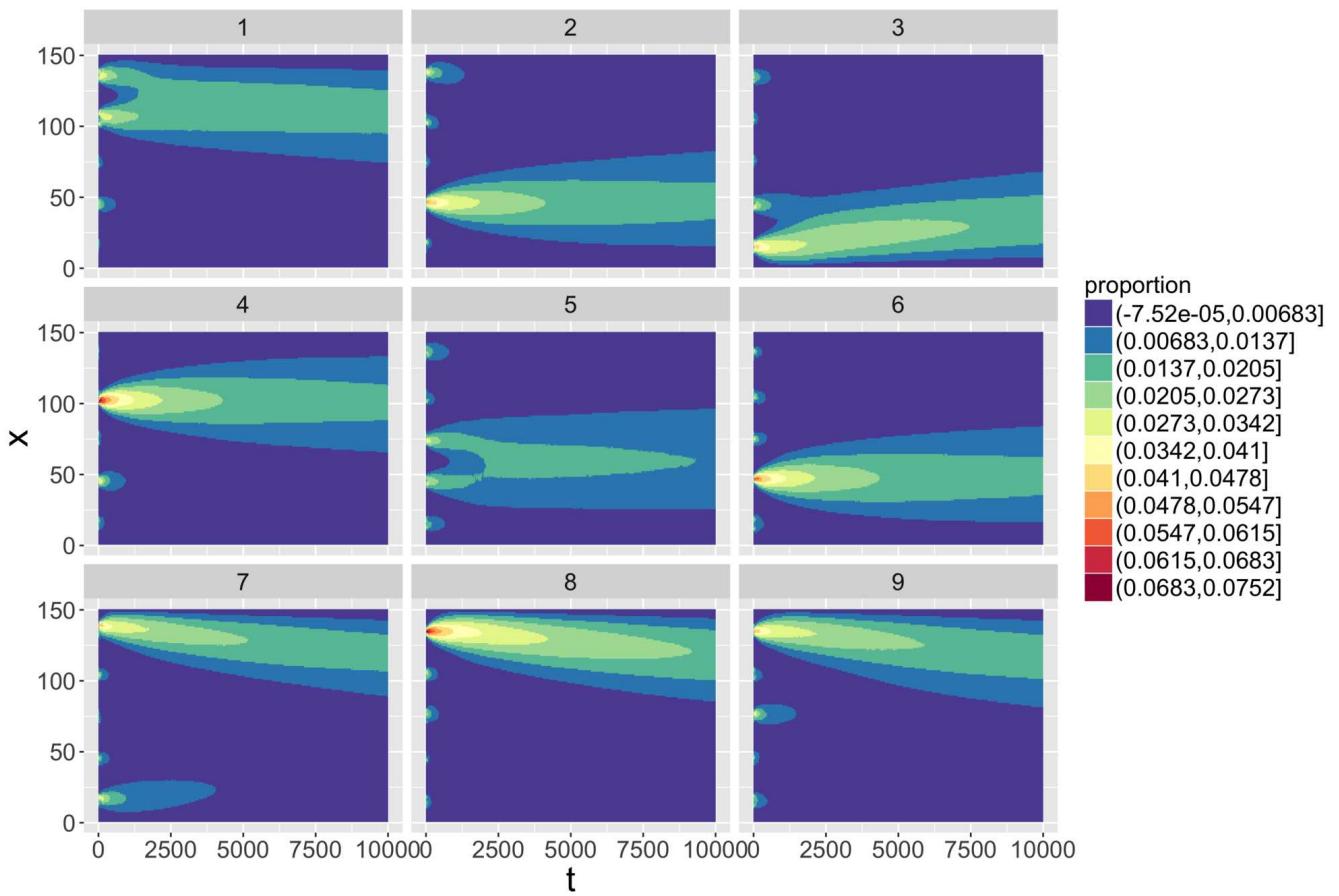


FIGURE 31 : Aléatoire et accidents figés. Nous montrons 9 réalisations aléatoires du système à une dimension avec des conditions initiales identiques, c'est à dire 5 cellules équidistantes peuplées également à l'instant initial. Les paramètres sont $\alpha = 1.4$, $\beta = 0.1$, $N_G = 10$. Chaque graphe montre le temps contre l'espace, le niveau de couleur donnant la proportion de population dans chaque cellule.

avec les valeurs réelles des indicateurs. Les scatterplots complets de chaque indicateur contre les autres, pour les configurations simulées et les réelles, sont donnés en A.7. 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 morphologique. Les plus grands écarts est 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 seule-

ment. Celles-ci représentent 85% de la variance cumulée. Les nuages de points projeté sur ces dimensions est montré en Fig. ???. 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 [PUMAIN et al., 2006]) 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, mais aussi que la fonction doit être relativement corrélée à la forme puisque la dimension fonctionnelle (avec une dimension économique supplémentaire dans la forme par exemple) n'est pas prise en compte dans le modèle.

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 prometteurs. Plus précisément, l'algorithme PSE [CHÉREL, COTTINEAU et REUILLOU, 2015] 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. 33 : 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 des indicateurs, celui-ci étant témoin de la réalité de structures urbaines même simulées). Cela vou-

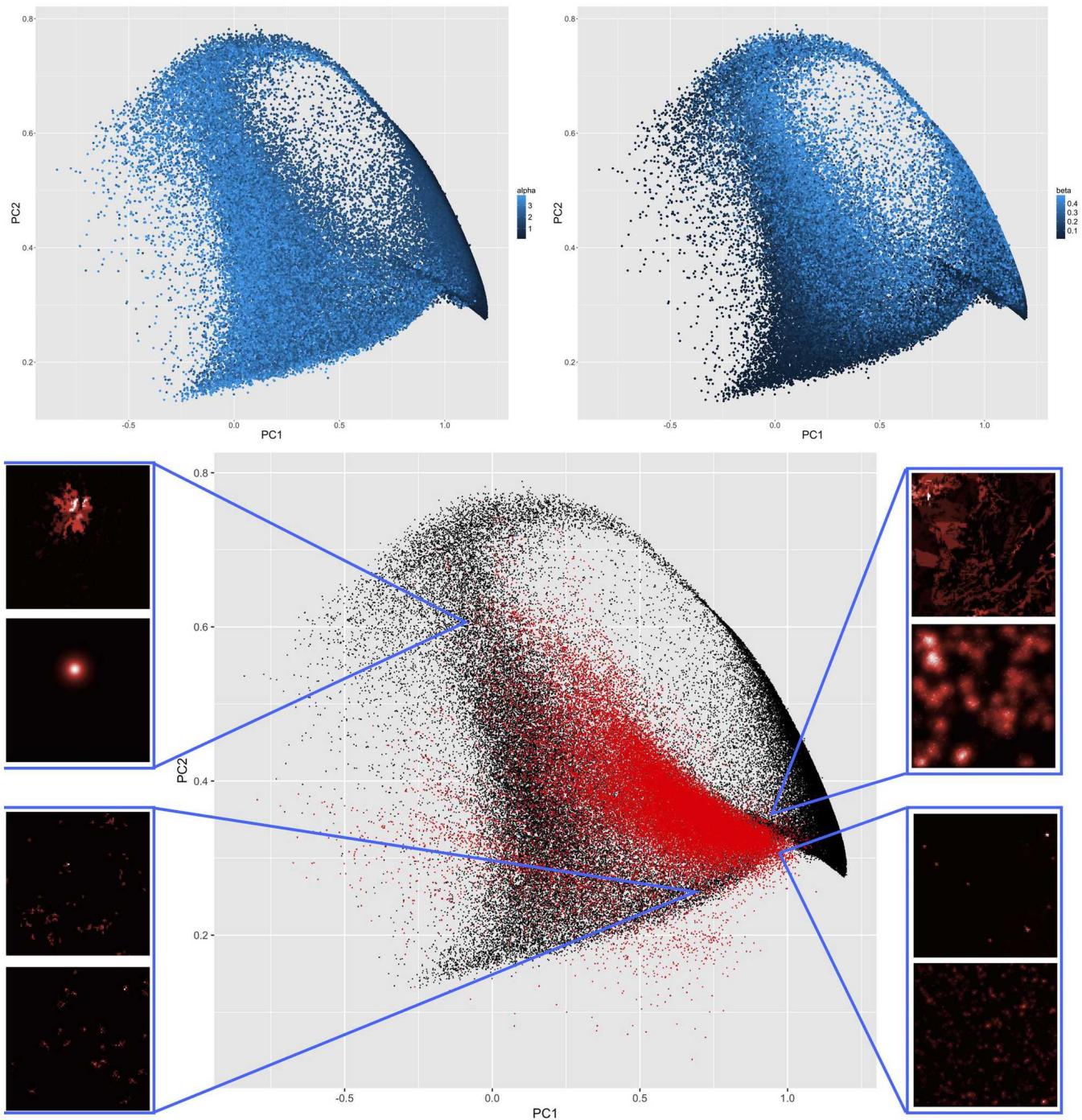


FIGURE 32 : Calibration du modèle. (Haut) Configurations simulées dans le plan des deux premières composantes principales, le niveau de couleur donnant l'influence de α (gauche) et de β (droite); (Bas) Points simulés dans le même espace (en noir) avec les configurations réelles (en rouge). Autour du graphe 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 $\alpha = 1.14, \beta = 0.108, N_G = 637, nd = 1, P_m = 13235$.

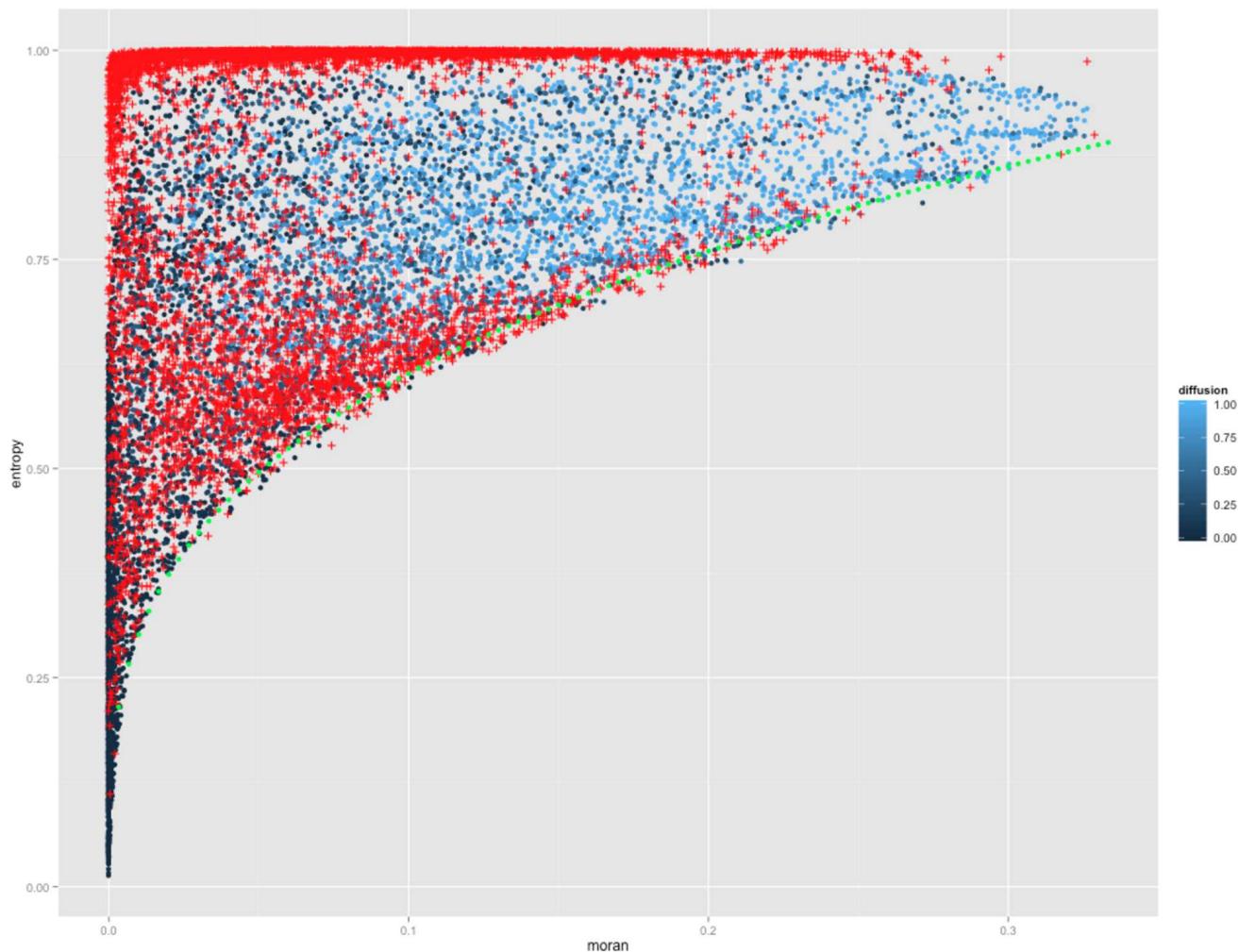


FIGURE 33 : Exploration par PSE. Scatterplot de l'entropie en fonction de Moran, les points bleus étant obtenus par LHS et les rouges par PSE. La ligne pointillée verte donne la borne inférieur faisable.

drait 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'ètre 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'aléatoire 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.

* * *

*

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 population synthétiques au sein de modèles basés agents, comme par exemple des modèles de mobilité, des modèles *LUTI* [PRITCHARD et MILLER, 2009]. On peut également citer des méthodes d'analyse spatiales 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, comme pour la Regression Géographique Pondérée [BRUNSDON, FOTHERINGHAM et CHARLTON, 1998], dans laquelle les noyaux de taille variables n'interpolent pas des données au sens propre mais extrapolent des variables abstraites représentant l'interaction entre variables explicites). 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 [SANDERS et al., 1997], pionnier d'une famille de modèles par la suite paramétrisés par des données réelles, 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 [SCHMITT, 2014], alors qu'il s'agit 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 [COTTINEAU et al., 2015b], 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 particu-

liers 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.

5.3.2 Modèle et Resultats

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 corrélations 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 [OFFNER et PUMAIN, 1996], mais toujours majoritairement ouvert, extrêmement complexe et difficile à quantifier [OFFNER, 1993]. Une modélisation dynamique des processus impliqués devrait apporter des connaissances sur ces interactions ([BRETAGNOLLE, 2009a], 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 corrélations selon la variation des paramètres des modèles restent à déterminer par l'exploration du modèle couplé.

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

$$V_{ij}(d) = \left[(1 - k_h) + k_h \cdot \left(\frac{P_i P_j}{P^2} \right)^\gamma \right] \cdot \exp \left(-\frac{d}{r_g(1 + d/d_0)} \right)$$

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 population dans le potentiel, γ régissant la forme de la hiérarchie selon les valeurs des populations, r_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 éventuelles créées par les nouveaux liens.

Notons que la construction du modèle de génération est heuristique, et que d'autres types de modèles comme un réseau biologique auto-généré [TERO et al., 2010], une génération par optimisation locale de contraintes géométriques [BARTHÉLEMY et FLAMMINI, 2008] 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 [COTTINEAU, CHAPRON et REUILLO, 2015], 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 couplé⁷ 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 $\vec{\alpha}_N = (N_C, k_h, \gamma, r_g, d_0)$. 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 $\vec{M} = (r, \bar{d}, \varepsilon, \alpha)$ donnant auto-corrélation spatiale (indice de Moran), distance moyenne, entropie, hiérarchie (voir [LE NÉCHET, 2015] 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.
- 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 $d_m = \sqrt{2}N$.

- Vitesse moyenne [BANOS et GENRE-GRANDPIERRE, 2012], 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 $\delta = \max_{i,j} d_N(i, j)$

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 35 comme

⁷ 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.

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 s'intéressera à 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 prend comme correlation associée la corrélation de Pearson estimée de la même façon.

Implémentation

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 [REUILLO, LECLAIRE et REY-COYREHOURCQ, 2013] 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 implémentée en langage basé-agent NetLogo [WILENSKY, 1999], ce qui facilite l'exploration interactive et construction heuristique interactive. Le code source est disponible pour reproductibilité sur le dépôt du projet⁸.

Résultats

L'étude du modèle de densité seul est développée dans la section précédente. Pour rappel, il est notamment calibré sur les données de 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 objectif par une Analyse en Composantes Principales (variance expliquée à deux composantes $\simeq 85\%$) 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.

A densité donnée, l'exploration de l'espace des paramètres du modèle de réseau suggèrent une assez bonne flexibilité sur des indicateurs globaux \vec{G} , ainsi que de bonnes propriétés de convergence. Pour

⁸ à l'adresse <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Synthetic>

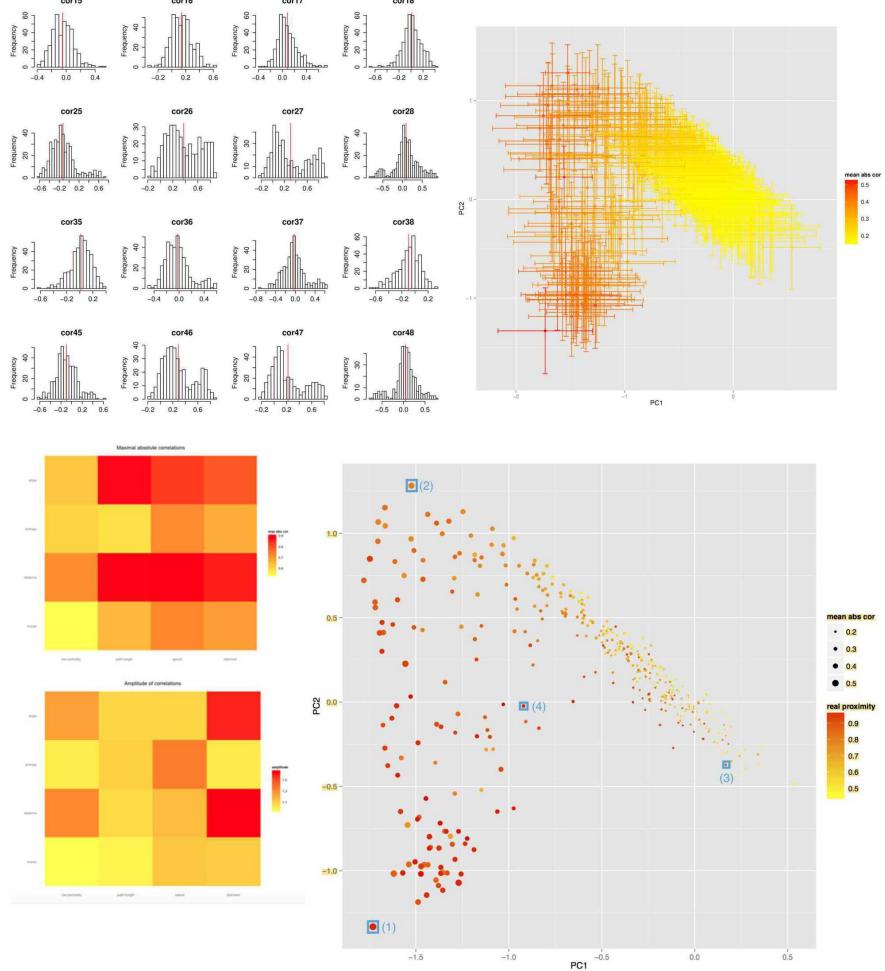


FIGURE 34 : Exploration de l'espace faisable des corrélations entre la morphologie urbaine et la structure du réseau. (Haut Gauche) Distribution des corrélations croisées entre les vecteurs \vec{M} des indicateurs morphologiques (dans l'ordre de numérotation Moran, distance, entropie et hiérarchie) et \vec{N} des mesures de réseau (centralité, longueur moyenne, vitesse, diamètre); (Haut Droite) Projection des matrices de correlations 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; (Bas Gauche) Amplitude des correlations, 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; (Bas Droite) Représentation dans le plan principal; l'échelle de couleur donnant la proximité aux données réelles définie par $1 - \min_r \|\vec{M} - \vec{M}_r\|$ où \vec{M}_r est l'ensemble des mesures morphologiques réelles; la taille des points donne la corrélation absolue moyenne.

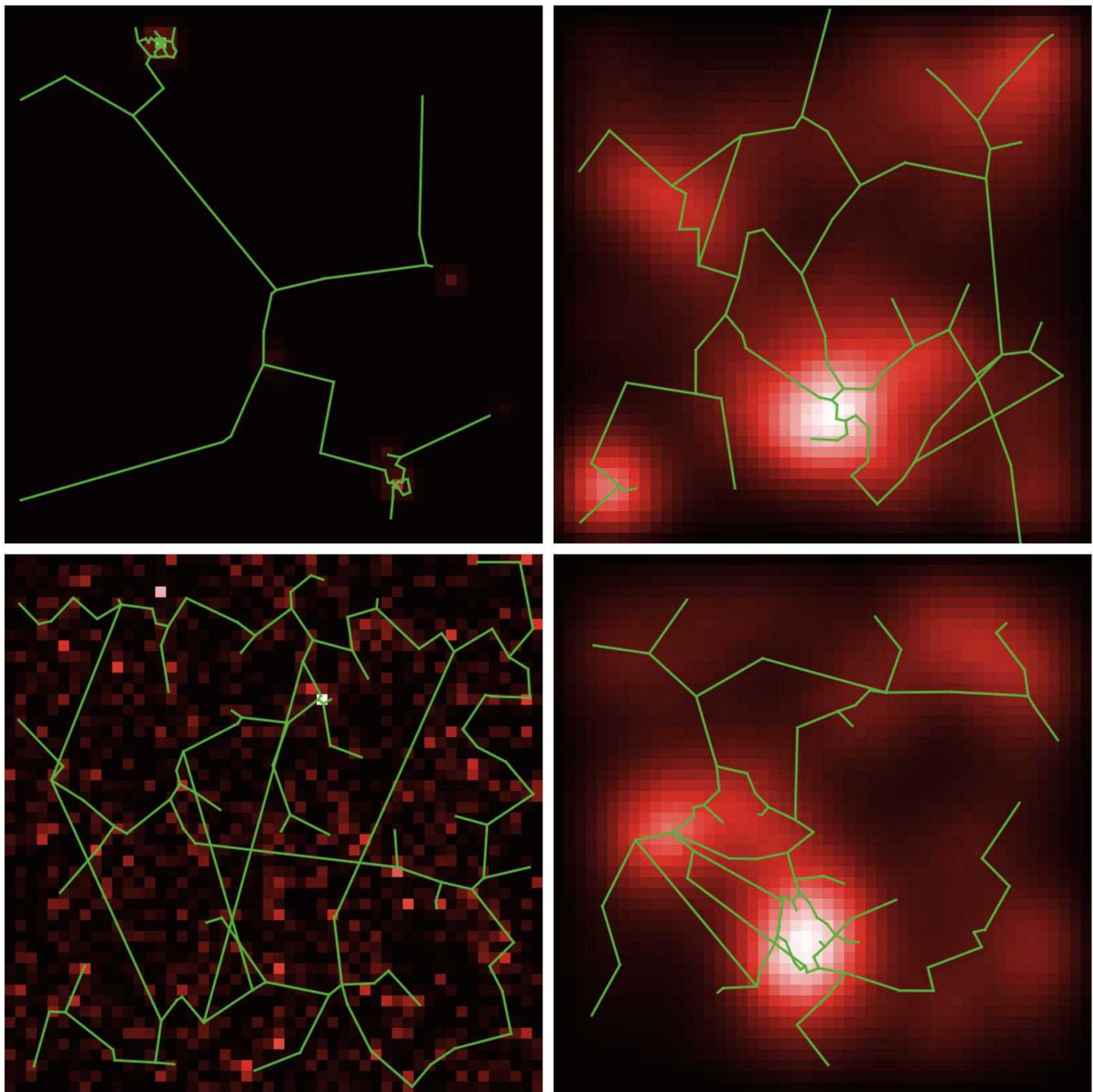


FIGURE 35 : Configurations obtenues pour les paramètres donnant les quatre points mis en évidence en 34 (d), dans l'ordre de gauche à droite et de haut en bas. Nous retrouvons des configurations de villes polycentriques (2 et 4), des établissements ruraux diffus (3) et une zone de densité agrégée faible (1). Se reporter à l'appendice A.8 pour les valuers exhaustives des paramètres, indicateurs, et corrélations correspondantes. Par exemple \bar{d} est fortement corrélé à \bar{l}, \bar{s} (≈ 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 : $\rho[\bar{d}, \bar{c}] \approx 0.34$ pour (4) mais ≈ -0.41 pour (2), ce qui est expliqué par un rôle plus fort de la hiérarchie de gravité dans (2) $\gamma = 3.9, k_h = 0.7$ (pour (4), $\gamma = 1.07, k_h = 0.25$), tandis que les paramètres de densité sont similaires.

une étude du comportement précis, voir l'appendice donnant les régressions traduisant le comportement du modèle couplé. Dans le but d'illustrer la méthode de génération de données synthétiques, l'exploration a été orientée vers l'étude des corrélations.

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 $N_C \in [50, 120]$, $r_g \in [1, 100]$, $d_0 \in [0.1, 10]$, $k_h \in [0, 1]$, $\gamma \in [0.1, 4]$, $N_L \in [4, 20]$. 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 corrélations, 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 34 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 corrélations 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 corrélation pour l'ensemble des indicateurs, la corrélation 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 corrélations) 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 montrent que des configurations de densité proches peuvent présenter des profils de corrélations très différents.

Le comportement statistique des indicateurs et des corrélations est donné en Appendice A.8.

5.3.3 Discussion

Positionnement Scientifique

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 statistiques 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 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 [CHÉREL, COTTINEAU et REUILLO, 2015].

Notre démarche s'inscrit dans un cadre épistémologique particulier. En effet, d'une part la volonté de multi-disciplinarité et d'autre part l'importance de la composante empirique couplée aux méthodes d'exploration computationnelles, en font une approche typique des sciences de la complexité, comme le rappelle la structure de la feuille de route pour les systèmes complexes [BOURGINE, CHAVALARIAS et AL., 2009] qui croise des grandes questions transversales aux disciplines à une intégration verticale de celles-ci, qui implique la construction de modèles multi-échelles hétérogènes présentant souvent les aspects précédent. Le croisement de connaissances empiriques issues de la fouille de données avec celles issues de la simulation est souvent central dans leur conception ou leur exploration, et les résultats présentés ici en sont un exemple typique pour le cas de l'exploration.

Applications Directes

En partant du deuxième exemple, qui s'est arrêté à la génération des données synthétiques, on peut proposer 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 réelle de réseau de transport (typiquement routier vu les formes heuristiques obtenues, il devrait par exemple être aisé d'utiliser les données ouvertes d'OpenStreetMap qui sont de qualité raisonnable pour l'Europe, du moins pour la France [GIRRES et TOUYA, 2010] 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 du 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 [PUMAIN, 2012b], 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 *corrélation 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 [COTTINEAU et al., 2015b]. 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 corrélations 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 [PIGOZZI, 1980], étendant le domaine d’étude vers une meilleure compréhension des corrélations dynamiques.

Généralisation

On s'est limité au contrôle des premiers 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 au nombre de paramètres relativement faibles. Par contre, l'étude de structures de dépendances non-linéaires comme celles utilisées dans [CHICHEPORTICHE et BOUCHAUD, 2013] 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 [BATTY, 2017] ? Si on se place à l'échelle d'un système de ville ou d'une mégarégion 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 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 à des activités 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.

* * *

*

Troisième partie

SYNTHÈSE : MODÈLES DE CO-ÉVOLUTION

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

C : à ce stade, expliquer lien entre les différents modèles : utiliser appendix unified framework urban growth

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 : 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 coevolution connaissances.

C : justifier ici pourquoi pas modèle très fin sur processus eco par exemple (// Levinson) : prix à payer pour être accross scales, disciplines et avoir vraiment de la coevolution ? pour ces premières étapes oui. à justifier

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 échelles⁹ : 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, C : (Florent) expliciter la différence avec ce que tu as fait jusque là 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

⁹ 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 7 :

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.

6

CO-ÉVOLUTION À L'ECHELLE MACROSCOPIQUE

Les dynamiques des systèmes territoriaux à l'échelle macroscopique peuvent être partiellement comprises par une approche par les interactions, comme montré en 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 des modèles économiques plus classiques et concerne d'autre types de processus, bases 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 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-il conduire à une modification des trajectoires économiques puis par conséquent à une modification des motifs de transports par une retroaction 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 régionale ou un choix de gouvernance plus difficile à cerner, et si-ils sont réalisés malgré le contexte politique plus difficile, 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 par la modélisation, en introduisant les processus de co-evolution dans les modèles d'interactions déjà développés. Pour donner une idée de la nature des conclusions qu'il est possible de tirer d'une telle approche, nous commençons en 6.1 par une exploration systématique du modèle SimpopNet, état de l'art en modélisation de la co-evolution au sein des systèmes de villes, comme établi en 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-evolution, qui est testé sur données synthétiques à deux niveaux de genericité pour le réseau, puis sur

le système de ville français de manière à pouvoir comparer avec les modèles statiques précédemment étudiés.

★ ★

★

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 étudiant les résultats produits par des modèles existants, ce qui permettra également d'introduire les méthodes et indicateurs d'exploration, ainsi que d'appréhender les questionnements typiques liés à ce type de modèles. En particulier, nous procémons à une exploration systématique du modèle SimpopNet [SCHMITT, 2014], initiative unique pour modéliser la co-evolution au sein d'un système de villes.

6.1.1 Contexte

Quelle différentiel 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 [SCHMITT, 2014], 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 recouplant. 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 $\mu_i^\gamma / \sum_j \mu_j^\gamma$) 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 $v(t)$, qui dépendra des technologies de transport courantes. La planarisation 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).

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 suggère 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 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 limite. L'introduction de l'effet tunnel dans le modèle, via les valeurs variables de $v(t)$ et le mécanisme de non-branchement, est exogène puisque spécifié dans les règles du modèle, contrairement au modèle d'interaction avec retroaction 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 contentons de regarder par exemple la hiérarchie des centralités qui en est un bon proxy.

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 métaparamètres spatiaux.

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.4 pour la notion de données synthétiques).

tiques, calibrées au premier et second ordre). Une nombre fixée 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 ajusté (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 voisins, 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 réseau 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.

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, 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.

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 traduiraient par inversions de rang, ainsi que sa capacité à capturer différents aspect des dynamiques co-évolutives, nous proposons un jeu d'indicateurs, incluant par exemple des mesures de correlation retardée en echo aux régimes de causalité exhibés en 4.2, ou une mesure de correlation 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 basiques : hiérarchie (pente du rang-taille) $\alpha(t)$, entropie $\varepsilon(t)$, statistiques descriptives $\mathbb{E}[\hat{X}](t)$, $\hat{\sigma}(t)$, 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, qui capture la diversité de forme des séries temporelles, avec $\tilde{X}_i(t) \in [0;1]$ les trajectoires mises à l'échelle individuellement,

$$\frac{2}{N \cdot (N-1)} \sum_{i < j} \left(\frac{1}{T} \int_t (\tilde{X}_i(t) - \tilde{X}_j(t))^2 \right)^{\frac{1}{2}}$$

- Complexité moyenne des trajectoires, la "complexité" d'une trajectoire étant prise simplement dans notre cas par son nombre de points d'inflexion
- 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.

$$\hat{\rho}_d [(X(\vec{x}_1, Y(\vec{x}_2)) | \| \vec{x}_1 - \vec{x}_2 \| \sim d]$$

- 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 sont à lire dans le sens des régimes potentiels, explorés en 4.2.

•

$$\hat{\rho}_\tau [\Delta X(t), \Delta Y(t-\tau)]$$

Ces indicateurs sont utilisés pour les populations $\mu_i(t)$, les centralités de proximité $c_i = \frac{1}{N-1} \sum_{i \neq j} \frac{1}{d_{ij}}$, et les accessibilités $Z_i = \frac{1}{\sum_k \mu_k} \sum_{i \neq j} P_j \exp(-d_{ij}/d_G)$.

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.

6.1.3 Résultats

Plan d'expérience

Etant donné une configuration spatiale initiale (c'est à dire une valeur des meta-paramètres), nous établissons les diagrammes de phase par l'exploration d'une grille de l'espace des paramètres. Le nombre

de paramètres étant restreint et les attentes étant d'obtenir 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 et paramètres, et 30 réplications aléatoires, ce qui correspond à 155,520 simulations.

Convergence

Pour quantifier la variabilité d'un indicateur X par rapport à la stochasticité, nous utilisons une mesure de type ratio de Sharpe, donnée par $\nu[X] = \mathbb{E}[\hat{X}] / \sigma[\hat{X}]$ avec les estimateurs basiques pour l'espérance et la deviation standard. Sur l'ensemble des exécutions, on obtient sur l'ensemble des indicateurs donnés précédemment, une médiane estimée sur l'ensemble des valeurs des paramètres, pour le ratio estimé au sein des réplications, 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.

Sensibilité à l'espace

La table 8 donne les valeurs de $\tilde{\alpha}$ pour 16 configurations des méta-paramètres, 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.

Motifs

La Fig. 36 rend compte du comportement du modèle selon les divers indicateurs donnés ci-dessus. Nous commentons une configuration spatiale particulière, donnée par les méta-paramètres $N_S = 80, \alpha_S = 0.5, d_S = 10, n_S = 30$. Les valeurs prises par l'entropie pour les centralités, en fonction du temps, pour $\gamma_N = 2.5$ et $v_0 = 110$, exhibent diffé-

TABLE 8 : Sensibilité à l'espace du modèle SimpopNet. Chaque colonne correspond à une instance du diagramme de phase, pour laquelle les meta-paramètres sont donnés, ainsi que la distance relative à un diagramme de référence arbitraire.

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	0.5	1.5	1.5	1.5	1.5
d_S	5	5	10	10	5	5	10	10	5	5	10	10	5	5	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

rents 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. 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, $\rho_d [d]$ a un comportement décroissant, 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 causalité circulaires, et donc pas dans une co-évolution à proprement parler comme nous l'avons défini. Cette exploration brève nous permet d'affirmer que ce modèle capture des trajectoires urbaines d'une certaine 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 est-il capable de capturer des dynamiques co-évolutives.

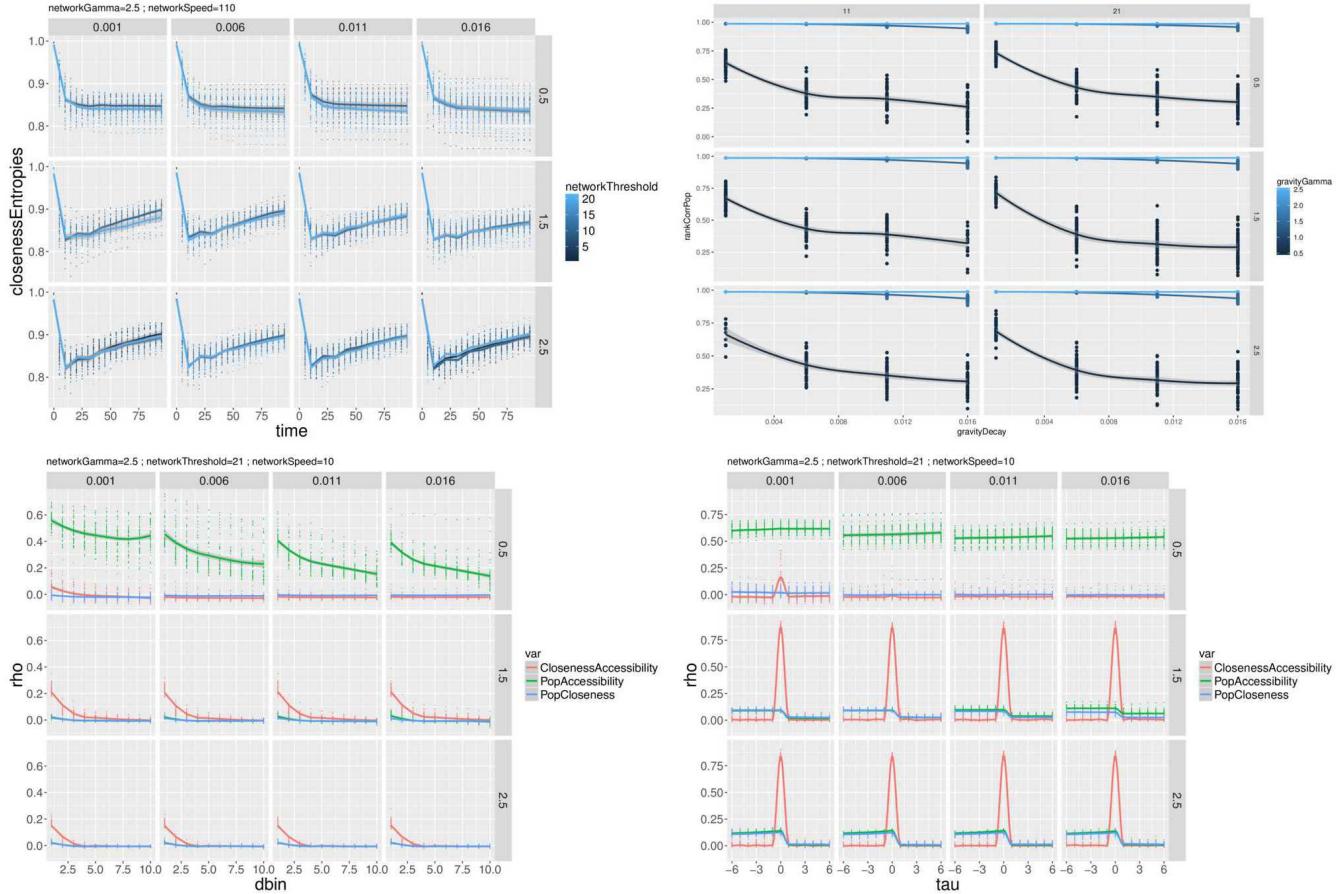


FIGURE 36 : Comportement du modèle. (*Haut Gauche*) Trajectoires temporelles de l'entropie des centralités de proximité; (*Haut Droite*) Correlation de rang pour la population; (*bas Gauche*) Correlations en fonction de la distance, pour $\gamma_S = 2.5$; (*Bas Droite*) Correlations retardées. Pour les interprétations, se référer au texte.

★ ★

★

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.

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 au systèmes de villes. Toujours dans un choix de simplicité, dont la relaxation pourra être explorée pour le cas d'application à la Chine avec l'ajout de variables économiques, nous restons ici à une description unidimensionnelle des villes par leur population. Concernant la croissance du réseau, nous proposons de se placer également à un niveau relativement agrégé et simplifié, en testant des heuristiques de croissance répondant à une demande, à différents niveaux d'abstraction. Par une forme de multi-modélisation, 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 induite par la demande pour le réseau.

Formulation Générique

Le système urbain est caractérisé par les populations $\mu_i(t)$ et le réseau /mathbf{N}(t) auquel on peut associer une matrice de distance $d_{ij}^N(t)$

CROISSANCE DU RÉSEAU Network growth heuristics are tested at different abstraction levels that are the time-distance matrix between cities, and physical network growth trying to satisfy greedy time-gain optimization criteria.

Given the flow ϕ in a link, its effective distance is updated following

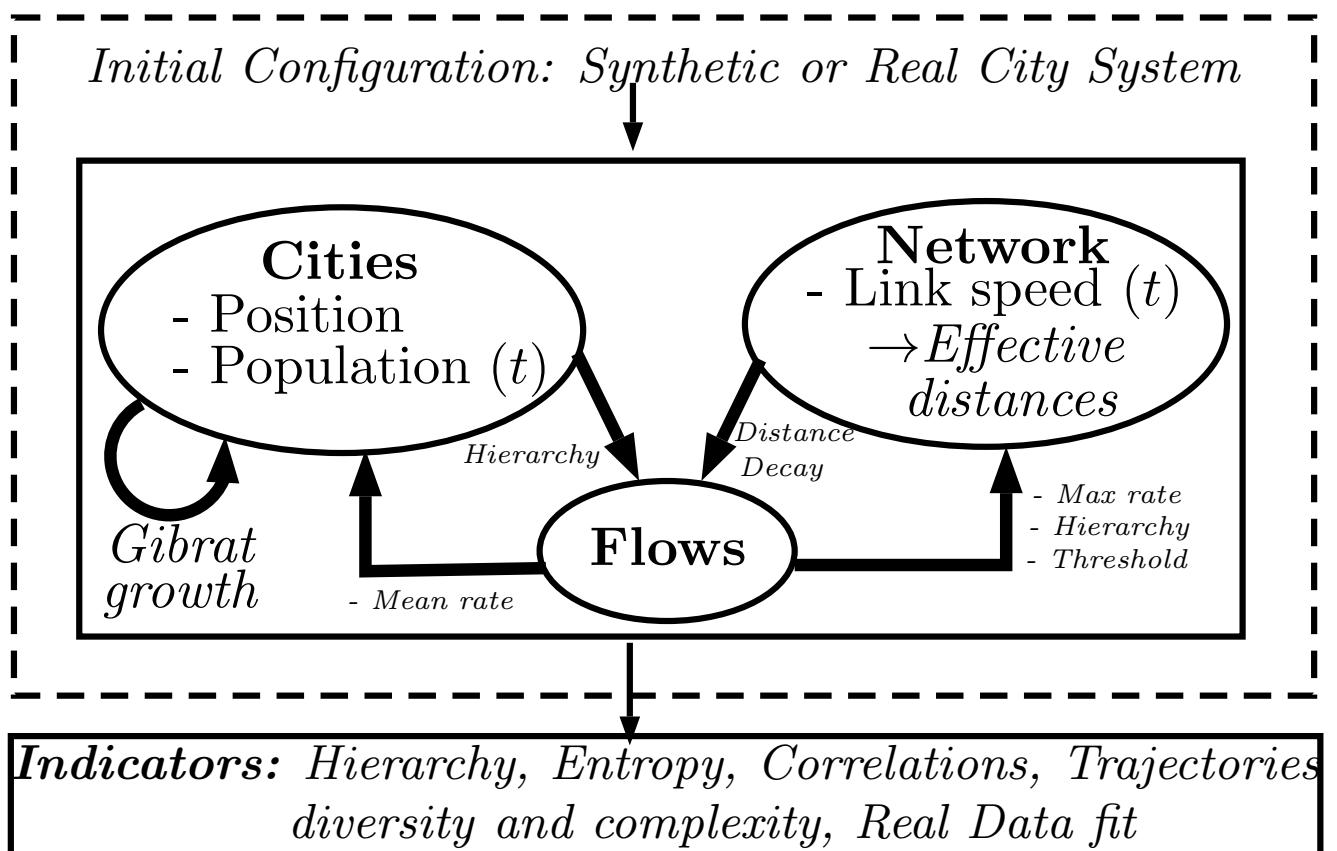


FIGURE 37 : Représentation abstraite des processus du modèle.

For the thresholded case

$$d(t+1) = d(t) \cdot \left(1 + g_{\max} \cdot \left[\frac{1 - \left(\frac{\phi}{\phi_0} \right)^{\gamma_s}}{1 + \left(\frac{\phi}{\phi_0} \right)^{\gamma_s}} \right] \right)$$

where γ_s is a hierarchy parameter, ϕ_0 a threshold parameter and g_{\max} the maximal growth rate easily adjustable to realistic values by computing $(1 + g_{\max})^{t_f}$

Implémentation

Le couplage du modèle d'interaction à la prise en compte plus fine des processus de réseau rend plus difficile l'intégration complète dans un plugin OpenMole comme c'était le cas pour le modèle étudié en 4.3. 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. 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.

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, c'est à dire sans explicitation spatiale du réseau et avec les règles d'évolution agissant directement sur d_{ij} selon les spécifications données précédemment.

DONNÉES SYNTHÉTIQUES Un système de villes synthétiques est généré de la façon la plus simple possible, en suivant l'heuristique utilisée dans la section précédente : (i) des villes en nombre N_S sont placées aléatoirement dans le plan euclidien¹; (ii) les populations sont attribuées aux villes selon une hiérarchie α_S , 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, le nombre de ville est fixé à $N_S = 30$ et la population maximale à $P_{\max} = 100000$.

¹ nous relâchons la possible contrainte de positionnement selon des logiques de places centrales

RÉSULTATS 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. 38, nous montrons l'évolution d'indicateurs dans le temps ainsi que des mesures agrégées, pour une grande partie de l'espace des paramètres couvert. L'évolution de la centralité de proximité moyenne dans le temps, visualisée pour $w_G = 0.001$, et à (d_G, γ_G, ϕ_0) variables, 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 correspond à des situations où l'ensemble des villes bénéficie d'accroissements d'accessibilité. En terme d'entropie des populations, l'ensemble des paramètre donne une entropie décroissante, c'est à dire des convergences des trajectoires dans le temps. 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 , consistant 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. On reconstruit ainsi une non-stationnarité spatiale, typiquement observée en 4.1, et rejoignons le concepts de niche écologique localisée dans l'espace (voir 9.2). Enfin, le comportement des correlations 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.

Tournons nous à présent vers les motifs de corrélation produits par le modèle, illustrés en Fig. 39. 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 ($d_G > 50$) décorrèlent totalement population avec centralité et accessibilité. 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 corrélation entre accessibilité et centralité est plus difficile à interpréter, et peut être dû aux phénomènes d'autocorrelation². Son niveau ne dépend pas de la distance mais de d_G , et est décroissant pour finir à une corrélation négative. Enfin, concer-

² 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 $\epsilon = \sigma_Y/\sigma_X \ll 1$ au premier ordre par $\rho[X + Y; Z] \simeq \rho[X; Z] \left(1 + \frac{1}{2} \rho[X; Y] \epsilon - \frac{\epsilon^2}{2} \right) + \epsilon \rho[Y; Z]$, mais cette hypothèse est trop restrictive pour être valable sur l'ensemble de la somme.

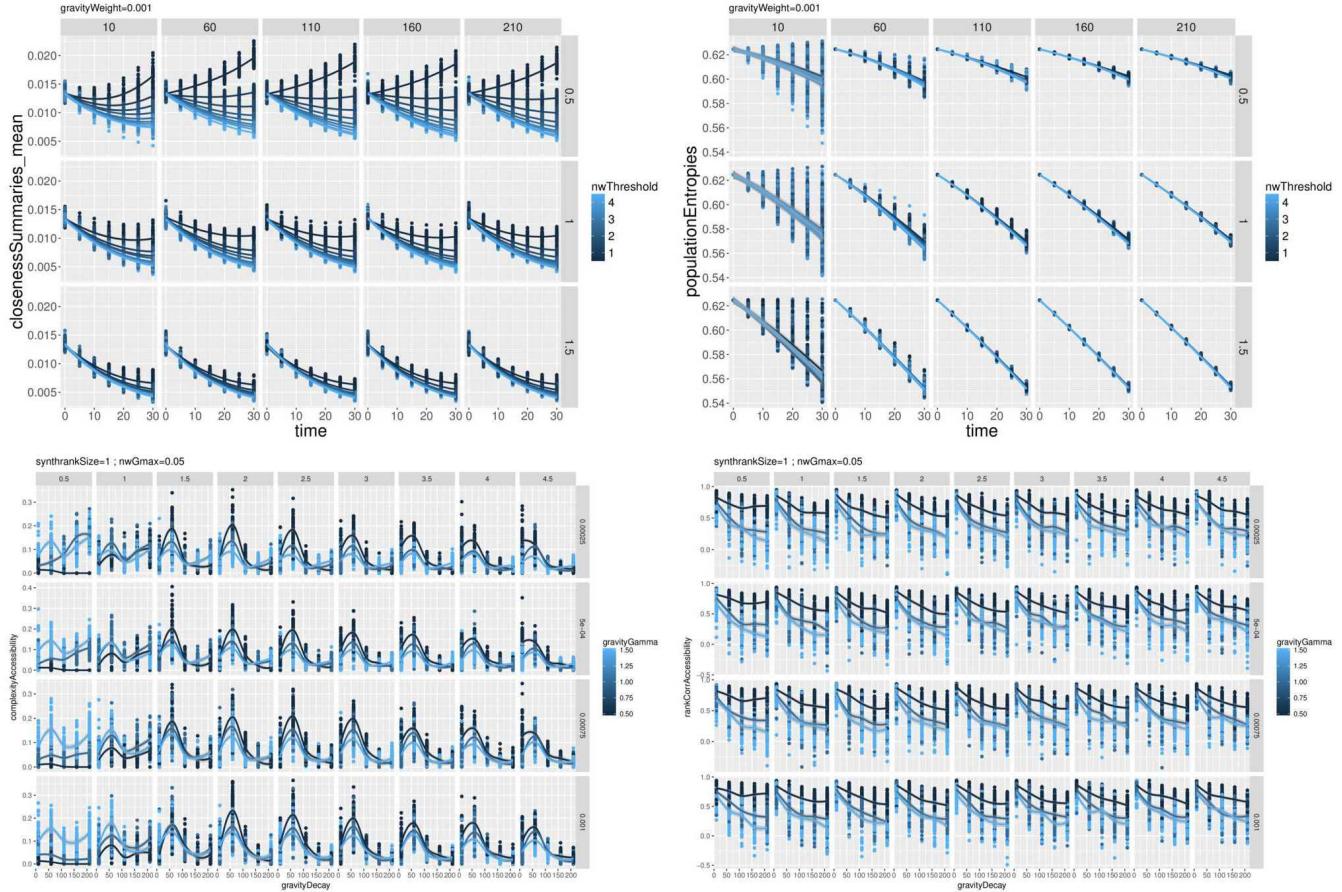


FIGURE 38 : Comportement du modèle de co-evolution avec réseau abstrait sur un système de villes synthétique.
(Haut Gauche) Moyenne des centralités de proximité, en fonction du temps, pour d_G (colonnes), γ_G (lignes) et ϕ_0 (couleur) variables, à $w_G = 0.001$ fixé; *(Haut Droite)* Entropie de populations, en fonction du temps, pour d_G (colonnes), γ_G (lignes) et ϕ_0 (couleur) variables, à $w_G = 0.001$ fixé; *(Bas Gauche)* Complexité des accessibilités, en fonction de d_G , pour ϕ_0 (colonnes), w_G (lignes) et γ_G variables, avec $\alpha_S = 1$ et $g_{\max} = 0.05$; *(Bas Droite)* Corrélations de rang des accessibilités, pour les mêmes paramètres. Se référer au texte pour l'interprétation.

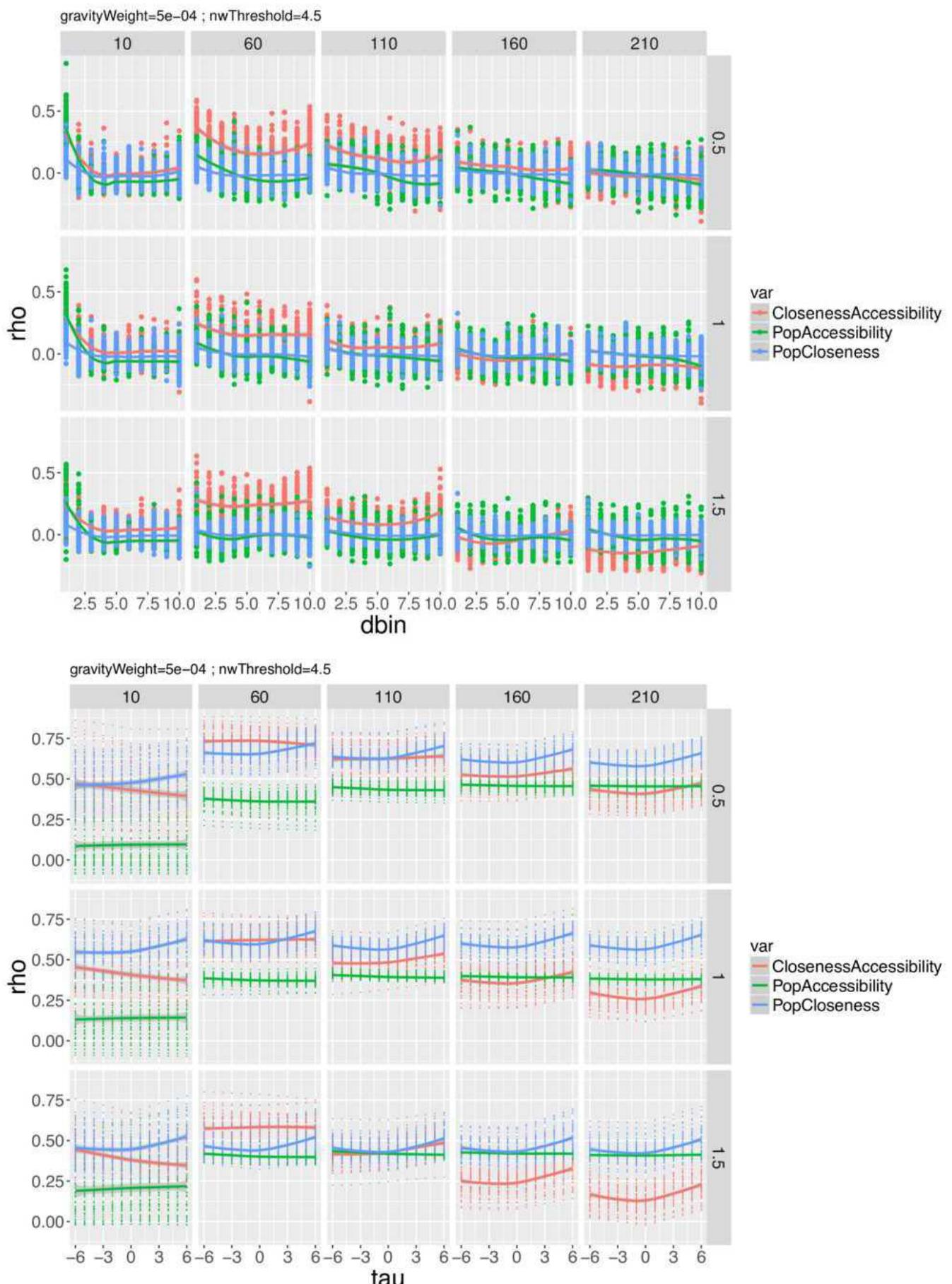


FIGURE 39 : **Motifs de corrélation dans le modèle abstrait.** (Haut) Corrélation ρ_d entre couples de variables (donnée par la couleur), en fonction de la distance d (discretisée en déciles), pour d_G variable (colonnes) et γ_G variable (lignes), à $w_G = 5e-4$ et $\phi_0 = 4.5$; (Bas) Correlations 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 $\phi_0 = 4.5$.

nant les corrélations retardées, on observe 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, 2009a]. 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 hiérarchie³.

SYNTHÈSE Les faits stylisés marquants qui ressortent de l'exploration du modèle synthétique sont les suivants :

1. On démontre 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 régime d'adaptation, un régime de co-évolution direct et un régime de co-évolution indirecte.

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, avec le réseau ferré dynamique de 1840 à 2000. 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 impliqués via les valeurs calibrées des paramètres.

³ Une analyse plus systématique par apprentissage non-supervisé comme en 4.2 et en 7.2 est laissée pour de futures développements.

Données de Réseau

Nous travaillons sur les données de réseau ferré construites par [THÉVENIN, SCHWARTZ et SAPET, 2013]. 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 ce moyen de transport a à toute période concrétisé l'implication d'acteurs publics et privés importants, tout en exhibant différents régimes 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. 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 matrice de distance-temps entre les villes considérées dans le modèle.

Faits stylisés

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.

Questions concrètes à poser au modèle, expériences ciblées :

1. le modèle calibre-t-il mieux les populations (en prenant en compte les paramètres supplémentaires)
2. motifs de calibration biobjectif réseau/populations pour le réseau abstrait

C (JR) : attention, expliquer le choix des indicateurs de réseau, il faut qu'ils soient adaptés à l'échelle : cf Mimeur nombre d'intersection - relève un peu de la modélisation procédurale.

INDICATEURS 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 cumulées, 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 simplement

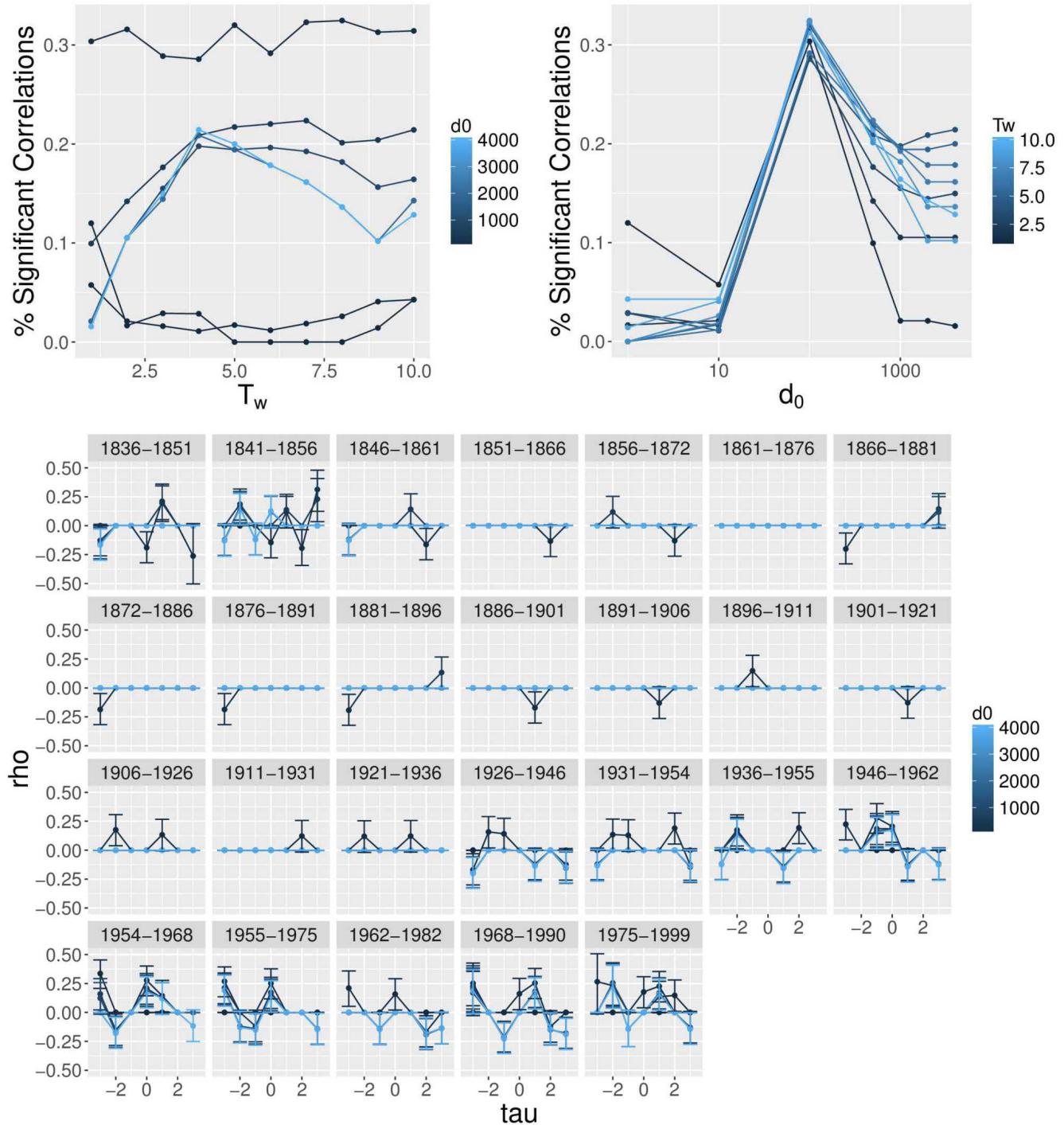


FIGURE 40 : Corrélations retardées empiriques pour le système de villes français.

$$\varepsilon_D = \log \left[\sum_t \sum_{i,j} (d_{ij}(t) - \tilde{d}_{ij}(t))^2 \right]$$

RÉSULTATS Nous procémons à une calibration non-stationnaire, sur les objectifs ($\varepsilon_P, \varepsilon_D$), 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 Fig. montre les fronts de Pareto obtenus, et la Fig. l'évolution dans le temps des valeurs des paramètres pour les solution optimales.

Modèle avec réseau physique

Q : reciprocally how can such coupled models produce realistic networks compared to more classical autonomous models of network growth.

La spécification avec réseau physique correspond en un sens à la combinaison de différentes échelles

Dans le cas d'un auto-renforcement, une specification similaire à celle utilisée précédemment suppose une croissance pour l'ensemble des liens :

$$d(t+1) = d(t) \cdot \left(1 + g_{max} \cdot \left[\frac{\phi}{\max \phi} \right]^{\gamma_s} \right)$$

la specification par seuil ne permettant pas une bonne convergence dans le temps.

Un exemple de configuration obtenue par cette spécification est donné en Fig. 43.

Développements possibles

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 [GALLOTTI et BARTHELEMY, 2014]. Chaque couche du réseau de transport devrait

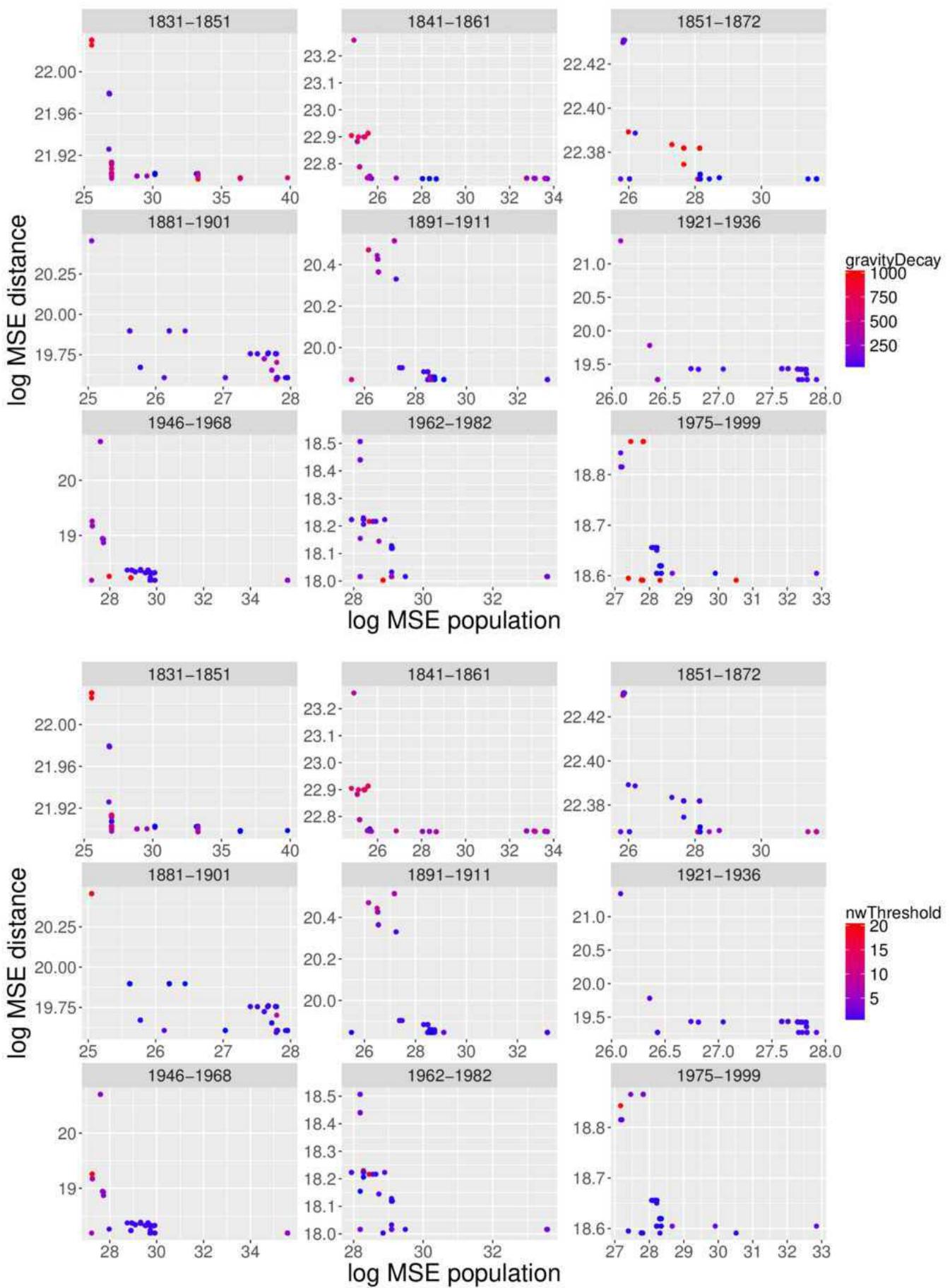


FIGURE 41 : Fronts de Pareto pour la calibration bi-objectif population et distance.
[24 octobre 2017 at 18:35 – Thesis version 3.3.2]

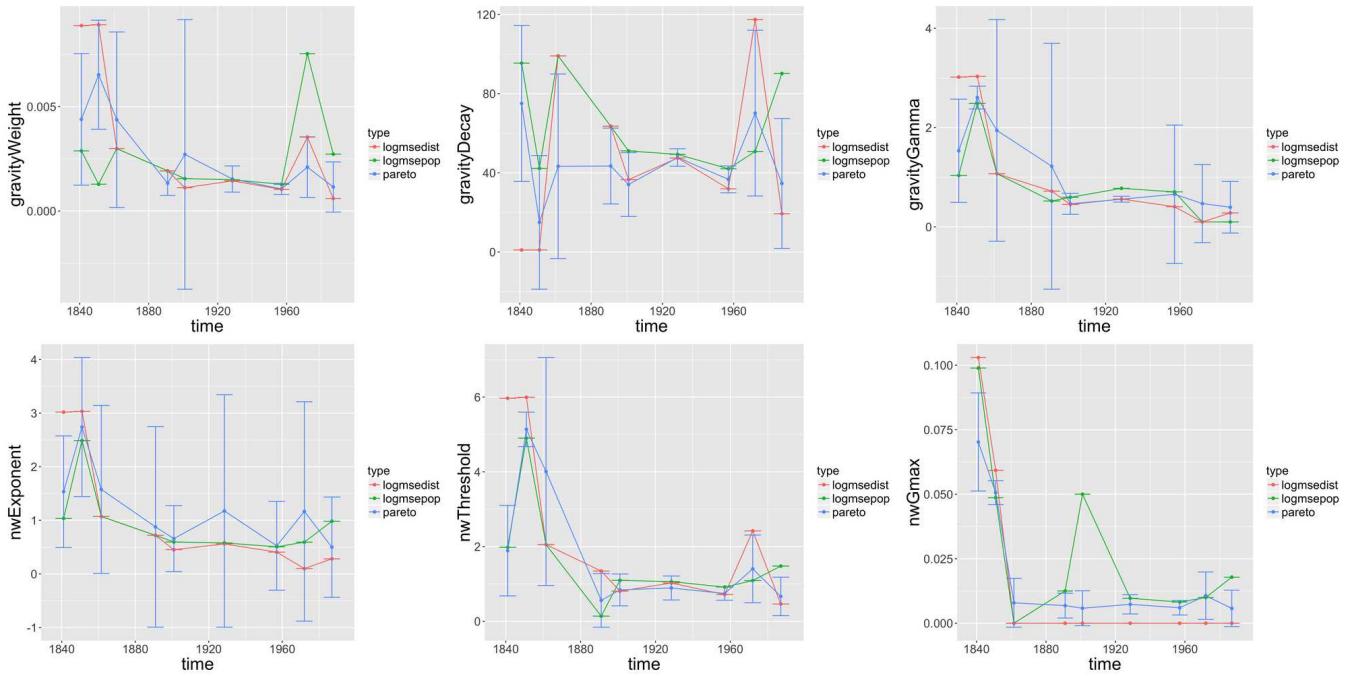


FIGURE 42 : Evolution temporelle des paramètres optimaux.

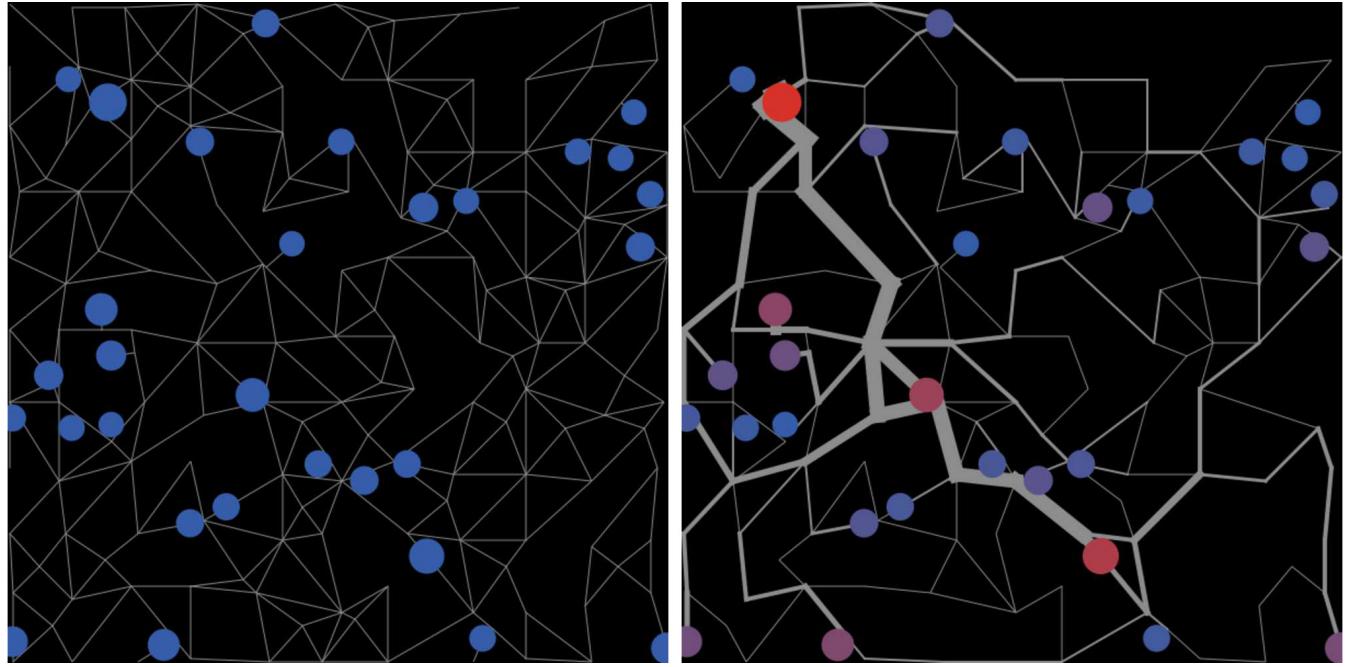


FIGURE 43 : Example de configuration obtenue avec réseau auto-renforçant. (Gauche) Configuration initiale aléatoire, capacités égales; (Droite) Configuration finale obtenue après 100 itérations.

avoir une dynamique co-evolutive avec les populations, avec possiblement l'existence d'une dynamique inter-couches.

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.

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 ferroviaire français a émergé de manière relativement auto-organisée (de par les multiples opérateurs), tandis que les autoroutes ont été fortement planifiées, à l'image du réseau ferré à grande vitesse Chinois, pour lequel un développement précis est envisagé.

CONCLUSION DU CHAPITRE

★ ★

★

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 petites échelle pour l'Afrique du Sud sont plus clairs que ceux pour les transactions immobilières et le Grand Paris. 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.

★ ★

★

Les résultats des deux premières sections de ce chapitre ont été présentés comme [] ; la troisième section correspond à une adaptation de []

7.1 MODÈLES DE CROISSANCE DE RÉSEAU

Nous proposons dans un premier temps de détailler la composante 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 [XIE et LEVINSON, 2009c] 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 réseau sous certaines contraintes. Comme déjà développé précédemment, 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 sur le temps long.

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 [WHITNEY, 2012] 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.

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.

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

Au debut de chaque étape de croissance du réseau :

1. Add new nodes preferentially to new population and connect them
2. Variable heuristic for new links, among : nothing, random, gravity-based deterministic breakdown, gravity-based random breakdown (from [SCHMITT, 2014]), cost-benefits (from [LOUF, JENSEN et BARTHELEMY, 2013]), biological network generation (based on [TERO et al., 2010])

Un nombre fixe n_N de nouveaux noeuds est ajouté. Séquentiellement, pour les patches tels que $d^{(r)} < d_0$, la probabilité de recevoir un nouveau noeud est donnée par

$$p_k = P_k / P_{\max} \cdot (d_M - d_k) / d_M \cdot \exp \left(-((d^{(r)} - d_0) / \sigma_r)^2 \right)$$

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).

General network generation parameters : growth time steps t_N , maximal additional links

Heuristiques de référence

Nous implémentons deux heuristiques de references pour mieux situer celles explorées par la suite : celle composée uniquement de la base décrite précédemment, qui produit des réseaux arborescents uniquement ; et la generation de réseau aléatoire, qui consiste à créer un nombre fixe n_L de nouveaux liens entre des sommets choisis aléatoirement, puis à planariser le réseau final.

Heuristique euclidienne

Nous appliquons la méthode développée en 5.3. Pour memoire, les paramètres sont Il est intéressant de faire le rapprochement avec la rupture de potentielle aléatoire du modele SimpopNet, que nous implementerons également.

Rupture de potentiel aléatoire

La rupture de potentielle aléatoire est celle utilisée par SimpopNet [SCHMITT, 2014], qui reprend le modele introduit par [BLUMENFELD-LIEBERTHAL et PORTUGALI, 2010]. A chaque pas de temps, deux villes sont tirées aléatoirement, la première selon une probabilité proportionnelle à $P_i^{Y_R}$ et la deuxième selon $V_{i_0 j}^{Y_R}$ sachant que i_0 est la première ville tirée et V_{ij} sont les potentiels gravitaires euclidiens. Si

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.

$d_N(i_0, j_0)/d(i_0, j_0) > \theta_R$, c'est à dire si le détour relatif par le réseau est supérieur à un paramètre de seuil, on crée un lien entre les deux villes².

Heuristique biologique

[RAIMBAULT et GONZALEZ, May 2015] explore des applications des modèles de croissance de réseau biologique, 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. Un aperçu des potentialités est donné en Appendice ???. Etant donné un réseau initial dont les liens ont des capacités uniformes, l'itération d'équilibres de pression successifs suivis d'une évolution des capacités, permet une convergence vers une distribution hiérarchique stable des capacités. Notre rationnelle 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 à la plupart des modèles d'investissement qui font évoluer uniquement les capacités dans le temps ; et (ii) elle traduit des processus bottom-up d'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 :

1. A partir du réseau existant auquel on ajoute un réseau en grille 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 itère pour k croissant ($k \in \{1, 2, 4\}$ en pratique) :
 - Etant donné la distribution de la population, on itère $k \cdot n_b$ fois le modèle de slime mould pour obtenir le réseau émergent par convergence des capacités
 - Les liens de capacité inférieure à θ_d sont supprimés
 - La plus grande composante connexe est conservée
3. Le réseau final est planarisé et simplifié

² 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

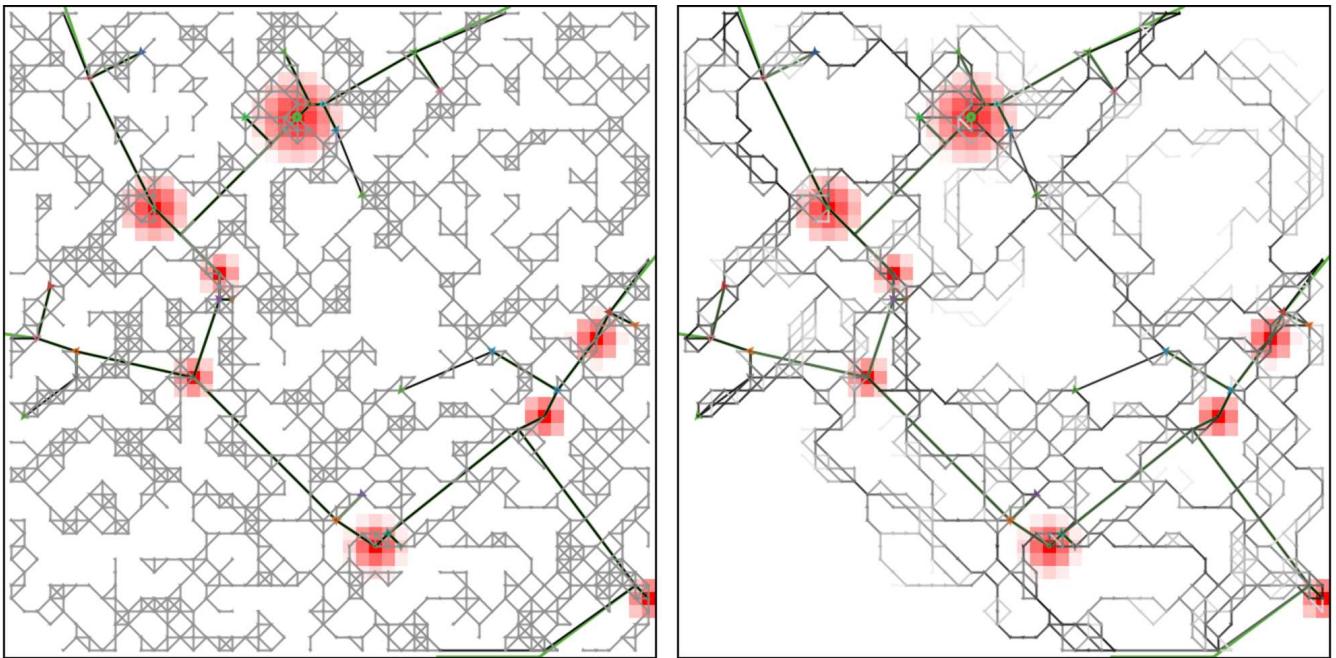


FIGURE 44 : Heuristique biologique pour la génération de réseau. Cet exemple de visualisation illustre les étapes intermédiaires pour l'ajout de lien. (*Gauche*) le réseau semi-aléatoire initial dans lequel le slime mould est lancé; (*Droite*) Même réseau après 80 itérations du slime mould, l'épaisseur des liens donnant la capacité.

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 réseau 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 [LOUF, JENSEN et BARTHELEMY, 2013], 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 minimal, avec un coût de la forme $d_{ij} - \lambda/d_{ij}^2$.

7.1.2 Résultats

Réseaux générés

Une illustration visuelle des différentes topologies générées est donnée en Fig 45. Les particularités de chacune des heuristiques sont immédiatement visibles

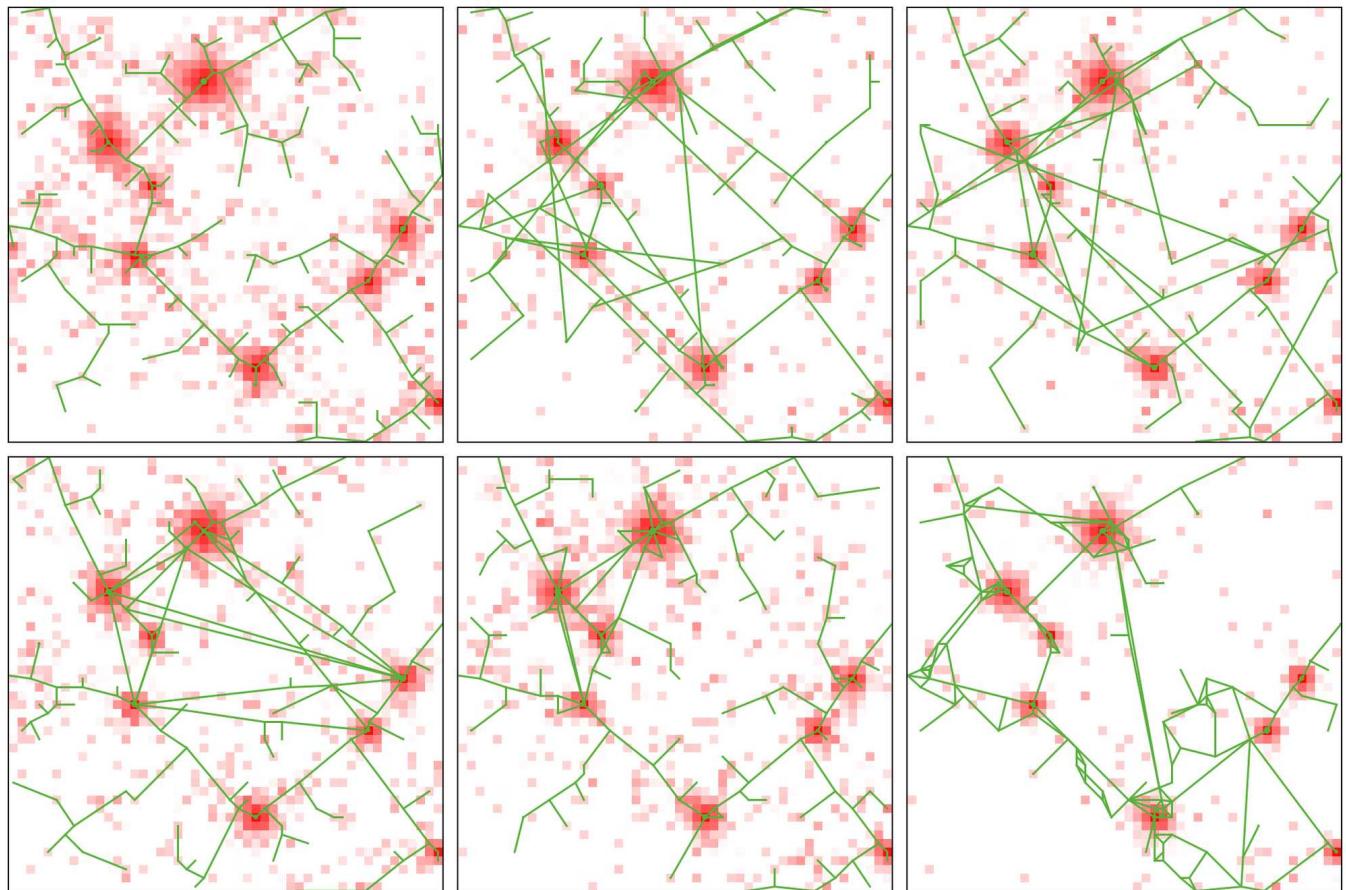


FIGURE 45 : Exemples de réseaux obtenus par les différentes heuristiques

Plan d'expérience

La longueur de réseau finale n'est pas directement contrôlée par les paramètres, nous fixons un critère d'arrêt de croissance du réseau à nombre de lien fixe pour pouvoir comparer les réseaux finaux. La génération de réseau est faite à densité de population constante, sur configurations réelles classifiées morphologiquement en 4.1.

Topologies obtenues

Les réseaux sont caractérisés par les indicateurs suivants : centralité de chemin et de proximité moyennes, diamètre, longueur moyenne de chemin, vitesse relative. Pour visualiser les espaces faisables et les comparer aux réseaux réels par la suite,

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 réelles.

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 [ADAMATZKY et JONES, 2010]. Nos résultats semblent confirmer ces analyses, puisque cette heuristique est la moins performante au sens de la distance aux réseaux réels.

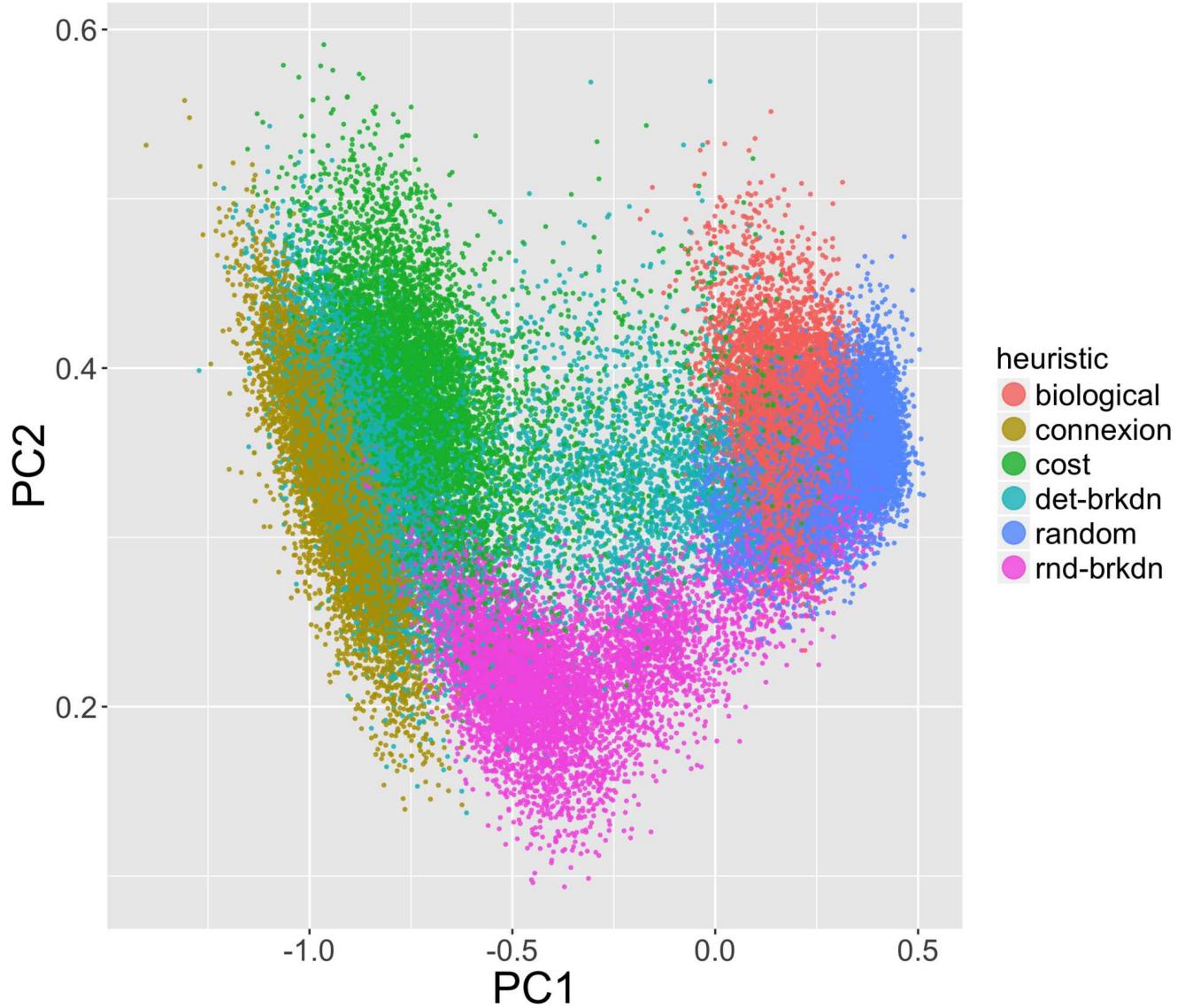


FIGURE 46 : Espace topologique faisable pour les différentes heuristiques de génération. La même figure conditionnée à la classe morphologique de densité est donnée en Appendice A.9.

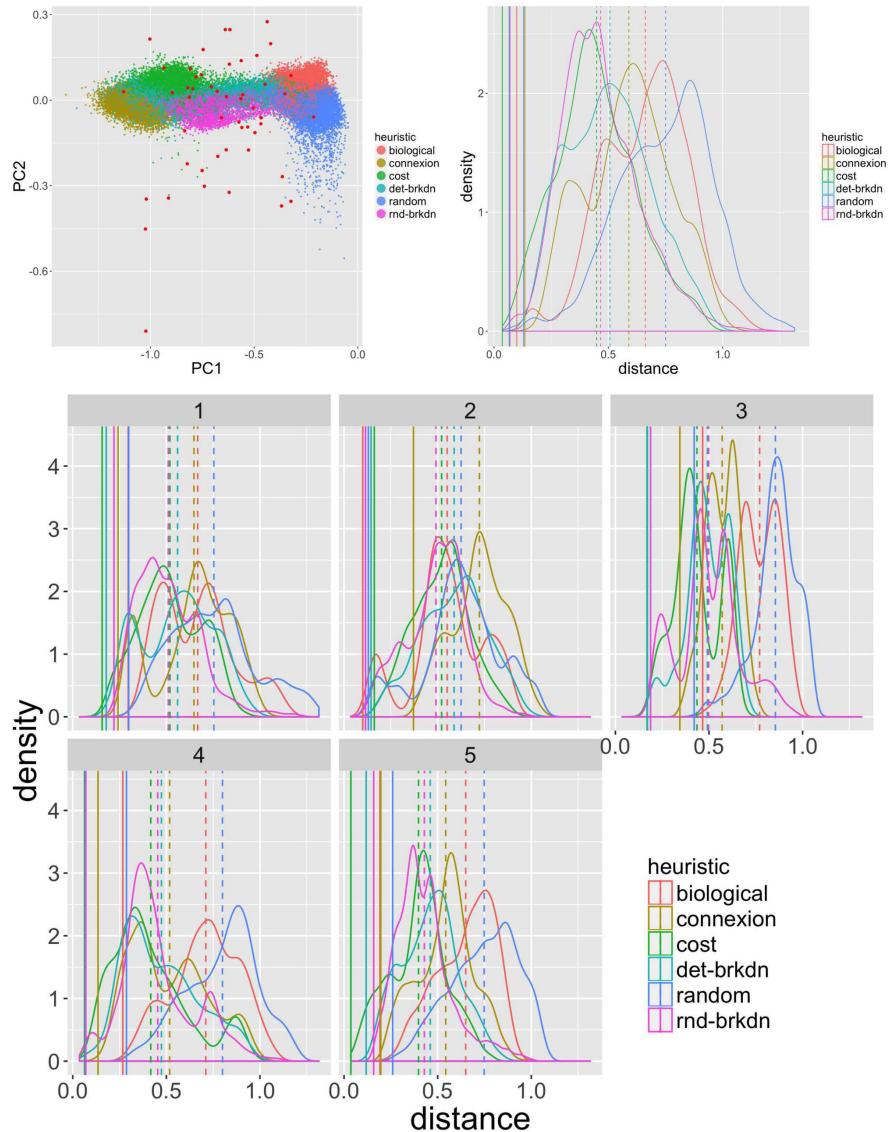


FIGURE 47 : Comparaison aux réseaux réels. (Haut Gauche)

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 un réseau routier croissant.

7.2.1 Modèle

RATIONNELLE 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. [RUI et BAN, 2014] 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 [RAIMBAULT, BANOS et DOURSAT, 2014]. 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. 48 résume la structure générale du modèle de morphogenèse.

FORMALISATION Patch utility given by $U_i = \sum_k w_k \cdot \tilde{x}_k$ with \tilde{x}_k normalized local variables among population, betweenness and closeness centrality, distance to roads, accessibility ; aggregation done with

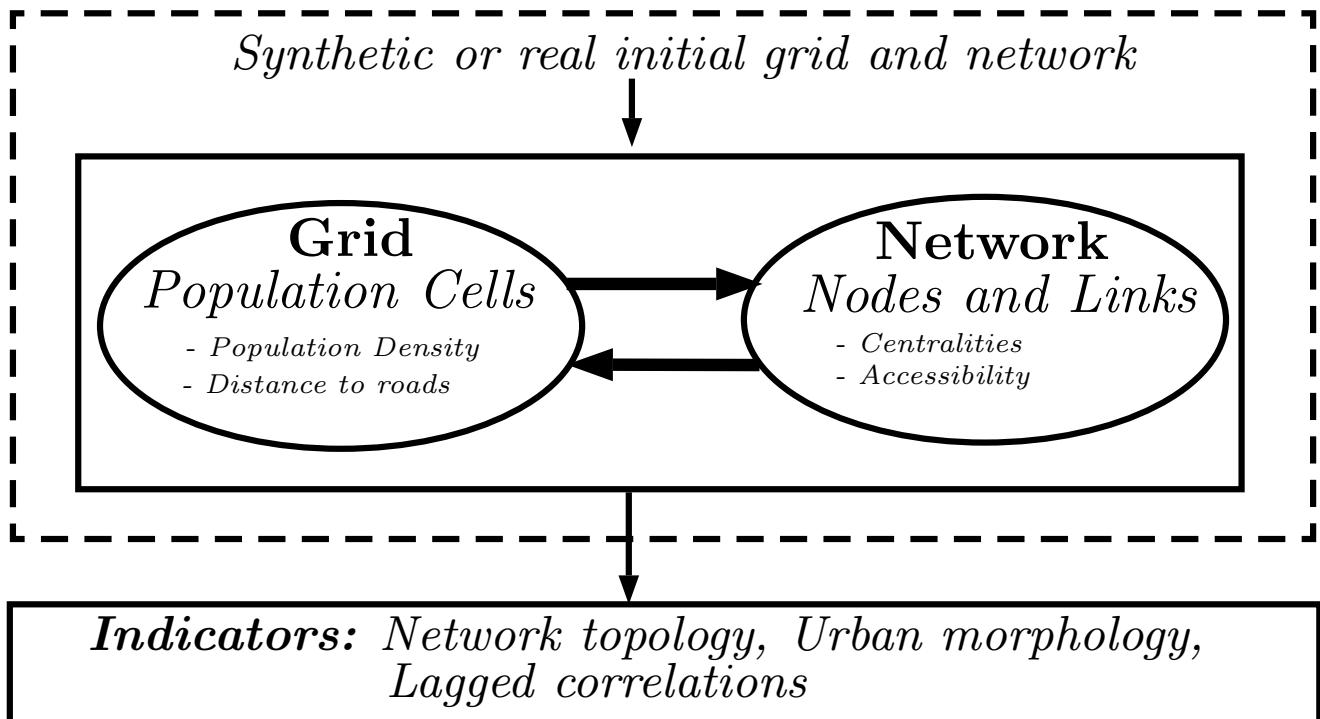


FIGURE 48 : Workflow du modèle de co-évolution à l'échelle mesoscopique.

probability $(U_i / \sum_k U_k)^\alpha$; diffusion among neighbors n_d times with strength β .

7.2.2 Résultats

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, basées sur les données de population raster Eurostat et le réseau routier issu d'OpenStreetMap. Le processus de calibration est le suivant :

La Fig. 49 résume les résultats de la calibration.

We obtain a full coverage of real configurations with simulation results in a principal component plan for indicators, for which most of them a close correlation structure is found. Both network heuristics are necessary for the full coverage. The model is thus able to reproduce existing urban form and networks, but also their *interaction* in the sense of correlations.

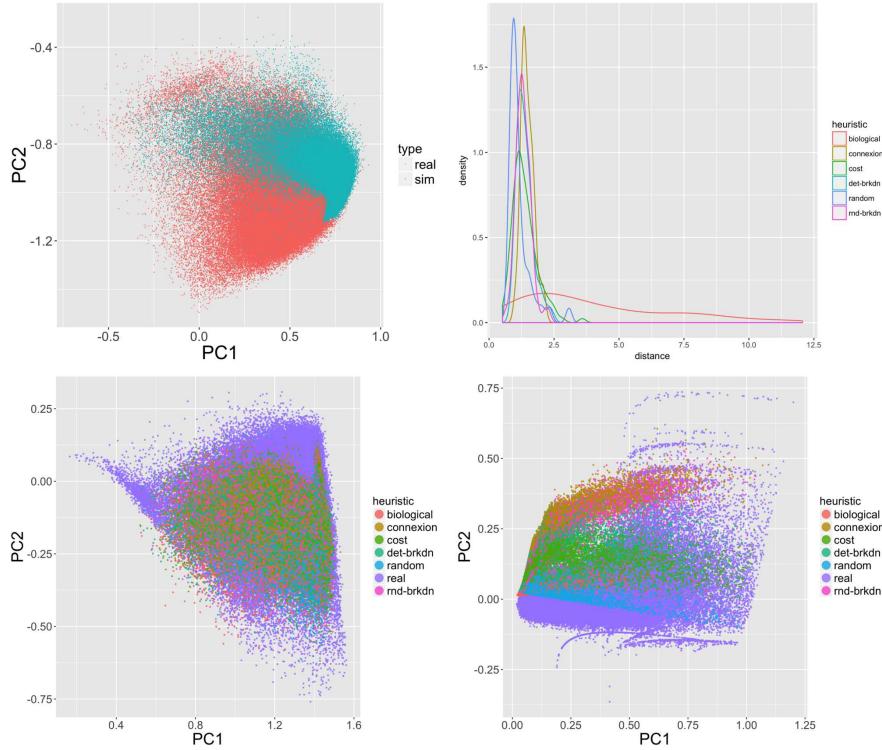


FIGURE 49 : Calibration du modèle de morphogenèse au premier et au second ordre.

Régimes de causalité

Nous étudions d'autre part les corrélations retardées dynamiques entre les retours normalisés de la population et des variables explicatives des patches, appliquant la méthode des régimes de causalité introduite en 4.2. Le modèle exhibe des régimes de causalité divers. Plus précisément, la Fig. 50 résume les résultats obtenus. Le nombre de classe 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

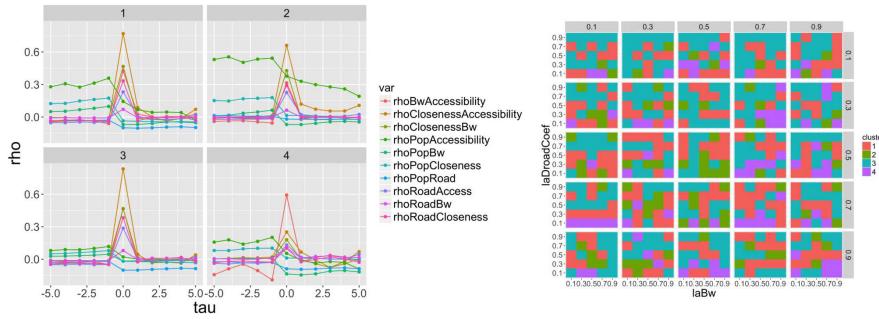


FIGURE 50 : Régimes de causalité.

7.2.3 Développements

C (JR) : [BLUMENFELD-LIEBERTHAL et PORTUGALI, 2010] : hybrid model (largely discussed by Clara); network growth induces migration; would be interesting to test its abilities to produce various causality regimes (note : may be one indicator of how a model captures co-evolution?)

C (JR) : develop on Elsa's comment, why just the network with prefatt? more subtle and want to reproduce coupled dynamics. could suit at the first order? surely as density-only model performs quite well on morphology only. question of territorial representations, what is necessary, what is an objective, what is both. dimension of urban system.

7.3 MODÉLISATION DE LA GOUVERNANCE DU SYSTÈME DE TRANSPORT

Cette section fait un pas supplémentaire vers des modèles plus complexes. Un modèle jouet incluant des processus de gouvernance est décrit. Cette exploration répond de manière logique à notre cadre théorique et aux études précédentes, en particulier pour essayer de valider l'hypothèse de nécessité des réseaux : si des processus non-linéaires sont montrés nécessaires pour la validation sur des faits stylisés, cela pousse à argumenter pour sa validité.

7.3.1 *Contexte*

7.3.2 *Le Modèle Lutecia*

C : (Florent) on DC module : c'est à dire ? comparer quoi avec qui ?

C : (Florent) Implémentation : développer les aspects méthodo "techniques" ce n'est pas sale, au contraire

7.3.3 *Application au Delta de la Rivière des Perles*

CONCLUSION DU CHAPITRE

* * *

*

Conclusion 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 Mémoire.

Feuille de Route

We give the following (non-exhaustive and provisory) roadmap for modeling explorations (theoretical and empirical domains being still explored conjointly) :

1. Complete the exploration of independent and weak coupled urban growth and network growth processes (all models presented in chapter ??), in order to know precisely involved mechanisms when they are virtually isolated, and to obtain morphogenesis scales.
2. Go further into the exploration of toy-model of non conventional processes such as governance network growth heuristic to pave the road for a possible integration of such modules in hybrid models.
3. Build a Marius-like generic infrastructure that implement the theory in a family of models that can be declined into diverse case studies.
4. Launch it and adapt it on these case studies.

Next steps would be too hypothetical if formulated, we propose thus to proceed iteratively in our construction of knowledge and naturally update this roadmap constantly.

- *La route est longue mais la voie est libre.*

Cas d'étude

→ potential application cases ?

Currently we expect to work on the following case studies to build these hybrid models :

- Dynamical data for Bassin Parisien should allow to parametrize and calibrate a model at this temporal and spatial scale.
- On larger scales, South African dataset of BAFFI will along empirical analysis also be used to parametrize hybrid co-evolution models.
- A possibility that is not currently set up (and that may however be difficult because of a disturbing closed-data policy among a frightening large number of scientists!) is the exploitation of French railway growth dataset (with population dataset) used in [BRETAGNOLLE, 2009a], that would also provide an interesting case study on other regimes, scales and transportation mode.

Quatrième partie

OUVERTURE

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 à l'existence de Macao **C (FL) : mal dit**, à une bifurcation géographique **C (FL) : trop simple, toutes les bifurcations sont à la fois géographique, historique, etc.** 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 précédent étudié empiriquement les manifestations morphologiques des interactions à l'échelle mesoscopique **C (FL) : c'est dans ce chapitre qu'il faut discuter de ces échelles, pas avant**, 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 8.1, 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 **C (FL) : et alors ? en quoi est-ce problématique ?**, et que les trajectoires microscopiques du système sont chaotiques. Cela nous permettra d'une part de conforter 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 8.2, en nous concentrons 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. Enfin, la dernière section ?? applique la méthode d'identification de causalités développée en 4.2 au différents projets de transport du Grand Paris et démontre des potentiels effets d'annonce des projets de transport sur la croissance de la population, confirmant la pertinence d'une échelle d'agrégation au moins mesoscopique et de se concentrer sur des variables territoriales relativement basiques.

★ ★

★

Ce chapitre est entièrement adapté de divers articles : la section 8.1 a été publiée en anglais comme [RAIMBAULT, 2017g]; la section 8.2 également en anglais en collaboration avec A. BERGEAUD comme [RAIMBAULT et BERGEAUD, 2017]; la section ?? correspond à la partie d'application de [RAIMBAULT, 2017f].

8.1 INVESTIGATION EMPIRIQUE DE L'EXISTENCE DE L'EQUILIBRE UTILISATEUR STATIQUE

L'Equilibre Utilisateur Statique est un cadre puissant pour l'étude théorique du trafic. Malgré l'hypothèse restreignante **C (FL) : de quel point de vue ?** 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.

8.1.1 Contexte

La modélisation du trafic a été largement étudiée depuis les travaux séminaux de Wardrop ([WARDROP, 1952]) : 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 [HAN, 2003]). 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 réalisables ([RASMUSSEN et al., 2015]). D'autres extensions prenant en compte le comportement de

l'utilisateur via des modèles de choix ont été proposé plus récemment, comme [ZHANG, MAHMASSANI et LU, 2013] qui inclut à la fois l'influence de la tarification 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 [MAHMASSANI et CHANG, 1987]. 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 [Guo et LIU, 2011]. D'autres modèles d'attribution de trafic inspirés d'autres domaines ont également été plus récemment proposés : dans [PUZIS et al., 2013], une définition étendue de la centralité de chemin qui combine linéairement la centralité des flots non-contraints 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 **C (FL) : ?** sur l'Equilibre Utilisateur Statique. La région parisienne utilise par exemple un modèle statique (MODUS) pour gérer et planifier le trafic. [LEURENT et BOUJNAH, 2014] 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 [ZHU et LEVINSON, 2010], 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) : smt (?) : tu ne dis pas ce que tu veux faire ni pourquoi.**

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.

8.1.2 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 **C (FL) : as tu le droit de l'ouvrir ? A1 : (JR) je le prends** 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 [BOUTEILLER et BERJOAN, 2013], 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*¹.

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 **C (FL) : sens ?**, 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 construite **C (FL) : mal dit** par inversion d'un fonction BPR **C (FL) : expliquer ce que**

¹ à l'adresse http://37.187.242.99/files/public/sytadin_latest.sqlite3

c'est 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 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 [GRIMM et al., 2005], nous nous intéressons au motifs **C (FL) : lesquelles et pourquoi?** 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 leaflet². 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 51 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 métaparamètres 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.

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 in-

² le code source de l'application et des analyses est disponible sur le dépôt ouvert du projet à <https://github.com/JusteRaimbault/TransportationEquilibrium>

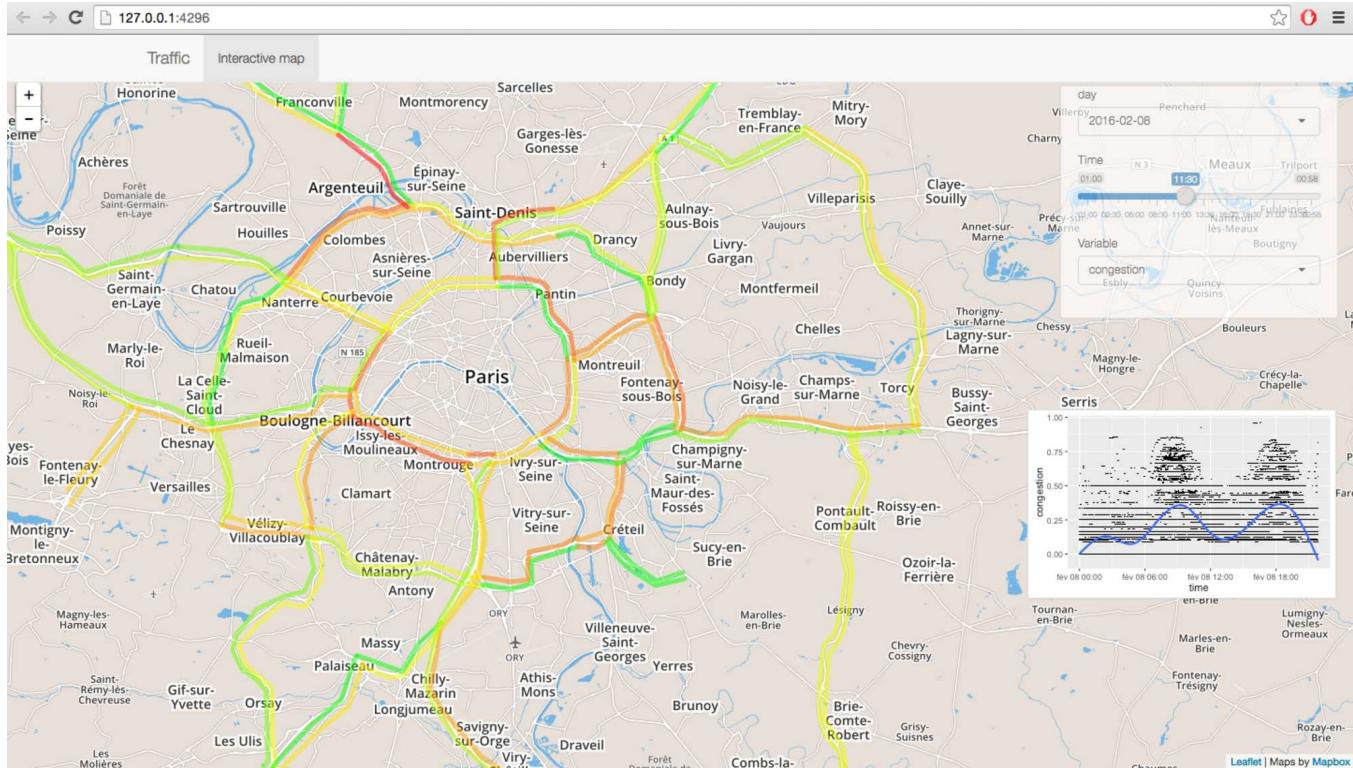


FIGURE 51 : Capture de l'application web **C (FL)** : développée par/pour (?) qui ? permettant l'exploration spatio-temporelle des données de traffic pour la région Parisienne. Il est possible de choisir date et heure (précision de 15min sur un mois, réduite par rapport au jeu de données initial pour 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).

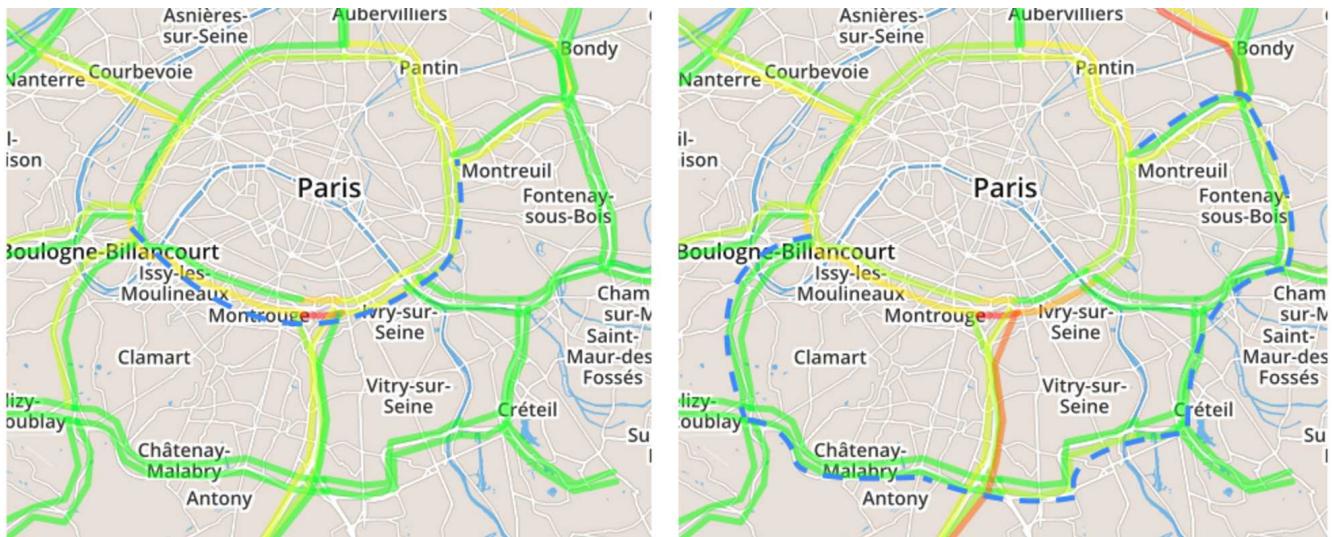


FIGURE 52 : Variabilité spatiale d'un plus court chemin en temps de trajet (trajet du plus court chemin en pointillé bleu). Dans un intervalle de seulement 10 minutes, entre le 11/02/2016 00 :06 (à gauche) et le 11/02/2016 00 :16 (à droite), le plus court chemin entre Porte d'Auteuil à l'ouest et Porte 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.

valider 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 variabilité 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

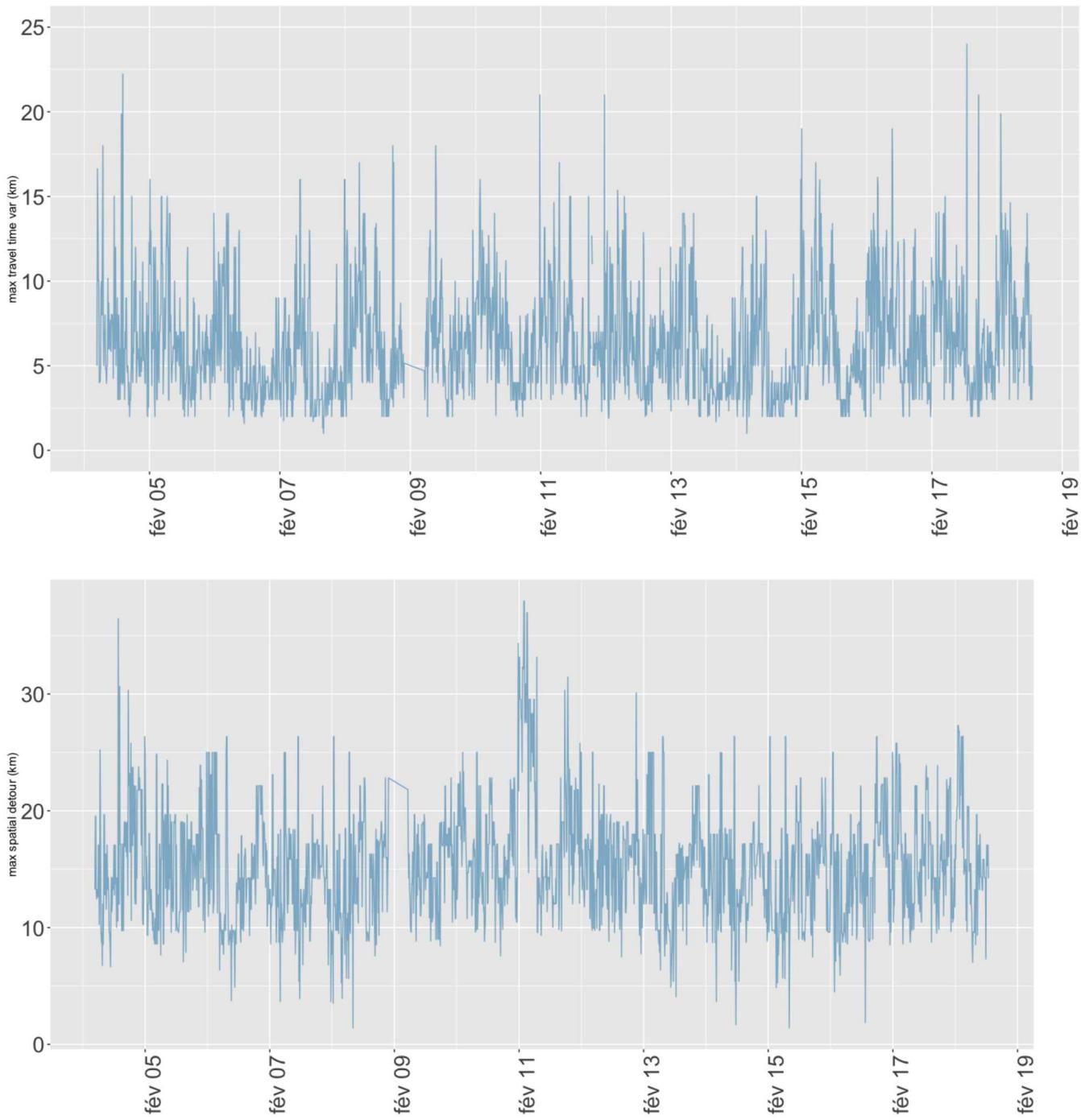


FIGURE 53 : Variabilité maximale du temps de trajet (en haut) en minutes et de la distance de trajet correspondante (en bas) pour un échantillon de deux semaines. Le graphe représente le maximum sur l'ensemble des paires Origine-Destination de la variabilité absolue entre deux pas de temps consécutifs. 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. **C (FL)** : pourquoi le choix de cet indicateur ? par ex. si le detour correspond a une OD tres tres minoritaire, en quoi est-ce pertinent ? **A1** : (JR) j'avais rajouté cet indicateur pour répondre à une remarque d'un reviewer, mais en effet il casse pas des briques. il faut faire une hypothèse que vu qu'on a une granularité assez grande, la plupart des OD seront réalisés (de plus en prenant les sous-trajets).

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 [BAVOUX et al., 2005]. La littérature en transports a développé des mesures de réseau élaborées et opérationnelles, comme des mesures de robustesse pour identifier 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 [SULLIVAN et al., 2010]).

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

$$b_i = \frac{1}{N(N-1)} \cdot \sum_{o \neq d \in V} \mathbb{1}_{i \in p(o \rightarrow d)} \quad (11)$$

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. 54 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

$$\Delta b(t) = \frac{|\max_i(b_i(t + \Delta t)) - \max_i(b_i(t))|}{\max_i(b_i(t))} \quad (12)$$

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 54) 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 [ZHU et LEVISON, 2010]), 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

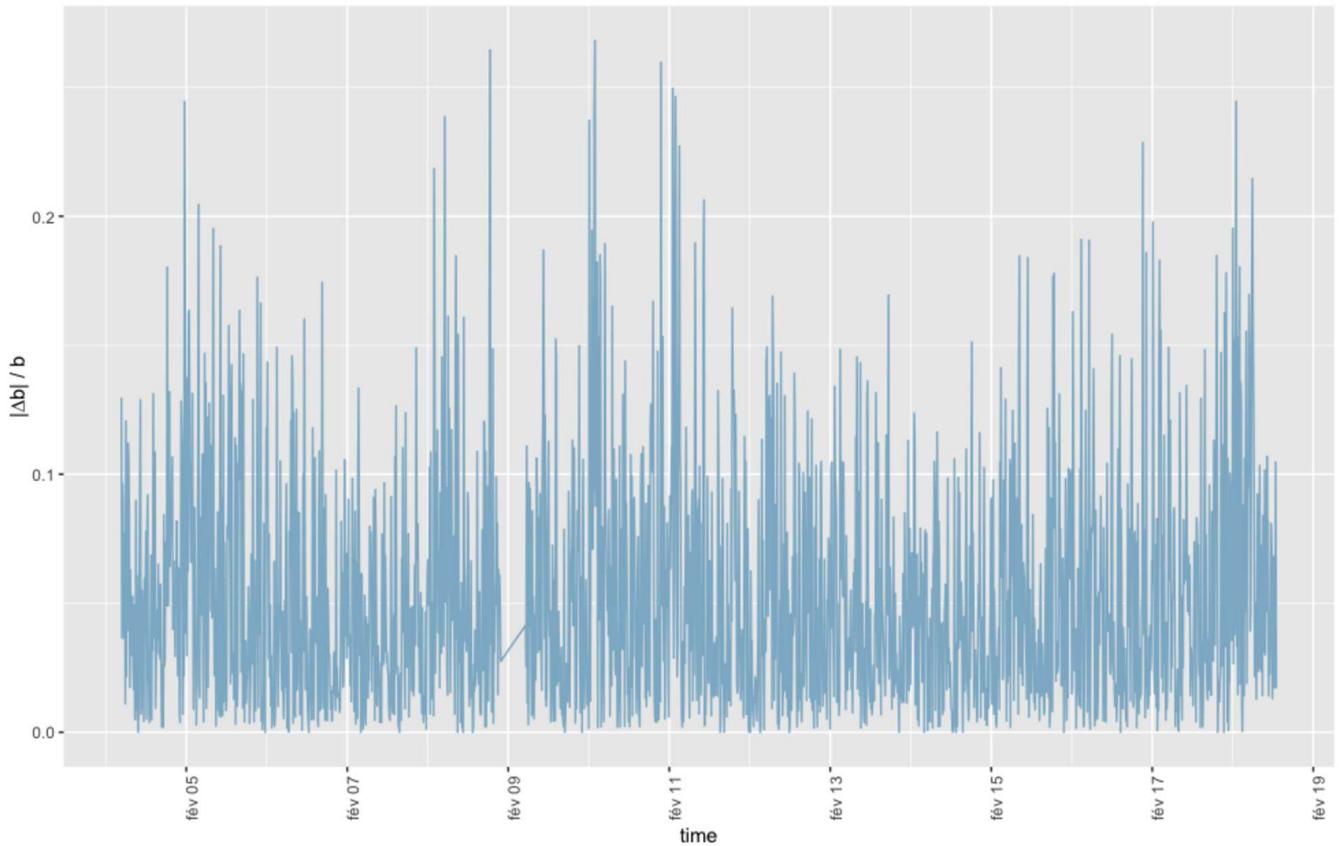


FIGURE 54 : Stabilité temporelle du maximum de la centralité de chemin. Le graphe montre dans le temps la dérivée normalisée du maximum de la centralité de chemin, qui capture ses variations relatives à chaque pas de temps. La valeur maximale de 25% correspond à de très fortes perturbations du réseau 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. **C (FL) : ce n'est pas possible à lire → synthétiser A1 : (JR) pas pertinent de lisser, car c'est justement les max locaux qui nous intéressent.**

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 [Tsai, 2005]). 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

$$\rho_i = \frac{1}{K} \cdot \sum_{i \neq j} w_{ij} \cdot (c_i - \bar{c})(c_j - \bar{c}) \quad (13)$$

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 55 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.

8.1.3 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 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

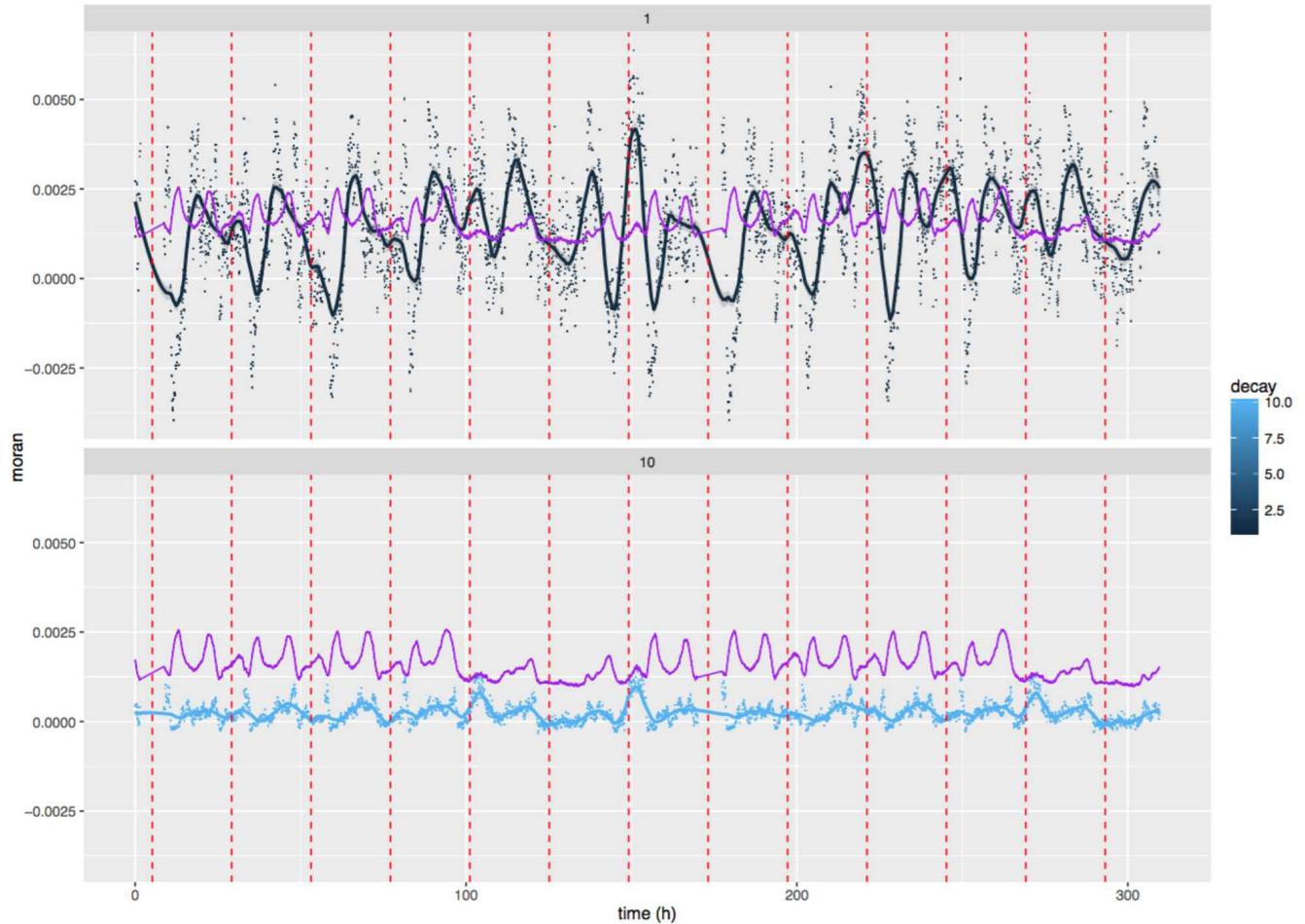


FIGURE 55 : Auto-corrélations spatiales pour les vitesses relatives sur deux semaines. Le graphe montre les valeurs de l'auto-corrélation dans le temps, pour des valeurs variables (1,10km) de la distance de décroissance. les valeurs intermédiaires de la distance de décroissance donnent une déformation 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.

taille et variés, et pour des objectifs d'application différents (prédiction du traffic, reproduction de faits stylisés, etc.), à l'image des revues systématiques qui sont la règle en évaluation thérapeutique par exemple. **C (FL) : source ?** 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. [KRYVOBOKOV et al., 2013] 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 à 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 **C (FL) : je ne pense pas que tu as prouve cela : tu ne dis pas comment sont utilisées ces modèles** que les modèles statiques ne devraient pas être utilisés pour la prédiction 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, [COMMENGES, 2013a] 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.

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 pour-

rait 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 évènements 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 traffic, par l'intermédiaire d'algorithmes numériques comme ceux décrits par [GOLDHIRSCH, SULEM et ORSZAG, 1987]. 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. [TORDEUX et LASSARRE, 2016] 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 prin-

cipal. 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.

* * *

*

8.2 TRANSPORT ROUTIER ET DÉTERMINANTS DES COÛTS

C (FL) : quel rapport avec la problematique ?

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 paramétrisées.

8.2.1 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, [RIETVELD, BRUINSMA et VAN VUREN, 2001] étudie l'impact de différences de prix transfrontalières et leur implications pour une taxation spatiale graduelle aux Pays-Bas. A l'échelle du pays, [RIETVELD et WOUDENBERG, 2005] fournit des modèles statistiques pour expliquer les variabilités des prix entre les pays Européens. [MACHARIS et al., 2010] 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, [GREGG et al., 2009] 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 [COMBES et LAFOURCADE, 2005] 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, [GAUTIER et SAOUT, 2015] 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 terri-

toire US métropolitain. Ces informations sont construites à partir des prix reportés par les utilisateurs et couvre pratiquement l'ensemble des station 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 ([TAN et al., 2013])) à la géographie (e.g. les nouvelles approches de la mobilité urbaine ou les perspectives de ville plus "intelligentes" ([BATTY, 2013a])) 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 [DUPUY et BENGUIGUI, 2015] 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, réparti 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 parsage 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.³ Nous avons construit

³ à <https://github.com/JusteRaimbault/EnergyPrice>

TABLE 9 : 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.⁴ 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).

⁴ 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).

8.2.2 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 ([TSAI, 2005]), 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 décroissance. 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,

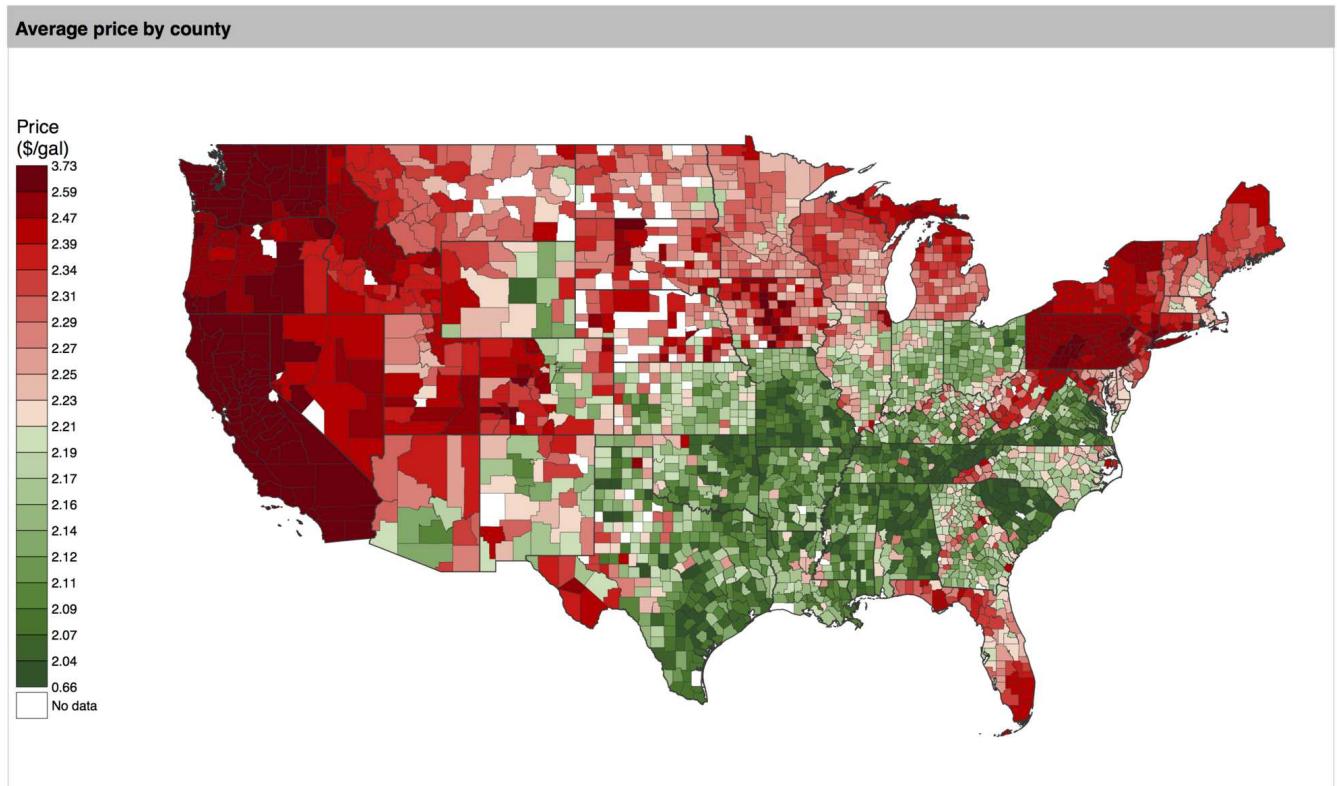


FIGURE 56 : Carte du prix moyen par comté, carburant régulier, moyenne temporelle prise sur l'ensemble de la période.

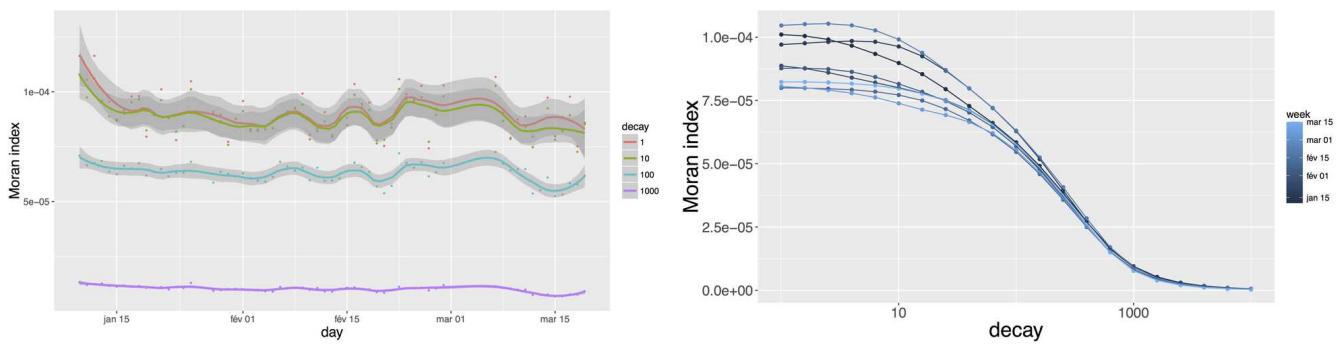


FIGURE 57 : Comportement de l'index d'autocorrelation spatiale de Moran. (Gauche) Evolution dans le temps de l'index de Moran, calculé sur des fenêtres journalières, pour différentes valeurs du paramètre de décroissance. (Droite) Index de Moran en fonction du paramètre de décroissance, calculé sur des 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 consistants sur des fenêtres temporelles à la semaine. Nous postulons que celles-ci correspondent aux 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 prix 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'accent 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 [BRUNSDON, FOTHERINGHAM et CHARLTON, 1996]

et a été utilisée 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,⁵ 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 surpasse 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.⁶ 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 c'est l'est), ce qui pourrait témoigner de différentes

⁵ on note que la forme du noyau n'a plus d'influence tant que des fonctions décroissantes graduellement sont utilisées.

⁶ 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 :

$$x_{i,s,c,t} = \beta_s + \varepsilon_{i,s,c,t} \quad (14)$$

$$x_{i,s,c,t} = \beta_c + \varepsilon_{i,s,c,t} \quad (15)$$

$$x_{i,s,c,t} = \beta_i + \varepsilon_{i,s,c,t} \quad (16)$$

$$(17)$$

Où $\varepsilon_{i,s,c,t}$ contient une erreur idiosyncratique **C (FL) : a expliquer** et un effet fixe jour. Cette première analyse confirme que la majorité de la variance **C (FL) : idem** 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 :

$$\log(x_i) = \beta_0 + X_i \beta_1 + \beta_{s(i)} + \varepsilon_i, \quad (18)$$

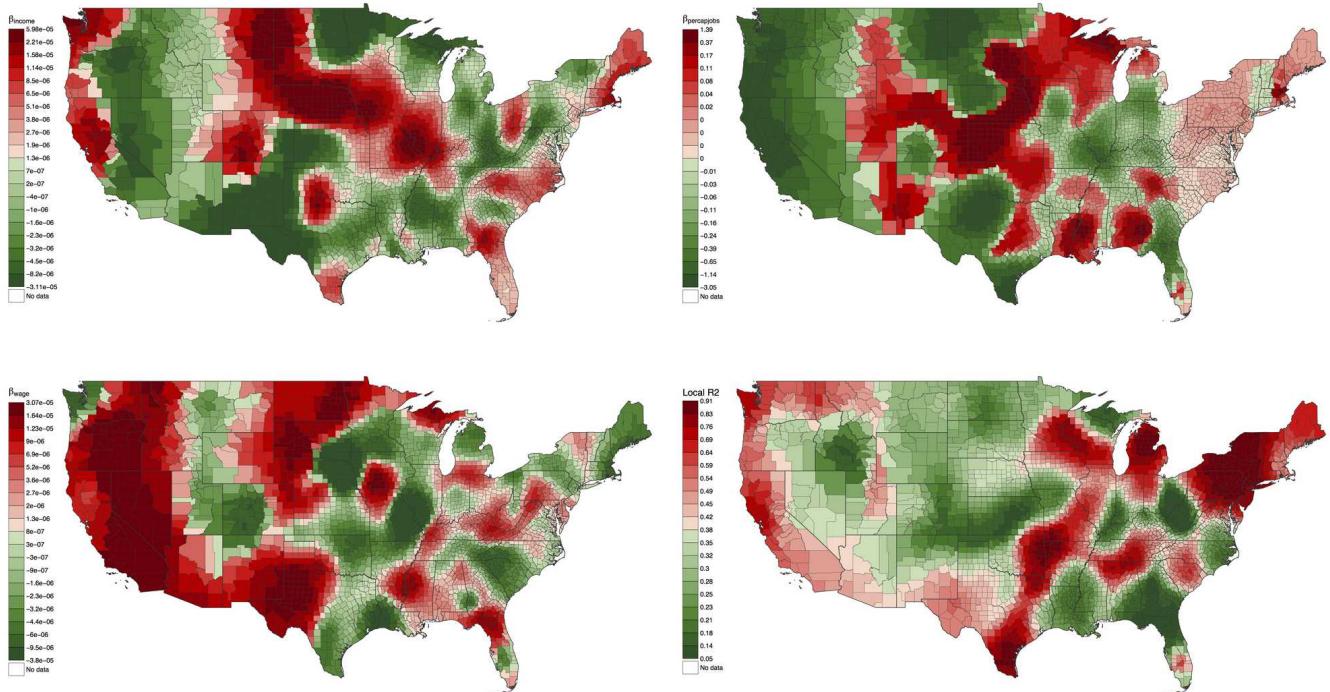


FIGURE 58 : **Résultats des analyses GWR.** Pour le meilleur modèle au sens de l'AICc, les cartes donnent la distribution spatiale des coefficients estimés, dans l'ordre de gauche à droite et de haut en bas, β_{income} , $\beta_{percapjobs}$, β_{wage} , et finalement les valeurs du R² local.

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 $Cov(\varepsilon_i, \varepsilon_j) = 0$ 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.

8.2.3 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 ([CHEN et TRUONG, 2012]), ou les ont comparées en terme de performance de modèle ou de robustesse ([LEE, KANG et KIM, 2009]). 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 géo-

TABLE 10 : Régressions au niveau du comté **C (FL) : quel rapport avec la these ?**

	(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 (18). 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 niveau de significativité 0.01, 0.05 and 0.1.

graphiques comme exogènes. Nous 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. [FULLERTON et WEST, 2002] montre qu'une taxation des carburants et des voitures peut être équivalente à une taxation des émissions. [BRAND, ANABLE et TRAN, 2013] 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 8.1, 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 8.2, 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 co-évolution, 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 7.3.

* * *

*

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 **C (FL) : affirmation gratuite** 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 [BOURGINE, CHAVALARIAS et AL., 2009] **C (FL) : dont il convient de rappeler les grands principes**, 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.

Nous proposons d'abord de construire une *Théorie Géographique* **C (FL) : éviter ce genre de phrase**, 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. **C (FL) : cela ressemble plus à des pensees qu'à un cadre d'analyse** 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 (B.7) un cadre

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.

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 (9.3) un cadre de connaissances appliquée 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 appliquée de manière réflexive à l'ensemble de notre travail.

Ce chapitre sera éventuellement le plus délicat à la lecture **C (FL)** : ne pas dire cela : l'éviter , 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 **C (FL) : digression inutile A1** : (JR) pas d'accord car aspect essentiel de mon positionnement sur la production de connaissance, c'est l'illustration récursive en quelque sorte), 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 appliquée 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. **C (FL) : B**

* * *

*

La première section de ce chapitre reprend un court passage de [RAIMBAULT, 2017b] ; la deuxième est entièrement inédite. La troisième a été proposée par [RAIMBAULT, 2017h] puis développée et appliquée dans [RAIMBAULT, 2017b], et son application réflexive a été présentée par [RAIMBAULT, 2017c].

9.1 CONTRIBUTIONS ET PERSPECTIVES

9.1.1 Définition de la co-évolution : résultats empiriques et méthodes

ÉCHELLES SPATIALES ET NON-STATIONNARITÉ

RÉGIMES DE CO-ÉVOLUTION

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 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 dévoile 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 intéressants 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 [PUMAIN, 1997], 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 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 génériques 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.

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 [ZHANG et al., 2013]. 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 [FAVARO et PUMAIN, 2011] 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.

9.2 UNE THÉORIE GÉOGRAPHIQUE

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 [PORTUGALI, 2011] 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 [MURPHY, 2012]. 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 [CHAMPOILLION, 2006]. 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. Cette théorie a été introduite initialement dans [PUMAIN, 1997] 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 [BERRY, 1964], 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 [PUMAIN, 2003]. 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 [PUMAIN et al., 2006], qui ont par ailleurs été mis en évidence de manière empirique pour plusieurs systèmes urbains [PUMAIN, PAULUS et VACCHIANI-MARCUZZO, 2009]. 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 [PUMAIN, 2010]. 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 [PUMAIN, 2010]. La Théorie Evolutive des Villes a été élaborée conjointement avec des modèles de systèmes urbains : par exemple le modèle Simpop2 introduit par [BRETAGNOLLE, DAUDÉ et PUMAIN, 2006] 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 [BRETAGNOLLE et PUMAIN, 2010b]. Les accomplissements les plus récents de la théorie évolutive reposent sur les productions du projet ERC GeoDiversity, présentées dans [PUMAIN et REUILLO, 2017b], qui incluent de progrès avancés à la fois techniques (logiciel OpenMole² [REUILLO, LECLAIRE et REY-COYREHOURCQ, 2013]), thématiques (connaissance issue des modèles SimpopLocal [SCHMITT, 2014] et Marius [COTTINEAU, 2014]) et méthodologiques (modélisation incrémentale [COTTINEAU, CHAPRON et REUILLO, 2015]). Pour une analyse épistémologique par méthode mixte 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. Ici, cette théorie nous permet d'interpréter les systèmes territoriaux comme systèmes complexes adaptatifs avec les implications listées ci-dessous.

² <http://openmole.org/>

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 quelle mesure ceux-ci contribuent à la construction de notre théorie. La morphogenèse a été particulièrement soulignée par TURING dans [TURING, 1952] 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 [DELILE, DOURSAT et PEYRIÉRAS, 2016]. Dans le cas des systèmes urbains, l'idée de morphogenèse urbaine, i.e. de mécanismes auto-cohérents qui produisent la forme urbaine, est plutôt utilisé dans les champs de l'architecture et de l'urbanisme [HACHI, 2013] (comme e.g. la grammaire générative du "Pattern Language" d'ALEXANDER), en relation avec des théories de la forme urbaine [MOUDON, 1997]. Cette idée peut être poussée jusqu'à de très petites échelles comme celle du bâtiment [WHITEHAND, MORTON et CARR, 1999] 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 [BONIN et HUBERT, 2012] décrit un modèle de morphogenèse urbaine avec différentiation qualitative, tandis que [MAKSE et al., 1998] 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 corrélations non intrinsèques [KOLCHINSKY, GATES et ROCHA, 2015] (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 [WANG, MATNI et DOYLE, 2017]. 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 [BOURGINE et STEWART, 2004], i.e. comme un réseau de processus qui s'auto-reproduisent³ 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 [HOLLAND, 2012]. 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 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 *co-évolution*. 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

³ qui ne sont toutefois pas cognitifs, ne rendant pas ces systèmes morphogénétiques vivants au sens de auto-poiétique et cognitif

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.

Definition 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.*

C (FL) : a mettre bien plus haut : c'est assez fondamental

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 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. **C (FL) : a discuter** 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. ■

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.

Assumption 1 - Imbrication des échelles et des sous-systèmes. Des réseaux complexes de retroaction existent à la fois entre et à l'intérieur des échelles [BEDAU, 2002]. 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.

Assumption 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 systèmes ferroviaire français par [BRETAGNOLLE, 2009a]. 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 [SANDERS, 1992], 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 [KNIGHT, 1975], tout processus stationnaire discret est un *Modèle de Markov Caché*. Etant donné la définition d'un état causal comme $P[\text{future}|A] = P[\text{future}|B]$, la partition des états du système par la relation d'équivalence correspondantes permet de produire un *Réseau Récurrent* qui est suffisant pour déterminer l'état suivant du système, puisqu'il s'agit d'une fonction *déterministe* des états précédents et des états cachés [SHALIZI et CRUTCHFIELD, 2001] : $(x_{t+1}, s_{t+1}) = F[(x_t, s_t)]$. L'estimation des états cachés et de la fonction récurrente capture ainsi entièrement par apprentissage pro-

fond 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 réglés 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" [KUHN, 1970] ou le positionnement de Feyerabend pour une diversité des approches [FEYERABEND, 1993], 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 [CRONIN et SUGIMOTO, 2014]. 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 buts différents. Par exemple, [DURANTIN et al., 2017] explore les potentialités de coupler l'ingénierie avec des paradigmes du design to encourage 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, [CARLILE, 2004] 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, [GEMINO et WAND, 2004] recommande l'utilisation de grammaires pour comparer les techniques de Modélisation

Conceptuelle. Les cadres de meta-modélisation peuvent aussi être compris comme des cadres de connaissance. [COTTINEAU et al., 2015a] décrit un cadre de multi-modélisation pour le test d'hypothèses dans la simulation des systèmes complexes socio-techniques. [GOLDEN, AIGUER et KROB, 2012] 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 [HOFSTADTER, 1980] 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. [MOULIN-FRIER et al., 2017] 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 [LIVET et al., 2010a], 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, 2017h], 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*⁴, 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 [LEGAVRE, 1996] 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 interviews⁵ 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, consulter le deuxième pilier de notre théorie géographique en 9.2, qui en donne la substantifique moelle.

⁴ 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 [PUMAIN et al., 2006], mais "Evolutive Urban Theory" convient aussi, mais il semble dans tous les cas difficile de transférer l'ambiguïté lors de la traduction.

⁵ 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> [RAIMBAULT, 2017e]. 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 [PUMAIN, 1997] 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 à 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 OpenMole. 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 sûr 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 [SCHMITT et al., 2014]. L'existence de multiples solution équivalentes pour les valeurs des paramètres est une barrière pour des question 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 [REUILLOU et al., 2015]. Ce cercle vertueux a été continué avec la famille de modèles Marius [COTTINEAU, 2014]

et l'algorithme Parameter Space Exploration [CHÉREL, COTTINEAU et REUILLOU, 2015]. 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 cœur du réseau de citations des publications principales construisant la Théorie Evolutive Urbaine. Nous construisons le réseau de citations comme décrit en Fig. 59, en utilisant l'outil de collection de données fournit par [RAIMBAULT, 2016b]⁶. Partant des deux publications séminales [PUMAIN, 1997] et [SANDERS et al., 1997], 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 cœur 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 deux aspects. Pour analyser la structure de manière plus fine, nous détectons

⁶ L'ensemble du code et des données pour cette analyse sont disponibles à <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/QuantEpistemo>

les communautés en utilisant l'algorithme de clustering de Louvain, et évaluons la modularité dirigée de la partition comme donnée par [NICOSIA et al., 2009]. Nous montrons en Fig. 59 une visualisation du réseau. Nous obtenons 7 communautés avec une valeur de modularité de 0.39. Pour assurer que cette valeur est significative, nous procémons 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 typologie 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 [BELMONTE et al., 2008]. L'utilisation et l'adaptation de méthodes particulières comme la modélisation basée-agent est cruciale pour le développement de transports autonomes innovants [BALBO, ADAM et MANDIAU, 2016]. Dans ce problème d'ingénierie, des solutions techniques comme les portes palières 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 [FOOT, 2005]. Mais ils peuvent aussi interagir avec d'autres aspects de la connaissance conceptuelle, comme le management et l'organisation au sein de l'opérateur [FOOT, 1994]. L'aspect multi-dimensionnel

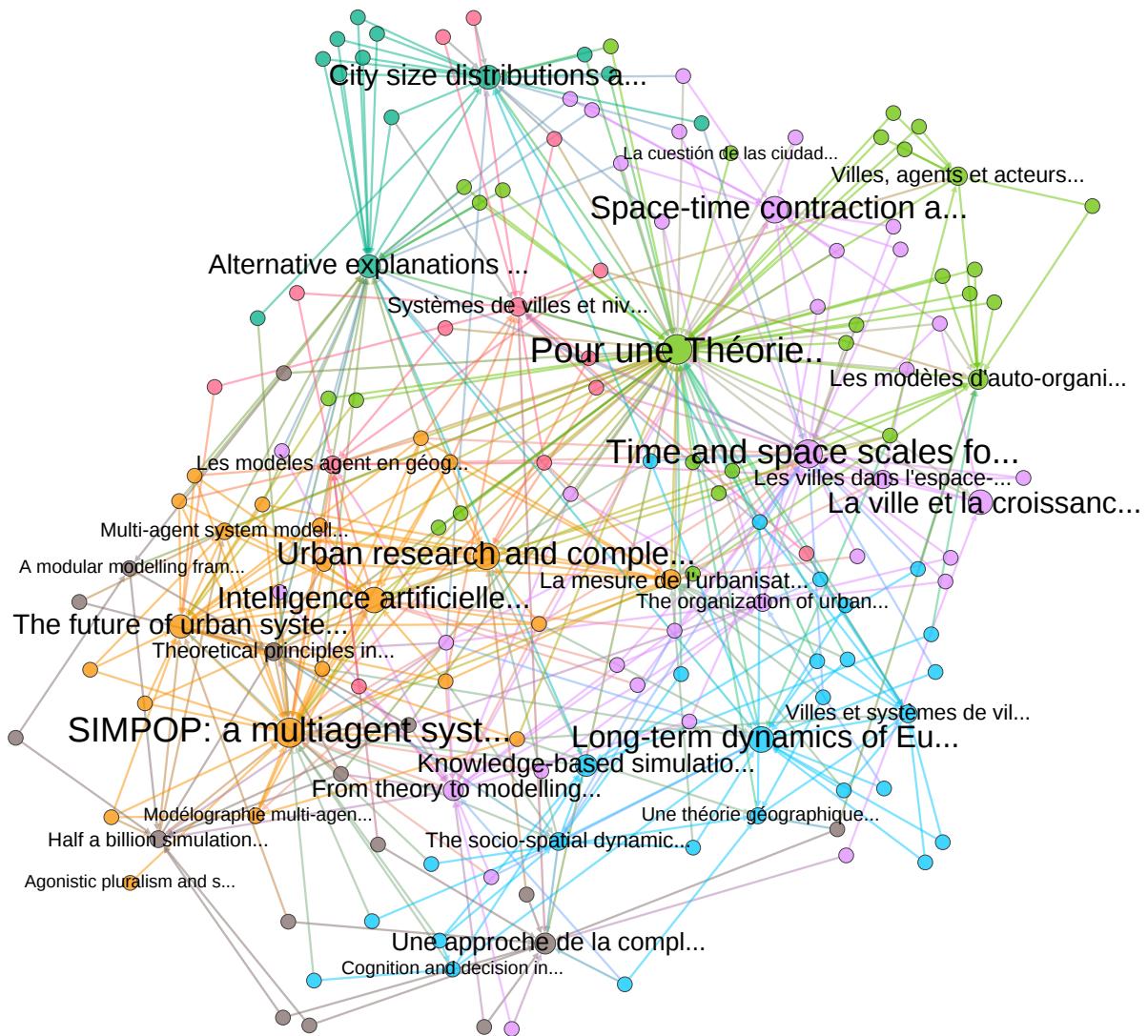


FIGURE 59 : Réseau de citations des publications principales de la Théorie Evolutive Urbaine. Le réseau est construit de la manière suivante : à partir des deux publications séminales [PUMAIN, 1997] and [SANDERS et al., 1997], nous récupérons les publications les citant, filtrons sous la condition d'un 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).

complexe de l'innovation pour de tels systèmes avait déjà été souligné depuis longtemps comme le montre [HATCHUEL, PALLEZ et PÉNY, 1988]. D'autres aspects techniques, comme des problèmes d'ingénierie civile [MORENO REGAN, 2016], 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'interprétation 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 [LIVET et al., 2010a], 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 production⁷
- 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* [GIERE, 2010c], composée des agents, des objets représentés, du but et du medium (le modèle). L'approche par agents est fondamentale pour la cohérence du cadre.

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.

⁷ 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 [VARENNE, 2010b] (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*⁸ 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 [HOLLAND, 2012], 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 [BEDAU, 2002] 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 la

⁸ Nous sommes convaincus que cet aspect intrigué 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 [BOLOTIN, 2014] 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.

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.

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 [VARENNE, 2010b] (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 voir 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 coévolution, 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. 6o 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.

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 B.7. Rappelons les éléments clés et comment

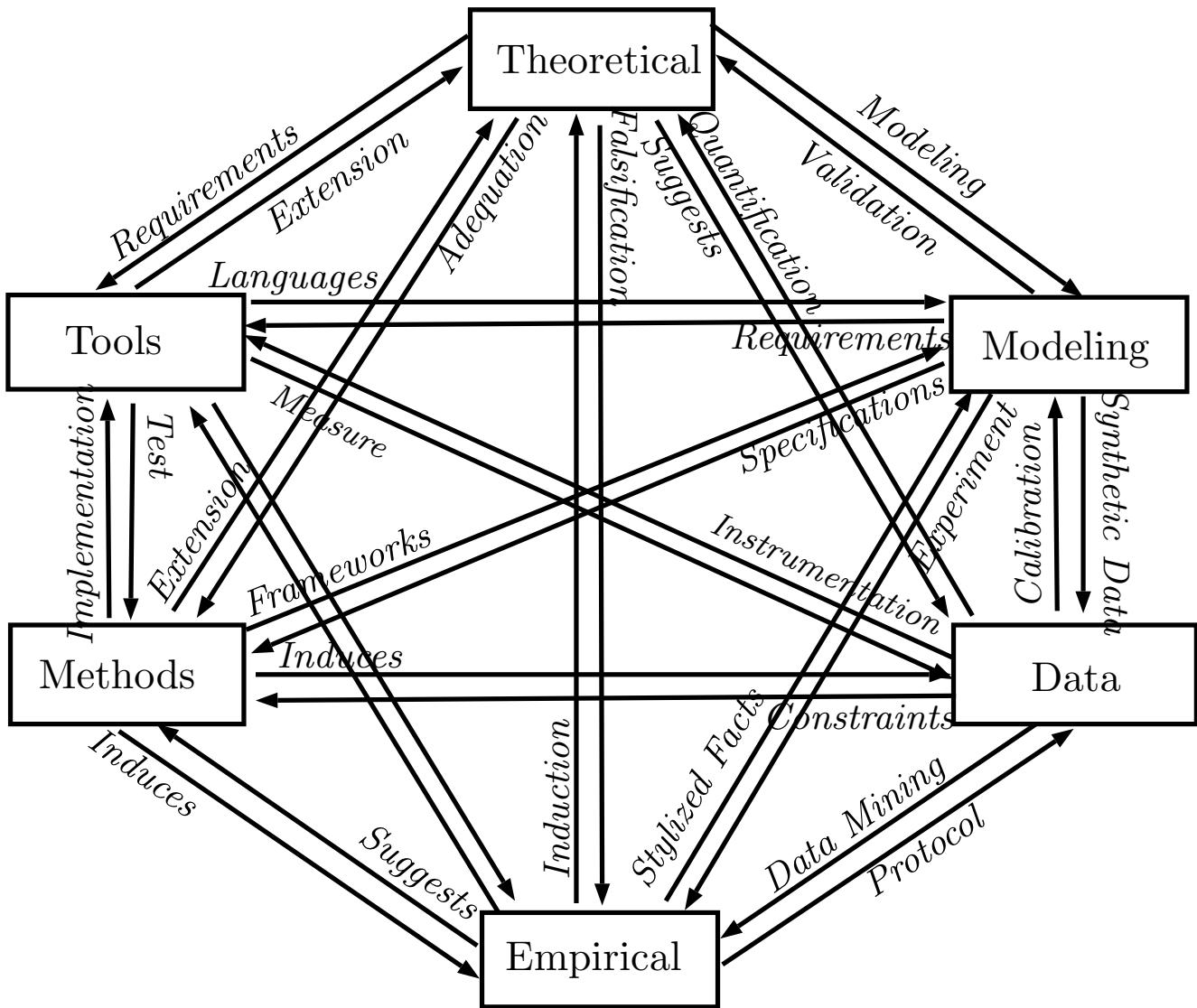


FIGURE 60 : **Projection d'une perspective comme un réseau complet des domaines de connaissance.** Pour illustrer les domaines et les processus d'interaction possibles entre ceux-ci, nous faisons l'exercice d'essayer de qualifier toutes les relations binaires possibles entre les domaines. Cela ne reflète 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.

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 [GOLDEN, AIGUER et KROB, 2012] qui donne un moyen pratique pour la représenter et pour introduire les échelles de temps et les

données, à laquelle est associée une ontologie O au sens de [LIVET et al., 2010a], 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 ($O_j \succ O_i$ 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 [GOLDEN, AIGUIER et KROB, 2012], 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 pos-

ture 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 imposée⁹. 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.

C : sur la communication pour l'extérieur [MARTINEZ-CONDE et MACKNIK, 2017]

La construction de théories géographiques, dans le cadre d'une Géographie Théorique et Quantitative, s'effectue par itérations dans une dynamique de co-évolution avec les efforts empiriques et de modélisation [LIVET et al., 2010a]. Parmi les nombreux exemples, on peut citer la théorie évolutive des villes (co-construite par un spectre de travaux s'étendant par exemple des premières propositions de [PUMAIN, 1997] jusqu'aux résultats matures présentés dans [PUMAIN, 2012a]), l'étude du caractère fractal des structures urbaines (par exemple de [FRANKHAUSER, 1998] à [FRANKHAUSER, 2008]) ou plus récemment le projet Transmondyn visant à enrichir la notion de transition des systèmes de peuplement (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

⁹ Ce qui relève bien sûr par ailleurs d'une problématique bien plus complexe que la simple audience [TARDY, 2004] et la richesse des pensées scientifiques permises par l'utilisation de différentes langues n'est pas discutable ainsi que la légitimité d'organisations comme l'ASRDLF. Mais c'est bien cette audience qui nous pose problème ici et dans ce cas il est quasiment aussi vieux jeu pour une école doctorale d'imposer le français comme langue d'écriture que le discours du consul et son snobisme d'énarque rapportés en 1.2.

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, 2009a], la littérature est très pauvre en modèles de simulation endogenisant cette co-évolution. Une exploration algorithmique de la littérature a été menée dans [RAIMBAULT, 2015a], suggérant un cloisonnement des domaines scientifiques s'intéressant à ce sujet. Des méthodes plus élaborées ainsi que les outils correspondants (collecte et analyse des données), couplant une analyse sémantique au réseau de citations, ont été développées pour renforcer ces conclusions préliminaires [RAIMBAULT, 2016b], 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 [RAIMBAULT, 2016a]. Cet aspect a aussi été mis en évidence dans l'espace et le temps à une échelle microscopique lors de l'étude des dynamiques d'un système de transport [RAIMBAULT, 2016c], conjointement avec l'hétérogénéité des processus pour un autre type de système [RAIMBAULT, 2015b]. 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 exer-

cice méthodologique explorant le possibilité de contrôle au second ordre de la structure de données synthétiques [RAIMBAULT, 2016], 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 [RAIMBAULT et GONZALEZ, May 2015], pour comparer par exemple des modèles de croissance de réseau routier basés sur l'optimisation locale à des modèles inspirés des réseaux biologiques : chacun présente une très grande variété de topologies générées. A l'échelle macroscopique, un modèle simple de croissance urbaine calibré dynamiquement sur les villes françaises de 1830 à 2000 (base Pumain-Ined) a permis de démontrer l'existence d'un effet réseau de par l'augmentation de pouvoir explicatif du modèle lors de l'ajout d'un effet des flux transitant par un réseau physique, tout en corrigeant le gain dû à l'ajout de paramètres par la construction d'un Critère d'Information d'Akaike empirique [RAIMBAULT, 2016d]. Cet ensemble de modèles se positionne avec un objectif de parcimonie et dans une perspective d'application en multi-modélisation. Dans une démarche basée-agent plus descriptive et donc par un modèle plus complexe, [LE NÉCHET et RAIMBAULT, 2015] 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 [ANTELOPE et al., 2016]. Pour notre part, la morphogénèse consiste en l'émergence de la forme et de la fonction, via des processus locaux autonomes dans un système qui exhibe alors une architecture auto-organisée. La présence d'une fonction et donc d'une architecture distingue les systèmes morphogénétiques de systèmes simplement auto-organisés (voir [DOURSAT, SAYAMA et MICHEL, 2012]). De plus, les notions d'autonomie et de localité s'appliquent

bien à des systèmes territoriaux, pour lesquels on essaye d'isoler les sous-systèmes et les échelles pertinentes. Les travaux sur la génération de forme urbaine calibrée par des processus autonomes, les premiers travaux sur la génération de réseaux par de multiples processus également autonomes, et des travaux plus anciens étudiant un modèle simple de morphogenèse urbaine qui suffisait à reproduire des motifs de forme stylisés [RAIMBAULT, BANOS et DOURSAT, 2014], 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 [RAFFESTIN, 1988], 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 [DUPUY, 1987]. 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 B.7, 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

The conclusion proposes to extract fundamental open questions that future research on territorial complex systems will have to tackle in the incoming decades.

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

C : sur l'evidence-based : même le subjectif est objectif en un sens ? question d'honnêteté 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.

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 spéculation 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 [BANOS, 2017]. 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 théo-

riques, les données, les outils), mais bien le développement conjoint de chaque composante. Nous étendons le cadre de connaissances de [LIVET et al., 2010b], qui consacre le triptyque des domaines empiriques, 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 [GIERE, 2010c], est une combinaison complexe des six domaines, les fronts de connaissance dans chacun étant en co-évolution. 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

C : lister les principaux contributeurs etc.; quoi est compatible avec quoi quest ce quon pourrait coupler etc; faire analyse epistemo quanti.

C : add somewhere something on the link “more systematic evidence-based”-politics in science - less dogmatism. or what place for evidence-based research in social science ? linked with quanti-quali : BEYOND classical separations, evidence-based and complex systems allow integration, socially responsible, but evidence-based and systematic..

C : in link “Complexity, Complexities, and Complex Knowledges”, importance of Nature of Complexity ?

C (JR) : evoquer ouverture des cours, formation interdisciplinaire etc. : pas ici, plutot en ouverture finale ?

NON-STATIONNARITÉ, NON-ERGODICITÉ ET DÉPENDANCE AU CHEMIN

COUPLAGE DES MODÈLES ET APPROCHES **C : different approaches to coupling / coupling to a certain degree using Kolmogorov etc : specific section or insert here ?**

CONSTRUIRE DES OUTILS DE VALIDATION POUR LES MODÈLES DE SIMULATION

PISTÉMOLOGIE QUANTITATIVE ET EXPÉRIMENTALE POUR UNE INTÉGRATION EFFECTIVE Le mantra du mariage entre qualitatif et quantitatif est asséné mécaniquement par de nombreux auteurs, mais lorsqu'il s'agit de mise en application, on peut se permettre de

soupçonner dans le meilleur des cas une naïveté, dans le pire des cas une hypocrisie. Quel sens à faire semblant de faire des analyses quantitatives en tartinant des pages de régression linéaires dont le R² ne dépasse pas 0.1 ? Quel sens à faire semblant de détenir une connaissance qualitative fine pour justifier la mise en place de modèles relevant de l'usine à gaz technocratique ?¹⁰

POUR UNE SCIENCE TOTALEMENT OUVERTE C (JR) : [FECHER et FRIESIKE, 2014] five schools in OpenScience : different dimensions or approaches

C : brosser ici directions vers lesquelles travailler; intégrer faits dans positionements

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

¹⁰ cette remarque est partiellement une auto-critique, puisqu'il faut rappeler le caractère très peu qualitatif de notre travail

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.

BIBLIOGRAPHIE

- AAGE, Niels et al. (2017). "Giga-voxel computational morphogenesis for structural design". In : *Nature* 550.7674, p. 84–86. URL : <http://dx.doi.org/10.1038/nature23911>.
- ABADIE, Alberto et al. (2010). "Synthetic control methods for comparative case studies : Estimating the effect of California's tobacco control program". In : *Journal of the American Statistical Association* 105.490.
- ABBAS, Assad et al. (2014). "A literature review on the state-of-the-art in patent analysis". In : *World Patent Information* 37, p. 3–13.
- ABERCROMBIE, Michael (1977). "Concepts in morphogenesis". In : *Proceedings of the Royal Society of London B : Biological Sciences* 199.1136, p. 337–344.
- ACEMOGLU Daron, Akcigit Ufuk et William KERR (2016). "Innovation Network". In : *Proceedings of the National Academy of Sciences (forthcoming)*.
- ACHIBET, Merwan et al. (2014). "A Model of Road Network and Buildings Extension Co-evolution". In : *Procedia Computer Science* 32, p. 828–833.
- ADAMATZKY, Andrew et Jeff JONES (2010). "Road planning with slime mould : if Physarum built motorways it would route M6/M74 through Newcastle". In : *International Journal of Bifurcation and Chaos* 20.10, p. 3065–3084.
- ADAMS, Stephen (2010). "The text, the full text and nothing but the text : Part 1 - Standards for creating textual information in patent documents and general search implications". In : *World Patent Information* 32.1, p. 22–29. URL : <https://ideas.repec.org/a/eee/worpat/v32y2010i1p22-29.html>.
- AGHION, Philippe et Peter HOWITT (1992). "A Model of Growth through Creative Destruction". In : *Econometrica* 60.2, p. 323–51. URL : <https://ideas.repec.org/a/ecm/emetrp/v60y1992i2p323-51.html>.
- AGHION, Philippe et al. (1998). *Endogenous growth theory*. MIT press.
- AGHION, Philippe et al. (2015). *Innovation and Top Income Inequality*.
- AKAIKE, Hirotugu (1998). "Information theory and an extension of the maximum likelihood principle". In : *Selected Papers of Hirotugu Akaike*. Springer, p. 199–213.
- AKCIGIT, Ufuk et al. (2013). *The Mechanics of Endogenous Innovation and Growth : Evidence from Historical US Patents*. Rapp. tech. Citeseer.
- ALBA, Martha de et D Miguel Ángel AGUILAR (2012). "Déplacements urbains et interaction sociale : le cas du système de transport collectif par métro dans la ville de Mexico". In : *Bulletin de psychologie* 1, p. 19–32.

- ALI, A. et al. (June 2014). *Les Eco-quartiers lus par la mobilité : vers une évaluation intégrée*. Rapp. tech. Ecole des Ponts ParisTech.
- ALLEN, Benjamin et al. (2017). "Multiscale Information Theory and the Marginal Utility of Information". In : *Entropy* 19.6, p. 273.
- ALLEN, P. et M. SANGLIER (1979). "A dynamic model of growth in a central place system". In : *Geographical Analysis* 11, p. 256–272.
- AMSDEN, Alice H (1994). "Why isn't the whole world experimenting with the East Asian model to develop ? : Review of the East Asian miracle". In : *World Development* 22.4, p. 627–633.
- ANAS, Alex et al. (1998). "Urban Spatial Structure". English. In : *Journal of Economic Literature* 36.3, pp. 1426–1464. ISSN : 00220515. URL : <http://www.jstor.org/stable/2564805>.
- ANDERSON, Philip W (1972). "More is different". In : *Science* 177.4047, p. 393–396.
- ANDERSSON, Claes et al. (2006). "A complex network approach to urban growth". In : *Environment and Planning A* 38.10, p. 1941.
- ANDERSSON, Claes et al. (2002). "Urban growth simulation from "first principles"". In : *Physical Review E* 66.2, p. 026204.
- ANGRIST, Joshua D et al. (1996). "Identification of causal effects using instrumental variables". In : *Journal of the American statistical Association* 91.434, p. 444–455.
- ANTELOPE, Chenling et al. (2016). "An Interdisciplinary Approach to Morphogenesis". In : *Forthcoming in Proceedings of Santa Fe Institute CSSS 2016*.
- ARCAUTE, E. et al. (2013). "Constructing cities, deconstructing scaling laws". In : ArXiv e-prints. arXiv : 1301.1674 [physics.soc-ph].
- ARCAUTE, Elsa et al. (2015). "Constructing cities, deconstructing scaling laws". In : *Journal of The Royal Society Interface* 12.102, p. 20140745.
- ARCHER, Margaret S. et Margaret S. ARCHER (1999). *Margaret Archer on structural and cultural morphogenesis*. JSTOR. (Visité le 02/01/2017).
- ARCHIBUGI, Daniele et Mario PIANTA (1992). "Specialization and size of technological activities in industrial countries : The analysis of patent data". In : *Research Policy* 21.1, p. 79 –93. ISSN : 0048-7333. DOI : [http://dx.doi.org/10.1016/0048-7333\(92\)90028-3](http://dx.doi.org/10.1016/0048-7333(92)90028-3). URL : <http://www.sciencedirect.com/science/article/pii/0048733392900283>.
- ARTHUR, W. Brian (2015). *Complexity and the Shift in Modern Science*. Conference on Complex Systems, Tempe, Arizona.
- ASH, J et D NEWTH (2007). "Optimizing complex networks for resilience against cascading failure". In : *Physica A : Statistical Mechanics and its Applications* 380, p. 673–683.
- AUDRETSCH, David B et Maryann P FELDMAN (1996). "R&D spillovers and the geography of innovation and production". In : *The American economic review* 86.3, p. 630–640.
- AXTELL, Robert L (2016). "120 million agents self-organize into 6 million firms : a model of the US private sector". In : *Proceedings*

- of the 2016 International Conference on Autonomous Agents & Multiagent Systems.* International Foundation for Autonomous Agents et Multiagent Systems, p. 806–816.
- BADARIOTTI, Dominique et al. (2007). "Conception d'un automate cellulaire non stationnaire à base de graphe pour modéliser la structure spatiale urbaine : le modèle Remus". In : *Cybergeo : European Journal of Geography*.
- BAFFI, Solène (2016). "Railways and city in territorialization processes in South Africa : from separation to integration ?" Theses. Université Paris 1 - Panthéon Sorbonne. URL : <https://halshs.archives-ouvertes.fr/tel-01389347>.
- BAIS, Sander et al. (2010). *In praise of science : curiosity, understanding, and progress*. MIT Press.
- BALBO, Flavien et al. (2016). "Positionnement des systèmes multi-agents pour les systèmes de transport intelligents". In : *Revue des Sciences et Technologies de l'Information-Série RIA : Revue d'Intelligence Artificielle* 30.3, p. 299–327.
- BALL, Stephen J (1990). "Self-doubt and soft data : social and technical trajectories in ethnographic fieldwork". In : *International Journal of Qualitative Studies in Education* 3.2, p. 157–171.
- BANOS, Arnaud (2001). "A propos de l'analyse spatiale exploratoire des données". In : *Cybergeo : European Journal of Geography*.
- (2013). "Pour des pratiques de modélisation et de simulation libérées en Géographies et SHS". In : *HDR. Université Paris 1*.
 - (2017). "Knowledge Accelerator' in Geography and Social Sciences : Further and Faster, but Also Deeper and Wider". In : sous la dir. de Denise PUMAIN et Romain REUILLOU. in *Urban Dynamics et Simulation Models*. Springer.
 - (Décembre 2013). "Pour des pratiques de modélisation et de simulation libérées en Géographie et SHS". In : *Thèse d'Habilitation à Diriger des Recherches, UMR CNRS 8504 Géographie-Cités, ISCPif*.
- BANOS, Arnaud et Cyrille GENRE-GRANDPIERRE (2012). "Towards new metrics for urban road networks : Some preliminary evidence from agent-based simulations". In : *Agent-based models of geographical systems*. Springer, p. 627–641.
- BAPTISTE, Hervé (1999). "Interactions entre le système de transport et les systèmes de villes : perspective historique pour une modélisation dynamique spatialisée". Thèse de doct. Centre d'études supérieures de l'aménagement (Tours).
- (2010). "Modeling the Evolution of a Transport System and its Impacts on a French Urban System". In : *Graphs and Networks : Multilevel Modeling, Second Edition*, p. 67–89.
- BARABASI, Albert-Laszlo (2002). "Linked : How everything is connected to everything else and what it means". In : *Plume Editors*.
- BARNDORFF-NIELSEN, Ole E et al. (2011). "Multivariate realised kernels : consistent positive semi-definite estimators of the covaria-

- tion of equity prices with noise and non-synchronous trading". In : *Journal of Econometrics* 162, p. 149–169.
- BARRICO, C. et C.H. ANTUNES (2006). "Robustness Analysis in Multi-Objective Optimization Using a Degree of Robustness Concept". In : *Evolutionary Computation, 2006. CEC 2006. IEEE Congress on*, p. 1887–1892. DOI : [10.1109/CEC.2006.1688537](https://doi.org/10.1109/CEC.2006.1688537).
- BARTHÉLEMY, Marc (2011). "Spatial networks". In : *Physics Reports* 499.1, p. 1–101.
- BARTHELEMY, Marc (2016). *The Structure and Dynamics of Cities*. Cambridge University Press.
- BARTHÉLEMY, Marc et Alessandro FLAMMINI (2008). "Modeling urban street patterns". In : *Physical review letters* 100.13, p. 138702.
- (2009). "Co-evolution of density and topology in a simple model of city formation". In : *Networks and spatial economics* 9.3, p. 401–425.
- BARTHELEMY, Marc et al. (2013). "Self-organization versus top-down planning in the evolution of a city". In : *Scientific reports* 3.
- BASTANI, O. et al. (2017). "Interpretability via Model Extraction". In : *ArXiv e-prints*. arXiv : [1706.09773 \[cs.LG\]](https://arxiv.org/abs/1706.09773).
- BASTIAN, Hilda et al. (2010). "Seventy-five trials and eleven systematic reviews a day : how will we ever keep up ?" In : *PLoS medicine* 7.9, e1000326.
- BATTISTON, F. et al. (2015). "Emergence of multiplex communities in collaboration networks". In : *ArXiv e-prints*. arXiv : [1506 . 01280 \[physics.soc-ph\]](https://arxiv.org/abs/1506.01280).
- BATTY, M. (2017). "The Age of the Smart City". In : *CASA Working Paper*, DOI : [10.13140/RG.2.2.34783.87204](https://doi.org/10.13140/RG.2.2.34783.87204).
- BATTY, Michael (1991). "Generating urban forms from diffusive growth". In : *Environment and Planning A* 23.4, p. 511–544.
- (2006). "Hierarchy in cities and city systems". In : *Hierarchy in natural and social sciences*. Springer, p. 143–168.
- (2007). *Cities and complexity : understanding cities with cellular automata, agent-based models, and fractals*. The MIT press.
- (2013a). "Big data, smart cities and city planning". In : *Dialogues in Human Geography* 3.3, p. 274–279.
- (2013b). *The new science of cities*. Mit Press.
- (2016). "Theoretical filters : Reducing explanations in cities to their very essence". In : *Environment and Planning B : Planning and Design* 43.5, p. 797–799.
- BATTY, Michael et Paul A LONGLEY (1986). "The fractal simulation of urban structure". In : *Environment and Planning A* 18.9, p. 1143–1179.
- (1994). *Fractal cities : a geometry of form and function*. Academic Press.

- BATTY, Michael et S MACKIE (1972). "The calibration of gravity, entropy, and related models of spatial interaction". In : *Environment and Planning A* 4.2, p. 205–233.
- BATTY, Michael et Yichun XIE (1994). "From cells to cities". In : *Environment and planning B : Planning and design* 21.7, S31–S48.
- BAVOUX, Jean-Jacques et al. (2005). *Géographie des transports*. Paris.
- BAZIN, Sylvie et al. (2007). "L'évolution des marchés immobiliers résidentiels dans l'aire urbaine de Reims : un effet de la Ligne à Grande Vitesse Est-européenne?" In : *Congress of the European Regional Science Association and ASRDLF, Parigi*.
- BAZIN, Sylvie et al. (2010). "Lignes ferroviaires à grande vitesse et dynamiques locales : une analyse comparée de la littérature". In : *Transport et développement des territoires*. Sous la dir. de Comité National de GÉOGRAPHIE. Le Havre, France, 21p. URL : <https://hal.archives-ouvertes.fr/hal-00615196>.
- BAZIN, Sylvie et al. (2011). "Grande vitesse ferroviaire et développement économique local : une revue de la littérature". In : *Recherche Transports Sécurité* 27.3, p. 215–238.
- BEAUCIRE, Francis et Matthieu DREVELLE (2013). "«Grand Paris Express» : un projet au service de la réduction des inégalités d'accèsibilité entre l'Ouest et l'Est de la région urbaine de Paris?" In : *Revue d'Économie Régionale & Urbaine* 3, p. 437–460.
- BEAUGUITTE, Laurent (2014). *Croissance et décroissance des réseaux*. Sous la dir. de Laurent BEAUGUITTE. Groupe fmr, p. 140. URL : <https://halshs.archives-ouvertes.fr/halshs-01068589>.
- BEDAU, Mark (2002). "Downward causation and the autonomy of weak emergence". In : *Principia : an international journal of epistemology* 6.1, p. 5–50.
- BELMONTE, Mylène et al. (2008). "Automatisation intégrale de la ligne 1 : étude et modélisation du trafic mixte". In : *Lambda-Mu*, Session-5B.
- BENGUIGUI, Lucien et Efrat BLUMENFELD-LIEBERTHAL (2007). "A dynamic model for city size distribution beyond Zipf's law". In : *Physica A : Statistical Mechanics and its Applications* 384.2, p. 613–627.
- BENNETT, Jonathan (2010). *OpenStreetMap*. Packt Publishing Ltd.
- BERGEAUD, Antonin et al. (2017). "Classifying patents based on their semantic content". In : *PloS one* 12.4, e0176310.
- BERNE, Laurence (2008). "Ouverture et fermeture de territoire par les réseaux de transports dans trois espaces montagnards (Bugey, Bauges et Maurienne)". Thèse de doct. Université de Savoie.
- BERNIER, Xavier (2007). "Les dynamiques réticulo-territoriales et la frontière en zone de montagne : approche typologique". In : *Flux* 4, p. 8–19.
- BERROIR, Sandrine et al. (2005). *La contribution des villes nouvelles au polycentrisme francilien*. UMR Géographie-cités.

- BERRY, Brian JL (1964). "Cities as systems within systems of cities". In : *Papers in Regional Science* 13.1, p. 147–163.
- BETTENCOURT, L. M. A. et J. LOBO (2015). "Urban Scaling in Europe". In : *ArXiv e-prints*. arXiv : 1510.00902 [physics.soc-ph].
- BETTENCOURT, Luís MA et al. (2008). "Why are large cities faster? Universal scaling and self-similarity in urban organization and dynamics". In : *The European Physical Journal B-Condensed Matter and Complex Systems* 63.3, p. 285–293.
- BETTENCOURT, Luís MA et al. (2007). "Growth, innovation, scaling, and the pace of life in cities". In : *Proceedings of the national academy of sciences* 104.17, p. 7301–7306.
- BIERNACKI, Christophe et al. (2000). "Assessing a mixture model for clustering with the integrated completed likelihood". In : *IEEE transactions on pattern analysis and machine intelligence* 22.7, p. 719–725.
- BIGOTTE, João F et al. (2010). "Integrated modeling of urban hierarchy and transportation network planning". In : *Transportation Research Part A : Policy and Practice* 44.7, p. 506–522.
- BIRD, Steven (2006). "NLTK : the natural language toolkit". In : *Proceedings of the COLING/ACL on Interactive presentation sessions*. Association for Computational Linguistics, p. 69–72.
- BITBOL, M. et P. L. LUISI (2004). "Autopoiesis with or without cognition : defining life at its edge". en. In : *Journal of The Royal Society Interface* 1.1, p. 99–107. ISSN : 1742-5689, 1742-5662. DOI : 10.1098/rsif.2004.0012. (Visité le 03/01/2017).
- BLANQUART, Corinne et Martin KONING (2017). "The local economic impacts of high-speed railways : theories and facts". In : *European Transport Research Review* 9.2, p. 12. ISSN : 1866-8887. DOI : 10.1007/s12544-017-0233-0. URL : <https://doi.org/10.1007/s12544-017-0233-0>.
- BLEI, David M et al. (2003). "Latent dirichlet allocation". In : *Journal of machine Learning research* 3.Jan, p. 993–1022.
- BLOCK-SCHACHTER, David (2012). "Hysteresis and urban rail : The effects of past urban rail on current residential and travel choices". Thèse de doct. Massachusetts Institute of Technology.
- BLOOM, Nicholas et al. (2013). "Identifying Technology Spillovers and Product Market Rivalry". In : *Econometrica* 81.4, p. 1347–1393. URL : <https://ideas.repec.org/a/ecm/emetrp/v81y2013i4p1347-1393.html>.
- BLUMENFELD-LIEBERTHAL, Efrat et Juval PORTUGALI (2010). "Network cities : A complexity-network approach to urban dynamics and development". In : *Geospatial Analysis and Modelling of Urban Structure and Dynamics*. Springer, p. 77–90.
- BOHANNON, John (2014). "Scientific publishing. Google Scholar wins raves—but can it be trusted?" In : *Science (New York, NY)* 343.6166, p. 14.

- BOLLEN, Johan et al. (2014). "From funding agencies to scientific agency". In : *EMBO reports* 15.2, p. 131–133.
- BOLÓN-CANEDO, Verónica et al. (2013). "A review of feature selection methods on synthetic data". In : *Knowledge and information systems* 34.3, p. 483–519.
- BOLOTIN, A. (2014). "Computational solution to quantum foundational problems". In : *ArXiv e-prints*. arXiv : 1403.7686 [quant-ph].
- BONANNO, G. et al. (2001). "Levels of complexity in financial markets". In : *Physica A Statistical Mechanics and its Applications* 299, p. 16–27. eprint : cond-mat/0104369.
- BONIN, Olivier, Jean-Paul HUBERT et al. (2012). "Modèle de morphogénèse urbaine : simulation d'espaces qualitativement différenciés dans le cadre du modèle de l'économie urbaine". In : *49è colloque de l'ASRDLF*.
- BONNAFOUS, Alain et François PLASSARD (1974). "Les méthodologies usuelles de l'étude des effets structurants de l'offre de transport". In : *Revue économique*, p. 208–232.
- BONNAFOUS, Alain et al. (1974). "La detection des effets structurants d'autoroute : Application à la Vallée du Rhône". English. In : *Revue économique* 25.2, pp. 233–256. ISSN : 00352764. URL : <http://www.jstor.org/stable/3500568>.
- BOSCH, F van den et al. (1990). "The velocity of spatial population expansion". In : *Journal of Mathematical Biology* 28.5, p. 529–565.
- BOTTINELLI, A. et al. (2016). "Balancing building and maintenance costs in growing transport networks". In : *ArXiv e-prints*. arXiv : 1609.06470 [physics.soc-ph].
- BOUCHAUD, J. P. et M. POTTERS (2009). "Financial Applications of Random Matrix Theory : a short review". In : *ArXiv e-prints*. arXiv : 0910.1205 [q-fin.ST].
- BOUCHAUD, J-P et al. (2000). "Apparent multifractality in financial time series". In : *The European Physical Journal B-Condensed Matter and Complex Systems* 13.3, p. 595–599.
- BOURGINE, P. et al. (2009). "French Roadmap for complex Systems 2008-2009". In : *ArXiv e-prints*. arXiv : 0907.2221 [nlin.AO].
- BOURGINE, Paul et Annick LESNE (2010). *Morphogenesis : origins of patterns and shapes*. Springer Science & Business Media.
- BOURGINE, Paul et John STEWART (2004). "Autopoiesis and cognition". In : *Artificial life* 10.3, p. 327–345.
- BOUTEILLER, Catherine et Sybille BERJOAN (2013). "Open data en transport urbain : quelles sont les données mises à disposition ? Quelles sont les stratégies des autorités organisatrices ?" In :
- BRAND, Christian et al. (2013). "Accelerating the transformation to a low carbon passenger transport system : The role of car purchase taxes, feebates, road taxes and scrappage incentives in the UK". In : *Transportation Research Part A : Policy and Practice* 49, p. 132–148.

- BRETAGNOLLE, Anne (2009a). "Villes et réseaux de transport : des interactions dans la longue durée, France, Europe, États-Unis". Français. HDR. Université Panthéon-Sorbonne - Paris I. URL : <http://tel.archives-ouvertes.fr/tel-00459720>.
- (2009b). "Villes et réseaux de transport : des interactions dans la longue durée, France, Europe, États-Unis". Français. HDR. Université Panthéon-Sorbonne - Paris I. URL : <http://tel.archives-ouvertes.fr/tel-00459720>.
- BRETAGNOLLE, Anne et al. (2006). "From theory to modelling : urban systems as complex systems". In : *CyberGeo : European Journal of Geography*.
- BRETAGNOLLE, Anne et al. (2002). "Time and space scales for measuring urban growth". In : *Cybergeo : European Journal of Geography*.
- BRETAGNOLLE, Anne et Denise PUMAIN (2010a). "Comparer deux types de systèmes de villes par la modélisation multi-agents". In : *Qu'appelle t-on aujourd'hui les sciences de la complexité ? Langages, réseaux, marchés, territoires*, p. 271–299.
- (2010b). "Simulating Urban Networks through Multiscalar Space-Time Dynamics : Europe and the United States, 17th-20th Centuries". In : *Urban Studies* 47.13, p. 2819–2839. DOI : [10.1177/0042098010377366](https://doi.org/10.1177/0042098010377366). eprint : <http://dx.doi.org/10.1177/0042098010377366>. URL : <http://dx.doi.org/10.1177/0042098010377366>.
- BRETAGNOLLE, Anne et al. (1998). "Space-time contraction and the dynamics of urban systems". In : *Cybergeo : European Journal of Geography*.
- BRETAGNOLLE, Anne et al. (2009). "The organization of urban systems". In : *Complexity perspectives in innovation and social change*. Springer, p. 197–220.
- BRETAGNOLLE, Anne et al. (2000). "Long-term dynamics of European towns and cities : towards a spatial model of urban growth". In : *Cybergeo : European Journal of Geography*.
- BRETAGNOLLE, Anne et al. (2016). "La ville à l'échelle de l'Europe- Apports du couplage et de l'expertise de bases de données issues de l'imagerie satellitaire". In : *Revue Internationale de Géomatique* 26.1, p. 55–78.
- BROWN, Matthew J (2009). "Models and perspectives on stage : remarks on Giere's scientific perspectivism". In : *Studies in History and Philosophy of Science Part A* 40.2, p. 213–220.
- BRUCK, Péter et al. (2016). "Recognition of emerging technology trends : class-selective study of citations in the US Patent Citation Network". In : *Scientometrics* 107.3, p. 1465–1475.
- BRUNSDON, Chris et al. (1996). "Geographically weighted regression : a method for exploring spatial nonstationarity". In : *Geographical analysis* 28.4, p. 281–298.

- BRUNSDON, Chris et al. (1998). "Geographically weighted regression". In : *Journal of the Royal Statistical Society : Series D (The Statistician)* 47.3, p. 431–443.
- BULCKE, Tim Van den et al. (2006). "SynTReN : a generator of synthetic gene expression data for design and analysis of structure learning algorithms". In : *BMC bioinformatics* 7.1, p. 43.
- BURKE, Nuala T (1972). "Dublin 1600-1800 : a study in urban morphogenesis". In :
- CANNING, DAVID et PETER PEDRONI (2008). "INFRASTRUCTURE, LONG-RUN ECONOMIC GROWTH AND CAUSALITY TESTS FOR COINTEGRATED PANELS". In : *The Manchester School* 76.5, p. 504–527. ISSN : 1467-9957. DOI : [10.1111/j.1467-9957.2008.01073.x](https://doi.org/10.1111/j.1467-9957.2008.01073.x). URL : <http://dx.doi.org/10.1111/j.1467-9957.2008.01073.x>.
- CAMERER, Colin F et al. (2016). "Evaluating replicability of laboratory experiments in economics". In : *Science*, aaf0918.
- CARLILE, Paul R (2004). "Transferring, translating, and transforming : An integrative framework for managing knowledge across boundaries". In : *Organization science* 15.5, p. 555–568.
- CARROUET, Guillaume et al. (2013). "Les stratégies d'anticipation des "effets" territoriaux des grands équipements de transport : le cas du TGV Rhin-Rhône". In : *Géotransports* 1-2, p. 35–50.
- CARVER, Stephen J (1991). "Integrating multi-criteria evaluation with geographical information systems". In : *International Journal of Geographical Information System* 5.3, p. 321–339.
- CECCARINI, Patrice (2001). "Essai de formalisation dynamique de la cathédrale gothique : morphogénèse et modélisation de la Basilique Saint-Denis : les relations entre théologie, sciences et architecture au XIIIème siècle à Saint-Denis". Thèse de doct. Paris, EHESS.
- CERQUEIRA, Eugênia Viana (2017). "Les inégalités d'accès aux ressources urbaines dans les franges périphériques de Belo Horizonte (Brésil) : quelles évolutions ?" In : *EchoGéo* 39.
- CHALIDABHONGSE, Junavit et CC Jay Kuo (1997). "Fast motion vector estimation using multiresolution-spatio-temporal correlations". In : *Circuits and Systems for Video Technology, IEEE Transactions on* 7.3, p. 477–488.
- CHAMPOLLION, Pierre (2006). "TERRITORY AND TERRITORIALIZATION : PRESENT STATE OF THE CAENTI THOUGHT". In : *International Conference of Territorial Intelligence*. INTI-International Network of Territorial Intelligence. Alba Iulia, Romania, p51–58. URL : <https://halshs.archives-ouvertes.fr/halshs-00999026>.
- CHANG, Justin S (2006). "Models of the Relationship between Transport and Land-use : A Review". In : *Transport Reviews* 26.3, p. 325–350.

- CHAPMAN, Michael J et Lynn MARGULIS (1998). "Morphogenesis by symbiogenesis". In : *International Microbiology* 1.4.
- CHASSET, Pierre-Olivier et al. (2016). *cybergeo20 v1.0*. DOI : [10.5281/zenodo.53905](https://doi.org/10.5281/zenodo.53905). URL : <http://dx.doi.org/10.5281/zenodo.53905>.
- CHAVALARIAS, David (2016). "What's wrong with Science?" In : *Scientometrics*, p. 1–23.
- CHAVALARIAS, David et Jean-Philippe COINTET (2013). "Phylomematic patterns in science evolution—the rise and fall of scientific fields". In : *Plos One* 8.2, e54847.
- CHAVALARIAS, David et al. (2005). "Nobel, Le Jeu De La Decouverte Scientifique". In :
- CHEN, Duan-Rung et Khoa TRUONG (2012). "Using multilevel modeling and geographically weighted regression to identify spatial variations in the relationship between place-level disadvantages and obesity in Taiwan". In : *Applied Geography* 32.2, p. 737–745.
- CHEN, Wenling et David M LEVINSON (2006). "Effectiveness of learning transportation network growth through simulation". In : *Journal of Professional Issues in Engineering Education and Practice* 132.1, p. 29–41.
- CHEN, Y. (2016). "Normalizing and Classifying Shape Indexes of Cities by Ideas from Fractals". In : *ArXiv e-prints*. arXiv : [1608.08839 \[physics.soc-ph\]](https://arxiv.org/abs/1608.08839).
- CHEN, Yanguang (2009). "Urban gravity model based on cross-correlation function and Fourier analyses of spatio-temporal process". In : *Chaos, Solitons & Fractals* 41.2, p. 603–614.
- (2010). "Characterizing growth and form of fractal cities with allometric scaling exponents". In : *Discrete Dynamics in Nature and Society* 2010.
- CHÉREL, Guillaume et al. (2015). "Beyond Corroboration : Strengthening Model Validation by Looking for Unexpected Patterns". In : *PLoS ONE* 10.9, e0138212. DOI : [10.1371/journal.pone.0138212](https://doi.org/10.1371/journal.pone.0138212). URL : <http://dx.doi.org/10.1371%2Fjournal.pone.0138212>.
- CHICHEPORTICHE, Rémy et Jean-Philippe BOUCHAUD (2013). "A nested factor model for non-linear dependences in stock returns". In : *arXiv preprint arXiv:1309.3102*.
- CHOAY, Françoise (2009). *Pour une anthropologie de l'espace*. Le Seuil.
- CHOI, Jinho et Yong-Sik HWANG (2014). "Patent keyword network analysis for improving technology development efficiency". In : *Technological Forecasting and Social Change* 83, p. 170–182.
- CHU, Dominique (2008). "Criteria for conceptual and operational notions of complexity". In : *Artificial Life* 14.3, p. 313–323.
- CLARKE, Keith C et Leonard J GAYDOS (1998). "Loose-coupling a cellular automaton model and GIS : long-term urban growth prediction for San Francisco and Washington/Baltimore". In : *International journal of geographical information science* 12.7, p. 699–714.

- CLARKE, Keith C et al. (2007). "A decade of SLEUTHing : Lessons learned from applications of a cellular automaton land use change model". In : *Classics in IJGIS : twenty years of the international journal of geographical information science and systems*, p. 413–427.
- CLAUSSET, Aaron et al. (2004). "Finding community structure in very large networks". In : *Physical review E* 70.6, p. 066111.
- CLAVAL, Paul (1985). "Causalité et géographie". In : *Espace géographique* 14.2, p. 109–115.
- (1987). "Réseaux territoriaux et enrangement". In : *Cahier/Groupe Réseaux* 3.7, p. 44–60.
- COLANDER, David (2003). *The complexity revolution and the future of economics*. Rapp. tech. Middlebury College, Department of Economics.
- COMBES, Pierre-Philippe et Miren LAFOURCADE (2005). "Transport costs : measures, determinants, and regional policy implications for France". In : *Journal of Economic Geography* 5.3, p. 319–349.
- COMMENGES, H (2013a). "The invention of daily mobility : Performative aspects of the instruments of economics of transportation." In : *Theses, Université Paris-Diderot-Paris VII*.
- COMMENGES, Hadrien (2013b). "The invention of daily mobility. Performative aspects of the instruments of economics of transportation." Theses. Université Paris-Diderot - Paris VII. url : <https://tel.archives-ouvertes.fr/tel-00923682>.
- COTTET, Nathanaël et al. (2017). "Observing a quantum Maxwell demon at work". In : *Proceedings of the National Academy of Sciences*, p. 201704827.
- COTTINEAU, C. (2016). "MetaZipf. (Re)producing knowledge about city size distributions". In : *ArXiv e-prints*. arXiv : 1606 . 06162 [physics.soc-ph].
- COTTINEAU, C et al. (2016). "Back to the Future of Multimodeling". In : *Royal Geographical Society-Annual Conference 2016-Session : Geocomputation, the Next 20 Years* (1).
- COTTINEAU, C. et al. (2015). "Paradoxical Interpretations of Urban Scaling Laws". In : *ArXiv e-prints*. arXiv : 1507 . 07878 [physics.soc-ph].
- COTTINEAU, Clementine (2014). "L'évolution des villes dans l'espace post-soviétique. Observation et modélisations." Thèse de doct. Université Paris 1 Panthéon-Sorbonne.
- COTTINEAU, Clémentine (2015). *Urban scaling : What cities are we talking about ?* Presentation of ongoing work at Quanturb seminar, April 1st 2015.
- COTTINEAU, Clémentine et al. (2015). "An incremental method for building and evaluating agent-based models of systems of cities". In :
- COTTINEAU, Clémentine et al. (2015a). "A modular modelling framework for hypotheses testing in the simulation of urbanisation". In : *Systems* 3.4, p. 348–377.

- COTTINEAU, Clémentine et al. (2015b). "Revisiting some geography classics with spatial simulation". In : *Plurimondi. An International Forum for Research and Debate on Human Settlements*. T. 7. 15.
- COTTINEAU, Clementine et al. (2017). "Initial spatial conditions in simulation models : the missing leg of sensitivity analyses?" In : *Geocomputation Conference*.
- COUCLELIS, Helen (1985). "Cellular worlds : a framework for modeling micro—macro dynamics". In : *Environment and planning A* 17.5, p. 585–596.
- COURTAT, Thomas et al. (2011). "Mathematics and morphogenesis of cities : A geometrical approach". In : *Physical Review E* 83.3, p. 036106.
- CRICK, T. et al. (2015). "Reproducibility in Research : Systems, Infrastructure, Culture". In : *ArXiv e-prints*. arXiv : 1503 . 02388 [cs.SE].
- CRONIN, Blaise et Cassidy R SUGIMOTO (2014). *Beyond bibliometrics : Harnessing multidimensional indicators of scholarly impact*. MIT Press.
- CROSATO, E. (2014). *Artificial Self-Assembly : Literature Review*. Rapp. tech. Vrije Universiteit Amsterdam.
- CROSATO, Emanuele et al. (2017). "Informative and misinformative interactions in a school of fish". In : *arXiv preprint arXiv :1705.01213*.
- CROSS, MC, PC HOHENBERG et al. (1994). "Spatiotemporal chaos". In : *Science-AAAS-Weekly Paper Edition-including Guide to Scientific Information* 263.5153, p. 1569–1569.
- CROZET, Yves et Francois DUMONT (2011a). "Retour sur les effets économiques du TGV. Les effets structurants sont un mythe (interview)". In : *Ville, Rail et Transports* 525, p. 48–51. URL : <https://halshs.archives-ouvertes.fr/halshs-01094554>.
- (2011b). "Retour sur les effets économiques du TGV. Les effets structurants sont un mythe (interview)". In : *Ville, Rail et Transports* 525, p. 48–51. URL : <https://halshs.archives-ouvertes.fr/halshs-01094554>.
- CURRAN, Clive-Steven et Jens LEKER (2011). "Patent indicators for monitoring convergence—examples from NFF and ICT". In : *Technological Forecasting and Social Change* 78.2, p. 256–273.
- CUSSAT-BLANC, Sylvain et al. (2012). "A synthesis of the Cell2Organ developmental model". In : *Morphogenetic Engineering*. Springer, p. 353–381.
- CUTHBERT, Angela L et al. (2005). "An empirical analysis of the relationship between road development and residential land development". In : *Canadian Journal of Regional Science* 28.1, p. 49–76.
- CUYALA, Sylvain (2014). "Analyse spatio-temporelle d'un mouvement scientifique. L'exemple de la géographie théorique et quantitative européenne francophone." Thèse de doct. Université Paris 1 Panthéon-Sorbonne.

- DAMM, David et al. (1980). "Response of urban real estate values in anticipation of the Washington Metro". In : *Journal of Transport Economics and Policy*, p. 315–336.
- DE DOMENICO, Manlio et al. (2015). "Ranking in interconnected multilayer networks reveals versatile nodes". In : *Nature communications* 6.
- DE LEON, FD et al. (2007). "NetLogo Urban Suite-Tijuana Border-towns model". In : *Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL*. URL : <http://ccl.northwestern.edu/netlogo/models/UrbanSuite-TijuanaBordertowns>.
- DE LUCA PICIONE, R. et M. FRANCESCA FREDA (2016). "The processes of meaning making, starting from the morphogenetic theories of René Thom". en. In : *Culture & Psychology* 22.1, p. 139–157. ISSN : 1354-067X, 1461-7056. DOI : [10.1177/1354067X15576171](https://doi.org/10.1177/1354067X15576171). (Visité le 02/01/2017).
- DE NADAI, Marco et al. (2016). "The Death and Life of Great Italian Cities : A Mobile Phone Data Perspective". In : *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, p. 413–423.
- DEB, Kalyanmoy et Himanshu GUPTA (2006). "Introducing robustness in multi-objective optimization". In : *Evolutionary Computation* 14.4, p. 463–494.
- DECELLE, Aurelien et al. (2011). "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications". In : *Physical Review E* 84.6, p. 066106.
- DECHEZLEPRÉ, Antoine et al. (2014). *Knowledge Spillovers from Clean and Dirty Technologies*. CEP Discussion Papers dp1300. Centre for Economic Performance, LSE. URL : <https://ideas.repec.org/p/cep/cepdps/dp1300.html>.
- DEFFUANT, Guillaume et al. (2015). "Visions de la complexité. Le démon de Laplace dans tous ses états". In : *Natures Sciences Sociétés* 23.1, p. 42–53.
- DELILE, Julien et al. (2016). "Chapitre 17. Modélisation multi-agent de l'embryogenèse animale". In : *Modélisations, simulations, systèmes complexes*, p. 581–624.
- DELONS, Jean et al. (2008). "PIRANDELLO an integrated transport and land-use model for the Paris area". URL : <https://halv3-preprod.archives-ouvertes.fr/hal-00319087>.
- DENG, Yi et Rongfang LIU (2007). "Potential Impact of Housing Policy on Transportation Infrastructure in Chinese Cities". In : *Transportation Research Record : Journal of the Transportation Research Board* 2038, p. 1–8.
- DESJARDINS, Xavier (2010). "la bataille du Grand Paris". In : *L'Information géographique* 74.4, p. 29–46.
- (2016). "Ce Grand Paris qui advient. Leçons pour la planification métropolitaine". In : *L'Information géographique* 80.4, p. 96–114.

- DESMARAIS, Gaëtan (1992). "Des prémisses de la théorie de la forme urbaine au parcours morphogénétique de l'établissement humain". In : *Cahiers de géographie du Québec* 36.98, p. 251–273.
- DI MEO, Guy (1998). "De l'espace aux territoires : éléments pour une archéologie des concepts fondamentaux de la géographie". In : *L'information géographique* 62.3, p. 99–110.
- DICK, Josef et Friedrich PILLICHSHAMMER (2010). *Digital nets and sequences : Discrepancy Theory and Quasi-Monte Carlo Integration*. Cambridge University Press.
- DIDEROT, Denis (1965). *Entretien entre d'Alembert et Diderot*. Garnier-Flammarion.
- DIETTERICH, Tom (1995). "Overfitting and undercomputing in machine learning". In : *ACM computing surveys (CSUR)* 27.3, p. 326–327.
- DING, Rui et al. (2017). "Heuristic urban transportation network design method, a multilayer coevolution approach". In : *Physica A : Statistical Mechanics and its Applications* 479, p. 71–83.
- DIRK, Lynn (1999). "A Measure of Originality The Elements of Science". In : *Social Studies of Science* 29.5, p. 765–776.
- DOBBIE, Melissa J et David DAIL (2013). "Robustness and sensitivity of weighting and aggregation in constructing composite indices". In : *Ecological Indicators* 29, p. 270–277.
- DOLLENS, Dennis (2014). "Alan Turing's Drawings, Autopoiesis and Can Buildings Think?" In : *Leonardo* 47.3, p. 249–254.
- DOLLFUS, O et F Durand DASTÈS (1975). "Some remarks on the notions of 'structure' and 'system' in geography". In : *Geoforum* 6.2, p. 83–94.
- DOURSAT, René (2008). "Programmable Architectures That Are Complex and Self-Organized-From Morphogenesis to Engineering." In : *ALIFE*, p. 181–188.
- DOURSAT, René et al. (2012). *Morphogenetic engineering : toward programmable complex systems*. Springer.
- (2013). "A review of morphogenetic engineering". In : *Natural Computing* 12.4, p. 517–535.
- DRAGOMIR, SS (1999). "The Ostrowski's integral inequality for Lipschitzian mappings and applications". In : *Computers & Mathematics with Applications* 38.11, p. 33–37.
- DRUMMOND, Chris (2009). "Replicability is not reproducibility : nor is it good science". In :
- DUCRUET, César et Laurent BEAUGUITTE (2014). "Spatial science and network science : Review and outcomes of a complex relationship". In : *Networks and Spatial Economics* 14.3-4, p. 297–316.
- DUDA, John (2013). "Cybernetics, anarchism and self-organisation". In : *Anarchist studies* 21.1, p. 52.

- DUPUY, Gabriel (1987). "Vers une théorie territoriale des réseaux : une application au transport urbain". In : *Annales de Géographie*. JSTOR, p. 658–679.
- DUPUY, Gabriel et Lucien Gilles BENGUIGUI (2015). "Sciences urbaines : interdisciplinarités passive, naïve, transitive, offensive". In : *Métropoles* 16.
- DUPUY, Gabriel et Vaclav STRANSKY (1996). "Cities and highway networks in Europe". In : *Journal of Transport Geography* 4.2, p. 107–121.
- DURAND-DASTES, François (2003). *Les géographes et la notion de causalité*.
- DURANTIN, Arnaud et al. (2017). "Disruptive Innovation in Complex Systems". In : *Complex Systems Design & Management*. Springer, p. 41–56.
- DURANTON, Gilles (1999). "Distance, land, and proximity : economic analysis and the evolution of cities". In : *Environment and Planning a* 31.12, p. 2169–2188.
- EUROSTAT (2014). *Eurostat Geographical Data*. url : <http://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units>.
- ELLIOTT, T. J. et M. GU (2017). "Occam's Vorpal Quantum Razor : Memory Reduction When Simulating Continuous-Time Stochastic Processes With Quantum Devices". In : *ArXiv e-prints*. arXiv : 1704.04231 [quant-ph].
- EPSTEIN, Joshua M et Robert L AXTELL (1996). *Growing artificial societies : Social science from the bottom up (complex adaptive systems)*. Brookings Institution Press MIT Press.
- FU, Jingying et al. (0001). *km grid population dataset of China (2005, 2010)*. Global Change Research Data Publishing and Repository, 2014. DOI : 10.3974/geodb.2014.01.06.v1.
- FAIVRE, Emmanuel (2003). "Infrastructures autoroutières, mobilité et dynamiques territoriales". 2003BESA1024. Thèse de doct., 620 p. URL : <http://www.theses.fr/2003BESA1024>.
- FARMER, J Doyne et Duncan FOLEY (2009). "The economy needs agent-based modelling". In : *Nature* 460.7256, p. 685–686.
- FATTORI, Michele et al. (2003). "Text mining applied to patent mapping : a practical business case". In : *World Patent Information* 25.4, p. 335–342.
- FAVARO, Jean-Marc et Denise PUMAIN (2011). "Gibrat Revisited : An Urban Growth Model Incorporating Spatial Interaction and Innovation Cycles." In : *Geographical Analysis* 43.3, p. 261–286.
- FEBRES, Gerardo et al. (2013). "Complexity measurement of natural and artificial languages". In : *arXiv preprint arXiv:1311.5427*.
- FECHER, Benedikt et Sascha FRIESIKE (2014). "Open science : one term, five schools of thought". In : *Opening science*. Springer, p. 17–47.
- FEYERABEND, Paul (1993). *Against method*. Verso.

- FOOT, Robin (1994). "RATP, un corporatisme à l'épreuve des voyageurs". In : *Travail* 31, p. 63–100.
- (2005). "Faut-il protéger le métro des voyageurs ? Ou l'appréhension du voyageur par les ingénieurs et les conducteurs". In : *Travailler* 2, p. 169–206.
- FRANCO, Jessica et al. (2009). "DiceDesign-package". In : *Designs of Computer Experiments*, p. 2.
- FRANK, Morgan R et al. (2014). "Constructing a taxonomy of fine-grained human movement and activity motifs through social media". In : *arXiv preprint arXiv :1410.1393*.
- FRANKHAUSER, Pierre (1998). "Fractal geometry of urban patterns and their morphogenesis". In : *Discrete Dynamics in Nature and Society* 2.2, p. 127–145.
- (2008). "Fractal geometry for measuring and modelling urban patterns". In : *The dynamics of complex urban systems*. Springer, p. 213–243.
- FREUD, Sigmund et al. (1989). *Totem and taboo : some points of agreement between the mental lives of savages and neurotics*. eng. The Standard edition of the complete psychological works of Sigmund Freud. New York : W.W. Norton. ISBN : 978-0-393-00143-3.
- FRIGG, Roman et Ioannis VOTSIS (2011). "Everything you always wanted to know about structural realism but were afraid to ask". In : *European journal for philosophy of science* 1.2, p. 227–276.
- FRITSCH, BERNARD (2007). "Infrastructures de transport, densification et étalement urbains : quelques enseignements de l'expérience nantaise". In : *Les Cahiers scientifiques du transport* 51, p. 37–60.
- FUJITA, Masahisa et al. (1999). "On the evolution of hierarchical urban systems". In : *European Economic Review* 43.2, p. 209–251.
- FUJITA, Masahisa et Jacques-François THISSE (1996). "Economics of agglomeration". In : *Journal of the Japanese and international economies* 10.4, p. 339–378.
- FULLERTON, Don et Sarah E WEST (2002). "Can taxes on cars and on gasoline mimic an unavailable tax on emissions?" In : *Journal of Environmental Economics and Management* 43.1, p. 135–157.
- FURLANELLO, C. et al. (2017). "Towards a scientific blockchain framework for reproducible data analysis". In : *ArXiv e-prints*. arXiv : 1707.06552.
- FURMAN, Jeffrey L. et Scott STERN (2011). "Climbing atop the Shoulders of Giants : The Impact of Institutions on Cumulative Research". In : *American Economic Review* 101.5, p. 1933–63. DOI : 10.1257/aer.101.5.1933. URL : <http://www.aeaweb.org/articles?id=10.1257/aer.101.5.1933>.
- FUSCO, Giovanni (2004). "La mobilité quotidienne dans les grandes villes du monde : application de la théorie des réseaux bayésiens". In : *Cybergeo : European Journal of Geography*.

- GABAIX, Xavier (1999). "Zipf's law for cities : an explanation". In : *Quarterly journal of Economics*, p. 739–767.
- GABAIX, Xavier et Yannis M. IOANNIDES (2004). "Chapter 53 The evolution of city size distributions". In : *Cities and Geography*. Sous la dir. de J. Vernon HENDERSON et Jacques-François THISSE. T. 4. Handbook of Regional and Urban Economics. Elsevier, p. 2341 –2378. DOI : [http://dx.doi.org/10.1016/S1574-0080\(04\)80010-5](http://dx.doi.org/10.1016/S1574-0080(04)80010-5). URL : <http://www.sciencedirect.com/science/article/pii/S1574008004800105>.
- GABORA, L. et M. STEEL (2017). "Autocatalytic networks in cognition and the origin of culture". In : *ArXiv e-prints*. arXiv : 1703.05917 [q-bio.NC].
- GALLOTTI, Riccardo et Marc BARTHELEMY (2014). "Anatomy and efficiency of urban multimodal mobility". In : *Scientific reports* 4.
- GAO, Jianxi et al. (2016). "Universal resilience patterns in complex networks". In : *Nature* 530.7590, p. 307–312. URL : <http://dx.doi.org/10.1038/nature16948>.
- GAO, Zhong-Ke et al. (2017). "Complex network analysis of time series". In : *EPL (Europhysics Letters)* 116.5, p. 50001.
- GAO, Zhong-Ke et al. (2015). "Multiscale complex network for analyzing experimental multivariate time series". In : *EPL (Europhysics Letters)* 109.3, p. 30005.
- GARGI CHAUDHURI AND KEITH C CLARKE (2015). "On the Spatiotemporal Dynamics of the Coupling between Land Use and Road Networks : Does Political History Matter ?" In : *Environment and Planning B : Planning and Design* 42.1, p. 133–156. DOI : {10.1068/b39089}. URL : {<http://dx.doi.org/10.1068/b39089>}.
- GAUTIER, Erwan et Ronan Le SAOUT (2015). "The dynamics of gasoline prices : Evidence from daily French micro data". In : *Journal of Money, Credit and Banking* 47.6, p. 1063–1089.
- GELL-MANN, Murray (1995). *The Quark and the Jaguar : Adventures in the Simple and the Complex*. Macmillan.
- GEMINO, Andrew et Yair WAND (2004). "A framework for empirical evaluation of conceptual modeling techniques". In : *Requirements Engineering* 9.4, p. 248–260.
- GERKEN, Jan M et Martin G MOEHRLE (2012). "A new instrument for technology monitoring : novelty in patents measured by semantic patent analysis". In : *Scientometrics* 91.3, p. 645–670.
- GERSHENSON, Carlos (2014). "Requisite Variety, Autopoiesis, and Self-organization". In : *arXiv :1409.7475 [nlin]*. arXiv : 1409.7475. URL : <http://arxiv.org/abs/1409.7475> (visité le 03/01/2017).
- (2015). "Requisite variety, autopoiesis, and self-organization". In : *Kybernetes* 44.6/7, p. 866–873.
- GIERE, Ronald N (2010a). "An agent-based conception of models and scientific representation". In : *Synthese* 172.2, p. 269–281.

- GIERE, Ronald N (2010b). *Explaining science : A cognitive approach*. University of Chicago Press.
- (2010c). *Scientific perspectivism*. University of Chicago Press, Chicago, ISBN : 9780226292137.
- GIERER, Alfred et Hans MEINHARDT (1972). "A theory of biological pattern formation". In : *Kybernetik* 12.1, p. 30–39.
- GILBERT, Scott F (2003). "The morphogenesis of evolutionary developmental biology". In : *International Journal of Developmental Biology* 47.7-8, p. 467.
- GILLI, Frédéric (2005). "Le Bassin parisien. Une région métropolitaine". In : *Cybergeo : European Journal of Geography*.
- (2014). *Grand Paris. L'émergence d'une métropole*. Presses de Sciences Po, coll. Nouveaux débats. 316p. ISBN : 978-2-7246-1516-6.
- GILLI, Frédéric et Jean-Marc OFFNER (2009). *Paris, métropole hors les murs : aménager et gouverner un Grand Paris*. Sciences Po, les presses.
- GIRAUT, Frédéric (2008). "Conceptualiser le territoire". In : *Historiens et géographes* 403, p. 57–68.
- GIRRES, Jean-François et Guillaume TOUYA (2010). "Quality assessment of the French OpenStreetMap dataset". In : *Transactions in GIS* 14.4, p. 435–459.
- GLEYZE, Jean-François (2005). "La vulnérabilité structurelle des réseaux de transport dans un contexte de risques". Thèse de doct. Université Paris-Diderot-Paris VII.
- GOFFMAN, Erving (1989). "On fieldwork". In : *Journal of contemporary ethnography* 18.2, p. 123–132.
- GOLDEN, Boris et al. (2012). "Modeling of complex systems ii : A minimalist and unified semantics for heterogeneous integrated systems". In : *Applied Mathematics and Computation* 218.16, p. 8039–8055.
- GOLDHIRSCH, Isaac et al. (1987). "Stability and Lyapunov stability of dynamical systems : A differential approach and a numerical method". In : *Physica D : Nonlinear Phenomena* 27.3, p. 311–337.
- GORYACHEV, Andrew B et Alexandra V POKHILKO (2008). "Dynamics of Cdc42 network embodies a Turing-type mechanism of yeast cell polarity". In : *FEBS letters* 582.10, p. 1437–1443.
- GOTTMANN, Jean (1964). *Megalopolis : the urbanized northeastern seaboard of the United States*. MIT Press Cambridge, MA.
- GREGG, Jay S et al. (2009). "The temporal and spatial distribution of carbon dioxide emissions from fossil-fuel use in North America". In : *Journal of Applied Meteorology and Climatology* 48.12, p. 2528–2542.
- GRIFFITH, Daniel A (1980). "Towards a theory of spatial statistics". In : *Geographical Analysis* 12.4, p. 325–339.
- (1992). "What is spatial autocorrelation ? Reflections on the past 25 years of spatial statistics". In : *Espace géographique* 21.3, p. 265–280.

- (2012). *Advanced spatial statistics : special topics in the exploration of quantitative spatial data series*. T. 12. Springer Science & Business Media.
- GRILICHES, Zvi (1990). *Patent Statistics as Economic Indicators : A Survey*. NBER Working Papers 3301. National Bureau of Economic Research, Inc. URL : <https://ideas.repec.org/p/nbr/nberwo/3301.html>.
- GRIMM, Volker et al. (2005). "Pattern-oriented modeling of agent-based complex systems : lessons from ecology". In : *science* 310.5750, p. 987–991.
- GUENÉE, Bernard (1968). "Espace et Etat dans la France du bas Moyen Age". In : *Annales. Histoire, Sciences Sociales*. T. 23. 4. JSTOR, p. 744–758.
- GUÉRIN-PACE, France et Denise PUMAIN (1990). "150 ans de croissance urbaine". In : *Economie et statistique* 230.1, p. 5–16.
- GUÉROIS, Marianne et Renaud LE GOIX (2009). "La dynamique spatio-temporelle des prix immobiliers à différentes échelles : le cas des appartements anciens à Paris (1990-2003)". In : *Cybergeo : European Journal of Geography*.
- GUÉROIS, Marianne et Fabien PAULUS (2002). "Commune centre, agglomération, aire urbaine : quelle pertinence pour l'étude des villes ?" In : *Cybergeo : European Journal of Geography*.
- GUÉROIS, Marianne et Denise PUMAIN (2008). "Built-up encroachment and the urban field : a comparison of forty European cities". In : *Environment and Planning A* 40.9, p. 2186–2203.
- GUILLOT, C. et T. LENCUIT (2013). "Mechanics of Epithelial Tissue". In : *Science* 340.June, p. 1185–1189. ISSN : 1095-9203. DOI : [10.1126/science.1235249](https://doi.org/10.1126/science.1235249).
- GUO, Xiaolei et Henry X LIU (2011). "Bounded rationality and irreversible network change". In : *Transportation Research Part B : Methodological* 45.10, p. 1606–1618.
- GURCIULLO, S. et al. (2015). "Complex Politics : A Quantitative Semantic and Topological Analysis of UK House of Commons Debates". In : *ArXiv e-prints*. arXiv : [1510.03797 \[physics.soc-ph\]](https://arxiv.org/abs/1510.03797).
- GUTMANN, Amy (2011). "The ethics of synthetic biology : guiding principles for emerging technologies". In : *Hastings Center Report* 41.4, p. 17–22.
- HACHI, R. (2013). *La fractalité comme indicateur de l'état de conservation du patrimoine urbain, Master Thesis Memoire*. Rapp. tech. Université Paris VII.
- HACKING, Ian (1999). *The social construction of what ?* Harvard university press, Cambridge, ISBN : 9780674004122.
- HAGGETT, Peter et Richard J CHORLEY (1970). "Network analysis in geography". In :

- HAKEN, Herman et Juval PORTUGALI (2003). "The face of the city is its information". In : *Journal of Environmental Psychology* 23.4, p. 385–408.
- HAKEN, Hermann (1980). "Synergetics". In : *Naturwissenschaften* 67.3, p. 121–128.
- HAKLAY, Mordechai (2010). "How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets". In : *Environment and planning B : Planning and design* 37.4, p. 682–703.
- HALL, Bronwyn H et al. (2001). *The NBER Patent Citations Data File : Lessons, Insights and Methodological Tools*. CEPR Discussion Papers 3094. C.E.P.R. Discussion Papers. URL : <https://ideas.repec.org/p/cpr/ceprdp/3094.html>.
- HALL, Peter Geoffrey et Kathy PAIN (2006). *The polycentric metropolis : learning from mega-city regions in Europe*. Routledge.
- HAMERLY, Greg, Charles ELKAN et al. (2003). "Learning the k in k-means". In : *NIPS*. T. 3, p. 281–288.
- HAN, Sangjin (2003). "Dynamic traffic modelling and dynamic stochastic user equilibrium assignment for general road networks". In : *Transportation Research Part B : Methodological* 37.3, p. 225–249.
- HARAN, EGP et Daniel R VINING (1973). "A MODIFIED YULE-SIMON MODEL ALLOWING FOR INTERCITY MIGRATION AND ACCOUNTING FOR THE OBSERVED FORM OF THE SIZE DISTRIBUTION OF CITIES*". In : *Journal of Regional Science* 13.3, p. 421–437.
- HART, Carolyn (2013). "Held in mind, out of awareness. Perspectives on the continuum of dissociated experience, culminating in dissociative identity disorder in children". en. In : *Journal of Child Psychotherapy* 39.3, p. 303–318. ISSN : 0075-417X, 1469-9370. DOI : [10.1080/0075417X.2013.846577](https://doi.org/10.1080/0075417X.2013.846577). (Visité le 02/01/2017).
- HATCHUEL, Armand et al. (1988). "Des stations de métro en mouvement : Station 2000, un scénario prospectif". In : *Les Annales de la recherche urbaine*. T. 39. 1. Persée-Portail des revues scientifiques en SHS, p. 35–42.
- HATTON, L. et G. WARR (2016). "Full Computational Reproducibility in Biological Science : Methods, Software and a Case Study in Protein Biology". In : *ArXiv e-prints*. arXiv : [1608.06897 \[q-bio.QM\]](https://arxiv.org/abs/1608.06897).
- HEDDEBAUT, Odile et Jean-Marie ERNECQ (2016). "Does the "tunnel effect" still remains in 2016?" In : *3e Colloque du programme Vingt années sous la Manche, et au-delà : "Régions accessibles, régions en croissance?"* Canterbury, United Kingdom. URL : <https://hal.archives-ouvertes.fr/hal-01355621>.
- HEISENBERG, Carl Philipp et Yohanns BELLAÏCHE (2013). "Forces in tissue morphogenesis and patterning". In : *Cell* 153.5. ISSN : 00928674. DOI : [10.1016/j.cell.2013.05.008](https://doi.org/10.1016/j.cell.2013.05.008).

- HIDALGO, C. A. (2015). "Disconnected! The parallel streams of network literature in the natural and social sciences". In : *ArXiv e-prints*. arXiv : [1511.03981 \[physics.soc-ph\]](#).
- HIJMANS, Robert J (2015). "Geographic data analysis and modeling". In :
- HILLIER, Bill (2016). "The Fourth Sustainability, Creativity : Statistical Associations and Credible Mechanisms". In : *Complexity, Cognition, Urban Planning and Design*. Springer, p. 75–92.
- HILLIER, Bill et Julienne HANSON (1989). *The social logic of space*. Cambridge university press.
- HOFSTADTER, Douglas H (1980). *Gödel, Escher, Bach : An Eternal Golden Braid ;[a Metaphoric Fugue on Minds and Machines in the Spirit of Lewis Carroll]*. Penguin Books.
- HOLLAND, John H (2012). *Signals and boundaries : Building blocks for complex adaptive systems*. Mit Press, Cambridge, ISBN : 9780262525930.
- HOLMES, C. et al. (2017). "Luria-Delbrück, revisited : The classic experiment does not rule out Lamarckian evolution". In : *ArXiv e-prints*. arXiv : [1701.05627 \[q-bio.PE\]](#).
- HOLZINGER, Andreas et al. (2014). "Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions". In : *BMC bioinformatics* 15.6, p. I1.
- HOMOCIANU, Marius (2009). "Transport-land use interaction modeling - Residential choices of households in urban area of Lyon". Theses. Université Lumière - Lyon II. URL : <https://tel.archives-ouvertes.fr/tel-00359302>.
- HUNG, Ling-Hong et al. (2016). "GUIDock : Using Docker Containers with a Common Graphics User Interface to Address the Reproducibility of Research". In : *PLoS ONE* 11.4, p. 1–14. DOI : [10.1371/journal.pone.0152686](#). URL : <http://dx.doi.org/10.1371%2Fjournal.pone.0152686>.
- Hypergeo*. <http://www.hypergeo.eu/spip.php?page=sommaire>.
- IACONO, Michael et al. (2008). "Models of transportation and land use change : a guide to the territory". In : *Journal of Planning Literature* 22.4, p. 323–340.
- IACOVACCI, Jacopo et al. (2015). "Mesoscopic Structures Reveal the Network Between the Layers of Multiplex Datasets". In : *arXiv preprint arXiv:1505.03824*.
- "Interdisciplinarity, Nature Special Issue". In :
- Issues in Neuroscience Research and Application* (2013). *Issues in Neuroscience Research and Application : 2013 Edition*. en. Google-Books-ID : dshjmDeDsWYC. ScholarlyEditions. ISBN : 978-1-4901-0778-3.
- JARROW, Robert A (1999). "In Honor of the Nobel Laureates Robert C. Merton and Myron S. Scholes : A Partial Differential Equation that Changed the World". In : *The Journal of Economic Perspectives*, p. 229–248.

- JÉGOU, Anne et al. (2012). "L'évaluation par indicateurs : un outil nécessaire d'aménagement urbain durable ?. Réflexions à partir de la démarche parisienne pour le géographe et l'aménageur". In : *Cybergeo : European Journal of Geography*.
- JUN, Joseph K et Alfred H HÜBLER (2005). "Formation and structure of ramified charge transportation networks in an electromechanical system". In : *Proceedings of the National Academy of Sciences of the United States of America* 102.3, p. 536–540.
- KONING, Martin et al. (2013). "Dessertes ferroviaires à grande vitesse et dynamisme économique local : Une analyse économétrique exploratoire sur les unités urbaines françaises". In : *ASRDLF 2013, 50ème colloque de l'Association des sciences régionales de langue française*. Belgium, 32p. URL : <https://hal.archives-ouvertes.fr/hal-00962384>.
- KAPLAN, Sarah et Keyvan VAKILI (2015). "The double-edged sword of recombination in breakthrough innovation". In : *Strategic Management Journal* 36.10, p. 1435–1457.
- KARATZOGLU, Alexandros et al. (2004). "kernlab – An S4 Package for Kernel Methods in R". In : *Journal of Statistical Software* 11.9, p. 1–20. URL : <http://www.jstatsoft.org/v11/i09/>.
- KASRAIAN, Dena et al. (2015). "Development of rail infrastructure and its impact on urbanization in the Randstad, the Netherlands". In : *Journal of Transport and Land Use* 9.1.
- KASRAIAN, Dena et al. (2016). "Long-term impacts of transport infrastructure networks on land-use change : an international review of empirical studies". In : *Transport Reviews* 36.6, p. 772–792. DOI : [10.1080/01441647.2016.1168887](https://doi.org/10.1080/01441647.2016.1168887). eprint : <http://dx.doi.org/10.1080/01441647.2016.1168887>. URL : <http://dx.doi.org/10.1080/01441647.2016.1168887>.
- KATZ, Michael L (1996). "Remarks on the economic implications of convergence". In : *Industrial and Corporate Change* 5.4, p. 1079–1095.
- KAY, Luciano et al. (2014). "Patent overlay mapping : Visualizing technological distance". In : *Journal of the Association for Information Science and Technology* 65.12, p. 2432–2443.
- KE, Yan et al. (2007). "Spatio-temporal shape and flow correlation for action recognition". In : *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, p. 1–8.
- KEERSMAECKER, Marie-Laurence et al. (2003). "Using fractal dimensions for characterizing intra-urban diversity : The example of Brussels". In : *Geographical analysis* 35.4, p. 310–328.
- KNIGHT, Frank B (1975). "A predictive view of continuous time processes". In : *The annals of Probability*, p. 573–596.
- KOCH, Christof et Gilles LAURENT (1999). "Complexity and the nervous system". In : *Science* 284.5411, p. 96–98.

- KOCH, Julian et Simon STISEN (2017). "Citizen science : A new perspective to advance spatial pattern evaluation in hydrology". In : *PLOS ONE* 12.5, p. 1–20. DOI : [10.1371/journal.pone.0178165](https://doi.org/10.1371/journal.pone.0178165). URL : <https://doi.org/10.1371/journal.pone.0178165>.
- KOLCHINSKY, A. et al. (2015). "Modularity and the Spread of Perturbations in Complex Dynamical Systems". In : *ArXiv e-prints*. arXiv : [1509.04386 \[physics.soc-ph\]](https://arxiv.org/abs/1509.04386).
- KONDO, Shigeru et Takashi MIURA (2010). "Reaction-diffusion model as a framework for understanding biological pattern formation". In : *science* 329.5999, p. 1616–1620.
- KRUGMAN, Paul (1998). "Space : the final frontier". In : *The Journal of Economic Perspectives* 12.2, p. 161–174.
- KRYVOBOKOV, Marko et al. (2013). "Comparison of Static and Dynamic Land Use-Transport Interaction Models". In : *Transportation Research Record : Journal of the Transportation Research Board* 2344.1, p. 49–58.
- KUHN, Thomas S (1970). *The structure of scientific revolutions*.
- KWAN, M.P. (2012). "The uncertain geographic context problem". In : *Annals of the Association of American Geographers* 102.5, p. 958–968.
- KWAN, Mei-Po (1998). "Space-time and integral measures of individual accessibility : a comparative analysis using a point-based framework". In : *Geographical analysis* 30.3, p. 191–216.
- L'ESPACE GÉOGRAPHIQUE (2014). *Les effets structurants des infrastructures de transport*, *L'Espace géographique* 2014/1 (Tome 43), p. 51-67.
- L'HOSTIS, Alain et al. (2014). "Contribution de la future ligne ferroviaire à grande vitesse Tours-Bordeaux au développement des réseaux des villes, une évaluation par le potentiel de contact". In :
- L'HOSTIS, Alain et al. (2016). "A Multicriteria approach for choosing a new public transport system linked to urban development : a method developed in the Bahn.Ville project for a tram-train scenario in the Saint-Étienne region". In : *Recherche Transports Sécurité* 2016.1-2, p. 17–25. URL : <https://hal.archives-ouvertes.fr/hal-01179934>.
- L'HOSTIS, Alain et al. (2012). "La ville orientée vers le rail". In : *Ville et mobilité*.
- LU, Yi et al. (2012). "The Chengdu-Guiyang High-Speed Rail Influence on the Location Advantage and Functional Positioning of Yibin City [J]". In : *Journal of Changsha University of Science & Technology (Social Science)* 5, p. 015.
- LAGESSE, C. (2015). "Read Cities through their Lines. Methodology to characterize spatial graphs". In : *ArXiv e-prints*. arXiv : [1512.01268 \[physics.soc-ph\]](https://arxiv.org/abs/1512.01268).
- LARIVIÈRE, Vincent et Yves GINGRAS (2014). "10 Measuring Interdisciplinarity". In : *Beyond bibliometrics : Harnessing multidimensional indicators of scholarly impact*, p. 187.

- LAUGHLIN, Robert B (2006). *A different universe : Reinventing physics from the bottom down*. Basic Books.
- LAUNER, Robert L et Graham N WILKINSON (2014). *Robustness in statistics*. Academic Press.
- LE NÉCHET, Florent (2015). "De la forme urbaine à la structure métropolitaine : une typologie de la configuration interne des densités pour les principales métropoles européennes de l'audit urbain". In : *Cybergeo : European Journal of Geography*. URL : <http://cybergeo.revues.org/26753> (visité le 21/04/2015).
- LE NÉCHET, Florent (2015). "De la forme urbaine à la structure métropolitaine : une typologie de la configuration interne des densités pour les principales métropoles européennes de l'Audit Urbain". In : *Cybergeo : European Journal of Geography*.
- LE NÉCHET, Florent et Juste RAIMBAULT (2015). "Modeling the emergence of metropolitan transport authority in a polycentric urban region". In : *Plurimondi. An International Forum for Research and Debate on Human Settlements* 7.15.
- LE TEXIER, Marion et Geoffrey CARUSO (2017). "Assessing geographical effects in spatial diffusion processes : The case of euro coins". In : *Computer, Environment and Urban Systems* 61.A, p. 81–93.
- LECHNER, Thomas et al. (2004). "Procedural modeling of land use in cities". In :
- LECHNER, Thomas et al. (2006). "Procedural modeling of urban land use". In : *ACM SIGGRAPH 2006 Research posters*. ACM, p. 135.
- LECUIT, Thomas et Pierre-françois LENNE (2007). "Cell surface mechanics and the control of cell shape, tissue patterns and morphogenesis". In : *Nat Rev Mol Cell Biol* 8.August, p. 633–644. DOI : [10.1038/nrm2222](https://doi.org/10.1038/nrm2222).
- LEE, SeongWoo et al. (2009). "Determinants of crime incidence in Korea : a mixed GWR approach". In : *World conference of the spatial econometrics association*, p. 8–10.
- LEEUW, Sander van der et al. (2009). "The Long-Term Evolution of Social Organization". In : *Complexity Perspectives in Innovation and Social Change*. Sous la dir. de David LANE et al. Dordrecht : Springer Netherlands, p. 85–116. ISBN : 978-1-4020-9663-1. DOI : [10.1007/978-1-4020-9663-1_4](https://doi.org/10.1007/978-1-4020-9663-1_4). URL : http://dx.doi.org/10.1007/978-1-4020-9663-1_4.
- LEFORT, Isabelle (2012). "Le terrain : l'Arlésienne des géographes?" In : *Annales de géographie*. 5. Armand Colin, p. 468–486.
- LEGAVRE, Jean Baptiste (1996). "La «neutralité» dans l'entretien de recherche. Retour personnel sur une évidence". In : *Politix* 9.35, p. 207–225.
- LERNER, Josh et Amit SERU (2015). "The use and misuse of patent data : Issues for corporate finance and beyond". In : *Booth/Harvard Business School Working Paper*.

- LEURENT, Fabien et Houda BOUJNAH (2014). "A user equilibrium, traffic assignment model of network route and parking lot choice, with search circuits and cruising flows". In : *Transportation Research Part C : Emerging Technologies* 47, p. 28–46.
- LEVINSON, David M (2011). "The coevolution of transport and land use : An introduction to the Special Issue and an outline of a research agenda". In : *Journal of Transport and Land Use* 4.2.
- LEVINSON, David Matthew et al. (2007). "The co-evolution of land use and road networks". In : *Transportation and traffic theory*, p. 839–859.
- LEVINSON, David (2008). "Density and dispersion : the co-development of land use and rail in London". In : *Journal of Economic Geography* 8.1, p. 55–77.
- LEVINSON, David et Wei CHEN (2005). "Paving new ground : a Markov chain model of the change in transportation networks and land use". In : *Access to destinations*. Emerald Group Publishing Limited, p. 243–266.
- LEVINSON, David et Ramachandra KARAMALAPUTI (2003). "Induced supply : a model of highway network expansion at the microscopic level". In : *Journal of Transport Economics and Policy (JTEP)* 37.3, p. 297–318.
- LEVINSON, David et al. (2012). "Forecasting and evaluating network growth". In : *Networks and Spatial Economics* 12.2, p. 239–262.
- LÉVY, Albert (2005). "Formes urbaines et significations : revisiter la morphologie urbaine". In : *Espaces et sociétés* 3, p. 25–48.
- L'HORTY, Yannick et Florent SARI (2013). "Le Grand Paris de l'emploi : l'extension des infrastructures de transport peut-elle avoir des effets positifs sur le chômage local ?" In : *Revue d'Économie Régionale & Urbaine* 3, p. 461–489.
- LI, Guan-Cheng et al. (2014). "Disambiguation and co-authorship networks of the US patent inventor database (1975–2010)". In : *Research Policy* 43.6, p. 941–955.
- LI, J et U WILENSKY (2009). *NetLogo Sugarscape 3 Wealth Distribution model*.
- LI, Tongfei et al. (2016). "Integrated co-evolution model of land use and traffic network design". In : *Networks and Spatial Economics* 16.2, p. 579–603.
- LIAO, Liao et Jean Pierre GAUDIN (2017). "L'ouverture au marché en Chine (années 1980-2000) et le développement économique local : une forme de gouvernance multi-niveaux ?" In : *Cybergeo : European Journal of Geography*.
- LIAW, Andy, Matthew WIENER et al. (2002). "Classification and regression by randomForest". In : *R news* 2.3, p. 18–22.
- LISSACK, Michael (2013). "Subliminal influence or plagiarism by negligence ? The Slodderwetenschap of ignoring the internet". In : *Journal of Academic Ethics*.

- LIU, Liu et L'Hostis ALAIN (2014). "Transport and Land Use Interaction : A French Case of Suburban Development in the Lille Metropolitan Area (LMA)". In : *Transportation Research Procedia* 4. Sustainable Mobility in Metropolitan Regions. mobil.TUM 2014. International Scientific Conference on Mobility and Transport. Conference Proceedings., p. 120 –139. ISSN : 2352-1465. DOI : <http://dx.doi.org/10.1016/j.trpro.2014.11.011>. URL : <http://www.sciencedirect.com/science/article/pii/S2352146514002944>.
- LIU, Wei et al. (2011). "Discovering spatio-temporal causal interactions in traffic data streams". In : *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, p. 1010–1018.
- LIU, Zhili et al. (2012). "Solving the last mile problem : Ensure the success of public bicycle system in Beijing". In : *Procedia-Social and Behavioral Sciences* 43, p. 73–78.
- LIVET, Pierre et al. (2010a). "Ontology, a Mediator for Agent-Based Modeling in Social Science". In : *Journal of Artificial Societies and Social Simulation* 13.1, p. 3. ISSN : 1460-7425. URL : <http://jasss.soc.surrey.ac.uk/13/1/3.html>.
- (2010b). "Ontology, a Mediator for Agent-Based Modeling in Social Science". In : *Journal of Artificial Societies and Social Simulation* 13.1, p. 3.
- LOI, Daniel (1985). "Une étude de la causalité dans la géographie classique française.[L'exemple des premières thèses régionales]". In : *Espace géographique* 14.2, p. 121–125.
- LOUAIL, Thomas et al. (2016). "Crowdsourcing the Robin Hood effect in cities". In : *arXiv preprint arXiv:1604.08394*.
- LOUF, R. et M. BARTHELEMY (2014). "How congestion shapes cities : from mobility patterns to scaling". In : *ArXiv e-prints*. arXiv : 1401.8200 [physics.soc-ph].
- LOUF, Rémi et Marc BARTHELEMY (2014a). "A typology of street patterns". In : *Journal of The Royal Society Interface* 11.101, p. 20140924.
- (2014b). "Scaling : lost in the smog". In : *arXiv preprint arXiv:1410.4964*.
- (2015). "Patterns of residential segregation". In : *arXiv preprint arXiv:1511.04268*.
- LOUF, Rémi et al. (2013). "Emergence of hierarchy in cost-driven growth of spatial networks". In : *Proceedings of the National Academy of Sciences* 110.22, p. 8824–8829.
- LOUF, Rémi et al. (2014). "Scaling in Transportation Networks". In : *PLoS ONE* 9.7, e102007. DOI : 10.1371/journal.pone.0102007. URL : <http://dx.doi.org/10.1371%2Fjournal.pone.0102007>.
- LUO, Qiang et al. (2013). "Spatio-temporal Granger causality : A new framework". In : *NeuroImage* 79, p. 241–263.
- LUZEAUX, Dominique (2015). "A formal foundation of systems engineering". In : *Complex Systems Design & Management*. Springer, p. 133–148.

- MACHARIS, Cathy et al. (2010). "A decision analysis framework for intermodal transport : Comparing fuel price increases and the internalisation of external costs". In : *Transportation Research Part A : Policy and Practice* 44.7, p. 550–561.
- MAHMASSANI, Hani S et Gang-Len CHANG (1987). "On boundedly rational user equilibrium in transportation systems". In : *Transportation science* 21.2, p. 89–99.
- MAINZER, Klaus et Leon O CHUA (2013). *Local activity principle*. World Scientific.
- MAKSE, Hernán A et al. (1998). "Modeling urban growth patterns with correlated percolation". In : *Physical Review E* 58.6, p. 7054.
- MAKSE, Hernán A et al. (1995). "Modelling urban growth". In : *Nature* 377.1912, p. 779–782.
- MANGIN, David (2014). *Le Grand Paris, où en est-on ?* Conférence de David Mangin le 14 mars 2014, ENPC et ENSAVT.
- MANGIN, David et Philippe PANERAI (1999). *Projet urbain*. Parenthèses.
- MANSON, Steven M (2001). "Simplifying complexity : a review of complexity theory". In : *Geoforum* 32.3, p. 405–414.
- (2008). "Does scale exist? An epistemological scale continuum for complex human–environment systems". In : *Geoforum* 39.2, p. 776–788.
- MANTEGNA, Rosario N et H Eugene STANLEY (1999). *Introduction to econophysics : correlations and complexity in finance*. Cambridge university press.
- MANTEGNA, Rosario Nunzio, Harry Eugene STANLEY et al. (2000). *An introduction to econophysics : correlations and complexity in finance*. T. 9. Cambridge university press Cambridge.
- MARCHIONNI, Caterina (2004). "Geographical economics versus economic geography : towards a clarification of the dispute". In : *Environment and Planning A* 36.10, p. 1737–1753.
- MARLER, R Timothy et Jasbir S ARORA (2004). "Survey of multi-objective optimization methods for engineering". In : *Structural and multidisciplinary optimization* 26.6, p. 369–395.
- MARTINEZ-CONDE, Susana et Stephen L. MACKNIK (2017). "Opinion : Finding the plot in science storytelling in hopes of enhancing science communication". In : *Proceedings of the National Academy of Sciences* 114.31, p. 8127–8129. DOI : [10.1073/pnas.1711790114](https://doi.org/10.1073/pnas.1711790114). eprint : <http://www.pnas.org/content/114/31/8127.full.pdf>. URL : <http://www.pnas.org/content/114/31/8127.short>.
- MASSON, Sophie (2000). "Les interactions entre système de transport et système de localisation en milieu urbain et leur modélisation". 2000LYO22027. Thèse de doct., 1 vol. (566 p.) URL : <http://theses.fr/2000LYO22027/document>.
- MASUCCI, A Paolo et al. (2013). "Gravity versus radiation models : On the importance of scale and heterogeneity in commuting flows". In : *Physical Review E* 88.2, p. 022812.

- MEHAFFY, Michael W (2007). "Notes on the genesis of wholes : Christopher Alexander and his continuing influence". In : *Urban Design International* 12.1, p. 41–49.
- MENDELEY (2015). *Mendeley Reference Manager*. <http://www.mendeley.com/>.
- MESOUDI, Alex (2017). "Pursuing Darwin's curious parallel : Prospects for a science of cultural evolution". In : *Proceedings of the National Academy of Sciences* 114.30, p. 7853–7860. DOI : [10.1073/pnas.1620741114](https://doi.org/10.1073/pnas.1620741114). eprint : <http://www.pnas.org/content/114/30/7853.full.pdf>. URL : <http://www.pnas.org/content/114/30/7853.abstract>.
- MILLER, Harvey J (1999). "Measuring space-time accessibility benefits within transportation networks : basic theory and computational procedures". In : *Geographical analysis* 31.1, p. 1–26.
- MIMEUR, Christophe (2016). "The traces of speed between space and network". Theses. Université de Bourgogne Franche-Comté. URL : <https://hal.archives-ouvertes.fr/tel-01451164>.
- MIN, Wanli et Laura WYNTER (2011). "Real-time road traffic prediction with spatio-temporal correlations". In : *Transportation Research Part C : Emerging Technologies* 19.4, p. 606–616.
- MOECKEL, Rolf et al. (2003). "Creating a synthetic population". In : *Proceedings of the 8th International Conference on Computers in Urban Planning and Urban Management* (CUPUM).
- MONOD, J. (1970). *Le hasard et la Nécessité*. Points, Paris.
- MOORE, Christopher et Stephan MERTENS (2011). *The nature of computation*. OUP Oxford.
- MOOSAVI, V. (2017). "Urban morphology meets deep learning : Exploring urban forms in one million cities, town and villages across the planet". In : *ArXiv e-prints*. arXiv : [1709.02939 \[cs.CV\]](https://arxiv.org/abs/1709.02939).
- MORENO REGAN, Omar (2016). "Etude du comportement des tunnels en maçonnerie du métro parisien". Thèse de doct. Paris Est.
- MORENO, Diego et al. (2012). "Un automate cellulaire pour expérimenter les effets de la proximité dans le processus d'étalement urbain : le modèle Raumulus". In : *Cybergeo : European Journal of Geography*.
- MORIN, Edgar (1976). *La méthode : la nature de la nature*. Le Seuil.
- (1991). "La méthode tome 4 : les idées". In : *Paris*, Seuil.
- MOUDON, Anne Vernez (1997). "Urban morphology as an emerging interdisciplinary field". In : *Urban morphology* 1.1, p. 3–10.
- MOULIN-FRIER, C. et al. (2017). "Embodied Artificial Intelligence through Distributed Adaptive Control : An Integrated Framework". In : *ArXiv e-prints*. arXiv : [1704.01407 \[cs.AI\]](https://arxiv.org/abs/1704.01407).
- MURPHY, Alexander B (2012). "Entente territorial : Sack and Raffestin on territoriality". In : *Environment and Planning D : Society and Space* 30.1, p. 159–172.
- NLTK (2015). *Natural Language Toolkit*, Stanford University.

- NAKAMASU, Akiko et al. (2009). "Interactions between zebrafish pigment cells responsible for the generation of Turing patterns". In : *Proceedings of the National Academy of Sciences* 106.21, p. 8429–8434.
- NELSON, Richard R et Sidney G WINTER (2009). *An evolutionary theory of economic change*. harvard university press.
- NEWMAN, M. E. J. (2013). "Prediction of highly cited papers". In : *ArXiv e-prints*. arXiv : 1310.8220 [physics.soc-ph].
- (2016). "Community detection in networks : Modularity optimization and maximum likelihood are equivalent". In : *ArXiv e-prints*. arXiv : 1606.02319.
- NEWMAN, MEJ (2011). "Complex systems : A survey". In : *arXiv preprint arXiv:1112.1440*.
- NEWMAN, Mark EJ (2003). "The structure and function of complex networks". In : *SIAM review* 45.2, p. 167–256.
- NICHOLS, Leah G (2014). "A topic model approach to measuring interdisciplinarity at the National Science Foundation". In : *Scientometrics* 100.3, p. 741–754.
- NICOSIA, Vincenzo et al. (2009). "Extending the definition of modularity to directed graphs with overlapping communities". In : *Journal of Statistical Mechanics : Theory and Experiment* 2009.03, P03024.
- NIEDERREITER, H (1972). "Discrepancy and convex programming". In : *Annali di matematica pura ed applicata* 93.1, p. 89–97.
- NIIZATO, Takayuki et al. (2010). "A model of network formation by Physarum plasmodium : interplay between cell mobility and morphogenesis". In : *Biosystems* 100.2, p. 108–112.
- NILSSON, Isabelle M et Oleg A SMIRNOV (2016). "Measuring the effect of transportation infrastructure on retail firm co-location patterns". In : *Journal of Transport Geography* 51, p. 110–118.
- NITSCH, Volker (2005). "Zipf zipped". In : *Journal of Urban Economics* 57.1, p. 86–100.
- NORUZI, Alireza (2005). "Google Scholar : The new generation of citation indexes". In : *Libri* 55.4, p. 170–180.
- OECD (2009). "OECD Patent Statistics Manual". In : doi : <http://dx.doi.org/10.1787/9789264056442-en>. URL : /content/book/9789264056442-en.
- O'SULLIVAN, David et Steven M MANSON (2015). "Do Physicists Have 'Geography Envy' ? And What Can Geographers Learn From It ?" In : *Annals of the Association of American Geographers*.
- O'BRIEN, Oliver et al. (2014). "Mining bicycle sharing data for generating insights into sustainable transport systems". In : *Journal of Transport Geography* 34, p. 262–273.
- OFFNER, Jean-Marc (1993). "Les "effets structurants" du transport : mythe politique, mystification scientifique". In : *Espace géographique* 22.3, p. 233–242.
- OFFNER, Jean-Marc et Denise PUMAIN (1996). "Réseaux et territoires-significations croisées". In :

- OLSEN, Sherry (1982). "Urban metabolism and morphogenesis". In : *Urban Geography* 3.2, p. 87–109.
- OMODEI, E. et al. (2016). "Evaluating the impact of interdisciplinary research : a multilayer network approach". In : *ArXiv e-prints*. arXiv : [1601.06075 \[physics.soc-ph\]](https://arxiv.org/abs/1601.06075).
- OPENSTREETMAP (2012). *OpenStreetMap*.
- OPENSHAW, S (1983). *FROM DATA CRUNCHING TO MODEL CRUNCHING – THE DAWN OF A NEW ERA*.
- OPENSHAW, Stan (1984). *The Modifiable Areal Unit Problem*. Norwich, UK : Geo Books.
- OSTROWETSKY, S. & al. (2004). "Les Villes Nouvelles, 30 ans après". In : *Espaces et Sociétés* n° 119, 4/2004.
- OTAMENDI, Javier et al. (2008). "Selection of the simulation software for the management of the operations at an international airport". In : *Simulation Modelling Practice and Theory* 16.8, p. 1103–1112.
- PADEIRO, Miguel (2013). "Transport infrastructures and employment growth in the Paris metropolitan margins". In : *Journal of Transport Geography* 31, p. 44 –53. ISSN : 0966-6923. DOI : <http://dx.doi.org/10.1016/j.jtrangeo.2013.05.007>. URL : <http://www.sciencedirect.com/science/article/pii/S0966692313000847>.
- PALCHYKOV, V. et al. (2016). "Ground truth ? Concept-based communities versus the external classification of physics manuscripts". In : *ArXiv e-prints*. arXiv : [1602.08451 \[cs.DL\]](https://arxiv.org/abs/1602.08451).
- PARK, Inchae et Byungun YOON (2014). "A semantic analysis approach for identifying patent infringement based on a product-patent map". In : *Technology Analysis & Strategic Management* 26.8, p. 855–874.
- PAULLEY, Neil J et F Vernon WEBSTER (1991). "Overview of an international study to compare models and evaluate land-use and transport policies". In : *Transport Reviews* 11.3, p. 197–222.
- PAULUS, Fabien (2004). "Coévolution dans les systèmes de villes : croissance et spécialisation des aires urbaines françaises de 1950 à 2000". Thèse de doct. Université Panthéon-Sorbonne-Paris I.
- PEREZ-RIVEROL, Yasset et al. (2016). "Ten Simple Rules for Taking Advantage of Git and GitHub". In : *PLoS Comput Biol* 12.7, p. 1–11. DOI : [10.1371/journal.pcbi.1004947](https://doi.org/10.1371/journal.pcbi.1004947). URL : <http://dx.doi.org/10.1371%2Fjournal.pcbi.1004947>.
- PÉTONNET, Colette (1982). "L'Observation flottante L'exemple d'un cimetière parisien". In : *l'Homme*, p. 37–47.
- PFAENDER, Fabien (2009). "Spatialisation de l'information". Thèse de doct. Compiègne.
- PHAN, Denis et Franck VARENNE (2010). "Agent-Based Models and Simulations in Economics and Social Sciences : from conceptual exploration to distinct ways of experimenting". In : *Journal of Artificial Societies and Social Simulation* 13.1, p. 5.

- PICHON RIVIÈRE, Enrique (2004). *Le processus groupal*. French. OCLC : 154119138. Ramonville-Saint-Agne (Haute-Garonne) : Érès. ISBN : 978-2-7492-0317-1.
- PICON, Antoine (2013). *Smart cities : théorie et critique d'un idéal auto-réalisateur*. B2.
- PIERS, Craig et al. (2007). *Self-Organizing Complexity in Psychological Systems*. en. Google-Books-ID : rRSyAAAAQBAJ. Jason Aronson, Incorporated. ISBN : 978-1-4616-3065-4.
- PIGOZZI, Bruce Wm (1980). "Interurban linkages through polynomially constrained distributed lags". In : *Geographical Analysis* 12.4, p. 340–352.
- PIKETTY, Thomas (2013). *Le capital au XXIe siècle*. Seuil.
- PINTEA, Camelia-M. et al. (2017). "The generalized traveling salesman problem solved with ant algorithms". In : *Complex Adaptive Systems Modeling* 5.1, p. 8. ISSN : 2194-3206. DOI : [10.1186/s40294-017-0048-9](https://doi.org/10.1186/s40294-017-0048-9). URL : <https://doi.org/10.1186/s40294-017-0048-9>.
- POHLE, J. et al. (2017). "Selecting the Number of States in Hidden Markov Models - Pitfalls, Practical Challenges and Pragmatic Solutions". In : *ArXiv e-prints*. arXiv : [1701.08673 \[stat.ME\]](https://arxiv.org/abs/1701.08673).
- PORTUGALI, Juval (2011). "SIRN—Synergetic Inter-Representation Networks". In : *Complexity, Cognition and the City*, p. 139–165.
- POTIRON, Y. (2016). "Estimating the integrated parameter of the locally parametric model in high-frequency data." In : *Working Paper*.
- POTIRON, Yoann et Per MYKLAND (2015). "Estimation of integrated quadratic covariation between two assets with endogenous sampling times". In : *arXiv preprint arXiv:1507.01033*.
- PRESCHITSCHEK, Nina et al. (2013). "Anticipating industry convergence : Semantic analyses vs IPC co-classification analyses of patents". In : *Foresight* 15.6. Sous la dir. de Tugrul DAIM, p. 446–464. ISSN : 1463-6689. DOI : [10.1108/fs-10-2012-0075](https://doi.org/10.1108/fs-10-2012-0075).
- PRIGOGINE, Ilya et Isabelle STENGERS (1997). *The end of certainty*. Simon et Schuster.
- PRITCHARD, David R et Eric J MILLER (2009). "Advances in agent population synthesis and application in an integrated land use and transportation model". In : *Transportation Research Board 88th Annual Meeting*. 09-1686.
- PUMAIN, Denise (1997). "Pour une théorie évolutive des villes". In : *Espace géographique* 26.2, p. 119–134.
- (2003). "Une approche de la complexité en géographie". In : *Géo-carrefour* 78.1, p. 25–31.
 - (2005). "Cumulativité des connaissances". In : *Revue européenne des sciences sociales. European Journal of Social Sciences* XLIII-131, p. 5–12.

- PUMAIN, Denise (2010). "Une théorie géographique des villes". In : *Bulletin de la Société géographie de Liège* 55, p. 5–15.
- (2012a). "Multi-agent system modelling for urban systems : The series of SIMPOP models". In : *Agent-based models of geographical systems*. Springer, p. 721–738.
 - (2012b). "Urban systems dynamics, urban growth and scaling laws : The question of ergodicity". In : *Complexity Theories of Cities Have Come of Age*. Springer, p. 91–103.
 - (2017). "Geography far from equilibrium". In : *Non-Equilibrium Social Science and Policy*. Springer, p. 71–80.
- PUMAIN, Denise et al. (2009). "Innovation cycles and urban dynamics". In : *Complexity perspectives in innovation and social change*, p. 237–260.
- PUMAIN, Denise et Romain REUILLO (2017a). "The SimpopLocal Model". In : *Urban Dynamics and Simulation Models*. Springer, p. 21–35.
- (2017b). *Urban Dynamics and Simulation Models*. Springer International, ISBN 978-3-319-46497-8.
- PUMAIN, Denise et Benoît RIANDEY (1986). "Le Fichier de l'Ined". In : *Espace, populations, sociétés* 4.2, p. 269–277.
- PUMAIN, Denise et Marie-Claire ROBIC (2002). "Le rôle des mathématiques dans une «révolution» théorique et quantitative : la géographie française depuis les années 1970". In : *Revue d'histoire des Sciences Humaines* 6.1, p. 123–144.
- PUMAIN, Denise et al. (1984). "Vers une modélisation de la dynamique intra urbaine". In : *Espace géographique* 13.2, p. 125–135.
- PUMAIN, Denise et Lena SANDERS (2013). "Theoretical principles in interurban simulation models : a comparison". In : *Environment and Planning A* 45.9, p. 2243–2260.
- PUMAIN, Denise et al. (2006). "An evolutionary theory for interpreting urban scaling laws". In : *Cybergeo : European Journal of Geography*.
- PUTMAN, Stephen H (1975). "Urban land use and transportation models : A state-of-the-art summary". In : *Transportation Research* 9.2, p. 187–202.
- PUZIS, Rami et al. (2013). "Augmented betweenness centrality for environmentally aware traffic monitoring in transportation networks". In : *Journal of Intelligent Transportation Systems* 17.1, p. 91–105.
- QGIS, DT (2011). "Quantum GIS geographic information system". In : *Open Source Geospatial Foundation Project*.
- R CORE TEAM (2015a). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL : <http://www.R-project.org/>.
- (2015b). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL : <http://www.R-project.org/>.

- RADDICK, M. J. et al. (2010). "Galaxy Zoo : Exploring the Motivations of Citizen Science Volunteers". In : *Astronomy Education Review* 9.1, p. 010103. DOI : [10.3847/AER2009036](https://doi.org/10.3847/AER2009036). arXiv : [0909.2925 \[astro-ph.IM\]](https://arxiv.org/abs/0909.2925).
- RAFFESTIN, Claude (1978). "Les construits en géographie humaine : notions et concepts". In :
- (1988). "Repères pour une théorie de la territorialité humaine". In :
- RAIMBAULT, J. (2015a). "Models Coupling Urban Growth and Transportation Network Growth : An Algorithmic Systematic Review Approach". In : *Forthcoming in ECTQG 2015 Proceedings*.
- (2016). "Generation of Correlated Synthetic Data". In : *Forthcoming in Actes des Journées de Rochebrune 2016*.
- RAIMBAULT, J. (2016). "Models Coupling Urban Growth and Transportation Network Growth : An Algorithmic Systematic Review Approach". In : *ArXiv e-prints*. arXiv : [1605.08888 \[cs.DL\]](https://arxiv.org/abs/1605.08888).
- RAIMBAULT, J. et al. (2014). "A hybrid network/grid model of urban morphogenesis and optimization". In : *Proceedings of the 4th International Conference on Complex Systems and Applications (ICCSA 2014), June 23-26, 2014, Université de Normandie, Le Havre, France ; M. A. Aziz-Alaoui, C. Bertelle, X. Z. Liu, D. Olivier, eds.* : pp. 51-60.
- RAIMBAULT, J. et J. GONZALEZ (May 2015). "Application de la Morphogénèse de Réseaux Biologiques à la Conception Optimale d'Infrastructures de Transport". In : *Rencontres du Labex Dynamites*.
- RAIMBAULT, Juste (2015b). "Hybrid Modeling of a Bike-Sharing Transportation System". In : *International Conference on Computational Social Science*.
- (2016a). "For a Cautious Use of Big Data and Computation". In : *Royal Geographical Society-Annual Conference 2016-Session : Geocomputation, the Next 20 Years (1)*.
 - (2016b). "Indirect Bibliometrics by Complex Network Analysis". In : *20e Anniversaire de Cybergeo*.
 - (2016c). "Investigating the Empirical Existence of Static User Equilibrium". In : *Forthcoming in Transportation Research Procedia, EWGT2016. arXiv preprint arXiv : 1608.05266*.
 - (2016d). "Models of growth for system of cities : Back to the simple". In : *Conference on Complex Systems 2016*.
 - (2017a). "A Discrepancy-Based Framework to Compare Robustness Between Multi-attribute Evaluations". In : *Complex Systems Design & Management*. Springer, p. 141–154.
 - (2017b). "An Applied Knowledge Framework to Study Complex Systems". In : *Forthcoming in CSDM2017 proceedings*.
 - (2017c). "Co-construire Modèles, Etudes Empiriques et Théories en Géographie Théorique et Quantitative : le cas des Interactions entre Réseaux et Territoires". In : *Treizièmes Rencontres de Théo-Quant*.

- RAIMBAULT, Juste (2017d). "Complexity, Complexities and Complex Knowledges". In : *forthcoming discussion at ERC Divercity Workshop, 12-13 October 2017*.
- (2017e). *Entretiens vo.2 [Data set]*. Zenodo. <http://doi.org/10.5281/zenodo.556331>.
 - (2017f). "Identification de Causalités dans des Données Spatio-temporelles". In : *submitted to SAGEO 2017*.
 - (2017g). "Investigating the Empirical Existence of Static User Equilibrium". In : *Transportation Research Procedia* 22C, p. 450–458. URL : DOI : [10.1016/j.trpro.2017.03.053](https://doi.org/10.1016/j.trpro.2017.03.053); arXiv preprint arXiv:1608.05266.
 - (2017h). "Un Cadre de Connaissances pour une Géographie Intégrée". In : *Journée des jeunes chercheurs de l'Institut de Géographie de Paris*. Paris, France. URL : <https://halshs.archives-ouvertes.fr/halshs-01505084>.
- RAIMBAULT, Juste et Antonin BERGEAUD (2017). "The Cost of Transportation : Spatial Analysis of Fuel Prices in the US". In : *Forthcoming in Transportation Research Procedia, EWGT2017*.
- RAM, Karthik (2013). "Git can facilitate greater reproducibility and increased transparency in science." In : *Source code for biology and medicine* 8.1, p. 7.
- RAMSEY, James B (2002). "Wavelets in economics and finance : Past and future". In : *Studies in Nonlinear Dynamics & Econometrics* 6.
- RASMUSSEN, Thomas Kjær et al. (2015). "Stochastic user equilibrium with equilibrated choice sets : Part II—Solving the restricted SUE for the logit family". In : *Transportation Research Part B : Methodological* 77, p. 146–165.
- READ, Dwight et al. (2009). "The innovation innovation". In : *Complexity perspectives in innovation and social change*. Springer, p. 43–84.
- REID, Chris R et al. (2016). "Decision-making without a brain : how an amoeboid organism solves the two-armed bandit". In : *Journal of The Royal Society Interface* 13.119, p. 20160030.
- RENFREW, Colin (1978). "Trajectory discontinuity and morphogenesis : the implications of catastrophe theory for archaeology". In : *American Antiquity*, p. 203–222.
- REUILLO, Romain et al. (2013). "OpenMOLE, a workflow engine specifically tailored for the distributed exploration of simulation models". In : *Future Generation Computer Systems* 29.8, p. 1981–1990.
- REUILLO, Romain et al. (2015). "A New Method to Evaluate Simulation Models : The Calibration Profile (CP) Algorithm". In : *Journal of Artificial Societies and Social Simulation* 18.1, p. 12. ISSN : 1460-7425. URL : <http://jasss.soc.surrey.ac.uk/18/1/12.html>.
- REY-COYREHOURCQ, Sébastien (2015). "Une plateforme intégrée pour la construction et Une plateforme intégrée pour la construction et l'évaluation de modèles de simulation en géographie". Thèse de doct. Université Paris 1 Panthéon-Sorbonne.

- REYMOND, Henri et Colette CAUVIN (2013). "La logique ternaire de Stéphane Lupasco et le raisonnement géocartographique bioculturel d'*Homo geographicus*. L'apport de la notion de couplage transdisciplinaire dans l'approche de l'agrégation morphologique des agglomérations urbaines". In : *Cybergeo : European Journal of Geography*.
- RIETVELD, Piet (1994). "Spatial economic impacts of transport infrastructure supply". In : *Transportation Research Part A : Policy and Practice* 28.4. Special Issue Transport Externalities, p. 329 –341. ISSN : 0965-8564. doi : [http://dx.doi.org/10.1016/0965-8564\(94\)90007-8](http://dx.doi.org/10.1016/0965-8564(94)90007-8). URL : <http://www.sciencedirect.com/science/article/pii/0965856494900078>.
- RIETVELD, Piet et al. (2001). "Spatial graduation of fuel taxes ; consequences for cross-border and domestic fuelling". In : *Transportation Research Part A : Policy and Practice* 35.5, p. 433–457.
- RIETVELD, Piet et Stefan van WOUDENBERG (2005). "Why fuel prices differ". In : *Energy Economics* 27.1, p. 79–92.
- RIPOLL, Fabrice (2017). "Géographie de l'alternatif, Géographies alternatives ? Grand Témoin." In : *Journée des Jeunes Chercheurs de l'Institut de Géographie*.
- ROBIC, Marie-Claire (1982). "Cent ans avant Christaller... une théorie des lieux centraux". In : *Espace géographique* 11.1, p. 5–12.
- ROCCA, Jean-Louis (2008). "Power of knowledge : The imaginary formation of the Chinese middle stratum in an era of growth and stability". In : *Patterns of middle class consumption in India and China*, p. 127–139.
- ROGERS, Katherine W et Alexander F SCHIER (2011). "Morphogen Gradients : From Generation to Interpretation". In : *Annu Rev Cell Dev Biol*. doi : [10.1146/annurev-cellbio-092910-154148](https://doi.org/10.1146/annurev-cellbio-092910-154148).
- ROMER, Paul M (1990). "Endogenous Technological Change". In : *Journal of Political Economy* 98.5, S71–102. URL : <https://ideas.repec.org/a/ucp/jpolc/v98y1990i5ps71-102.html>.
- ROTH, Camille (2009). "Reconstruction Failures : Questioning Level Design". In : *Epistemological Aspects of Computer Simulation in the Social Sciences*. Springer, p. 89–98.
- ROZENFELD, Hernán D et al. (2008). "Laws of population growth". In : *Proceedings of the National Academy of Sciences* 105.48, p. 18702–18707.
- RUBNER, Yossi et al. (2000). "The earth mover's distance as a metric for image retrieval". In : *International journal of computer vision* 40.2, p. 99–121.
- RUCKER, Gerta (2012). "Network meta-analysis, electrical networks and graph theory". In : *Research Synthesis Methods* 3.4, p. 312–324.
- RUI, Yikang (2013). "Urban Growth Modeling Based on Land-use Changes and Road Network Expansion". In :

- RUI, Yikang et Yifang BAN (2011). "Urban growth modeling with road network expansion and land use development". In : *Advances in Cartography and GIScience. Volume 2*. Springer, p. 399–412.
- (2014). "Exploring the relationship between street centrality and land use in Stockholm". In : *International Journal of Geographical Information Science* 28.7, p. 1425–1438. DOI : [10.1080/13658816.2014.893347](https://doi.org/10.1080/13658816.2014.893347). eprint : <http://dx.doi.org/10.1080/13658816.2014.893347>. URL : <http://dx.doi.org/10.1080/13658816.2014.893347>.
- RUI, Yikang et al. (2013). "Exploring the patterns and evolution of self-organized urban street networks through modeling". In : *The European Physical Journal B* 86.3, p. 1–8.
- RUSSO, Francesco et Giuseppe MUSOLINO (2012). "A unifying modelling framework to simulate the Spatial Economic Transport Interaction process at urban and national scales". In : *Journal of Transport Geography* 24, p. 189–197.
- SDRIF (2013). *Île-de-France 2030. ORIENTATIONS RÉGLEMENTAIRES ET CARTE DE DESTINATION GÉNÉRALE DES DIFFÉRENTES PARTIES DU TERRITOIRE*.
- STIF (2010). *ArcExpress, débat public sur le métro de rocade. Dossier du Maître d’Ouvrage*. archived at http://archive.wikiwix.com/cache/?url=http%3A%2F%2Fwww.debatpuarcexpress.org%2F_script%2Fntsp-document-file_download.php%3Fdocument_id%3D92%26document_file_id%3D106.
- SAMANIEGO, Horacio et Melanie E MOSES (2008). "Cities as organisms : Allometric scaling of urban road networks". In : *Journal of Transport and Land use* 1.1.
- SANDERS, Lena (1992). *Système de villes et synergétique*. Economica.
- SANDERS, Lena et al. (1997). "SIMPOP : a multiagent system for the study of urbanism". In : *Environment and Planning B* 24, p. 287–306.
- SARIGÖL, E. et al. (2014). "Predicting Scientific Success Based on Co-authorship Networks". In : *ArXiv e-prints*. arXiv : [1402.7268 \[physics.soc-ph\]](https://arxiv.org/abs/1402.7268).
- SAYAMA, Hiroki (2009). "Swarm chemistry". In : *Artificial life* 15.1, p. 105–114.
- SCHMICKL, Thomas et al. (2016). "How a life-like system emerges from a simplistic particle motion law". In : *Scientific Reports* 6.1. ISSN : 2045-2322. DOI : [10.1038/srep37969](https://doi.org/10.1038/srep37969). URL : <http://dx.doi.org/10.1038/srep37969>.
- SCHMID, Helmut (1994). "Probabilistic part-of-speech tagging using decision trees". In : *Proceedings of the international conference on new methods in language processing*. T. 12. Citeseer, p. 44–49.
- SCHMITT, Clara (2014). "Modélisation de la dynamique des systèmes de peuplement : de SimpopLocal à SimpopNet." Thèse de doct. Paris 1.

- SCHMITT, Clara et al. (2014). "Half a billion simulations : Evolutionary algorithms and distributed computing for calibrating the SimpopLocal geographical model". In :
- SCHWARZ, Nina (2010). "Urban form revisited—Selecting indicators for characterising European cities". In : *Landscape and Urban Planning* 96.1, p. 29 –47. ISSN : 0169-2046. DOI : <http://dx.doi.org/10.1016/j.landurbplan.2010.01.007>. URL : <http://www.sciencedirect.com/science/article/pii/S0169204610000320>.
- SEIDL, David (2004). "Luhmann's theory of autopoietic social systems". In : *Ludwig-Maximilians-Universität München-Munich School of Management*. (Visité le 01/01/2017).
- SERVAIS, Marc et al. (2004). "Polycentrisme urbain : une réalité spatialement mesurable ?" In : *Cybergeo : European Journal of Geography*.
- SETO, Karen C et al. (2012). "Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools". In : *Proceedings of the National Academy of Sciences* 109.40, p. 16083–16088.
- SHALIZI, Cosma Rohilla et James P CRUTCHFIELD (2001). "Computational mechanics : Pattern and prediction, structure and simplicity". In : *Journal of statistical physics* 104.3-4, p. 817–879.
- SHENZHENSHI (2016). 关于地铁2号线东延线、地铁3号线西延线线站位初步方案. Urban Planning Commission.
- SHERIDAN DODDS, P. et al. (2016). "Simon's fundamental rich-gets-richer model entails a dominant first-mover advantage". In : *ArXiv e-prints*. arXiv : [1608.06313 \[physics.soc-ph\]](https://arxiv.org/abs/1608.06313).
- SHIBATA, Naoki et al. (2008). "Detecting emerging research fronts based on topological measures in citation networks of scientific publications". In : *Technovation* 28.11, p. 758–775.
- SIMON, Herbert A. (1955). "On a Class of Skew Distribution Functions". English. In : *Biometrika* 42.3/4, pp. 425–440. ISSN : 00063444. URL : <http://www.jstor.org/stable/2333389>.
- SORENSEN, Olav et al. (2006). "Complexity, networks and knowledge flow". In : *Research policy* 35.7, p. 994–1017.
- SOUAMI, Taoufik (2012). *Ecoquartiers : secrets de fabrication*. Scrineo.
- STANLEY, H Eugene et al. (1999). "Econophysics : Can physicists contribute to the science of economics ?" In : *Physica A : Statistical Mechanics and its Applications* 269.1, p. 156–169.
- STEVENS, Forrest R. et al. (2015). "Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data". In : *PLoS ONE* 10.2, p. 1–22. DOI : [10.1371/journal.pone.0107042](https://doi.org/10.1371/journal.pone.0107042). URL : <http://dx.doi.org/10.1371%2Fjournal.pone.0107042>.
- STODDEN, Victoria (2010). "The scientific method in practice : Reproducibility in the computational sciences". In :

- STORPER, Michael et Allen J SCOTT (2009). "Rethinking human capital, creativity and urban growth". In : *Journal of economic geography* 9.2, p. 147–167.
- SULLIVAN, JL et al. (2010). "Identifying critical road segments and measuring system-wide robustness in transportation networks with isolating links : a link-based capacity-reduction approach". In : *Transportation Research Part A : Policy and Practice* 44.5, p. 323–336.
- SWERTS, Elfie et Eric DENIS (2015). "Megacities : The Asian Era". In : *Urban Development Challenges, Risks and Resilience in Asian Mega Cities*. Springer, p. 1–28.
- TADMOR, Eitan (2012). "A review of numerical methods for nonlinear partial differential equations". In : *Bulletin of the American Mathematical Society* 49.4, p. 507–554.
- TAN, Wei et al. (2013). "Social-network-sourced big data analytics". In : *IEEE Internet Computing* 17.5, p. 62–69.
- TARDY, Christine (2004). "The role of English in scientific communication : lingua franca or Tyrannosaurus rex ?" In : *Journal of English for academic purposes* 3.3, p. 247–269.
- TAYLOR, Peter J (2016). "A Polymath in City Studies". In : *Sir Peter Hall : Pioneer in Regional Planning, Transport and Urban Geography*. Springer, p. 11–20.
- TEAM, R Core (2000). *R Language Definition*.
- TEPLITSKIY, M. et al. (2015). "Amplifying the Impact of Open Access : Wikipedia and the Diffusion of Science". In : *ArXiv e-prints*. arXiv : 1506.07608 [cs.DL].
- TERO, Atsushi et al. (2006). "Physarum solver : a biologically inspired method of road-network navigation". In : *Physica A : Statistical Mechanics and its Applications* 363.1, p. 115–119.
- TERO, Atsushi et al. (2010). "Rules for Biologically Inspired Adaptive Network Design". In : *Science* 327.5964, p. 439–442. DOI : 10.1126/science.1177894. eprint : <http://www.sciencemag.org/content/327/5964/439.full.pdf>. URL : <http://www.sciencemag.org/content/327/5964/439.abstract>.
- THÉVENIN, Thomas et al. (2013). "Mapping the Distortions in Time and Space : The French Railway Network 1830–1930". In : *Historical Methods : A Journal of Quantitative and Interdisciplinary History* 46.3, p. 134–143.
- THOM, René (1974). "Stabilité structurelle et morphogénèse". In : *Poetics* 3.2, p. 7–19.
- THOMAS, Isabelle et al. (2017). "City delineation in European applications of LUTI models : review and tests". In : *Transport Reviews* 0.0, p. 1–27. DOI : 10.1080/01441647.2017.1295112. eprint : <http://www.tandfonline.com/doi/pdf/10.1080/01441647.2017.1295112>. URL : <http://www.tandfonline.com/doi/abs/10.1080/01441647.2017.1295112>.

- THOMPSON, Darcy Wentworth et al. (1942). "On growth and form." In : *On growth and form*.
- TILMAN, David et Peter M KAREIVA (1997). *Spatial ecology : the role of space in population dynamics and interspecific interactions*. T. 30. Princeton University Press.
- TIMMERMANS, Harry (2003). "The saga of integrated land use-transport modeling : how many more dreams before we wake up ?" In : *Key-note paper, Moving through nets : The Physical and social dimension of travel, 10th International Conference on Travel Behaviour Research, Lucerna, www. ict. baug. ethz. ch/allgemein/pdf/timmermans. pdf*.
- TIVADAR, Mihai et al. (2014). "OASIS–un Outil d'Analyse de la Ségrégation et des Inégalités Spatiales". In : *Cybergeo : European Journal of Geography*.
- TOBLER, Waldo (2004). "On the first law of geography : A reply". In : *Annals of the Association of American Geographers* 94.2, p. 304–310.
- TORDEUX, Antoine et Sylvain LASSARRE (2016). "Jam avoidance with autonomous systems". In : *arXiv preprint arXiv :1601.07713*.
- TORRICELLI, Gian Paolo (2002). "Traversées alpines, ville et territoire : le paradoxe de la vitesse". In : *Revue de géographie alpine* 90.3, p. 25–36.
- Tošić, Predrag T et Carlos ORDONEZ (2017). "Boolean Network Models of Collective Dynamics of Open and Closed Large-Scale Multi-agent Systems". In : *International Conference on Industrial Applications of Holonic and Multi-Agent Systems*. Springer, p. 95–110.
- TSAI, Yu-Hsin (2005). "Quantifying urban form : compactness versus' sprawl'". In : *Urban studies* 42.1, p. 141–161.
- TSAY, Ruey S. (2015). *MTS : All-Purpose Toolkit for Analyzing Multivariate Time Series (MTS) and Estimating Multivariate Volatility Models*. R package version 0.33. URL : <http://CRAN.R-project.org/package=MTS>.
- TSEKERIS, Theodore et Nikolas GEROLIMINIS (2013). "City size, network structure and traffic congestion". In : *Journal of Urban Economics* 76.0, p. 1 –14. ISSN : 0094-1190. DOI : <http://dx.doi.org/10.1016/j.jue.2013.01.002>. URL : <http://www.sciencedirect.com/science/article/pii/S0094119013000119>.
- TSENG, Yuen-Hsien et al. (2007). "Text mining techniques for patent analysis". In : *Information Processing & Management* 43.5, p. 1216–1247.
- TUMMINELLO, Michele et al. (2005). "A tool for filtering information in complex systems". In : *Proceedings of the National Academy of Sciences of the United States of America* 102, p. 10421–10426.
- TURING, Alan Mathison (1952). "The chemical basis of morphogenesis". In : *Philosophical Transactions of the Royal Society of London B : Biological Sciences* 237.641, p. 37–72.

- VALLES-CATALA, Toni et al. (2016). "Multilayer stochastic block models reveal the multilayer structure of complex networks". In : *Physical Review X* 6.1, p. 011036.
- VARENNE, Franck (2010a). "Framework for M&S with Agents in Regard to Agent Simulations in Social Sciences". In : *Activity-Based Modeling and Simulation*, p. 53–84.
- (2010b). "Les simulations computationnelles dans les sciences sociales". In : *Nouvelles Perspectives en Sciences Sociales* 5.2, p. 17–49.
- VARENNE, Franck, Marc SILBERSTEIN et al. (2013). *Modéliser & simuler. Epistémologies et pratiques de la modélisation et de la simulation, tome 1.*
- VARET, Suzanne (2010). "Développement de méthodes statistiques pour la prédiction d'un gabarit de signature infrarouge". Thèse de doct. Université Paul Sabatier-Toulouse III.
- VATTAY, Gabor et al. (2015). "Quantum Criticality at the Origin of Life". In : *arXiv preprint arXiv :1502.06880*.
- VEIGA, Allan Koch et al. (2017). "A conceptual framework for quality assessment and management of biodiversity data". In : *PLOS ONE* 12.6, p. 1–20. DOI : [10.1371/journal.pone.0178731](https://doi.org/10.1371/journal.pone.0178731). URL : <https://doi.org/10.1371/journal.pone.0178731>.
- VERLINDE, E. P. (2016). "Emergent Gravity and the Dark Universe". In : *ArXiv e-prints*. arXiv : [1611.02269 \[hep-th\]](https://arxiv.org/abs/1611.02269).
- VILLANI, C. (2017). *Mathématique de la Chauve-Souris*. Conférence à la Maison des Metallos, 16 janvier 2017.
- VISSEUR, Hans et T DE NIJS (2006). "The map comparison kit". In : *Environmental Modelling & Software* 21.3, p. 346–358.
- WAL, Anne L. J. Ter et Ron BOSCHMA (2011). "Co-evolution of Firms, Industries and Networks in Space". In : *Regional Studies* 45.7, p. 919–933. DOI : [10.1080/00343400802662658](https://doi.org/10.1080/00343400802662658). eprint : <http://dx.doi.org/10.1080/00343400802662658>. URL : <http://dx.doi.org/10.1080/00343400802662658>.
- WANG, Jiang-Jiang et al. (2009). "Review on multi-criteria decision analysis aid in sustainable energy decision-making". In : *Renewable and Sustainable Energy Reviews* 13.9, p. 2263–2278.
- WANG, Y.-S. et al. (2017). "Separable and Localized System Level Synthesis for Large-Scale Systems". In : *ArXiv e-prints*. arXiv : [1701.05880 \[math.OC\]](https://arxiv.org/abs/1701.05880).
- WARD, Douglas P et al. (2000). "A stochastically constrained cellular model of urban growth". In : *Computers, Environment and Urban Systems* 24.6, p. 539–558.
- WARDROP, John Glen (1952). "Some theoretical aspects of road traffic research." In : *Proceedings of the institution of civil engineers* 1.3, p. 325–362.
- WATSON, Benjamin et al. (2008). "Procedural urban modeling in practice". In : *IEEE Computer Graphics and Applications* 3, p. 18–26.

- WEE, Bert (2015). "Viewpoint : Toward a new generation of land use transport interaction models". In : *Journal of Transport and Land Use* 8.3. ISSN : 1938-7849. URL : <https://www.jtlu.org/index.php/jtlu/article/view/611>.
- WEGENER, Michael et Franz FÜRST (2004). "Land-use transport interaction : state of the art". In : Available at SSRN 1434678.
- WEGENER, Michael et al. (1991). "One city, three models : comparison of land-use/transport policy simulation models for Dortmund". In : *Transport Reviews* 11.2, p. 107–129.
- WEIBULL, Jörgen W (1976). "An axiomatic approach to the measurement of accessibility". In : *Regional Science and Urban Economics* 6.4, p. 357–379.
- WEST, Geoffrey (2017). *Scaling, the universal laws of growth, innovation, sustainability, and the pace of life in organisms, cities, economies, and companies*. Penguin Press.
- WHITEHAND, JWR et al. (1999). "Urban morphogenesis at the microscale : how houses change". In : *Environment and Planning B : Planning and Design* 26.4, p. 503–515.
- WHITNEY, D. E. (2012). "Growth Patterns of Subway/Metro Systems Tracked by Degree Correlation". In : ArXiv e-prints. arXiv : 1202.1747 [physics.soc-ph].
- WIENER, Norbert (1948). *Cybernetics*. Hermann Paris.
- WILENSKY, Uri (1999). "NetLogo". In :
- WILSON, A. (1981). *Catastrophe theory and bifurcation : Application to Urban and Regional System*. London : Croom Helm.
- WILSON, G et al. (2017). "Good enough practices in scientific computing". In : *PLoS Comput Biol* 13.6, e1005510.
- WOLFRAM, Stephen (2002). *A new kind of science*. T. 5. Wolfram media Champaign.
- WOLPERT, L (2011). "Positional information and patterning revisited". In : *J Theor Biol* 269.1, p. 359–365. DOI : 10.1016/j.jtbi.2010.10.034. URL : http://www.ncbi.nlm.nih.gov/pubmed/21044633http://ac.els-cdn.com/S0022519310005795/1-s2.0-S0022519310005795-main.pdf?{_}tid=5a5ab764-a7fa-11e5-a834-00000aacb35d{\&}acdnat=1450713120{_}44ebcc31cb7683da9cfaa05d97abdb37http://ac.els-cdn.com/S0022519310005795/1-s2.
- WOLPERT, Lewis (1969). "Positional Information and the Spatial Pattern of Cellular Differentiation". In : *J Theor Biol*, p. 1–47.
- WU, Jianjun et al. (2017). "City expansion model based on population diffusion and road growth". In : *Applied Mathematical Modelling* 43, p. 1–14.
- XIE, Feng et David LEVINSON (2009a). "How streetcars shaped suburbanization : a Granger causality analysis of land use and transit in the Twin Cities". In : *Journal of Economic Geography*, lbp031.
- (2009b). "Jurisdictional control and network growth". In : *Networks and Spatial Economics* 9.3, p. 459–483.

- XIE, Feng et David LEVINSON (2009c). "Modeling the growth of transportation networks : A comprehensive review". In : *Networks and Spatial Economics* 9.3, p. 291–307.
- (2011). *Evolving transportation networks*. Springer Science & Business Media.
- XIE, Yichun (1996). "A Generalized Model for Cellular Urban Dynamics". In : *Geographical Analysis* 28.4, p. 350–373. ISSN : 1538-4632. DOI : [10.1111/j.1538-4632.1996.tb00940.x](https://doi.org/10.1111/j.1538-4632.1996.tb00940.x). URL : <http://dx.doi.org/10.1111/j.1538-4632.1996.tb00940.x>.
- XIE, Yihui (2013). "knitr : A general-purpose package for dynamic report generation in R". In : *R package version 1.7*.
- YAMASAKI, Kazuko et al. (2006). "Preferential attachment and growth dynamics in complex systems". In : *Physical Review E* 74.3, p. 035103.
- YAMINS, Daniel et al. (2003). "Growing urban roads". In : *Networks and Spatial Economics* 3.1, p. 69–85.
- YANG, Yiming et al. (2000). "Improving text categorization methods for event tracking". In : *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, p. 65–72.
- YE, Lin (2014). "State-led metropolitan governance in China : Making integrated city regions". In : *Cities* 41, Part B.o. Chinese Cities in a Globalizing Context, p. 200 –208. ISSN : 0264-2751. DOI : [http://dx.doi.org/10.1016/j.cities.2014.03.001](https://doi.org/10.1016/j.cities.2014.03.001). URL : <http://www.sciencedirect.com/science/article/pii/S0264275114000365>.
- YE, Xin (2011). "Investigation of Underlying Distributional Assumption in Nested Logit Model Using Copula-Based Simulation and Numerical Approximation". In : *Transportation Research Record : Journal of the Transportation Research Board* 2254, p. 36–43.
- YERRA, Bhanu M et David M LEVINSON (2005). "The emergence of hierarchy in transportation networks". In : *The Annals of Regional Science* 39.3, p. 541–553.
- YOON, Byungun et Yongtae PARK (2004). "A text-mining-based patent network : Analytical tool for high-technology trend". In : *The Journal of High Technology Management Research* 15.1, p. 37–50.
- YOON, Janghyeok et Kwangsoo KIM (2011). "Detecting signals of new technological opportunities using semantic patent analysis and outlier detection". In : *Scientometrics* 90.2, p. 445–461.
- YOUN, Hyejin et al. (2015). "Invention as a combinatorial process : evidence from US patents". In : *Journal of The Royal Society Interface* 12.106. ISSN : 1742-5689. DOI : [10.1098/rsif.2015.0272](https://doi.org/10.1098/rsif.2015.0272).
- ZEMBRI, Pierre (1997). "Les fondements de la remise en cause du Schéma Directeur des liaisons ferroviaires à grande vitesse : des faiblesses avant tout structurelles". In : *Annales de géographie. JSTOR*, p. 183–194.
- (2008). "La contribution de la grande vitesse ferroviaire à l'interrégionalité en France.(High-speed rail and inter-regionality in

- France)". In : *Bulletin de l'Association de géographes français* 85.4, p. 443–460.
- (2010). "The new purposes of the French high-speed rail system in the framework of a centralized network : a substitute to the domestic air transport market?" In :
- ZHANG, Kuilin et al. (2013). "Dynamic pricing, heterogeneous users and perception error : Probit-based bi-criterion dynamic stochastic user equilibrium assignment". In : *Transportation Research Part C : Emerging Technologies* 27, p. 189–204.
- ZHANG, Lei et David M LEVINSON (2016). "A model of the rise and fall of roads". In :
- ZHANG, Lei et David LEVINSON (2007). "The economics of transportation network growth". In : *Essays on transport economics*. Springer, p. 317–339.
- ZHANG, Yingjia et al. (2015). "Density and diversity of OpenStreetMap road networks in China". In : *Journal of Urban Management* 4.2, p. 135–146.
- ZHANG, Zhonghao et al. (2013). "Identifying determinants of urban growth from a multi-scale perspective : A case study of the urban agglomeration around Hangzhou Bay, China". In : *Applied Geography* 45, p. 193–202.
- ZHENG, Shudan et Jianghua ZHENG (2014). "Assessing the completeness and positional accuracy of OpenStreetMap in China". In : *Thematic Cartography for the Society*. Springer, p. 171–189.
- ZHOU, Suhong (2016). "The Development of the PRD and the New Pathways for Sustainable Urban Development of Zhuhai". In : *Medium Seminar - Urban Sustainable Development in Zhuhai*. Sun Yat-Sen University, Guangzhou, China.
- ZHU, Liping et al. (2013a). "Amoeba-based computing for traveling salesman problem : Long-term correlations between spatially separated individual cells of *Physarum polycephalum*". In : *Biosystems* 112.1, p. 1–10.
- ZHU, Shanjiang et David LEVINSON (2010). "Do people use the shortest path ? An empirical test of Wardrop's first principle". In : *91th annual meeting of the Transportation Research Board, Washington*. T. 8. Citeseer.
- ZHU, Yaojia et al. (2013b). "Scalable text and link analysis with mixed-topic link models". In : *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, p. 473–481.
- ZILSEL (2015). "LA POSITION DE LA REVUE SOCIÉTÉS DANS L'ESPACE DISCURSIF DE LA SOCIOLOGIE FRANÇAISE". In : <http://zilsel.hypotheses.org/category/canular>.
- ZIMAN, John (2003). *Technological innovation as an evolutionary process*. Cambridge University Press.
- TEAM, Osmosis (2016). OSMOSIS. <http://wiki.openstreetmap.org/wiki/Osmosis>.

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.

A

INFORMATIONS SUPPLÉMENTAIRES

This appendix gathers various supplementary materials, necessary for the robustness but not necessary to the main argument. It includes for example more precise model explorations, generally needed to support conclusions in main text but too long or repetitive to be included.

A.1 ELEMENTS DE TERRAIN

A.1.1 Carnet de Terrain

Nous listons ici de manière synthétique les différentes sorties de terrain alimentant la section 1.3. S'il n'est a priori pas standard de fournir de manière brute et ouverte le contenu des carnets de terrain, [GOFFMAN, 1989] souligne que celui-ci peut être un matériau de recherche en lui-même. Les compte-rendus bruts et les photos sont disponibles de manière ouverte à .

29 / 10 / 2016 Sortie à Zhuhai, Xiangzhou et Gongbei, journée. Expérience de la nature en ville.

07 / 11 / 2016

16 / 01 / 2017 Tentative de relier Tangjia à Guangzhou par bus de ville, journée. Itinéraire final Tangjia-Zhongshan-Xiaolan-Zhuahaibei.

11 / 12 / 2016 De Pekin à Guangzhou à Shenzhen par Dongguan.

19 / 06 / 2017 Visite de terrain officielle (Conférence Medium, SYSU), Guangzhou.

8 / 06 / 2017 Hong-Kong → Zhuhai → Tangjia

11 / 07 / 2017 Aller-retour Tangjia-Guangzhou.

24 / 07 / 2017 Sortie à Tangjia.

31 / 07 / 2017 Xiangzhou ; test du Tramway, Ligne 2.

09 / 08 / 2017 Sortie TOD : terminus ouest Tram ; bus pour la gare de Tangjia le long de la ligne à grande vitesse.

13 / 08 / 2017 Train à Grande Vitesse : Yangshuo → Guangzhou-Nan

17 / 08 / 2017 Bureau du Planning Committee de la Zhuhai High Tech Zone.

20 / 08 / 2017 Traversée de Leshan en bus, aller-retour.

21 / 08 / 2017 Guangzhou Baiyun → Zhongshan Daxue Nan Qi → Tangjia

A.1.2 *Entretiens*

A.2 EPISTÉMOLOGIE QUANTITATIVE

A.2.1 Revue systématique algorithmique

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 projet¹. Les requêtes au catalogue, qui consistent à récupérer un ensemble de références à partir d'un ensemble de mots-clés, sont faites via l'API du logiciel Mendeley [MENDELEY, 2015] 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 [CHAVALARIA et COINTET, 2013], via un script Python qui utilise [BIRD, 2006].

CONVERGENCE ET ANALYSE DE SENSIBILITÉ **C : (Florent)** avec quels mots clés as tu validé empiriquement la convergence de l'algo ?

Une preuve formelle de convergence de l'algorithme n'est guère envisageable puisque qu'elle dépendra de la structure empirique inconnue 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 figure ??, 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.

A.2.2 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. ??.

* * *

*

¹ à l'adresse <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/QuantEpistemo/AlgoSP>

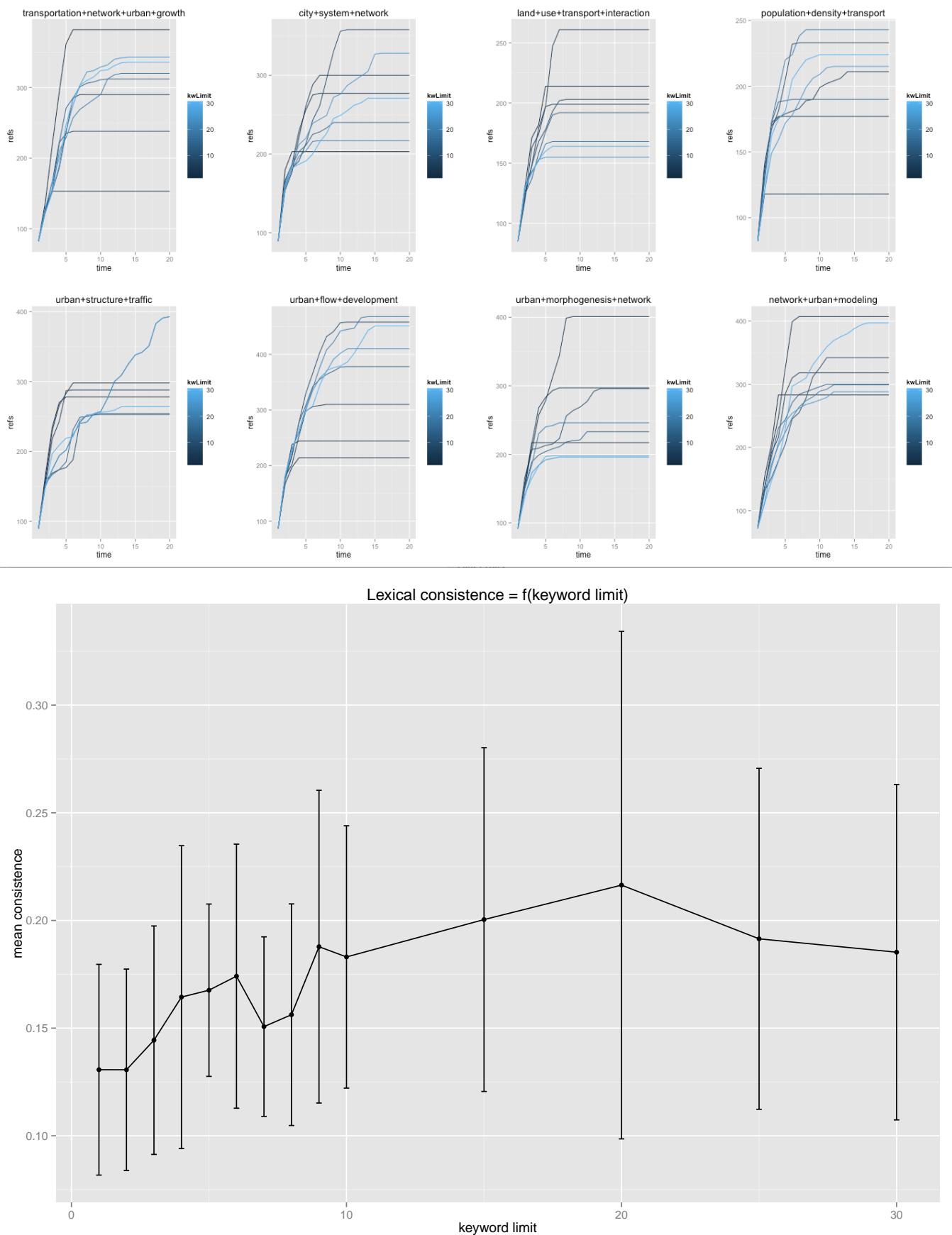


FIGURE 61: Convergence et analyse de sensibilité de l'algorithme de revue systématique. C (FL) : illisible

[24 octobre 2017 at 18:35 – Thesis version 3.3.2]

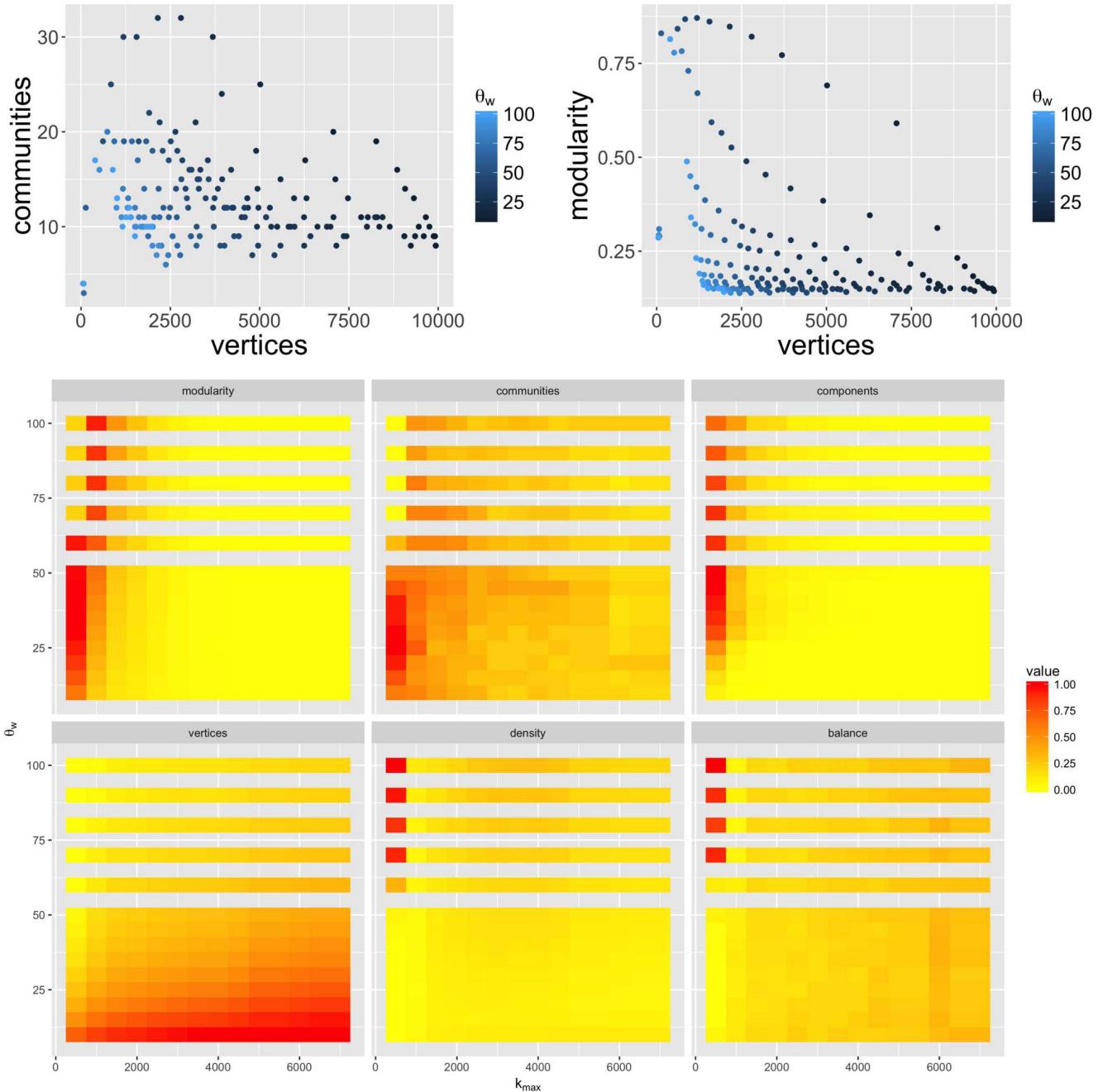


FIGURE 62: Analyse de sensibilité des propriétés modulaires du réseau sémantique en fonction des paramètres de filtrage.

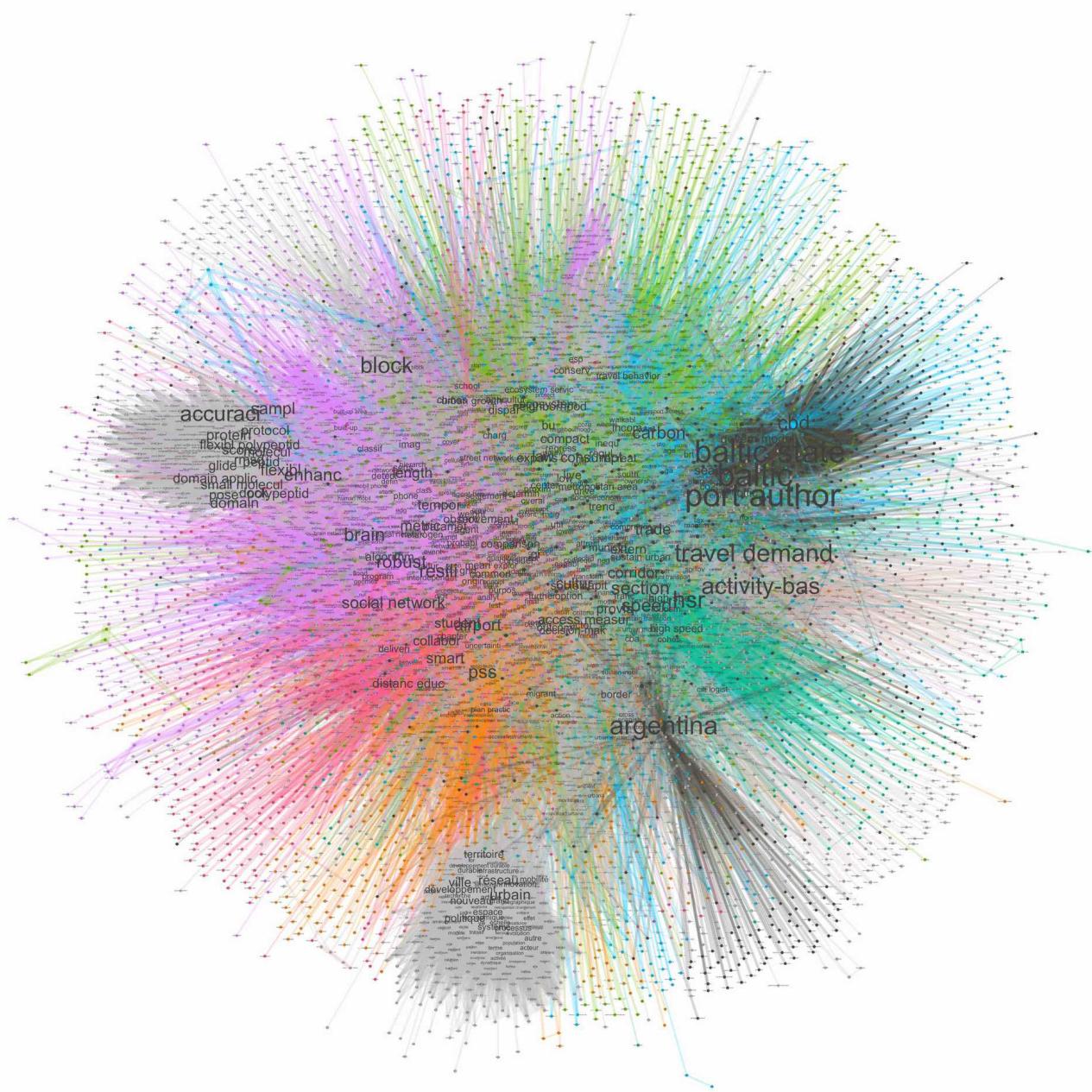


FIGURE 63: Réseau sémantique des domaines

A.3 MODÉOGRAPHIE

A.3.1 Méthodologie de la revue systématique

Choix possible : extraire les mots-clés pertinents par sous-communautés du réseau de citations, puis prendre les plus pertinents ensuite. ou (a priori ce que l'on fait) extraire sur le corpus complet, puis récupérer par sous-communautés. Pour un petit corpus, deuxième plus souhaitable, notion de pertinence moins importante que pour du big-data (mentionner Patents et Cybergeo, ressemblances et différences).

REMARQUES SUR LA CLASSIFICATION MANUELLE Journals → disciplines jtgeo : geography, jtlu : transportation, transportation research, epb : geography...

geography includes urbanisme, etudes urbaines si pas trop proche du planning (urban durable).

A.3.2 Meta-analyse

Nous donnons ici les résultats numériques complets des analyses statistiques reliant caractéristiques de modèles et variables explicatives. Uniquement le meilleur modèle, choisi par critère d'AIC parmi l'ensemble des modèles possibles, est donné.

ECHELLE DE TEMPS L'échelle de temps est ajustée selon le modèle linéaire suivant :

TEMPSCALE ~YEAR + CITCOM + TYPE + SPATSCALE + FMETHOD
+ DISCIPLINE + INTERDISC + SEMCOM

avec un R² ajusté de 0.969.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.382e+04	2.379e+03	5.809	0.01015 *
YEAR	-6.181e+00	1.109e+00	-5.572	0.01141 *
CITCOMInfra Planning	1.117e+02	2.234e+01	4.999	0.01540 *
CITCOMLUTI	1.140e+02	1.956e+01	5.827	0.01007 *
CITCOMNetworks	7.977e+01	2.243e+01	3.556	0.03793 *
TYPEterritory	-2.234e+02	2.233e+01	-10.004	0.00213 **
SPATSCALE	9.248e-02	1.807e-02	5.119	0.01443 *
FMETHODsem	-1.130e+02	2.640e+01	-4.281	0.02341 *
FMETHODsim	-1.212e+03	1.962e+02	-6.180	0.00853 **
FMETHODstat	-1.029e+03	1.968e+02	-5.228	0.01361 *
DISCIPLINEgeography	-2.285e+00	1.253e+01	-0.182	0.86692
DISCIPLINEphysics	-6.824e+01	1.111e+01	-6.140	0.00869 **
DISCIPLINEplanning	9.928e+02	1.826e+02	5.437	0.01222 *
DISCIPLINEtransportation	-2.360e+01	1.128e+01	-2.092	0.12750
INTERDISC	-1.635e+02	3.757e+01	-4.353	0.02239 *
SEMCOMinfra planning	-1.171e+02	2.690e+01	-4.353	0.02240 *
SEMCOMnetworks	-8.645e+01	2.009e+01	-4.303	0.02309 *
SEMCOMtod	-1.027e+03	1.905e+02	-5.393	0.01249 *

ECHELLE D'ESPACE L'échelle spatiale est ajustée selon le modèle linéaire suivant :

SPATSCALE ~YEAR + CITCOM + TYPE + TEMPSCALE +
FMETHOD + DISCIPLINE + INTERDISC + SEMCOM

avec un R² ajusté de 0.999.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.440e+05	1.890e+04	-7.622	0.00469 **
YEAR	6.436e+01	9.165e+00	7.022	0.00593 **
CITCOMInfra Planning	-1.188e+03	1.354e+02	-8.773	0.00312 **
CITCOMLUTI	-1.188e+03	1.549e+02	-7.667	0.00461 **
CITCOMNetworks	-8.778e+02	1.358e+02	-6.466	0.00751 **
TYPEterritory	2.087e+03	5.872e+02	3.555	0.03796 *
TEMPSCALE	9.702e+00	1.895e+00	5.119	0.01443 *
FMETHODsem	1.210e+03	1.771e+02	6.834	0.00641 **
FMETHODsim	1.287e+04	4.487e+02	28.685	9.30e-05 ***
FMETHODstat	1.110e+04	1.581e+02	70.218	6.37e-06 ***
DISCIPLINEgeography	7.562e+01	1.214e+02	0.623	0.57749
DISCIPLINEphysics	6.551e+02	1.810e+02	3.620	0.03626 *
DISCIPLINEplanning	-1.067e+04	1.944e+02	-54.871	1.33e-05 ***
DISCIPLINEtransportation	2.622e+02	9.951e+01	2.635	0.07799 .
INTERDISC	1.644e+03	4.269e+02	3.852	0.03090 *
SEMCOMinfra planning	1.233e+03	2.203e+02	5.597	0.01127 *
SEMCOMnetworks	9.091e+02	1.680e+02	5.411	0.01238 *
SEMCOMtod	1.105e+04	1.945e+02	56.805	1.20e-05 ***

INTERDISCIPLINARITÉ Le niveau d'interdisciplinarité est expliqué par l'année uniquement :

INTERDISC ~ YEAR

avec un R² ajusté de 0.047.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.218932	2.314180	2.687	0.00849 **
YEAR	-0.002781	0.001153	-2.413	0.01772 *

ANNÉE L'année est ajustée selon le modèle :

YEAR ~ CITCOM + TYPE + TEMPSCALE + SPATSCALE + FMETHOD
+ DISCIPLINE + INTERDISC + SEMCOM

	Estimate	Std. Error	t value Pr(> t)	
(Intercept)	2.227e+03	2.598e+01	85.729	3.5e-06 ***
CITCOMInfra Planning	1.779e+01	2.373e+00	7.498	0.00492 **
CITCOMLUTI	1.806e+01	1.940e+00	9.307	0.00263 **
CITCOMNetworks	1.310e+01	2.326e+00	5.633	0.01107 *
TYPEterritory	-3.217e+01	7.996e+00	-4.024	0.02758 *
TEMPSCALE	-1.475e-01	2.648e-02	-5.572	0.01141 *
SPATSCALE	1.465e-02	2.086e-03	7.022	0.00593 **
FMETHODsem	-1.838e+01	2.372e+00	-7.749	0.00447 **
FMETHODsim	-1.903e+02	2.334e+01	-8.151	0.00386 **
FMETHODstat	-1.633e+02	2.143e+01	-7.622	0.00469 **
DISCIPLINEgeography	-8.011e-01	1.890e+00	-0.424	0.70027
DISCIPLINEphysics	-1.008e+01	2.474e+00	-4.076	0.02667 *
DISCIPLINEplanning	1.570e+02	2.057e+01	7.631	0.00467 **
DISCIPLINEtransportation	-3.701e+00	1.704e+00	-2.172	0.11821
INTERDISC	-2.499e+01	6.192e+00	-4.036	0.02735 *
SEMCOMinfra planning	-1.835e+01	3.752e+00	-4.891	0.01635 *
SEMCOMnetworks	-1.355e+01	2.811e+00	-4.820	0.01702 *
SEMCOMtod	-1.625e+02	2.153e+01	-7.547	0.00482 **

A.4 CORRELATIONS STATIQUES

A.4.1 Mesures morphologiques

A.4.2 Algorithme de Simplification du Réseau

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.
- 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. **C : (Florent)** y'a t'il un effet de bord dans les carrés 50x50 qui se trouvent à la frontière de 2 pays **A1** : pas avec nouvelle parallelisation pas par pays mais par split and merge (TODO rewrite nouvel algo)

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 preprocesing.

A.4.3 Résilience des Réseaux

La description complète d'un réseau suppose la donnée d'une grande quantité d'information, puisque la moitié de la matrice d'adjacence est nécessaire dans le cas non-dirigé (matrice complète dans le cas dirigé). L'établissement de typologies, c'est à dire de sortes de classes d'équivalence topologiques au sens large, est une façon de voir les enjeux de la recherche actuelle sur les réseaux : existe-t-il des formes typiques de réseau, et comment les représenter dans une dimension réduite ? Relativement importante est la dimension épistémologique de

cette interrogation fondamentale, qui peut sous certaines hypothèses être ramenée à l'opposition d'un réductionisme à une vision intrinsèque de la complexité. Si une classification systématique réductrice existe pour tout système complexe, alors les niveaux d'émergence supérieurs n'ont pas de signification propre. Il est paradoxal d'observer dans ce cas la position ambiguë de certains travaux de physiciens qui tiennent ce réductionisme comme un dogmatisme mais prétendent s'attaquer à des problèmes complexes typiques des systèmes socio-techniques.

Les mesures de réseau globales comme nous avons vu en cours sont une façon de répondre partiellement à cette question de réduction dimensionnelle. En pratique, on veut être également capable de relier ces mesures à des propriétés pratiques du réseau, qui seront cruciales pour le design et le management de réseaux réels (par exemple réseaux techniques : transport d'électricité, internet ; réseaux de transport ; réseaux sociaux ; réseaux de villes ; etc.) : par exemple le coût, la résilience, la performance. Cet exercice propose de survoler ces deux questions fondamentales pour différents exemples de réseaux réels et synthétiques : illustrer dans un premier temps la signification concrète de différentes valeurs en relation à la topologie "apparente" du réseau ; dans un second temps explorer des liens potentiels entre mesures et propriétés afin de donner une idée d'une manière de caractériser la résilience.

Ce qui est demandé pourra paraître relativement simple mais est central pour une appréhension juste de la complexité des processus géographiques tels qu'ils occurront dans toute leur réalité. La table 11 donne les valeurs d'indicateurs, pour certains types de réseaux. On fournit leur valeur moyenne et leur écart-type dans le cas de réseaux synthétiques aléatoires pour lesquels les mesures auront alors été estimées sur $b = 500000$ répétitions des réseaux aléatoires.² Pour les réseaux réels ou synthétiques non variables, une valeur seule est donnée, sachant que l'estimation de paramètres moyens dans des situations réelles est directement liée à la stationnarité spatio-temporelle des processus³, ce qui est également une question ouverte concernant les systèmes spatiaux complexes. Les figures ?? et ?? illustrent des exemples des réseaux considérés. Les questions suivantes portent sur une interprétation simple des mesures. Parmi les réseaux considérés, on étudie des réseaux routiers réels, dont la localisation spatiales est présentée en figure ??, ainsi que des réseaux synthétiques.

² qu'on fixe comme arbitraire. Le problème du nombre de répétitions nécessaires pour une convergence raisonnable des indicateurs statistiques et hors du champ de ce devoir, et d'ailleurs un problème ouvert pour la plupart des modèles de simulation complexes puisque qu'on sait établir des intervalles de confiances soit sous certaines hypothèses théoriques de distribution statistique soit par simulation (*bootstrap*) ce qui ne réduit pas la complexité intrinsèque.

³ et donc à leur ergodicité, i.e. à l'équivalence entre moyenne spatiale et moyenne temporelle

Les réseaux étudiés sont donc les suivants⁴ :

- Réseau routiers réels (simplifiés à une résolution de 100m) : Paris, Ile-de-France, La Courtine (Creuse), Grand Lyon, London Metropolitan Area, Randstad
- Aléatoire (probabilité fixe d'établir un lien entre chaque paire de noeud)
- Attachement préférentiel (type Barabasi-Albert : les liens sont établis itérativement avec une probabilité proportionnelle au degré des noeuds)
- Grille perturbée (grille régulière dont on retire une proportion fixée de liens)
- Arbre (au nombre de feuilles par branche fixe)

Les mesures calculées pour un réseau $N = (V, E)$ sont :

- Statistiques descriptives : nombre de noeuds $|V|$ et nombre de liens $|E|$
- Densité γ
- Degré moyen \bar{d}
- Diamètre⁵ δ
- Centralité d'intermédiairité b : s'agissant d'une mesure locale (associée ici aux liens), on considère sa moyenne $\langle b \rangle$ et son niveau de hiérarchie⁶ $\alpha [b]$
- Centralité de proximité c : de même on calcule $\langle c \rangle$ et $\alpha [c]$
- Efficacité $e = \frac{2}{n \cdot (n-1)} \sum_{i < j} \frac{1}{d_{ij}}$ avec d_{ij} distance topologique entre i et j
- Coefficient de clustering t , qui donne la probabilité que les voisins d'un noeud soient connectés
- Modularité μ qui donne une mesure plus générale de la structure en communauté du graphe

⁴ chaque générateur synthétique a des paramètres propres, pour lesquels nous choisissons les valeurs par défaut suivantes : probabilité aléatoire d'Erdos-Renyi $p = 0.005$; proportion de liens de la grille conservés 65%; attachement préférentiel : nouveau liens $m = 10$, exposant $\alpha = 1$, exposant de vieillissement $\beta = -2$, pas de vieillissement 100; nombre de feuille par branche de l'arbre $f = 3$.

⁵ pour toutes les mesures liées au plus courts chemins, les distances topologiques et non pondérées ont été prises en compte, afin de permettre la comparabilité des réseaux réels et des réseaux synthétiques. Pour une comparabilité des réseaux réels ayant des couvertures géographiques d'étendue significativement différentes, il faut normaliser par le diamètre.

⁶ donné par la pente de la regression linéaire d'un fit brutal d'une loi rang-taille : $\log b_i = \beta + \alpha \cdot \log i$ où les b_i sont triés par ordre décroissants.

Réseau	$ V $	$ E $	γ	\bar{d}	δ	$\langle b \rangle$
Aléatoire	1000	2498 ± 50	$0.005 \pm 1 \cdot 10^{-4}$	5 ± 0.1	9 ± 0.59	$0.0018 \pm 5.4 \cdot 10^{-5}$
Att. Préf.	1000	4579 ± 21	$0.0092 \pm 4.3 \cdot 10^{-5}$	9.2 ± 0.043	7 ± 0.18	$0.00084 \pm 8.6 \cdot 10^{-6}$
Grille	499 ± 3.7	624	$0.005 \pm 7.5 \cdot 10^{-5}$	2.5 ± 0.019	53 ± 4.5	0.03 ± 0.0027
Arbre	1000	999	0.002	2	12	0.0099
Ile-de-France	$ V $	$ E $	γ	\bar{d}	δ	$\langle b \rangle$
Paris	$ V $	$ E $	γ	\bar{d}	δ	$\langle b \rangle$
Grand Lyon	$ V $	$ E $	γ	\bar{d}	δ	$\langle b \rangle$
La Courtine	$ V $	$ E $	γ	\bar{d}	δ	$\langle b \rangle$
Randstad	$ V $	$ E $	γ	\bar{d}	δ	$\langle b \rangle$
London	$ V $	$ E $	γ	\bar{d}	δ	$\langle b \rangle$
Réseau	$\alpha [b]$	$\langle c \rangle$	$\alpha [c]$	e	t	μ
Aléatoire	-0.29 ± 0.0092	0.46 ± 0.68	-0.076 ± 0.032	0.24 ± 0.003	0.005 ± 0.0011	0.45 ± 0.0072
Att. Préf.	-0.63 ± 0.015	0.26 ± 0.0021	-0.062 ± 0.0022	0.28 ± 0.0018	0.079 ± 0.0028	0.66 ± 0.0082
Grille	-1.2 ± 0.077	7.3 ± 3.5	-0.99 ± 0.31	0.072 ± 0.0032	0	0.87 ± 0.0058
Arbre	-1	0.1	-0.094	0.11	0	0.93

TABLE 11 : Valeurs de mesures de réseaux pour différents exemples typiques réels et synthétiques

FORMES URBAINES ET INDICATEURS DE RÉSEAUX Commentez qualitativement les différentes formes des systèmes territoriaux observées. On pourra par exemple se poser la question de la polycentricité d'une Méga-région Urbaine. Reliez ces observations aux valeurs prises par les indicateurs que vous jugez pertinents.

INTERPRÉTATION DES INDICATEURS Pour les réseaux synthétiques, commentez les valeurs prises par les indicateurs. Lesquelles étaient intuitivement attendues ? Dans quelle mesure est-on capable de caractériser et discriminer chaque type de réseau. De quel réseau synthétique s'attend-on à ce que les réseaux réels soient les plus proches ? Peut-on le confirmer par les indicateurs ?

Nous proposons à présent de tenter une caractérisation de la résilience des réseau. La définition utilisée prend en compte la capacité du réseau à rester performant face à la rupture de lien, comme proposé par [ASH et NEWTH, 2007]. On considère la rupture aléatoire d'une proportion fixée α de liens, et on note N_α le réseau résultant de la suppression à partir du réseau N . L'indicateur de résilience au niveau α est alors défini par $r = \mathbb{E} [\Delta_\alpha e] = \mathbb{E} [e(N_\alpha) - e(N)]$. On peut définir de même les variations des autres indicateurs. La structure de covariance des différentes variations, et en particulier la correlation

Réseau	$ V $	$ E $	γ	\bar{d}	δ	$\langle b \rangle$	$\alpha [b]$	$\langle c \rangle$	$\alpha [c]$	e	t	μ
Aléatoire	$ V $	$ E $	γ	\bar{d}	δ	$\langle b \rangle$	$\alpha [b]$	$\langle c \rangle$	$\alpha [c]$	e	t	μ
Attachement Préférentiel	$ V $	$ E $	γ	\bar{d}	δ	$\langle b \rangle$	$\alpha [b]$	$\langle c \rangle$	$\alpha [c]$	e	t	μ
Grille Perturbée	$ V $	$ E $	γ	\bar{d}	δ	$\langle b \rangle$	$\alpha [b]$	$\langle c \rangle$	$\alpha [c]$	e	t	μ

TABLE 12 : Correlations estimées

avec la résilience, est un moyen indirect de caractériser la résilience, au sens de quel type de propriété permet au réseau d'augmenter sa résilience. La table 12 donne pour chaque réseau aléatoire les corrélations estimées $\rho[X] = \hat{\rho}[\Delta_\alpha e, \Delta_\alpha X]$ où l'estimateur $\hat{\rho}$ est calculé en pratique sur une plage de valeurs pour α (de 0.5 à 0.95 par 0.05) et sur un nombre de répétitions fixé ($b =$, en faisant l'hypothèse que répéter sur le réseau est équivalent à répéter sur la suppression des liens).

CARACTÉRISATION DE LA RÉSILIENCE Commentez à partir de la table des corrélations, pour les différents types de réseau, les facteurs influençant la résilience. Quel enseignement peut-on en tirer pour la conception de réseaux techniques par exemple ?

RÉSILIENCE DYNAMIQUE ET TOPOLOGIE Les approches à la notion de résilience sont diverses et complémentaires. L'aspect dynamique, au sens par exemple du temps nécessaire au système pour retrouver son état initial après perturbation, est particulièrement intéressant pour les systèmes urbains. Dans le cas de réseaux où les noeuds ont leur dynamique propre, les solutions pour une définition robuste et universelle sont très récentes, comme celle proposée par [gao2016universal].

Vous semble-t-il simple, dans le cas de dynamiques couplées au sein d'un réseau, d'isoler la contribution de la topologie du réseau de celle des dynamiques propres à la dynamique générale ? Dans quelle mesure serait-il alors complexe de quantifier la résilience dynamique dans une dimension réduite ?

A.4.4 Indicateurs de réseau

A.4.5 Corrélations Spatiales

★ ★

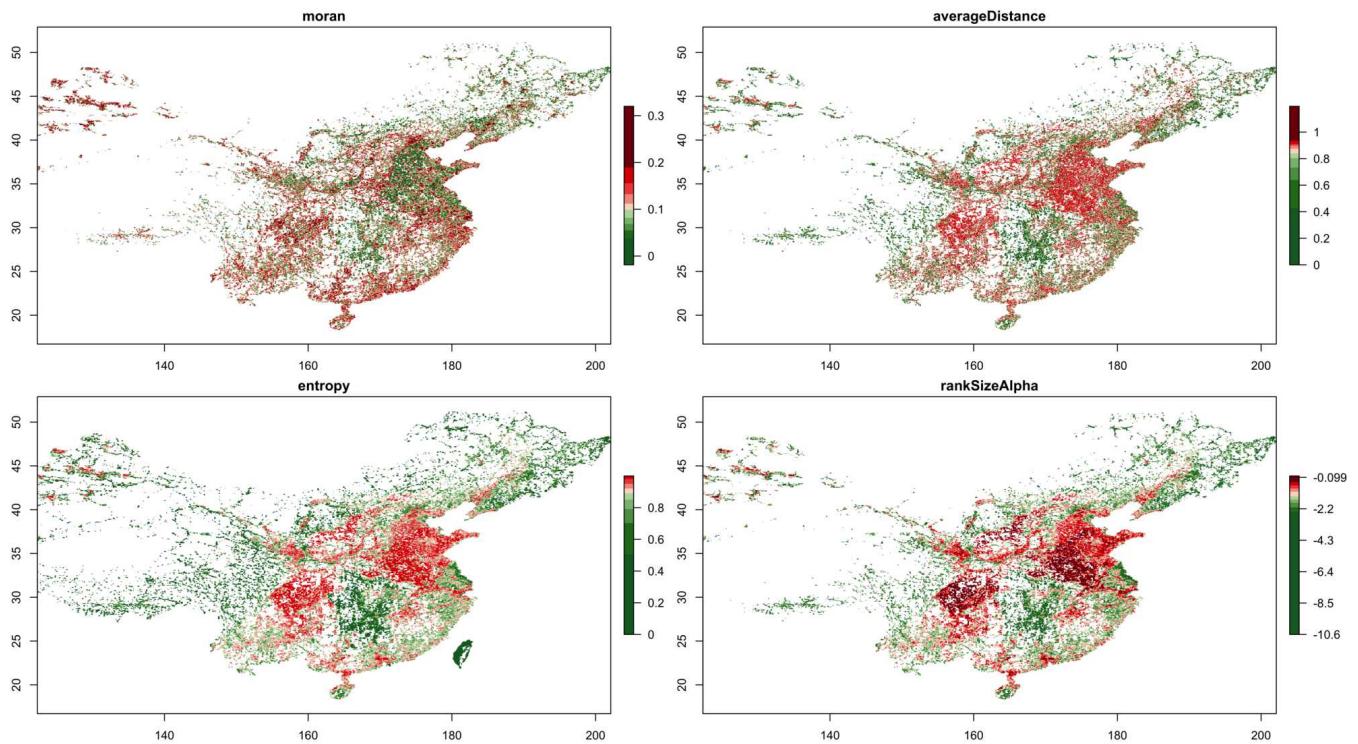


FIGURE 64: Indicateurs morphologiques pour la Chine.

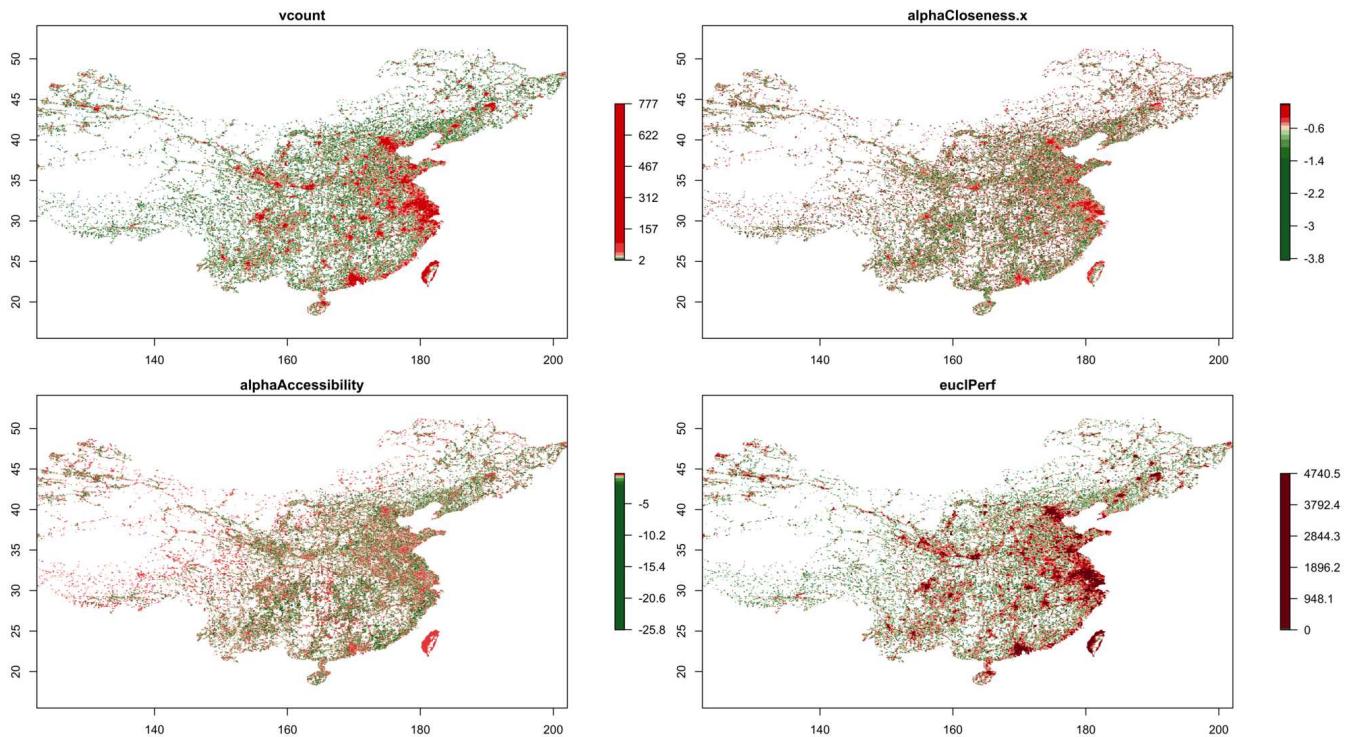


FIGURE 65: Indicateurs de réseau pour la Chine.

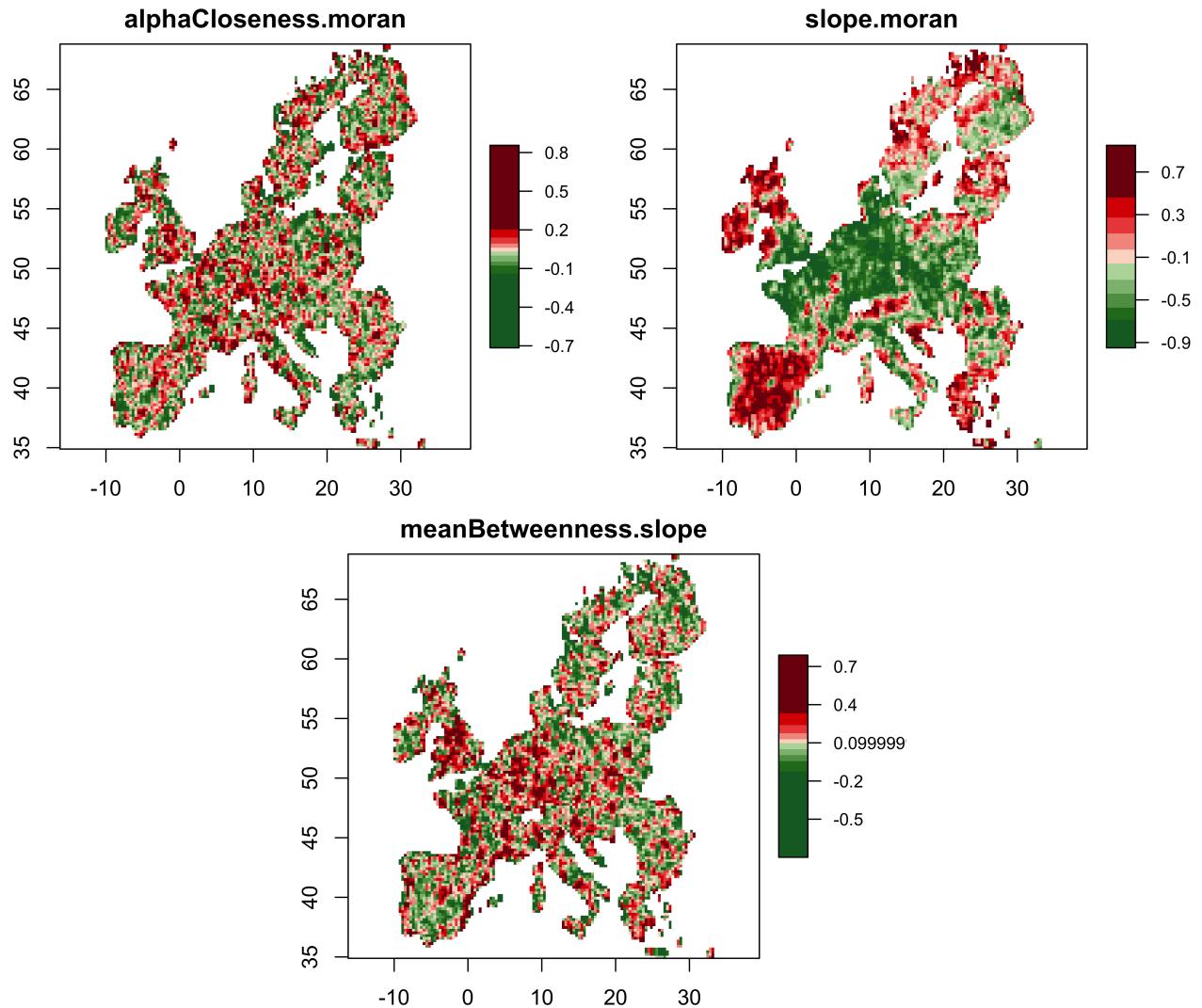


FIGURE 66: Correlations spatiales pour l'Europe.

*

A.5 RÉGIMES DE CAUSALITÉ

Formalisation

We assume a dynamic transportation network $n(\vec{x}, t)$ within a dynamic territorial landscape $\vec{T}(\vec{x}, t)$, whose components are to simplify population $p(\vec{x}, t)$ and employments $e(\vec{x}, t)$. Data is structured the following way :

- Observation of territorial variables are discretized in space and in time, i.e. the spatial field \vec{T} is summarized by $T = (\vec{T}(\vec{x}_i, t_j^{(T)}))_{i,j}$ with $1 \leq i \leq N$ and $1 \leq j \leq T$. They concretely correspond to census on administrative units (*communes* in our case) at different dates.
- Network has a continuous spatial position but is represented by the vector of network distances N **C : (Florent) vol d'oiseau/-distance temps ? second faisable et à privilégier je pense**

Sur l'accessibilité

The notion of accessibility has been central to regional science since its introduction and systematization in planning around 1970.

As already introduced in the first chapter, we question the notion of accessibility : *Is the notion of accessibility crucial for statistical analysis ?*

Weibull has proposed an axiomatic approach to accessibility [WEIBULL, 1976], deriving a canonical decomposition for any *attraction-accessibility* function $A(a, d)$, assuming expected thematic axioms among others technical ones that are :

1. A is invariant regarding the order of the configuration
2. A decrease with distance at fixed attraction and increase with attraction at fixed distance
3. A is invariant when adding null attractions and constant configurations

Then A verifies these if and only if it is of the form

$$A[(a_i, d_i)] = T \left(\bigoplus_i z(d_i, a_i) \right)$$

where T is increasing with null origin, z is a *distance substitution function* (i.e. verifying axiom 2) and \oplus a *standard composition* associating two attractions at zero distance to the corresponding unique one.

It means that well suited matrices of autocorrelation should capture accessibility in regressions; **C : (Florent) pas sur de comprendre, à discuter** or it must be captured by non-linear regression on N . It

may reveal some kind of intrinsic accessibility that is related to real phenomena (that we expect to fit with calibrated functions of accessibility based on Hedonic models e.g.) Seeing accessibility as a potential field is an equivalent vision : given any stationary dynamic for n, \vec{T} , Helmholtz theorem states that it derives from a potential (can be adapted to non-stationary dynamics with a time-varying potential).

Données

We will work on a novel dataset provided by LE NECHET, that consists in main road infrastructures with their opening dates and train network for network dynamics, and in population and employments of communes at census dates, for Bassin Parisien on the last fifty year. The temporal granularity due to census temporal step may be an obstacle to obtain good dynamical statistics. **C : (Florent) enfin c'est surtout INSEE, IGN, et Wiki[?] qu'il faut citer (c'est vrai qu'il y a du formatage, mais en tout cas il faut citer les sources de première main)**

Tests Statistiques

The following large set of analysis are to be tested (non exhaustive) :

C : (Florent) interprétation ? si O/N

- On raw data :

- Multivariate models

$$\mathcal{L} [\mathbf{T}, \mathbf{N}] \sim \varepsilon$$

- Autocorrelated univariate models

$$(\mathbf{I} - \Sigma RW)\mathbf{X} \sim \varepsilon$$

- Autocorrelated multivariate models

$$(\mathcal{L}' - \Sigma RW) [\mathbf{T} + \mathbf{N}] \sim \varepsilon$$

- Geographically Weighted Regression [BRUNSDON, FOTHERINGHAM et CHARLTON, 1998]

$$\mathcal{L} [\mathcal{G} (\mathbf{T}, \mathbf{N})] \sim \varepsilon$$

- Granger causality tests : [XIE et LEVINSON, 2009a] use for example Granger causality to link transit with land-use changes.

- On data returns :

- Autoregressive multivariate models

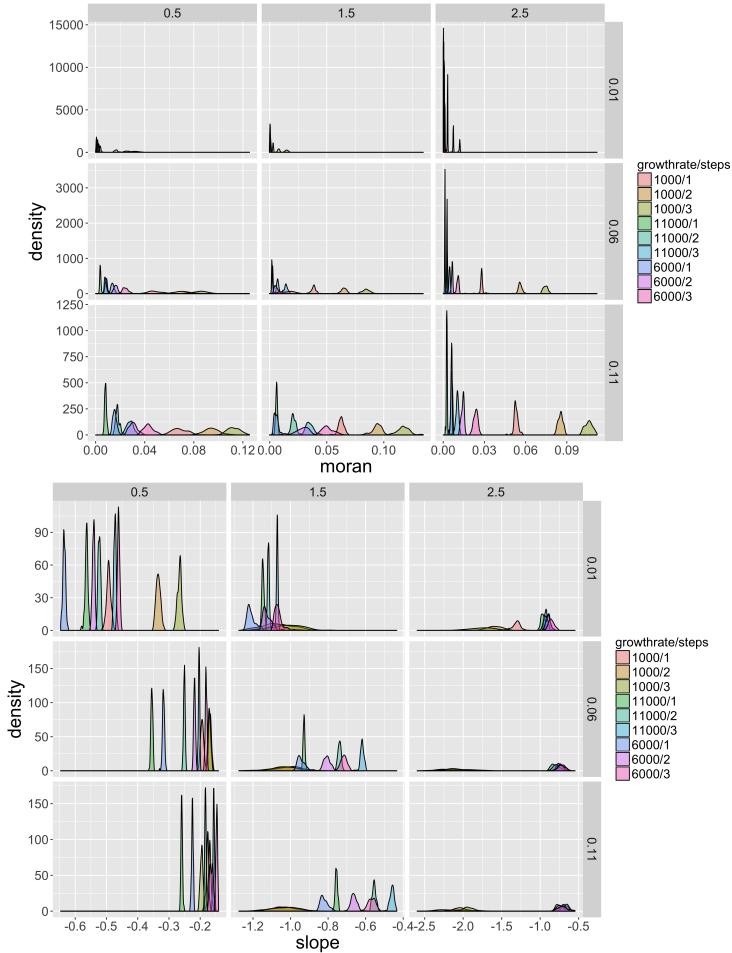
$$\mathcal{L} [(\Delta \mathbf{T}(t_{j'}))_{j' \leq j}, (\Delta \mathbf{N}(t_{j'}))_{j' \leq j}] \sim \varepsilon$$

- Autoregressive autocorrelated multivariate models : idem with spatial autocorrelation term.
- Synthetic Instrumental Variables : static territory and/or network?

★ ★

★

A.6 EFFETS DE RÉSEAU



A.7 MORPHOGENÈSE PAR AGRÉGATION-DIFFUSION

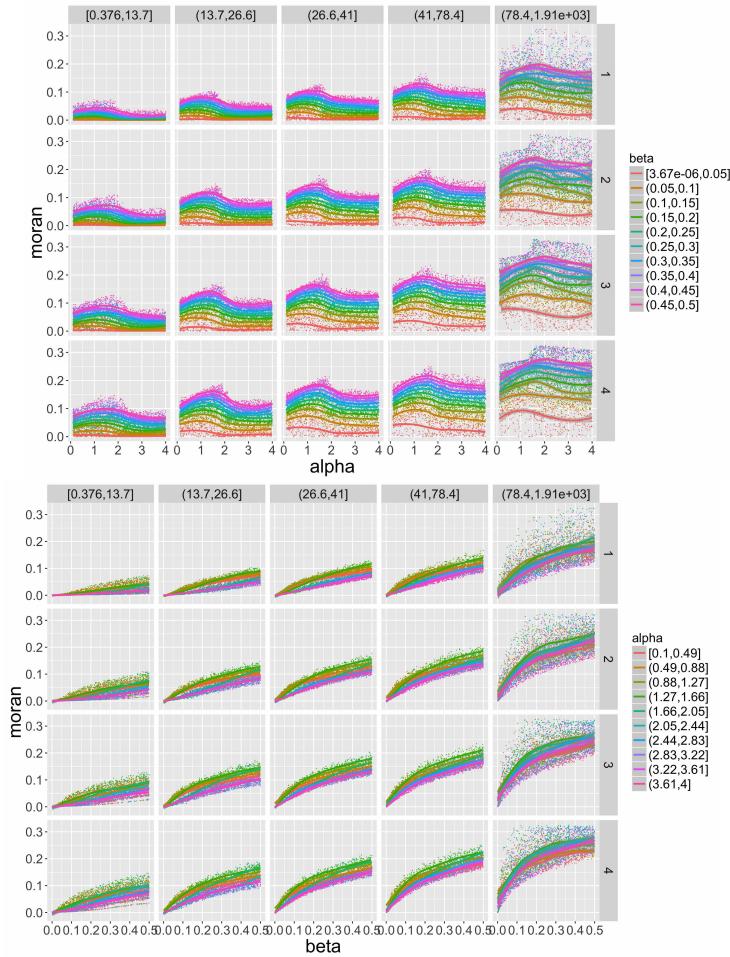
A.7.1 Figures supplémentaires pour l'exploration du modèle

Convergence

Histograms for the 81 parameters points for which we did 100 repetitions are given in Fig. ??, for Moran index and slope indicators. Other indicators showed similar convergence patterns. The visual exploration of histograms confirms the numerical analysis done in main text for statistical convergence.

Indicateurs

FIGURE 67: We show in Fig. to Fig. 10 the full behavior of all indicators, with all parameters varying, obtained through the extensive exploration, from which the plots in main text have been extracted. Because of the complex nature of emergent urban form, one can not



predict output values without referring to this “exhaustive” parameter sweep.

FIGURE 71:

Scatterplot des indicateurs

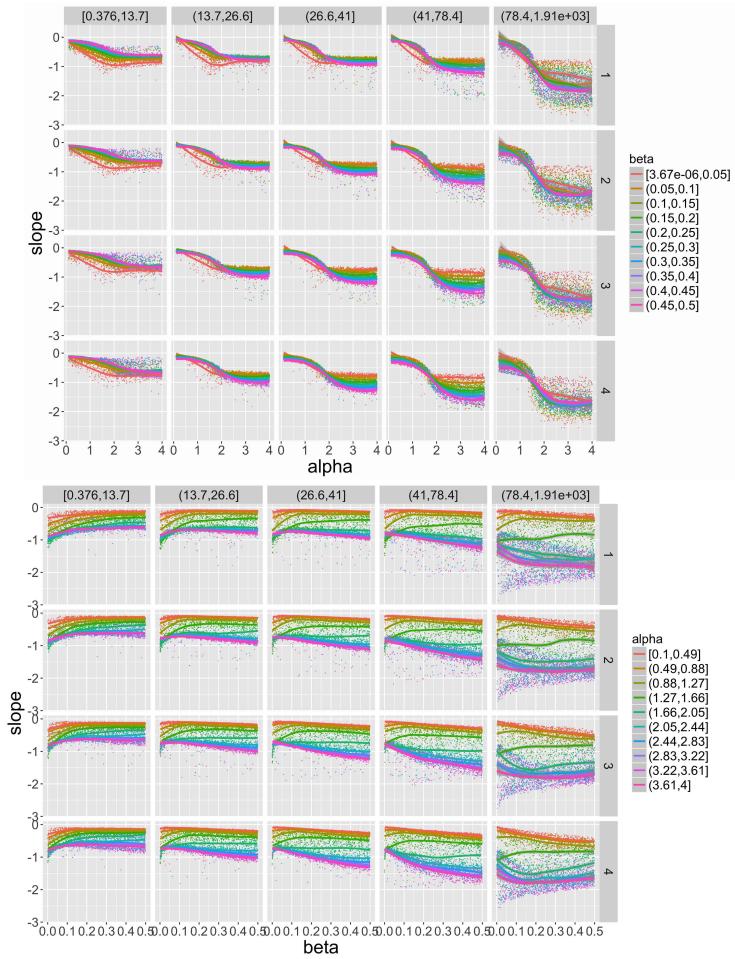
We show finally the full scatterplots of indicators, with real data points, in Fig. ?? . These are preliminary step of the calibration on principal components, and we can see on these on which dimensions the model fails relatively to fit real data (in particular average distance).

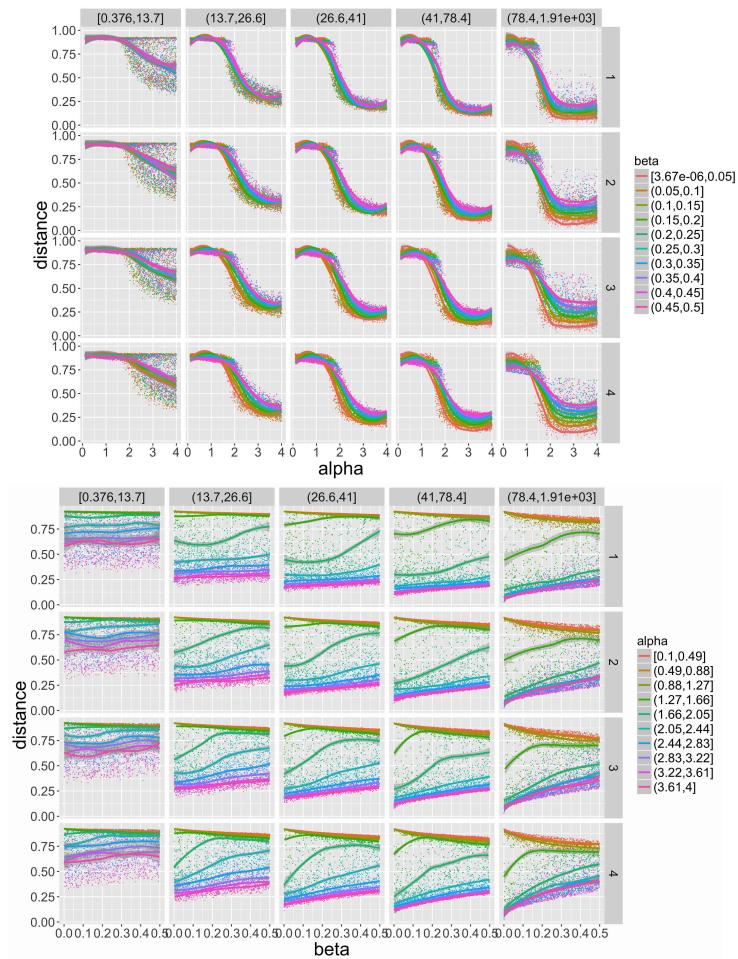
A.7.2 Analyse semi-analytique du modèle simplifié

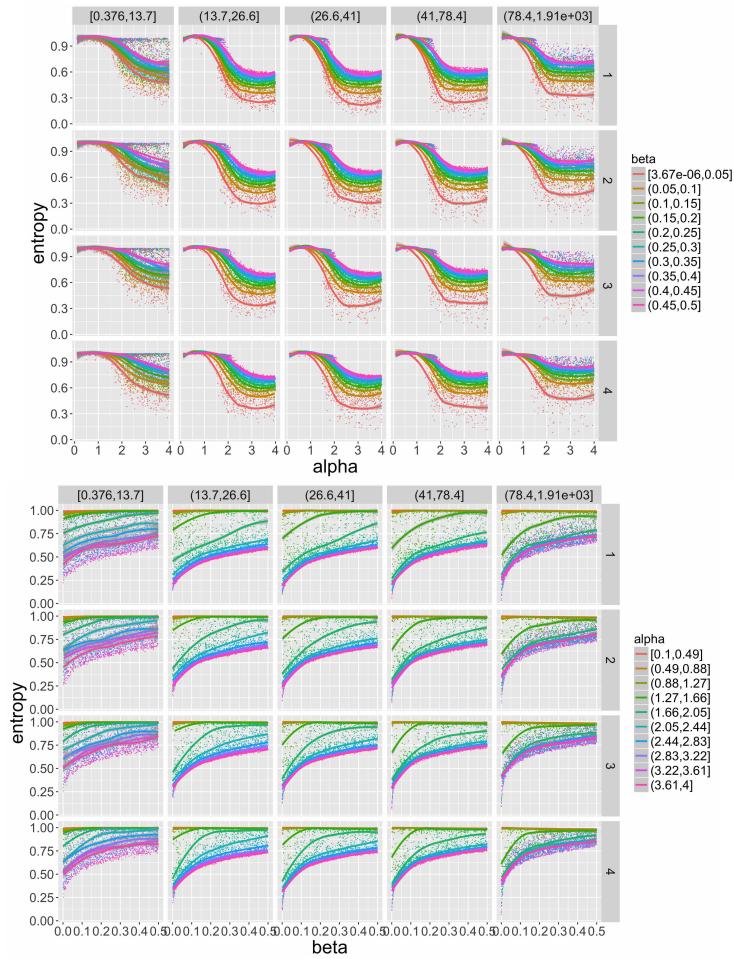
FIGURE 72:

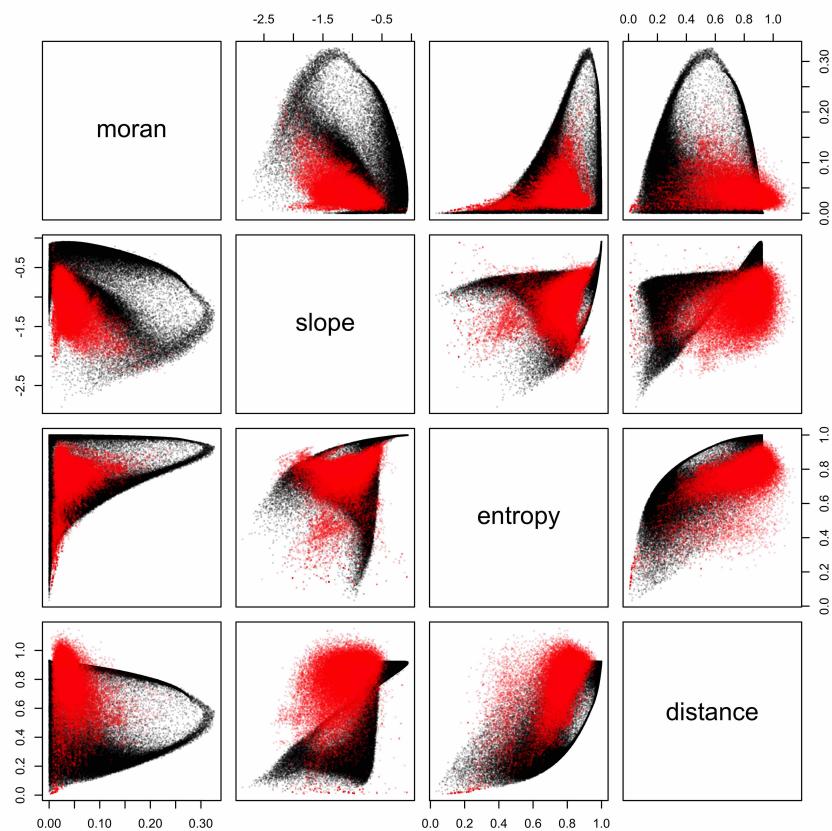
Equation aux dérivées partielles

We propose to derive the PDE in a simplified setting. To recall the configuration given in main text, the system has one dimension, such









that $x \in \mathbb{R}$ with $1/\delta x$ cells of size δx , and we use the expected values of cell population $p(x, t) = \mathbb{E}[P(x, t)]$. We furthermore take $n_d = 1$. Larger values would imply derivatives at an order higher than 2 but the following results on the existence of a stationary solution should still hold.

Denoting $\tilde{p}(x, t)$ the intermediate populations obtained after the aggregation stage, we have

$$\tilde{p}(x, t) = p(x, t) + N_g \cdot \frac{p(x, t)^\alpha}{\sum_x p(x, t)^\alpha}$$

since all populations units are added independently. If $\delta x \ll 1$ then $\sum_x p^\alpha \simeq \int_x p(x, t)^\alpha dx$ and we write this quantity $P_\alpha(t)$. We furthermore write $p = p(x, t)$ and $\tilde{p} = \tilde{p}(x, t)$ in the following for readability.

The diffusion step is then deterministic, and for any cell not on the border ($0 < x < 1$), if δt is the interval between two time steps, we have

$$\begin{aligned} p(x, t + \delta t) &= (1 - \beta) \cdot \tilde{p} + \frac{\beta}{2} [\tilde{p}(x - \delta x, t) + \tilde{p}(x + \delta x, t)] \\ &= \tilde{p} + \frac{\beta}{2} [(\tilde{p}(x + \delta x, t) - \tilde{p}) - (\tilde{p} - \tilde{p}(x - \delta x, t))] \end{aligned}$$

Assuming the partial derivatives exist, and as $\delta x \ll 1$, we make the approximation $\tilde{p}(x + \delta x, t) - \tilde{p} \simeq \delta x \cdot \frac{\partial \tilde{p}}{\partial x}(x, t)$, what gives

$$(\tilde{p}(x + \delta x, t) - \tilde{p}) - (\tilde{p} - \tilde{p}(x - \delta x, t)) = \delta x \cdot \left(\frac{\partial \tilde{p}}{\partial x}(x, t) - \frac{\partial \tilde{p}}{\partial x}(x - \delta x, t) \right)$$

and therefore at the second order

$$p(x, t + \delta t) = \tilde{p} + \frac{\beta \delta x^2}{2} \cdot \frac{\partial^2 \tilde{p}}{\partial x^2}$$

Substituting \tilde{p} gives

$$\begin{aligned} \frac{\partial^2 \tilde{p}}{\partial x^2} &= \frac{\partial^2 p}{\partial x^2} + \frac{N_G}{P_\alpha} \cdot \frac{\partial}{\partial x} \left[\alpha \frac{\partial p}{\partial x} p^{\alpha-1} \right] \\ &= \frac{\partial^2 p}{\partial x^2} + \alpha \frac{N_G}{P_\alpha} \left[\frac{\partial^2 p}{\partial x^2} p^{\alpha-1} + (\alpha - 1) \left(\frac{\partial p}{\partial x} \right)^2 p^{\alpha-2} \right] \end{aligned}$$

By supposing that $\frac{\partial p}{\partial t}$ exists and that δt is small, we have $p(x, t + \delta t) - p(x, t) \simeq \delta t \frac{\partial p}{\partial t}$, what finally yields , by combining the results above, the partial differential equation

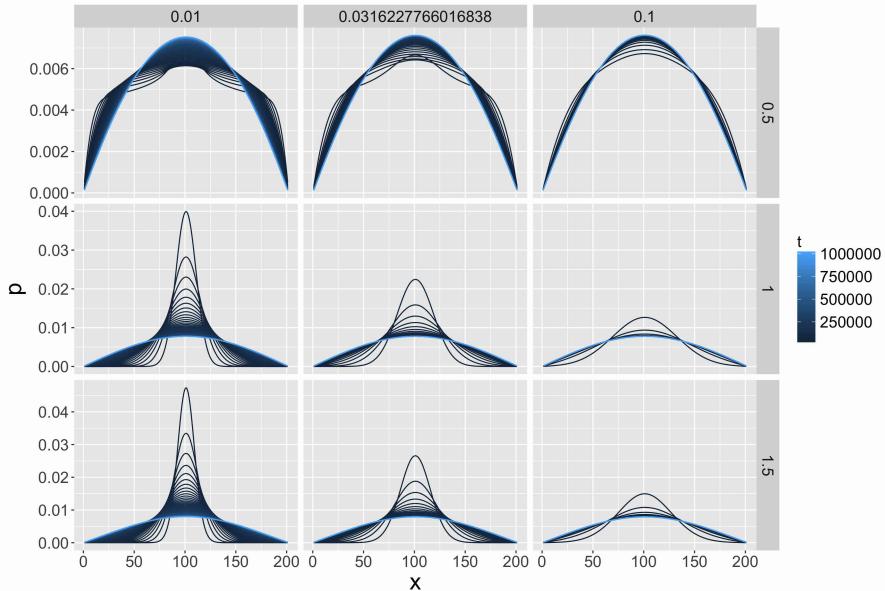
$$\delta t \cdot \frac{\partial p}{\partial t} = \frac{N_G \cdot p^\alpha}{P_\alpha(t)} + \frac{\alpha \beta (\alpha - 1) \delta x^2}{2} \cdot \frac{N_G \cdot p^{\alpha-2}}{P_\alpha(t)} \cdot \left(\frac{\partial p}{\partial x} \right)^2 + \frac{\beta \delta x^2}{2} \cdot \frac{\partial^2 p}{\partial x^2} \cdot \left[1 + \alpha \frac{N_G p^{\alpha-1}}{P_\alpha(t)} \right]$$

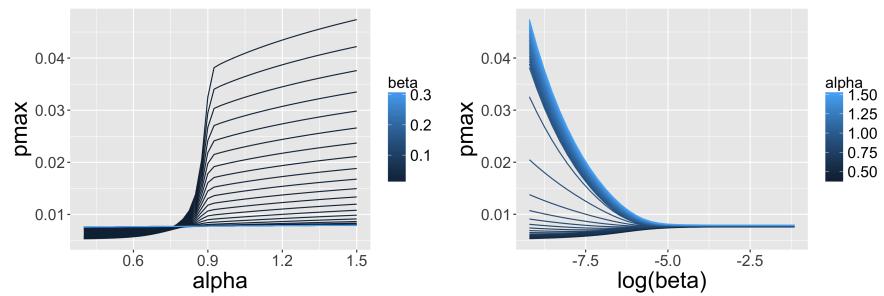
(19)

Initial conditions should be specified as $p_0(x) = p(x, t_0)$. To have a well-posed problem similar to more classical PDE problems, we need to assume a domain and boundary conditions. A finite support is expressed by $p(x, t) = 0$ for all t and x such that $|x| > x_m$.

Solution stationnaire pour la densité

The non-linearity and the integral terms making the equation above out of the scope for analytical resolution, we study its behavior numerically in some cases. Taking a simple initial condition $p_0(0) = 1$ and $p_0(x) = 0$ for $x \neq 0$, we show that on a finite domain, density $d(x, t)$ always converge to a stationary solution for large t , for a large set of values of (α, β) with fixed $N_G = 10$ ($\alpha \in [0.4, 1.5]$ varying with step 0.025 and $\log \beta \in [-1, -0.5]$ with step 0.1). We show in Fig. ?? the corresponding trajectories on a typical subset. The variation of the asymptotic distribution as a function of α and β are not directly visible, as they depend on very low values of the outward flows at boundaries. We give in Fig. ?? their behavior, by showing the value of the maximum of the distribution. Low values of β give an inversion in the effect of α , whereas high values of β give comparable values for all α .





A.8 DONNÉES SYNTHÉTIQUES CORRÉLÉES

A.9 HEURISTIQUES DE GÉNÉRATION DE RÉSEAU

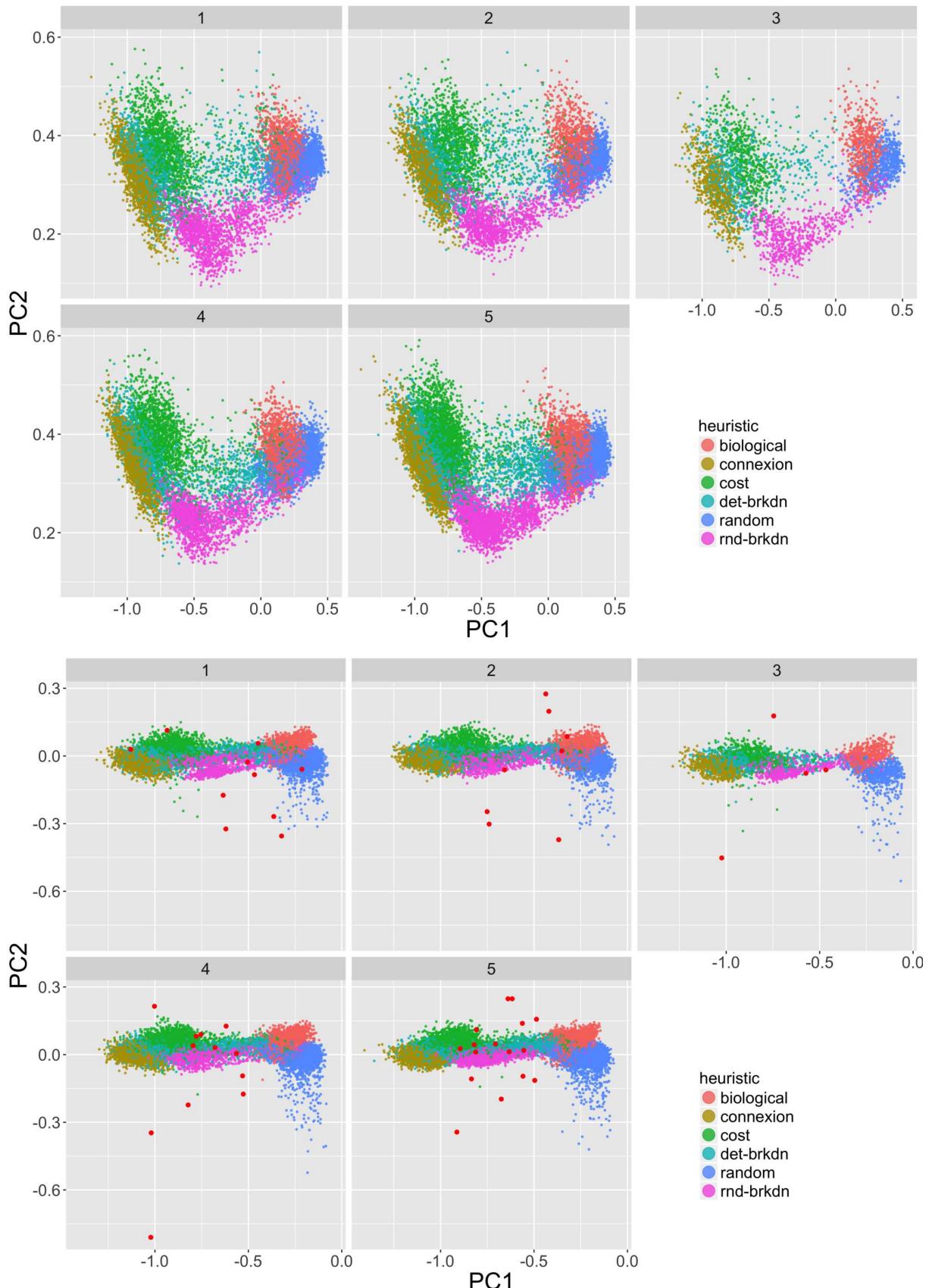


FIGURE 75: Espace topologique faisable pour les différentes heuristiques de génération, conditionné à la classe morphologique de densité.

B

DÉVELOPPEMENTS MÉTHODOLOGIQUES

We are now building a rigorous Science of Cities, contrarily to what was done before.

- MARC BARTHÉLÉMY EMCSS
Fall 2014, Network Course Introduction

This chapter gathers various methodological and technical developments, that have the common points to be not essential to the core of the thesis and difficult to digest.

Such a shocking phrase **C : (Florent) je crois que si tu t'appuies explicitement sur la mise en exergue alors ce n'est plus une mise en exergue** was pronounced during the introduction of a *Network* course for students of Complex System Science. Besides the fact that the spirit of CSS **C : (Florent) pas mettre trop d'acronymes que tu ne réutiliseras pas** is precisely the opposite, i. e. the construction of integrative disciplines (vertical integration that is necessarily founded on the existing body of knowledge of concerned fields) that answer transversal questions (horizontal integration that imply interdisciplinarity) - see e. g. the roadmap for CS [BOURGINE, CHAVALARIAS et AL., 2009], it reveals how methodological considerations shape the perceptions of disciplines. From a background in Physics, **C : (Florent) soit on connaît ton background ?** “rigorous” implies the use of tools and methods judged more rigorous (analytical derivations, large datasets statistics, etc.). **C : (Florent) je ne suis pas sûr que cela soit ça la rigueur physicienne. ce serait plutôt un raisonnement sans trou du début à la fin sur des objets clairement définis ; en sciences sociales il y a fréquemment des trous** But what is rigorous for someone will not be for an other discipline¹, depending on the purpose of each piece of research (perspectivism [GIERE, 2010c] poses the *model*, that includes methods, as the articulating core of research enterprises). Thus the full role of methodology aside and not beside theory and experiments. We go in this chapter into various methodological developments which may be precisely used later or contribute to the global background.

¹ a funny but sad anecdote told by a friend comes to mind : defending his PhD in statistics, he was told at the end by economists how they were impressed by the mathematical rigor of his work, whereas a mathematician judged that “he could have done everything on the back of an enveloppe”. **C : (Florent) ce n'est pas lié à la rigueur**

B.1 UN CADRE UNIFIÉ POUR LES MODÈLES STOCHASTIQUES DE CROISSANCE URBAINE

B.1.1 *Introduction*

Various stochastic models aiming to reproduce population patterns on large temporal and spatial scales (city systems) have been discussed across various fields of the literature, from economics to geography, including models proposed by physicists. We propose here a general framework that allows to include different famous models (in particular Gibrat, Simon and Preferential Attachment model) within an unified vision. It brings first an insight into epistemological debates on the relevance of models. Furthermore, bridges between models lead to the possible transfer of analytical results to some models that are not directly tractable.

Seminal models of urban growth are Simon [SIMON, 1955] (later generalized as e.g. [HARAN et VINING, 1973]) and Gibrat models. **C :** (Florent) à détailler davantage, c'est une matière basique de la thèse Many examples can be given across disciplines. [BENGUIGUI et BLUMENFELD-LIEBERTHAL, 2007] give an equation-based dynamical model, whereas [GABAIX, 1999] solves a stationary model. **C :** (Florent) après tu es dans l'implémentation **A1** : non a priori, variantes et extensions [GABAIX et IOANNIDES, 2004] reviews urban growth approaches in economics. A model adapted from evolutive urban theory is solved in [FAVARO et PUMAIN, 2011] and improves Gibrat models. The question of empirical scales at which it is consistent to study urban growth was also tackled in the particular case of France [BRETAGNOLLE, PAULUS et PUMAIN, 2002]. We stay to a certain level of tractability to include models as essence of our approach is links between models but do not make ontologic assumptions **C (JR) : sens ?**.

B.1.2 *Cadre de Travail*

What we propose as a framework can be understood as a meta-model in the sense of [COTTINEAU, CHAPRON et REUILLO, 2015], i.e. an modular general modeling process within each model can be understood as a limit case or as a specific case of another model. More simply it should be a diagram of formal relations between models. **C :** (Florent) à ce stade on ne sait pas si tu vas faire 1 ou N modèles, c'est un choix qu'il te faut défendre avant d'en arriver là The ontological aspect is also tackled by embedding the diagram into an ontological state space (which discretization corresponds to the “bricks” of the incremental construction of [COTTINEAU, CHAPRON et REUILLO, 2015]). It constructs a sort of model classification or modelography. **C :** (Florent) PAS UTILE ICI JE PENSE

We are still at the stage of different derivations of links between models that are presented hereafter.

B.1.3 Dérivations

Généralisation de l'Attachement Préférentiel

[YAMASAKI et al., 2006] give a generalization of the classical Preferential Attachment Network Growth model, as a birth and death model with evolving entities. More precisely, network units gain and lose population (equivalent to links connexions) at fixed probabilities, and new unit can be created at a fixed rate.

Lien entre Gibrat et Attachement Préférentiel

C : (Florent) est-ce standard d'introduire de la stochasticité dans Gibrat : $P_{t+1}=RP_t$ **A1 : c'est la formulation standard a priori** Considérons un modèle de croissance strictement positive de Gibrat donnée par $P_i(t) = R_i(t) \cdot P_i(t-1)$ avec $R_i(t) > 1$, $\mu_i(t) = \mathbb{E}[R_i(t)]$ et $\sigma_i(t) = \mathbb{E}[R_i(t)^2]$. **C : (Florent) expliquer le sens des PR etc.** 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 $m > 0$. **C : (Florent) quelle est l'équation $P_{t+1} = P_t \cdot m \cdot \lambda$** 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. **C : (Florent) on a déjà discuté de cette eq Gibrat/att pref mais tu ne peux pas faire l'économie d'expliquer pourquoi tu t'es posé la question, i.e. à quoi cela va te servir ensuite**

Lemma 1 *The limit of a Preferential Attachment model when $\lambda \ll 1$ is a linear-growth Gibrat model, with limit parameters $\mu_i(t) = 1 + \frac{\lambda}{m \cdot (t-1)}$.*

La preuve est donnée en Annexe ??.

C : (Florent) certain limit : à qualifier plus précisément

C : (Florent) je n'arrive pas à te suivre : si tu as besoin d'être relu sur ces développements, il faut convenir d'un rendez-vous pour que tu m'expliques le cheminement

Lien entre Simon et Attachement Préférentiel

A rewriting of Simon model yields a particular case of the generalized preferential attachment, in particular by vanishing death probability.

Lien entre Favarro-Pumain et Gibrat

[FAVARO et PUMAIN, 2011] generalizes Gibrat models with innovation propagation dynamics, being therefore a generalization of that model. Theoretically, a process-based model equivalent to the Favarro-Pumain should then fill the missing case in model classification at the corresponding discretization. Simpop models do not fill that case as they stay at the scale of city systems, as for Marius models [COTTINEAU, 2014]. These must also have their counterparts in discrete microscopic formulation.

C : (Florent) la encore tu parles de modèles que tu ne décris pas par ailleurs ; or Personnes connaissant FavaraoPumain \cap Personnes connaissant Gibrat $\cap \dots =$ quelques personnes sur terre !

Lien entre Bettencourt-West et Pumain

We are considering to study Bettencourt-West model for urban scaling laws [BETTENCOURT, LOBO et WEST, 2008] as entering the stochastic urban growth framework as stationary component of a random growth model, but investigation are still ongoing.

Autres modèles

[GABAIX, 1999] develops an economic model giving a Simon equivalent formulation. They in particular find out that in upper tail, proportional growth process occurs. We find the same result as a consequence of the derivation of the link between Gibrat and Preferential attachment models.

C : (Florent) je pense que tu as intérêt soit à présenter moins de modèles, mais plus en détails, soit à partir d'angles d'attaque précis et faire des typologies de modèles

B.1.4 Dérivations pour les modèles de croissance urbaine

Lemma 2 *The limit of a Preferential Attachment model when $\lambda \ll 1$ is a linear-growth Gibrat model, with limit parameters $\mu_i(t) = 1 + \frac{\lambda}{m \cdot (t-1)}$.*

Proof Starting with first moment, we denote $\bar{P}_i(t) = \mathbb{E}[P_i(t)]$. Independence of Gibrat growth rate yields directly $\bar{P}_i(t) = \mathbb{E}[R_i(t)] \cdot \bar{P}_i(t-1)$. Starting for the preferential attachment model, we have $\bar{P}_i(t) = \mathbb{E}[P_i(t)] = \sum_{k=0}^{+\infty} k \mathbb{P}[P_i(t) = k]$. But

$$\{P_i(t) = k\} = \bigcup_{\delta=0}^{\infty} (\{P_i(t-1) = k-\delta\} \cap \{P_i \leftarrow P_i + 1\}^{\delta})$$

where the second event corresponds to city i being increased δ times between $t-1$ and t (note that events are empty for $\delta \geq k$). Thus,

being careful on the conditional nature of preferential attachment formulation, stating that $\mathbb{P}[\{P_i \leftarrow P_i + 1\} | P_i(t-1) = p] = \lambda \cdot \frac{p}{P(t-1)}$ (total population $P(t)$ assumed deterministic), we obtain

$$\begin{aligned}\mathbb{P}[\{P_i \leftarrow P_i + 1\}] &= \sum_p \mathbb{P}[\{P_i \leftarrow P_i + 1\} | P_i(t-1) = p] \cdot \mathbb{P}[P_i(t-1) = p] \\ &= \sum_p \lambda \cdot \frac{p}{P(t-1)} \mathbb{P}[P_i(t-1) = p] = \lambda \cdot \frac{\bar{P}_i(t-1)}{P(t-1)}\end{aligned}$$

It gives therefore, knowing that $P(t-1) = P_0 + m \cdot (t-1)$ and denoting $q = \lambda \cdot \frac{\bar{P}_i(t-1)}{P_0 + m \cdot (t-1)}$

$$\begin{aligned}\bar{P}_i(t) &= \sum_{k=0}^{\infty} \sum_{\delta=0}^{\infty} k \cdot \left(\lambda \cdot \frac{\bar{P}_i(t-1)}{P_0 + m \cdot (t-1)} \right)^{\delta} \cdot \mathbb{P}[P_i(t-1) = k - \delta] \\ &= \sum_{\delta'=0}^{\infty} \sum_{k'=0}^{\infty} (k' + \delta') \cdot q^{\delta'} \cdot \mathbb{P}[P_i(t-1) = k'] \\ &= \sum_{\delta'=0}^{\infty} q^{\delta'} \cdot (\delta' + \bar{P}_i(t-1)) = \frac{q}{(1-q)^2} + \frac{\bar{P}_i(t-1)}{(1-q)} \\ &= \frac{\bar{P}_i(t-1)}{1-q} \left[1 + \frac{1}{\bar{P}_i(t-1)} \frac{q}{(1-q)} \right]\end{aligned}$$

As it is not expected to have $\bar{P}_i(t) \ll P(t)$ (fat tail distributions), a limit can be taken only through λ . Taking $\lambda \ll 1$ yields, as $0 < \bar{P}_i(t)/P(t) < 1$, that $q = \lambda \cdot \frac{\bar{P}_i(t-1)}{P_0 + m \cdot (t-1)} \ll 1$ and thus we can expand in first order of q , what gives $\bar{P}_i(t) = \bar{P}_i(t-1) \cdot \left[1 + \left(1 + \frac{1}{\bar{P}_i(t-1)} \right) q + o(q) \right]$

$$\bar{P}_i(t) \simeq \left[1 + \frac{\lambda}{P_0 + m \cdot (t-1)} \right] \cdot \bar{P}_i(t-1)$$

It means that this limit is equivalent in expectancy to a Gibrat model with $\mu_i(t) = \mu(t) = 1 + \frac{\lambda}{P_0 + m \cdot (t-1)}$.

For the second moment, we can do an analog computation. We have still

$$\mathbb{E}[P_i(t)^2] = \mathbb{E}[R_i(t)^2] \cdot \mathbb{E}[P_i(t-1)^2]$$

and

$$\mathbb{E}[P_i(t)^2] = \sum_{k=0}^{+\infty} k^2 \mathbb{P}[P_i(t) = k]$$

We obtain the same way

$$\begin{aligned}
\mathbb{E}[P_i(t)^2] &= \sum_{\delta'=0}^{\infty} \sum_{k'=0}^{\infty} (k' + \delta')^2 \cdot q^{\delta'} \cdot \mathbb{P}[P_i(t-1) = k'] \\
&= \sum_{\delta'=0}^{\infty} q^{\delta'} \cdot \left(\mathbb{E}[P_i(t-1)^2] + 2\delta' \bar{P}_i(t-1) + \delta'^2 \right) \\
&= \frac{\mathbb{E}[P_i(t-1)^2]}{1-q} + \frac{2q \bar{P}_i(t-1)}{(1-q)^2} + \frac{q(q+1)}{(1-q)^3} \\
&= \frac{\mathbb{E}[P_i(t-1)^2]}{1-q} \left[1 + \frac{q}{\mathbb{E}[P_i(t-1)^2]} \left(\frac{2\bar{P}_i(t-1)}{1-q} + \frac{(1+q)}{(1-q)^2} \right) \right]
\end{aligned}$$

We have therefore an equivalence between the Gibrat model as a continuous formulation of a Preferential Attachment (or Simon model) in a certain limit. ■

B.2 SENSIBILITÉ DES LOIS D'ECHELLE URBAINES À L'ETENDUE SPATIALE

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. **C : (Florent) présenté comme cela ce n'est pas évident de comprendre le rapport avec ta thèse**

Les lois d'échelle ont été montrées universelles des systèmes urbains à de nombreuses échelles et pour différents indicateurs. **C : (Florent) pas très précis** 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 simulations numériques confirment les résultats théoriques et révèle que les résultats sont raisonnablement indépendants du noyau spatial utilisé pour distribuer la densité.

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 [GABAIX, 1999], **C : (Florent) déjà dit** ont été montrées être 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 [GABAIX et IOANNIDES, 2004] 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 [BETTENCOURT, LOBO et WEST, 2008] mais sont difficiles à ajuster empiriquement. Une proposition intéressante par PUMAIN est qu'elles sont intrinsèquement dues au caractère évo-

lutionnaire des systèmes de villes, où l'émergence complexe par les interactions entre villes génère de telles distributions globales [PUMAIN et al., 2006]. Même si un parallèle tentant peut être fait avec les systèmes biologiques auto-organisés **C : possibly here make a link with morphogenesis - depending if introduced before or not**, PUMAIN insiste sur le fait que l'hypothèse d'ergodicité **C : (Florent) préciser ce que cela signifie** pour de tels systèmes n'est pas raisonnable dans le cas de système géographiques et que l'analogie est difficilement exploitable [PUMAIN, 2012b]. 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 [LOUF et BARTHELEMY, 2014] 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 [FAVARO et PUMAIN, 2011] fournissent une approximation au premier ordre des systèmes exhibant des lois d'échelles, 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 Favar-Pumain pour la propagation des cycles d'innovation proposé dans [FAVARO et PUMAIN, 2011], qui généralise un modèle de Gibrat pour la croissance des villes françaises avec une ontologie similaire à celle des modèles Simpop. **C : (Florent) ok : modèles qui reproduisent scaling, est-ce un des critères de validation des modèles que tu vas développer ?**

C : IDEE - take the FavaroPumain again, try to fit/compare with the IntGib-network model? – sort of benchmark, should be easy to implement

C : (Florent) qu'est ce que ça veut dire, blind application of models?

The derivations in the simple case of exponential mixture density, are done in Appendix ??.

C : mention way of fitting; golden standard to fit power laws? check thèse d'Olivier pour voir si le cutoff est appliqué?

We formalize the simple theoretical context in which we will derive the sensitivity of scaling to city definition. Let consider a polycentric city system, which spatial density distributions can be reasonably constructed as the superposition of monocentric fast-decreasing spatial kernels, such as an exponential mixture model [ANAS, ARNOTT et SMALL, 1998]. Taking a geographical space as \mathbb{R}^2 , we take for any $\vec{x} \in \mathbb{R}^2$ **C : (Florent) attention à la sensibilité de certains géographes** the density of population as

$$d(\vec{x}) = \sum_{i=1}^N d_i(\vec{x}) = \sum_{i=1}^N d_i^0 \cdot \exp\left(\frac{-\|\vec{x} - \vec{x}_i\|}{r_i}\right) \quad (20)$$

where r_i are spread parameters of kernels, d_i^0 densities at origins, \vec{x}_i positions of centers. We furthermore assume the following constraints :

1. To simplify, cities are monocentric, in the sense that for all $i \neq j$, we have $\|\vec{x}_i - \vec{x}_j\| \gg r_i$.
2. It allows to impose structural scaling in the urban system by the simple constraint on city populations P_i . One can compute by integration that $P_i = 2\pi d_i^0 r_i^2$, what gives by injection into the scaling hypothesis $\ln P_i = \ln P_{\max} - \alpha \ln i$, the following relation between parameters : $\ln [d_i^0 r_i^2] = K' - \alpha \ln i$.

To study scaling relations, we consider a random scalar spatial variable $a(\vec{x})$ representing one aspect of the city, that can be everything but has the dimension of a spatial density, such that the indicator $A(D) = \mathbb{E}[\iint_D a(\vec{x}) d\vec{x}]$ represents the expected quantity of a in area D . We make the assumption that $a \in \{0; 1\}$ ("counting" indicator) and that its law is given by $P[a(\vec{x}) = 1] = f(d(\vec{x}))$. Following the empirical work done in [COTTINEAU, 2015], the integrated indicator on city i as a function of θ is given by

$$A_i(\theta) = A(D(\vec{x}_i, \theta))$$

where $D(\vec{x}_i, \theta)$ is the area centered in \vec{x}_i where $d(\vec{x}) > \theta$. Assumption 1 ensures that the area are roughly disjoint circles. We take furthermore a simple amenity such that it follows a local scaling law in the sense that $f(d) = \lambda \cdot d^\beta$. It seems a reasonable assumption since it was shown that many urban variable follow a fractal behavior at the intra-urban scale [KEERSMAECKER, FRANKHAUSER et THOMAS, 2003] and that it implies necessarily a power-law distribution [CHEN, 2010]. We make the additional assumption that $r_i = r_0$ does not depend on i , what is reasonable if the urban system is considered from a large scale. This assumption should be relaxed in numerical simulations. The estimated scaling exponent $\alpha(\theta)$ is then the result of the log-regression of $(A_i(\theta))_i$ against $(P_i(\theta))_i$ where $P_i(\theta) = \iint_{D(\vec{x}_i, \theta)} d$.

B.2.1 Dérivation Analytique de la Sensibilité

With above notations, let derive the expression of estimated exponent for quantity a as a function of density threshold parameter θ . The quantity computed for a given city i is, thanks to the monocentric assumption and in a spatial range and a range for θ such that $\theta \gg \sum_{j \neq i} d_j(\vec{x})$, allowing to approximate $d(\vec{x}) \simeq d_i(\vec{x})$ on $D(\vec{x}_i, \theta)$, is computed by

$$\begin{aligned} A_i(\theta) &= \lambda \cdot \iint_{D(\vec{x}_i, \theta)} d^\beta = 2\pi \lambda d_i^{0\beta} \int_{r=0}^{r_0 \ln \frac{d_i^0}{\theta}} r \exp\left(-\frac{r\beta}{r_0}\right) dr \\ &= \frac{2\pi d_i^{0\beta} r_0^2}{\beta^2} \left[1 + \beta \ln \frac{\theta}{d_i^0} \left(\frac{\theta}{d_i^0}\right)^\beta - \left(\frac{\theta}{d_i^0}\right)^\beta \right] \end{aligned}$$

We obtain in a similar way the expression of $P_i(\theta)$

$$P_i(\theta) = 2\pi d_i^0 r_0^2 \left[1 + \ln \left[\frac{\theta}{d_i^0} \right] \frac{\theta}{d_i^0} - \frac{\theta}{d_i^0} \right]$$

The Ordinary-Least-Square estimation, solving the problem $\inf_{\alpha, C} \|(\ln A_i(\theta) - C - \alpha \ln P_i(\theta))_i\|^2$, gives the value $\alpha(\theta) = \frac{\text{Cov}[(\ln A_i(\theta))_i, (\ln P_i(\theta))_i]}{\text{Var}[(\ln P_i(\theta))_i]}$. As we work on city boundaries, threshold is expected to be significantly smaller than center density, i.e. $\theta/d_i^0 \ll 1$. We can develop the expression in the first order of θ/d_i^0 and use the global scaling law for city sizes, what gives $\ln A_i(\theta) \simeq K_A - \alpha \ln i + (\beta - 1) \ln d_i^0 + \beta \ln \frac{\theta}{d_i^0} \left(\frac{\theta}{d_i^0} \right)^\beta$ and $\ln P_i(\theta) = K_P - \alpha \ln i + \ln \left[\frac{\theta}{d_i^0} \right] \frac{\theta}{d_i^0}$. Developing the covariance and variance gives finally an expression of the scaling exponent as a function of θ , where k_j, k_j' are constants obtained in the development :

$$\alpha(\theta) = \frac{k_0 + k_1 \theta + k_2 \theta^\beta + k_3 \theta^{\beta+1} + k_4 \theta \ln \theta + k_5 \theta^\beta \ln \theta + k_6 \theta^\beta (\ln \theta)^2 + k_7 \theta^{\beta+1} (\ln \theta)^2 + k'_0 + k'_1 \ln \theta + k'_2 \theta \ln \theta + k'_3 \theta^2 + k'_4 \theta^2 \ln \theta + k'_5 \theta^2 (\ln \theta)^2}{k'_0 + k'_1 \ln \theta + k'_2 \theta \ln \theta + k'_3 \theta^2 + k'_4 \theta^2 \ln \theta + k'_5 \theta^2 (\ln \theta)^2} \quad (21)$$

This rational fraction predicts the evolution of the scaling exponent when the threshold varies. We study numerically its behavior in the next section, among other numerical experiments.

B.2.2 Simulations Numériques

IMPLÉMENTATION **C :** (Florent) définir ton champ d'investigation (des grilles carrées de taille prédéfinies, ce n'est pas du tout standard)

We implement empirically the density model given in section ???. Centers are successively chosen such that in a given region of space only one kernel dominates in the sense that the sum of other contributions are above a given threshold θ_e . **C :** (Florent) est-ce toujours possible, y'a t-il unicité du centre? Par quelle méthode précise détermine tu le centre? In practice, adapting N to world size allows to respect the monocentric condition. Population are distributed in order to follow the scaling law with fixed α and r_i (arbitrary choice) by computing corresponding d_i^0 . Technical details of the implementation done in R [R CORE TEAM, 2015b] and using the package kernlab for efficient kernel mixture methods [KARATZOGLU et al., 2004] are given as comments in source code². **C :** (Florent) cela ne suffit pas, il faut en dire plus sur la méthode **A1** : sure, surtout qu'on formule cette requête dans la partie méthodologique précédente, tout cela est un peu contradictoire.. We show in figure ?? example of synthetic density distributions on which the numerical study is conducted.

² available at <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Scaling>

The validation of theoretical results on these experimental mixtures must still be conducted, along with sensitivity tests to random perturbations, influence of kernel type, and two-parameters phase diagram when adding in the computational model functional density distribution and associated cut-off threshold.

PERTURBATIONS ALÉATOIRES The simple model used is quite reducing for maximal densities and radius distribution. We aim to proceed to an empirical study of the influence of noise in the system by fixing d_i^0 and r_i the following way :

- d_i^0 follows a reversed log-normal distribution with maximal value being a realistic maximal density
- Radiiuses are computed to respect rank-size law and then perturbed by a white noise. **C : (Florent) pourquoi ?**

TYPE DE NOYAU We shall test the influence of the type of spatial kernel used on results. We can test gaussian kernels and quadratic kernels with parameters within reasonable ranges analog to the exponential kernel.

B.3 CORRELATIONS SPATIO-TEMPORELLES

L'espace et le temps sont cruciaux pour l'étude des systèmes géographiques quand on cherche à comprendre les *processus* (par définition dynamiques [*Hypergeo*]) **C : (Florent) c'est déjà une lecture, certes processus renvoie à une évolution, mais les échelles de temps du modèle/processus ne sont pas nécessairement les mêmes qui évoluent dans une structure spatiale au sens de [DOLLFUS et DASTÈS, 1975]. C : (Florent) citer Cottineau ?**

[CROSS et HOHENBERG, 1994] : spatio-temporal chaos

The capture of neighborhood effects in statistical models is a wisely used practice in spatial statistics, as the technique of Geographically Weighted Regression illustrates [BRUNSDON, FOTHERINGHAM et CHARLTON, 1998]. A possible interpretation among many definitions of spatial autocorrelation [GRIFFITH, 1992] yields that by estimating a plausible characteristic distance for spatial correlations or autocorrelations, one can isolate independent effects between variables from effects due to neighborhood interactions³. The study of the spatial covariance structure is a cornerstone of advanced spatial statistics that was early formulated [GRIFFITH, 1980]. **C : (Florent) cela semble tout de même loin du sujet ou alors il faut que tu expliques clairement**
We propose now to study possible links between spatial and temporal correlations, using spatio-temporal covariance structure to infer information on dynamical processes.

B.3.1 Notations

We consider a multivariate spatio-temporal stochastic process denoted by $\vec{Y}[\vec{x}, t]$. At a given point \vec{x}_0 in space, we can define temporal covariance structure by

$$\mathbf{C}_t(\vec{x}_0) = \text{Var}[\vec{Y}[\vec{x}_0, \cdot]]$$

and spatial covariance structure at fixed time by

$$\mathbf{C}_x(t) = \text{Var}[\vec{Y}[\cdot, t]]$$

It is clear that these quantities will be in practice first ill-defined because of the difficulty in interpreting such a process by a spatio-temporal random variable, secondly highly non-stationary in space and time. We stay however at a theoretical level to gain structural knowledge, **C : (Florent) sens ?** reviewing simple cases in which a formal link can be established.

³ note that the formal link between models of spatial autocorrelation (see e.g. [GRIFFITH, 2012]) is not clear and should be further investigated

B.3.2 Equation des Ondes

C : (Florent) pourquoi aborder cela ? A1 : cas idéal des STARMA, ondes d'innovation etc. : approche fondamentalement liée à l'analyse spatiale, mais bien plus complexe qu'une simple équation. justifie cette approche de lien spatio-temporel ?

In the case of propagating waves, there is an immediate link. Let assume that a wave equation if verified by "deterministic" parts of components

$$c^2 \cdot \partial_t^2 \bar{Y}_i = \Delta \bar{Y}_i \quad (22)$$

with $Y_i = \bar{Y}_i + \varepsilon_i$. If errors are uncorrelated and processes are stationary, we have then directly

$$\mathbf{C}_t [\partial_t^2 Y_i, \partial_t^2 Y_j] = \frac{1}{c^2} \cdot \mathbf{C}_x [\Delta Y_i, \Delta Y_j] \quad (23)$$

This gives us however few insight on real systems as local diffusion, stationary assumptions and uncorrelated noises are far from being verified in empirical situations.

B.3.3 Equation de Fokker-Planck

An other interesting approach may when the process verifies a Fokker-Planck equation on probabilities of the state of the system when it is given by its position (diffusion of particles in that case)

$$\partial_t P(x_i, t) = -d \cdot \partial_x P(x_i, t) + \frac{\sigma^2}{2} \partial_x^2 P(x_i, t) \quad (24)$$

With no cross-correlation terms in the Fokker-Planck equation, covariance between processes vanish. We have finally in that case only a relation between averaged spatial and temporal variances that brings no information to our question.

B.3.4 Equation Maitresse

In the case of a master equation on probabilities of discrete states of the system

$$\partial_t \vec{P} = W \vec{P} \quad (25)$$

we have then for state i , $\partial_t P_i = \sum_j W_{ij} P_j$. As this relation is at a fixed time we can average in time to obtain an equation on temporal covariance. It is not clear how to make the link with spatial covariance as these will depend on spatial specification of discrete states. This question is still under investigation.

B.3.5 Echantillonnage spatial cohérent

In a more empirical way, we propose to not assume any constraint of process dynamics but to however investigate how the computation of spatial correlations can inform on temporal correlations. We try to formulate easily verifiable assumptions under which this is possible.

We make the following assumptions on the spatio-temporal stochastic processes $Y_i[\vec{x}, t]$:

1. Local spatial autocorrelation is present and bounded by l_ρ (in other words the processes are continuous in space) : at any \vec{x} and t , $|\rho_{\|\Delta\vec{x}\| < l_\rho} [Y_i(\vec{x} + \Delta\vec{x}, t), Y_i(\vec{x}, t)]| > 0$. **C : (Florent) je ne comprends pas ce qui est écrit, qu'une abs soit > 0 ok, donc c'est autre mais quoi? A1 : c'est le strict > 0 qui compte, c'est une façon de postuler que les processus sont continus à une certaine échelle fine**
2. Processes are locally parametrized : $Y_i = Y_i[\alpha_i]$, where $\alpha_i(\vec{x})$ varies with l_α , with $l_\alpha \gg l_\rho$.
3. Spatial correlations between processes have a sense at an intermediate scale l such that $l_\alpha \gg l \gg l_\rho$.
4. Processes covariance stationarity times scale as \sqrt{l} .
5. Local ergodicity is present at scale l and dynamics are locally chaotic.

Assumptions one to three can be tested empirically and allow to compare spatial correlation estimated on spatial samplings at scale l . Assumption four is more delicate as we are precisely constructing this methodology because we have no temporal information on processes. It is however typical of spatial diffusion processes, and population or innovation diffusion should verify this assumption. **C : (Florent) cela devrait être un point de départ (expliquerait pourquoi ces modèles ; te ferait peut être en considérer d'autres** The last assumption can be tested if feasible space is known, by checking cribbing on image space on the spatial sample. Under these conditions, local spatial sampling is equivalent to temporal sampling and spatial correlation estimators provide estimator of temporal correlations.

B.4 GÉNÉRATION DE DONNÉES SYNTHÉTIQUES CORRÉLÉES

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 et permet l'étude de l'interférence entre composantes à différentes fréquences sur la performance d'un modèle prédictif, en fonction des correlations entre composantes à différentes échelles. On présente ensuite 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, qui sont calibrées selon des objectifs morphologiques sur l'ensemble de l'Europe. L'exploration intensive du modèle permet l'obtention d'un large spectre de valeurs pour la matrice de correlation entre mesures morphologiques et mesures du réseau. On démontre ainsi les possibilités d'applications variées et les potentialités de la méthode.

B.4.1 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 [ABADIE, DIAMOND et HAINMUELLER, 2010], l'étude des systèmes territoriaux [MOECKEL, SPIEKERMANN et WEGENER, 2003; PRITCHARD et MILLER, 2009], l'apprentissage statistique [BOLÓN-CANEDO, SÁNCHEZ-MAROÑO et ALONSO-BETANZOS, 2013] ou la bio-informatique [BULCKE et al., 2006]. 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 à des contraintes plus faibles de valeurs pour des indica-

teurs 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 coeur de l'étude des systèmes complexes de par la maturation de nouvelles approches computационnelles [ARTHUR, 2015], 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 [YE, 2011] 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 [NEWMAN, 2003] 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.4.2 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 , $\|f(X) - f(\tilde{X})\| < \varepsilon$
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, sans rapport thématique avec la thèse, est présenté en Appendice C.2. 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.

UNE VUE ALTERNATIVE : DONNÉES SYNTHÉTIQUES Let M_m a stochastic model of simulation, which inputs are to simplify initial conditions D_0 and parameters $\vec{\alpha}$, and output $M_m[\vec{\alpha}, D_0](t)$ at a given time t . We assume that it is partially data-driven in the sense that D_0 is supposed to represent a real situation at a given time, and model performance is measured by the distance of its output at final time to the real situation at the corresponding time, i.e. error function is of the form $\|\mathbb{E}[\vec{g}(M_m[\vec{\alpha}, D_0](t_f))] - \vec{g}(D_f)\|$ where \vec{g} is a deterministic field corresponding to given indicators.

Evaluating the model on real data is rapidly limited in control possibilities, being restricted to the search of datasets allowing natural control groups. Furthermore, statistical behaviors are generally poorly characterized because of the small number of realizations. Working with synthetic data first allows to solve this issue of robustness of statistics, and then gives possibilities of control on some “meta-parameters” in the sense described before.

C : link between synthetic data and model coupling ?

B.5 UN CADRE BASÉ SUR LA DISCRÉPANCE

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 méthode. 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.5.1 *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 [MARLER et ARORA, 2004], 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 [NEWMAN, 2011] 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 [HAKEN et PORTUGALI, 2003], 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 *ciborgs* [PICON, 2013]), et cumulent ainsi la potentialité d'être à l'origine de problèmes multi-objectifs⁴. La notion récente d'*éco-quartier* [SOUAMI,

⁴ 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

[\[2012\]](#) 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 [BAVOUX et al., [2005](#)]. 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 [[CARVER, 1991](#)]. 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 [[JÉGOU et al., 2012](#)].

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'estimation de modèles statistiques donne divers indicateurs de la consistance des résultats [[LAUNER et WILKINSON, 2014](#)]. 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 [[DEB et GUPTA, 2006](#)] où les fronts de Pareto robustes sont définis comme des solutions insensibles aux petites perturbations. Dans [[BARRICO et ANTUNES, 2006](#)], 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 don-

pour un design de système, une évaluation multi-objectif d'un système existant, une évaluation multi-attributs ; notre cadre particulier correspondra au dernier cas).

nées 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 [BRUNSDON, FOTHERINGHAM et CHARLTON, 1998], 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.

OBJECTIFS AGRÉGÉS LINÉAIREMENT Une seconde hypothèse que nous faisons est que l'évaluation multi-objectifs est effectuée par agrégation linéaire des objectifs, 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, [DOBBIE et DAIL, 2013] 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.5.2 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 [ALI et al., June 2014] 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 [MANGIN et PANERAI, 1999] - 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 [DICK et PILLICHSHAMMER, 2010].

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 géographiquement disjoints, **C : Q pourquoi nécessaire des les avoir spatially disjoints, could be different indicators on the same area ? maybe makes less sense ? missing point for comparability ?** 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 coincident si et seulement si les deux variables ont même ontologie au sens de [LIVET et al., 2010a], 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, **C : not that free on the computation here !** 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

$$(\mathcal{X}, \mathcal{Y}) \underset{\text{def}}{=} \left(\prod \tilde{\mathcal{X}}_c \right) \times \left(\prod \tilde{\mathcal{Y}}_c \right) = \left(\prod_{\mathcal{X}_{i,k} \in \mathcal{D}_X} \mathbb{R}^{p_{i,k}^X} \right) \times \left(\prod_{\mathcal{Y}_{i,k} \in \mathcal{D}_Y} \mathbb{R}^{p_{i,k}^Y} \right) \quad (26)$$

avec $\mathcal{D}_X = \{\mathcal{X}_{i,k} | I(i, k) \text{ distincts, } n_{i,k}^X \text{ maximal}\}$ (de même pour \mathcal{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 $(\tilde{\mathcal{X}}_c, \tilde{\mathcal{Y}}_c)$, tel que pour tout $h \in H_c$:

1. h est "suffisamment" régulière (distribution tempérée par exemple)
2. $q_c = \int_{(\tilde{\mathcal{X}}_c, \tilde{\mathcal{Y}}_c)} h$ 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 [WANG et al., 2009] 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. $q_c \in [0, 1]$, pour une construction plus simple des poids relatifs. **C : indeed $h_c \in [0, 1]$ is the right assumption**
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 $w_{i,c}^L = \frac{\hat{q}_{i,c}}{\sum_i \hat{q}_{i,c}}$ 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 contraignante 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 [VARET, 2010], puisque la forme intégrée des indicateurs permet justement de bénéficier de tels résultats théoriquement puissant. Soit $X_{i,c} = (\vec{X}_{i,c,l})_{1 \leq l \leq n_{i,c}}$ et $D_{i,c} = \text{Disc}_{\vec{X}_c, L^2}(X_{i,c})$ le discrépance du jeu de données⁵ [NIEDERREITER, 1972]. 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

$$\left\| \int \sum w_{i,c} h_c - \frac{1}{n_{i,c}} \sum_l w_{i,c} h_c(\vec{X}_{i,c,l}) \right\| \leq K \sum_c |w_{i,c}| \|h_c\| \cdot D_{i,c}$$

⁵ La discrépance est définie comme la norme-L2 de la discrépance locale qui est pour des points de données normalisés $X = (x_{ij}) \in [0, 1]^d$, 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.

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

$$R_{i,i'} = \frac{\sum_c w_{i,c} \cdot D_{i,c}}{\sum_c w_{i',c} \cdot D_{i',c}} \quad (27)$$

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.5.3 Résultats

IMPLÉMENTATION Le pré-traitement des données géographiques est fait via QGIS [QGIS, 2011] pour des raisons de performances. **C : plutôt ergonomie ?** L'implémentation du cœur est faite en R [TEAM, 2000] pour la flexibilité de la gestion des données et du traitement statistique. De plus, le package DiceDesign [FRANCO et al., 2009] 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 ou-

verte sur le dépôt git du projet⁶ pour permettre la reproductibilité et la réutilisation [RAM, 2013].

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 [BENNETT, 2010] 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 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 implementons 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),

⁶ à <https://github.com/JusteRaimbault/RobustnessDiscrepancy>

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 pédestre.

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 simulations donnent des résultats constants. Les indicateurs obtenus en fixant un ratio fixe montrent 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 concrète 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'INSEE⁷. Les données sont associées à l'étendue géographique des unités statistiques, permettant le calcul d'indicateurs d'analyse spatiale.

⁷ <http://www.insee.fr>

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

TABLE 13 : 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.

C : wrong à l'oral 15th block size ? A cause de la petite taille du 1er arrondissement, on s'attend que la majorité des arrondissements aient un ratio plus petit que 1, ce qui se confirme dans les résultats même lorsque l'on ajoute la déviation standard.

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'in-

tensité 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 $i \neq j$ C :
typo eng paper 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 [LOUF et BARTHELEMY, 2015]. Le choix est ici arbitraire, afin d'illustrer la méthode avec un nombre raisonnable de dimensions.

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 [GILLI et OFFNER, 2009] met en évidence des interrogations sur sa pertinence, notamment sur ses capacités d'atténuer les inégalités spatiales. On peut voir en Fig. ?? 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 [GUÉROIS et LE GOIX, 2009] 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

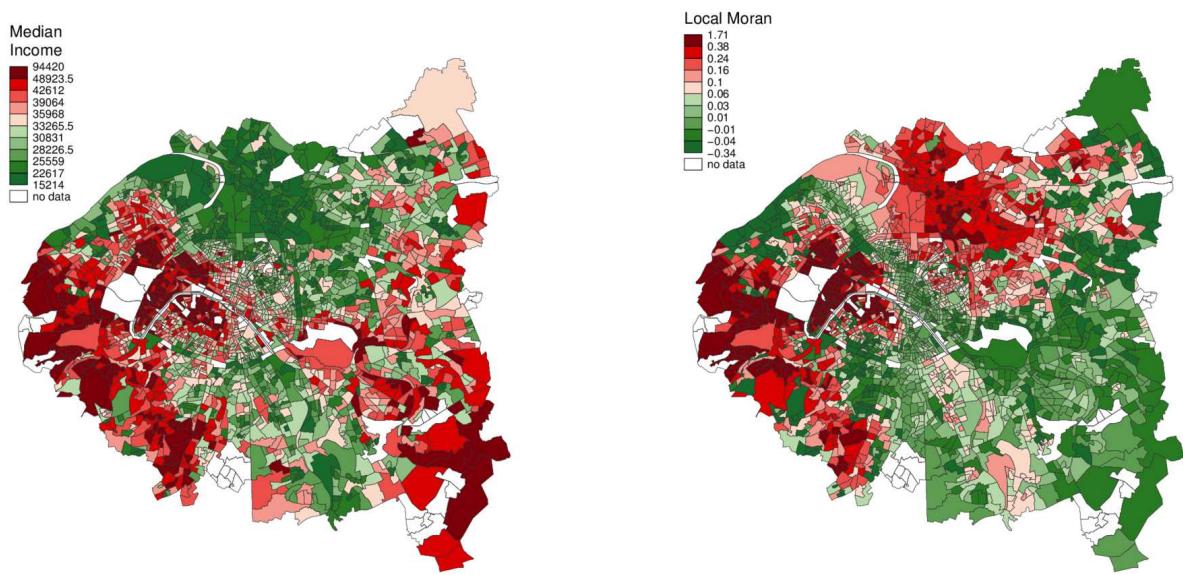


FIGURE 76: Cartes de ségrégation métropolitaine. Les cartes montrent le revenu annuel médian pour les unités statistiques élémentaires (IRIS) pour les trois départements correspondant globalement à la métropole du Grand Paris, et l'index local d'autocorrelation spatiale de Moran correspondant, défini 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.

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.5.4 *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 peut 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 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 [TIVADAR et al., 2014] qui calcule divers indices de ségrégation urbaine, comme on l'a déjà illustré pour l'application à la ségrégation métropolitaine.

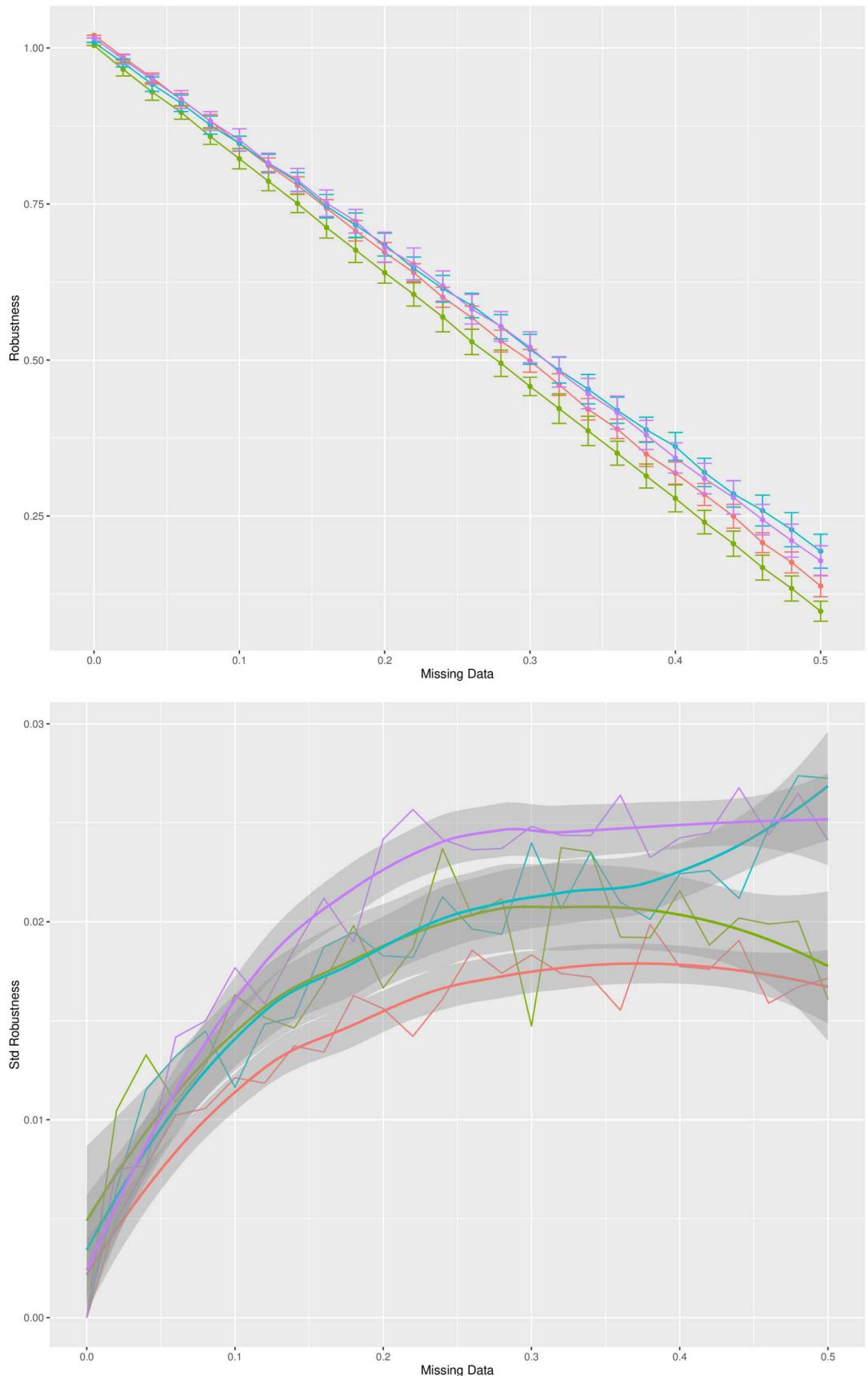


FIGURE 77: Sensibilité de la robustesse aux données manquantes. Gauche. Pour chaque département, des simulations de Monte-Carlo ($N=75$ répétitions) sont utilisées pour déterminer l'impact des données manquantes sur la robustesse de l'évaluation de la ségrégation. Les ratios de robustesse sont tous calculés 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. Détermination de la robustesse standardisée. Les différences entre les quatre trajectoires sont moins marquées que celles de la robustesse standardisée.

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 à 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 [KWAN, 1998] le souligne au sujet de l’accessibilité, 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 [BANOS, Décembre 2013], 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 [DRAGOMIR, 1999].

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 l'intégration spatiale. **C : confusion spatial / kernel ? -> idem à l'oral ?**

B.6 EXPLORATION DES MOTIFS D'INTERDISCIPLINARITÉ POUR UN JOURNAL GÉNÉRALISTE

We use the python library `nltk` [BIRD, 2006] that provides state-of-the-art operations in Natural Language Processing. A particular treatment is required for language detection with *stop-words* and a specific tagger TreeTagger is used for other languages than english ([SCHMID, 1994]). More precisely, we go through the following steps :

1. Language detection using *stop-words*
2. Parsing and tokenizing / pos-tagging (word functions) / stemming done differently depending on language :
 - English : `nltk` built-in pos-tagger, combined to a *PorterStemmer*
 - French or other : use of TreeTagger [SCHMID, 1994]
3. Selection of potential *n-grams* (with $1 \leq n \leq 4$) following the given patterns : for English $\cap\{\text{NN} \cup \text{VBG} \cup \text{JJ}\}$, and for French $\cap\{\text{NOM} \cup \text{ADJ}\}$
4. Database insertion for instantaneous utilisation (reducing effective time from 10 days to 2 minutes)
5. Estimation of *n-grams* relevance, following co-occurrences statistical distribution

B.7 UN CADRE POUR LES SYSTÈMES SOCIO-TECHNIQUES

Après avoir introduit une cadre théorique sur le plan thématique, nous développons un cadre plus général au sein duquel le précédent peut entrer. Il vise à contextualiser les directions générales de recherche à un niveau épistémologique mais formalisé, essayant d'obtenir une certaine structure algébrique pour capturer certaines propriétés des processus de modélisation.

B.7.1 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 : [HIDALGO, 2015] 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*⁸ 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)⁹, 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 [PUMAIN, 2005] 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 sont naturellement bénéfiques pour la diversité scien-

⁸ 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 [LAUGHLIN, 2006].

⁹ nous prenons ici le parti que soluble analytiquement implique la simplicité, puisque le système n'exhibe alors pas d'émergence faible (voir 3.3).

tifique, 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 [DUPUY et BENGUI-GUI, 2015]. La Science des Systèmes Complexes¹⁰ est présentée par certains comme "un nouveau type de science" [WOLFRAM, 2002], et serait au moins symptomatique d'un changement de paradigme des pratiques, des approches analytiques "exactes" vers des approches computationnelles et *evidence-based* [ARTHUR, 2015], 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 [BOURGINE, CHAVALARIAS et AL., 2009]. 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 naturel¹¹, 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 [LOUF et BARTHELEMY, 2014b]), mais aussi par des chercheurs venant des sciences sociales comme BATTY [BATTY, 2013b]), 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).

¹⁰ 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.

¹¹ les systèmes géographiques au sens de [DOLLFUS et DASTÈS, 1975] 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 objets¹².
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.7.2 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 [GIERE, 2010c]. Pour résumer, cette position interprète toute démarche scientifique comme une perspective, au sein de laquelle chacun poursuit

¹² comme déjà expliqué précédemment, ce positionnement combiné à l'importance de la structure pourrait être relié au *Réalisme Structurel Ontologique* dans des approfondissements.

certains 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 [VARENNE, 2010a] 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 [GOLDEN, AIGUIER et KROB, 2012], 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 [LIVET et al., 2010a], 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 est 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 :

Definition 2 *Une perspective sur un système est donnée par une Dataflow Machine $M = (i, o, \mathbb{T})$ 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 :

Definition 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.

Graphe 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 [FRIGG et VOTSI, 2011], 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 complexe¹³. Nous prenons pour cet aspect le positionnement de BEAU sur les différents types d'émergence déjà présenté plusieurs fois, en particulier sa définition de l'émergence faible donnée dans [BEAU, 2002]. Rappelons brièvement les définitions que nous utiliserons par la suite. BEAU commence par définir les propriétés émergentes puis étend le concept aux phénomènes, entités, etc. De la même manière, notre cadre n'est pas restreint aux objets ou propriétés et inclut ainsi les définitions généralisées comme lien entre ontologies. Nous appliquons la notion d'émergence sous les deux formes suivantes¹⁴ :

- *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".

¹³ ce qui bien sûr ne peut être formulé comme une affirmation prouvable car cela dépendra de la définition d'un système, etc.

¹⁴ la troisième forme rappelée par BEAU, *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 :

Assumption 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 $\mathcal{P}(O)$:*

- *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.

Definition 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 réduits¹⁵ par : si $o \in O, O'$ et $(O' \preccurlyeq O$ ou $O' \sqsubseteq O)$ mais pas $(O \preccurlyeq O'$ 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 :

Definition 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.

Assumption 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 [GOLDEN, AIGUIER et KROB, 2012] 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 préférons être moins restrictifs et postuler l'existence de la

¹⁵ la procédure de réduction vise à supprimer la redondance, gardant une entité au plus haut niveau où elle existe.

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.*

Proof (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 des 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 confirmions 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 [MANSON, 2008], 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'échelle¹⁶ $\|\cdot\|_s$ 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.7.3 Applications et discussion

Le cas particulier des systèmes géographiques

Dans [DOLLFUS et DASTÈS, 1975], DURAND-DASTÈS introduit une définition des systèmes et structures géographiques, la structure étant le contenant 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, complétée 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

¹⁶ 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 [CHÉREL, COTTINEAU et REUILLO, 2015] pourrait être fait dans le cas où ils ajouteraient la couche bibliographique, qui correspondrait à la reconstruction des perspectives. [REYMOND et CAUVIN, 2013] 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. [LUZEAUX, 2015] 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'émergence 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'émergence 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 [8.1](#), les prix des carburants [8.2](#), les formes urbaines et de réseau [4.1](#).

★ ★

★

C

DÉVELOPPEMENTS THÉMATIQUES

C.1 PONTS ENTRE GÉOGRAPHIE ET ECONOMIE : LEÇONS DES PERSPECTIVES DE MODÉLISATION

★ ★

★

C.2 GENERATION OF CORRELATED SYNTHETIC DATA

Application : Séries temporelles financières

Contexte

Un premier domaine d'application proposé pour notre méthode est celui des séries temporelles financières, signaux typiques de systèmes complexes hétérogènes et multiscalaires [MANTEGNA et STANLEY, 2000] 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 correlations pour un grand nombre d'actifs échantillonnés à faible fréquence (retours journaliers par exemple) [BOUCHAUD et POTTERS, 2009]. De même, l'analyse de réseaux complexes construits à partir des corrélations, selon des méthodes type arbre couvrant minimal [BONANNO, LILLO et MANTEGNA, 2001] ou des extensions raffinées pour cette application précise [TUMMINELLO et al., 2005], 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 [BARNDORFF-NIELSEN et al., 2011]. 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, [POTIRON et MYKLAND, 2015] 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 $(X_i(t))_{1 \leq i \leq N}$ échantillonés à haute fréquence (typiquement 1s). On se place dans un cadre multi-scalaire (utilisé par exemple dans les approches par ondelettes [RAMSEY, 2002] ou analyses multifractales du signal [BOUCHAUD, POTTERS et MEYER, 2000]) 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 problème 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 comparaison¹. 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 [JARROW, 1999] 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.

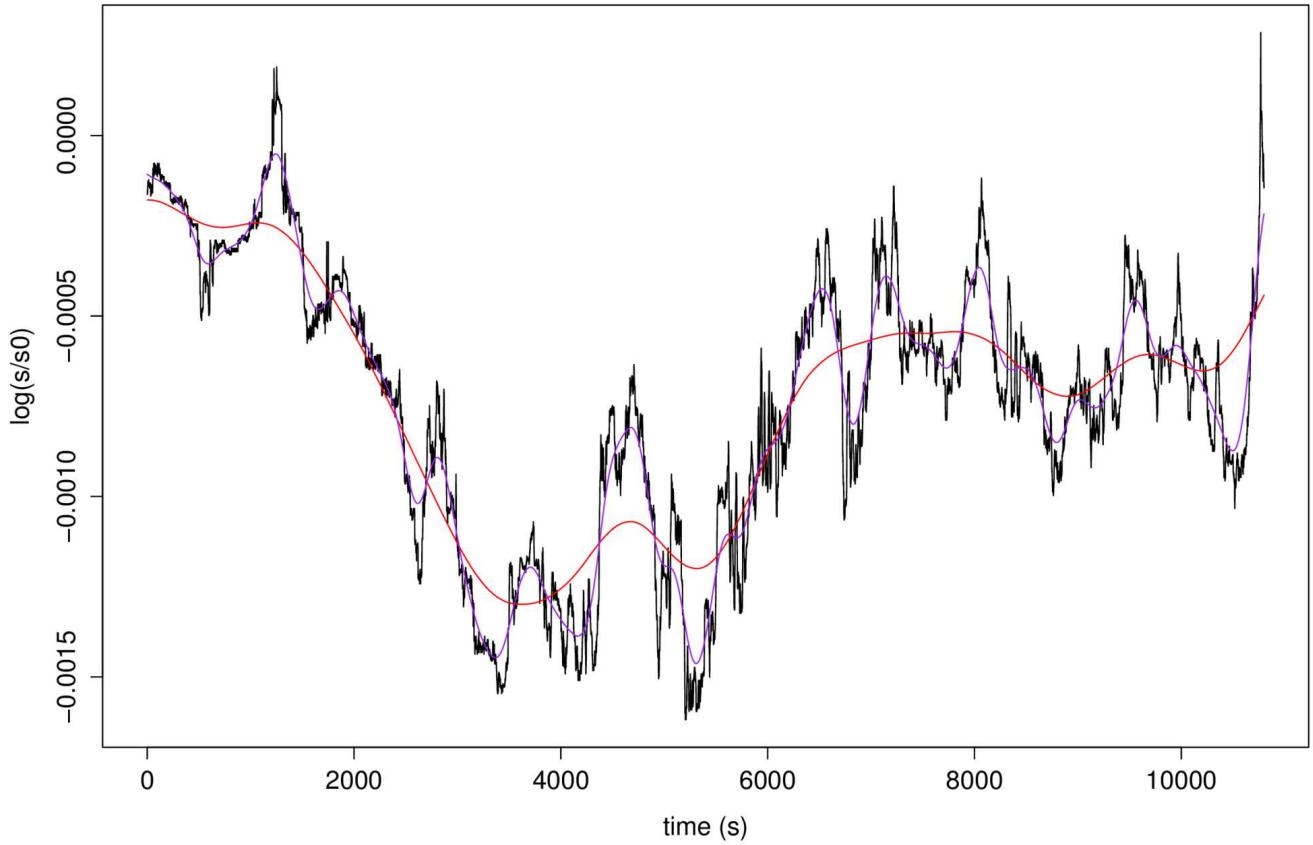
GÉNÉRATION DES DONNÉES Il est alors aisé de générer \tilde{X}_i tel que $\text{Var}[\tilde{X}_i^{\omega_1}] = \Sigma R$ (Σ variance estimée et R matrice de corrélation fixée), par la simulation de processus de Wiener au niveau de corrélation fixé et tel que $X_i^{\omega \leq \omega_0} = \tilde{X}_i^{\omega \leq \omega_0}$ (critère de proximité au données : les composantes à plus basse fréquence qu'une fréquence fondamentale $\omega_0 < \omega_1$ sont identiques). En effet, si $dW_1 \perp\!\!\!\perp dW_1^{\perp}$ (et $\sigma_1 < \sigma_2$ pour fixer les idées, quitte à échanger les actifs), alors $W_2 = \rho_{12}W_1 + \sqrt{1 - \frac{\sigma_1^2}{\sigma_2^2} \cdot \rho_{12}^2} W_1^{\perp}$ est tel que $\rho(dW_1, dW_2) = \rho_{12}$. Les signaux suivants sont construits de la même manière par ortho-normalisation 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 $\tilde{X}_i = T_i^{\omega_0} + \tilde{X}_i^{\omega_1}$.

Implémentation et 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 données², originellement échantillonnées 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 extrémités 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 simi-

¹ voir [GELL-MANN, 1995] pour une discussion étendue sur la construction de *schema* pour l'étude de systèmes complexes adaptatifs (par des systèmes complexes adaptatifs).

² 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 78: Exemple de la nature multi-scalaire des signaux.**

laires entre actifs). On étudie alors les *log-prix* et *log-retours*, 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 ?? pour un exemple de la structure du signal à ce différentes échelles.

Il est crucial de noter l'interférence entre les fréquences ω_0 et ω_1 dans le signal construit : la correlation 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 $\rho_0 = \rho [\Delta T_1^{\omega_0}, \Delta T_2^{\omega_0}]$, $\rho = \rho [\tilde{X}_1^{\omega_1}, \tilde{X}_2^{\omega_1}]$, et $\varepsilon_i =$

$\frac{\sigma(\Delta T_i^{\omega_0})}{\sigma(\Delta \tilde{X}_i^{\omega_1})}$, la correction sur la correlation effective due aux interférences : la correlation effective est alors au premier ordre

$$\rho_e = [\varepsilon_1 \varepsilon_2 \rho_0 + \rho] \cdot \left[1 - \frac{1}{2} (\varepsilon_1^2 + \varepsilon_2^2) \right] \quad (28)$$

ce qui donne l'expression de la correlation que l'on pourra effectivement simuler dans les données synthétiques.

La correlation est estimée par méthode de Pearson, avec l'estimateur de la covariance au biais corrigé, c'est à dire $\hat{\rho}[X_1, X_2] = \frac{\hat{C}[X_1, X_2]}{\sqrt{\hat{V}\text{ar}[X_1]\hat{V}\text{ar}[X_2]}}$, où $\hat{C}[X_1, X_2] = \frac{1}{(T-1)} \sum_t X_1(t)X_2(t) - \frac{1}{T \cdot (T-1)} \sum_t X_1(t) \sum_t X_2(t)$ et $\hat{V}\text{ar}[X] = \frac{1}{T} \sum_t X^2(t) - \left(\frac{1}{T} \sum_t X(t)\right)^2$.

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 correlation retardée, on ne s'attend donc pas à de grand ordre d'auto-regression, 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 adaptative³. 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 language R, utilisant en particulier la bibliothèque MTS [TSAY, 2015] pour les modèles de séries temporelles. Les données nettoyées et le code source sont disponibles de manière ouverte sur le dépôt git du projet⁴.

RÉSULTATS La figure ?? donne les correlations 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 $\varepsilon_1 \simeq 0.3$ 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 prédictive par 28 et les valeurs observées, et une déviation pour de plus grandes valeurs absolues, d'autant plus grande que ω_1 est petit : cela confirme 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 correlation des données. Les résultats pour $\omega_1 =$

³ 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 [POTIRON, 2016] qui suppose une dynamique localement paramétrique, mais pour lequel on fixe les méta-paramètres 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.

⁴ at <https://github.com/JusteRaimbault/SynthAsset>

1h, 1h30, 2h sont présentés en figure ???. Le résultat a priori contre-intuitif d'une performance maximale à correlation nulle pour l'un des actifs confirme l'intérêt d'une génération de données hybrides : l'étude des correlations 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 correlation : plus les bruits sont corrélés, plus le modèle les prendra en compte et se trompera plus à cause du caractère markovien des browniens simulés⁵.

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.

⁵ en théorie le modèle utilisé n'a aucun pouvoir prédictif sur des browniens purs

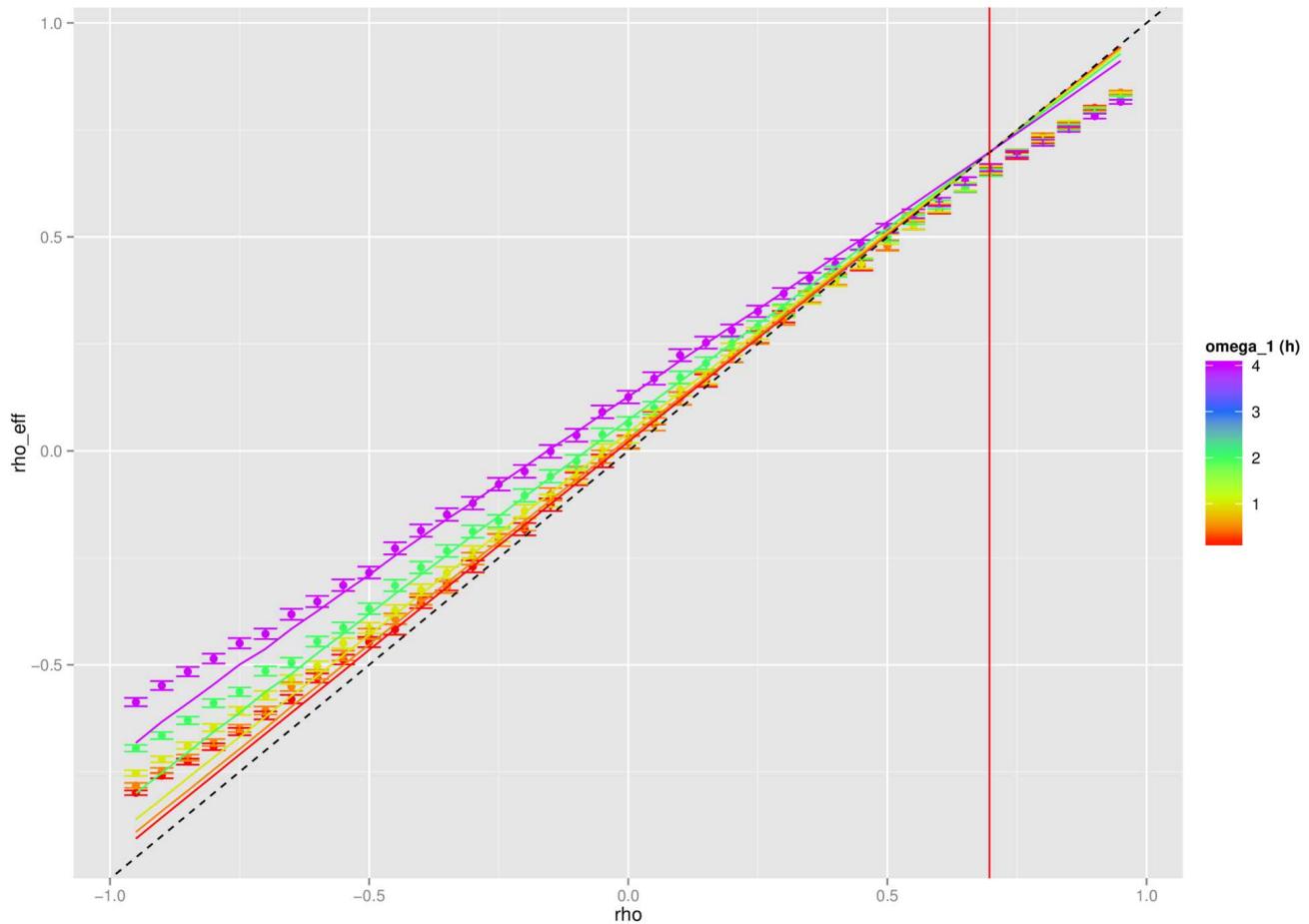


FIGURE 79: Corrélations Effectives obtenues sur données synthétiques.

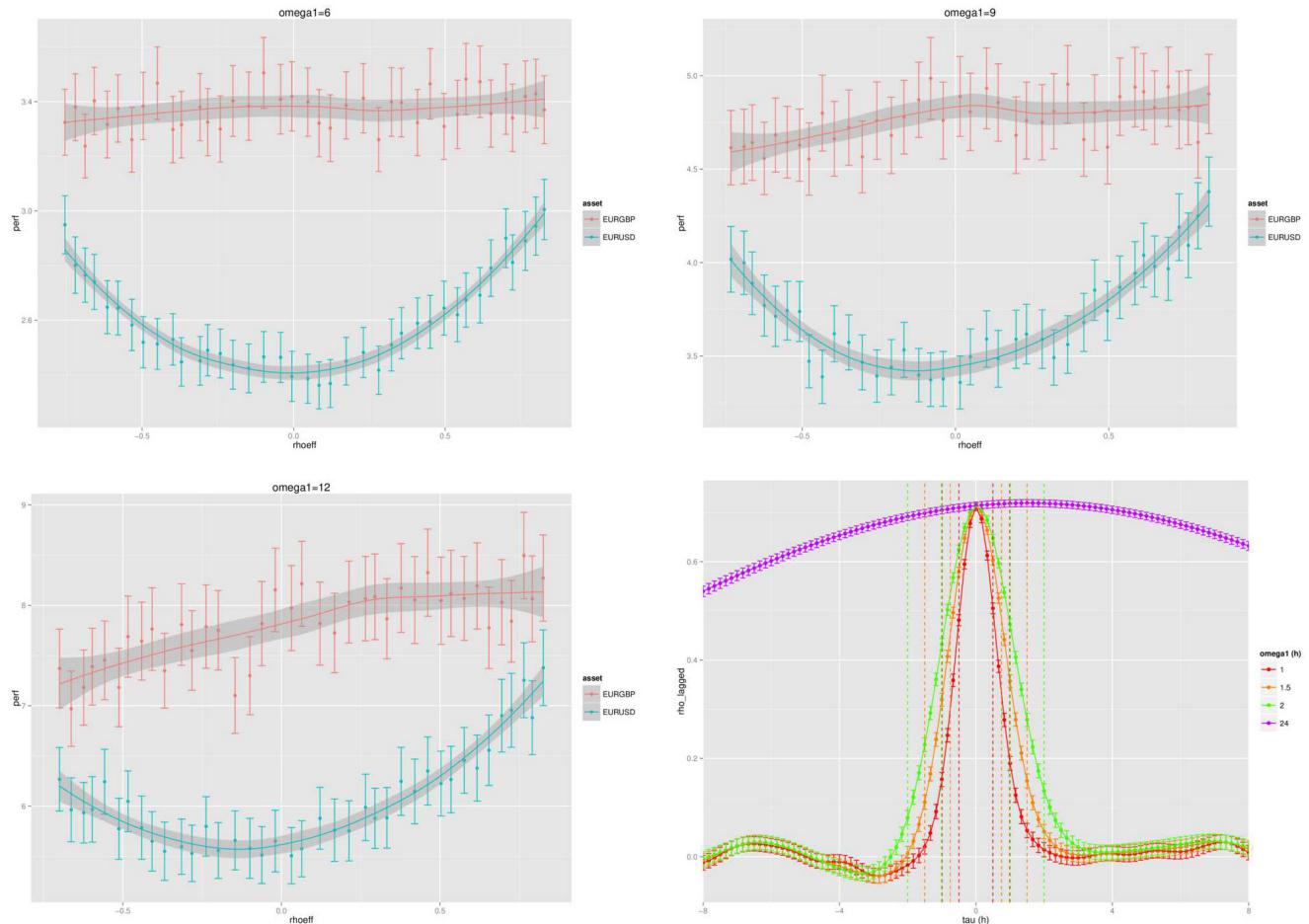


FIGURE 80: Performance d'un modèle prédictif en fonction des corrélations simulées.

C.3 CLASSIFYING PATENTS BASED ON THEIR SEMANTIC CONTENT

In this paper, we extend some usual techniques of classification resulting from a large-scale data-mining and network approach. This new technology, which in particular is designed to be suitable to big data, is used to construct an open consolidated database from raw data on 4 million patents taken from the US patent office from 1976 onward. To build the pattern network, not only do we look at each patent title, but we also examine their full abstract and extract the relevant keywords accordingly. We refer to this classification as *semantic approach* in contrast with the more common *technological approach* which consists in taking the topology when considering US Patent office technological classes. Moreover, we document that both approaches have highly different topological measures and strong statistical evidence that they feature a different model. This suggests that our method is a useful tool to extract endogenous information.

Introduction

Innovation and technological change have been described by many scholars as the main drivers of economic growth as in [AGHION et HOWITT, 1992] and [ROMER, 1990]. [GRILICHES, 1990] advertised the use of patents as an economic indicator and as a good proxy for innovation. Subsequently, the easier availability of comprehensive databases on patent details and the increasing number of studies allowing a more efficient use of these data (e.g. [HALL, JAFFE et TRAJTENBERG, 2001]) have opened the way to a very wide range of analysis. Most of the statistics derived from the patent databases relied on a few key features : the identity of the inventor, the type and identity of the rights owner, the citations made by the patent to prior art and the technological classes assigned by the patent office post patent's content review. Combining this information is particularly relevant when trying to capture the diffusion of knowledge and the interaction between technological fields as studied in [YOUN et al., 2015]. With methods such as citation dynamics modeling discussed in [NEWMAN, 2013] or co-authorship networks analysis in [SARIGÖL et al., 2014], a large body of the literature such as [SORENSEN, RIVKIN et FLEMING, 2006] or [KAY et al., 2014] has studied patents citation network to understand processes driving technological innovation, diffusion and the birth of technological clusters. Finally, [BRUCK et al., 2016] look at the dynamics of citations from different classes to show that the laser/ink-jet printer technology resulted from the recombination of two different existing technologies.

Consequently, technological classification combined with other features of patents can be a valuable tool for researchers interested in studying technologies throughout history and to predict future inno-

vations by looking at past knowledge and interaction across sectors and technologies. But it is also crucial for firms that face an ever changing demand structure and need to anticipate future technological trends and convergence (see, e.g., [CURRAN et LEKER, 2011]) to adapt to the resulting increase in competition discussed in [KATZ, 1996] and to maintain market share. Curiously, and in spite of the large number of studies that analyze interactions across technologies [FURMAN et STERN, 2011], little is known about the underlying “innovation network” (e.g. [ACEMOGLU et KERR, 2016]).

In this monograph, we propose an alternative classification based on semantic network analysis from patent abstracts and explore the new information emerging from it. In contrast with the regular technological classification which results from the choice of the patent reviewer, semantic classification is carried automatically based on the content of the patent abstract. Although patent officers are experts in their fields, the relevance of the existing classification is limited by the fact that it is based on the state of technology at the time the patent was granted and cannot anticipate the birth of new fields. To correct for this, the USPTO regularly make changes in its classification in order to adapt to technological change (for example, the “nanotechnology” class (977) was established in 2004 and retroactively to all relevant previously granted patents). In contrast we don’t face this issue with the semantic approach. The semantic links can be clues of one technology taking inspiration from another and good predictors of future technology convergence (e.g. [PRESCHITSCHER et al., 2013] study semantic similarities from the whole text of 326 US-patents on *phytosterols* and show that semantic analysis have a good predicting power of future technology convergence). One can for instance consider the case of the word *optic*. Until more recently, this word was often associated with technologies such as photography or eye surgery, while it is now almost exclusively used in a context of semi-transistor design and electro-optic. This semantic shift did not happen by chance but contains information on the fact that modern electronic extensively uses technologies that were initially developed in optic.

Previous research has already proposed to use semantic networks to study technological domains and detect novelty. [YOUN et PARK, 2004] was one of the first to enhance this approach with the idea of visualizing keywords network illustrated on a small technological domain. The same approach can be used to help companies identifying the state of the art in their field and avoid patent infringement as in [PARK et YOUN, 2014] and [YOUN et KIM, 2011]. More closely related to our methodology, [GERKEN et MOEHRLE, 2012] develop a method based on patent semantic analysis of patent to vindicate the view that this approach outperform others in the monitoring of technology and in the identification of novelty innovation. Semantic analysis has al-

ready proven its efficiency in various fields, such as in technology studies (e.g. [CHOI et HWANG, 2014] and [FATTORI, PEDRAZZI et TURRA, 2003]) and in political science (e.g. [GURCIULLO et al., 2015]).

Building on such previous research, we make several contributions by fulfilling some shortcomings of existing studies, such as for example the use of frequency-selected single keywords. First of all, we develop and implement a novel fully-automatized methodology to classify patents according to their semantic abstract content, which is to the best of our knowledge the first of its type. This includes the following refinements for which details can be found in Section ?? : (i) use of multi-stems as potential keywords; (ii) filtering of keywords based on a second-order (co-occurrences) relevance measure and on an external independent measure (technological dispersion); (iii) multi-objective optimization of semantic network modularity and size. The use of all this techniques in the context of semantic classification is new and essential from a practical perspective.

Furthermore, most of the existing studies rely on a subsample of patent data, whereas we implement it on the full US Patent database from 1976 to 2013. This way, a general structure of technological innovation can be studied. We draw from this application promising qualitative stylized facts, such as a qualitative regime shift around the end of the 1990s, and a significant improvement of citation modularity for the semantic classification when comparing to the technological classification. These thematic conclusions validate our method as a useful tool to extract endogenous information, in a complementary way to the technological classification.

Finally, the statistical model introduced in Section seems to indicate that patents tend to cite more similar patents in the semantic network when fitted to data. In particular, this propensity is shown to be significantly bigger than the corresponding propensity for technological classes, and this seems to be consistent over time. On the account of this information, we believe that patent officers could benefit very much from looking at the semantic network when considering potential citation candidates of a patent in review.

The paper is organized as follows. Section ?? presents the patent data, the existing classification and provide details about the data collection process. Section ?? explains the construction of the semantic classes. Section ?? tests their relevance by providing exploratory results. Finally, section ?? discusses potential further developments and conclude. More details, including robustness checking, figures and technical derivations can be found in ??, ?? and ??.

Background

In our analysis, we will consider all utility patents granted in the United States Patent and Trademark Office (USPTO) from 1976 to 2013.

A clearer definition of utility patent is given in [??](#). Also, additional information on how to correctly exploit patent data can be found in [HALL, JAFFE et TRAJTENBERG, [2001](#)] and [LERNER et SERU, [2015](#)].

An existing classification : the USPC system

Each USPTO patent is associated with a non-empty set of technological classes and subclasses. There are currently around 440 classes and over 150,000 subclasses constituting the United State Patent Classification (USPC) system. While a technological class corresponds to the technological field covered by the patent, a subclass stands for a specific technology or method used in this invention. A patent can have multiple technological classes, on average in our data a patent has 1.8 different classes and 3.9 pairs of class/subclass. At this stage, two features of this system are worth mentioning : (i) classes and subclasses are not chosen by the inventors of the patent but by the examiner during the granting process based on the content of the patent; (ii) the classification has evolved in time and continues to change in order to adapt to new technologies by creating or editing classes. When a change occurs, the USPTO reviews all the previous patents so as to create a consistent classification.

A bibliographical network between patents : citations

As with scientific publications, patents must give reference to all the previous patents which correspond to related prior art. They therefore indicate the past knowledge which relates to the patented invention. Yet, contrary to scientific citations, they also have an important legal role as they are used to delimit the scope of the property rights awarded by the patent. One can consult [OECD, [2009](#)] for more details about this. Failing to refer to prior art can lead to the invalidation of the patent (e.g. [DECHEZLEPRÂTRE, MARTIN et MOHNEN, [2014](#)]). Another crucial difference is that the majority of the citations are actually chosen by the examiners and not by the inventors themselves. From the USPTO, we gather information of all citations made by each patent (backward citations) and all citations received by each patent as of the end of 2013 (forward citations). We can thus build a complete network of citations that we will use later on in the analysis.

Turning to the structure of the lag between the citing and the cited patent in terms of application date, we see that the mean of this lag is 8.5 years and the median is 7 years. This distribution is highly skewed, the 95th percentile is 21 years. We also report 164,000 citations with a negative time lag. This is due to the fact that some citations can be added during the examination process and some patents require more time to be granted than others.

In what follows, we choose to restrict attention to pairs of citations with a lag no larger than 5 years. We impose this restriction for two

reasons. First, the number of citations received peaks 4-5 years after application. Second, the structure of the citation lag is necessarily biased by the truncation of our sample : the more recent patents mechanically receive less citations than the older ones. As we are restricting to citations received no later than 5 years after the application date, this effect will only affect patents with an application date after 2007.

Data collection and basic description

Each patent contains an abstract and a core text which describe the invention. To see what a patent looks like in practice, one can refer to the USPTO patent full-text database <http://patft.uspto.gov/netahtml/PTO/index.html> or to Google patent which publishes USPTO patents in pdf format at <https://patents.google.com>. Although including the full core texts would be natural and probably very useful in a systematic text-mining approach as done in [TSENG, LIN et LIN, 2007], they are too long to be included and thus we consider only the abstracts for the analysis. Indeed, the semantic analysis counts more than 4 million patents, with corresponding abstracts with an average length of 120.8 words (and a standard deviation of 62.4), a size that is already challenging in terms of computational burden and data size. In addition, abstracts are aimed at synthesizing purpose and content of patents and must therefore be a relevant object of study (see [ADAMS, 2010]). The USPTO defines a guidance stating that an abstract should be “a summary of the disclosure as contained in the description, the claims, and any drawings; the summary shall indicate the technical field to which the invention pertains and shall be drafted in a way which allows the clear understanding of the technical problem, the gist of the solution of that problem through the invention, and the principal use or uses of the invention” (PCT Rule 8).

We construct from raw data a unified database. Data is collected from USPTO patent redbook bulk downloads, that provides as raw data (specific dat or xml formats) full patent information, starting from 1976. Detailed procedure of data collection, parsing and consolidation are available in ?? . The latest dump of the database in Mongodb format is available at <http://dx.doi.org/10.7910/DVN/BW3ACK>. Collection and homogenization of the database into a directly usable database with basic information and abstracts was an important task as USPTO raw data formats are involved and change frequently.

We count 4,666,365 utility patents with an abstract granted from 1976 to 2013. A very small number of patents have a missing abstract, these are patents that have been withdrawn and we do not consider them in the analysis. The number of patents granted each year increases from around 70,000 in 1976 to about 278,000 in 2013. When distributed by the year of application, the picture is slightly different. The number of patents steadily increase from 1976 to 2000 and re-

mains constant around 200,000 per year from 2000 to 2007. Restricting our sample to patent with application date ranging from 1976 to 2007, we are left with 3,949,615 patents. These patents cite 38,756,292 other patents with the empirical lag distribution that has been extensively analyzed in [HALL, JAFFE et TRAJTENBERG, 2001]. Conditioned on being cited at least once, a patent receives on average 13.5 citations within a five-year window. 270,877 patents receive no citation during the next five years following application, 10% of patents receive only one citation and 1% of them receive more than 100 citations. A within class citation is defined as a citation between two patents sharing at least one common technological class. Following this definition, 84% of the citations are within class citations. 14% of the citations are between two patents that share the exact same set of technological classes.

Towards a Complementary Classification

Potentialities of text-mining techniques as an alternative way to analyze and classify patents are documented in [TSENG, LIN et LIN, 2007]. The author's main argument, in support of an automatic classification tool for patent, is to reduce the considerable amount of human effort needed to classify all the applications. The work conducted in the field of natural language processing and/or text analysis has been developed in order to improve search performance in patent databases, build technology map or investigate the potential infringement risks prior to developing a new technology (see [ABBAS, ZHANG et KHAN, 2014] for a review). Text-mining of patent documents is also widely used as a tool to build networks which carry additional information to the simplistic bibliographic connections model as argued in [YOUN et PARK, 2004]. As far as the authors know, the use of text-mining as a way to build a global classification of patents remains however largely unexplored. One notable exception can be found in [PRESCHITSCHET AL., 2013] where semantic-based classification is shown to outperform the standard classification in predicting the convergence of technologies even in small samples. Semantic analysis reveals itself to be more flexible and more quickly adaptable to the apparition of new clusters of technologies. Indeed, as argued in [PRESCHITSCHET AL., 2013], before two distinct technologies start to clearly converge, one should expect similar words to be used in patents from both technologies.

Finally, a semantic classification where patents are gathered based on the fact that they share similar significant keywords has the advantage of including a network feature that cannot be found in the USPC case, namely that each patent is associated with a vector of probability to belong to each of the semantic classes (more details on this feature can be found in Section C.3). Using co-occurrence of keywords, it is then possible to construct a network of patents and

to study the influence of some key topological features. As reviewed previously, the use of co-occurrences is the usual way to construct a semantic network. Other hybrid technique such as bipartite semantic/authors networks, do not have the nice feature of relying solely on endogenous semantic information contained in data.

Semantic Classification Construction

In this section, we describe methods and empirical analysis leading to the construction of semantic network and the corresponding classification.

Keywords extraction

Let \mathcal{P} be the set of patents, we first assign to a patent $p \in \mathcal{P}$ a set of potentially significant keywords $K(p)$ from its text $A(p)$ (that corresponds to the concatenation of its own title and abstract). $K(p)$ are extracted through a similar procedure as the one detailed in [CHAVALARIAS et COINTET, 2013] :

1. Text parsing and Tokenization : we transform raw texts into a set of words and sentences, reading it (parsing) and splitting it into elementary entities (words organized in sentences).
2. Part-of-speech tagging : attribution of a grammatical function to each of the tokens defined previously.
3. Stem extraction : families of words are generally derived from a unique root called stem (for example compute, computer, computation all yield the same stem comput) that we extract from tokens. At this point the abstract text is reduced to a set of stems and their grammatical functions.
4. Multi-stems construction : these are the basic semantic units used in further analysis. They are constructed as groups of successive stems in a sentence which satisfies a simple grammatical function rule. The length of the group is between 1 and 3 and its elements are either nouns, attributive verbs or adjectives. We choose to extract the semantics from such nominal groups in view of the technical nature of texts, which is not likely to contain subtle nuances in combinations of verbs and nominal groups.

Text processing operations are implemented in python in order to use built-in functions `nltk` library [NLTK, 2015] for most of above operations. This library supports most of state-of-the-art natural language processing operations. Source code is openly available on the repository of the project at <https://github.com/JusteRaimbault/PatentsMining>.

Keywords relevance estimation

RELEVANCE DEFINITION Following the heuristic in [CHAVALARIAS et COINTET, 2013], we estimate relevance score in order to filter multi-stem. The choice of the total number of keywords to be extracted, which we shall denote K_w , is important, too small a value would yield similar network structures but including less information whereas very large values tend to include too many irrelevant keywords. We choose to set this parameter to $K_w = 100,000$. We first consider the filtration of $k \cdot K_w$ (with $k = 4$) to keep a large set of potential keywords but still have a reasonable number of co-occurrences to be computed. This step has only very marginal effects on the nature of the final keywords but is necessary for computational purposes. The filtration is done on the *unithood* u_i , defined for keyword i as $u_i = f_i \cdot \log(1 + l_i)$ where f_i is the multi-stem's number of apparitions over the whole corpus and l_i its length in words. A second filtration of K_w keywords is done on the *termhood* t_i , where the formal definition can be found in (29). It is computed as a chi-squared score on the distribution of the stem's co-occurrences and then compared to a uniform distribution within the whole corpus. Intuitively, uniformly distributed terms will be identified as plain language and they are thus not relevant for the classification. More precisely, we compute the co-occurrence matrix (M_{ij}) , where M_{ij} is defined as the number of patents where stems i and j appear together. The *termhood* score t_i is defined as

$$t_i = \sum_{j \neq i} \frac{(M_{ij} - \sum_k M_{ik} \sum_k M_{jk})^2}{\sum_k M_{ik} \sum_k M_{jk}}. \quad (29)$$

MOVING WINDOW ESTIMATION The previous scores are estimated on a moving window with fixed time length following the idea that the present relevance is given by the most recent context and thus that the influence vanishes when going further into the past. Consequently, the co-occurrence matrix is chosen to be constructed at year t restricting to patent which applied during the time window $[t - T_0; t]$. Note that the causal property of the window is crucial as the future cannot play any role in the current state of keywords and patents. This way, we will obtain semantic classes which are exploitable on a T_0 time span. For example, this enables us to compute the modularity of classes in the citation network as in section C.3. In the following, we take $T_0 = 4$ (which corresponds to a five year window) consistently with the choice of maximum time lag for citations made in Section ???. Accordingly, the sensitivity analysis for $T_0 = 2$ can be found in Appendix ??.

Construction of the semantic network

We keep the set of most relevant keywords \mathcal{K}_W and obtain their co-occurrence matrix as defined in Section C.3. This matrix can be directly interpreted as the weighted adjacency matrix of the semantic network. At this stage, the topology of raw networks does not allow the extraction of clear communities. This is partly due to the presence of hubs that correspond to frequent terms common to many fields (e.g. method, apparat) which are wrongly filtered as relevant. We therefore introduce an additional measure to correct the network topology : the concentration of keywords across technological classes, defined as :

$$c_{\text{tech}}(s) = \sum_{j=1}^{N^{(\text{tec})}} \frac{k_j(s)^2}{(\sum_i k_i(s))^2},$$

where $k_j(s)$ is the number of occurrences of the s th keyword in each of the j th technological class taken from one of the $N^{(\text{tec})}$ USPC classes. The higher c_{tech} , the more specific to a technological class the node is. For example, the terms **semiconductor** is widely used in electronics and does not contain any significant information in this field. We use a threshold parameter, defined as θ_c , and keep nodes with $c_{\text{tech}}(s) > \theta_c$. Likewise, edges with low weights correspond to rare co-occurrences and are considered to be noise. To account for this we define the threshold parameter for edges θ_w , and we filter edges with a weight below θ_w , following the rationale that two keywords are not linked “by chance” if they appear simultaneously a minimal number of time. To control for size effect, we normalize by taking $\theta_w = \theta_w^{(0)} \cdot N_p$ where N_p is the number of patents in the corpus ($N_p = |\mathcal{P}|$). $\theta_w^{(0)}$ is thus a varying parameter interpreted as a noise threshold *per patent*. Communities are then extracted using a standard modularity maximization procedure as described in [CLAUSET, NEWMAN et MOORE, 2004] to which we add the two constraints captured by θ_w and θ_c , namely that edges must have a weight greater than θ_w and nodes a concentration greater than θ_c . At this stage, both parameters θ_c and $\theta_w^{(0)}$ are unconstrained and their choice is not straightforward. Indeed, many optimization objectives are possible, such as the modularity, network size or number of communities. We find that modularity is maximized at a roughly stable value of θ_w across different θ_c for each year, corresponding to a stable $\theta_w^{(0)}$ across years, which leads us to choose $\theta_w^{(0)} = 4.1 \cdot 10^{-5}$. Then for the choice of θ_c , different candidates points lie on a Pareto front for the bi-objective optimization on number of communities and network size. There is a priori no reason to choose any specific point among the different optima. Consequently, we have tried the analysis with all the candidate values for θ_c and found that the results are the most

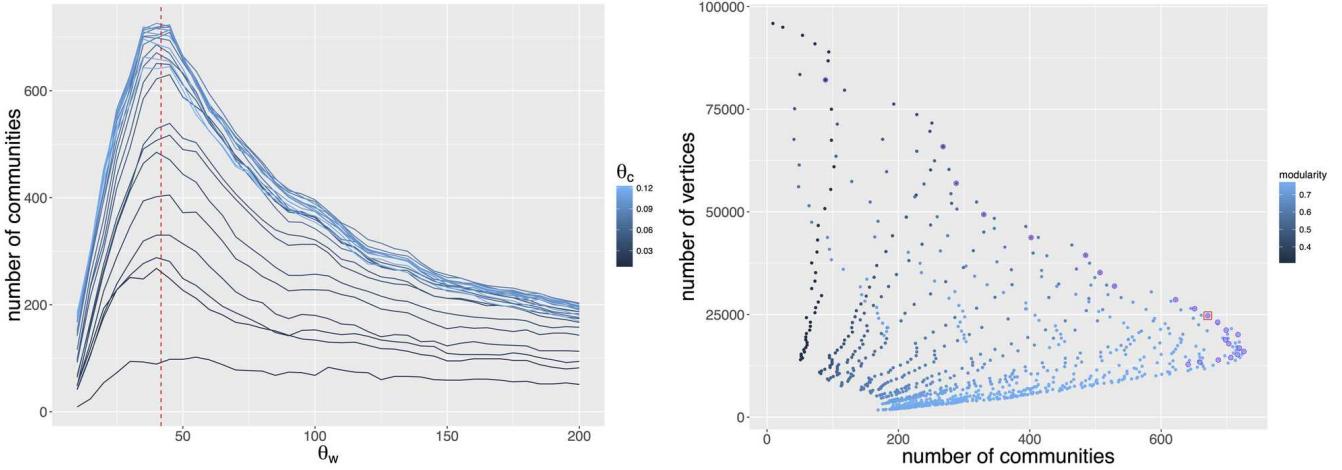


FIGURE 81: Sensitivity analysis of network community structure to filtering parameters. We consider a specific window 2000-2004 and the obtained plots are typical. (*Left panel*) We plot the number of communities as a function of the edge threshold parameter θ_w for different values of the node threshold parameter θ_c . The maximum is roughly stable across θ_c (dashed red line). (*Right panel*) To choose θ_c , we do a Pareto optimization on communities and network size : the compromise point (red overline) on the Pareto front (purple overline : possible choices after having fixed $\theta_w^{(0)}$; blue level gives modularity) corresponds to $\theta_c = 0.06$.

reasonable when taking $\theta_c = 0.06$ (see Fig. ??). We show in Fig. ?? an example of semantic network visualization.

Characteristics of Semantic Classes

For each year t , we define as $N_t^{(sem)}$ the number of semantic classes which have been computed by clustering keywords from patents appeared during the period $[t - T_0, t]$ (we recall that we have chosen $T_0 = 4$). Each semantic class $k = 1, \dots, N_t^{(sem)}$ is characterized by a set of keywords $K(k, t)$ which is a subset of \mathcal{K}_W selected as described in previous sections. The cardinal of $K(k, t)$ distribution across each semantic class k is highly skewed with a few semantic classes containing over 1,000 keywords, most of them with roughly the same number of keywords. In contrast, there are also many semantic classes with only two keywords. There are around 30 keywords by semantic class on average and the median is 2 for any t . Fig. ?? shows that the average number of keywords is relatively stable from 1976 to 1992 and then picks around 1996 prior to going down.

TITLE OF SEMANTIC CLASSES USPC technological classes are defined by a title and a highly accurate definition which help retrieve patents easily. The title can be a single word (e.g. : class 101 : “Printing”) or more complex (e.g. : class 218 : “High-voltage switches with

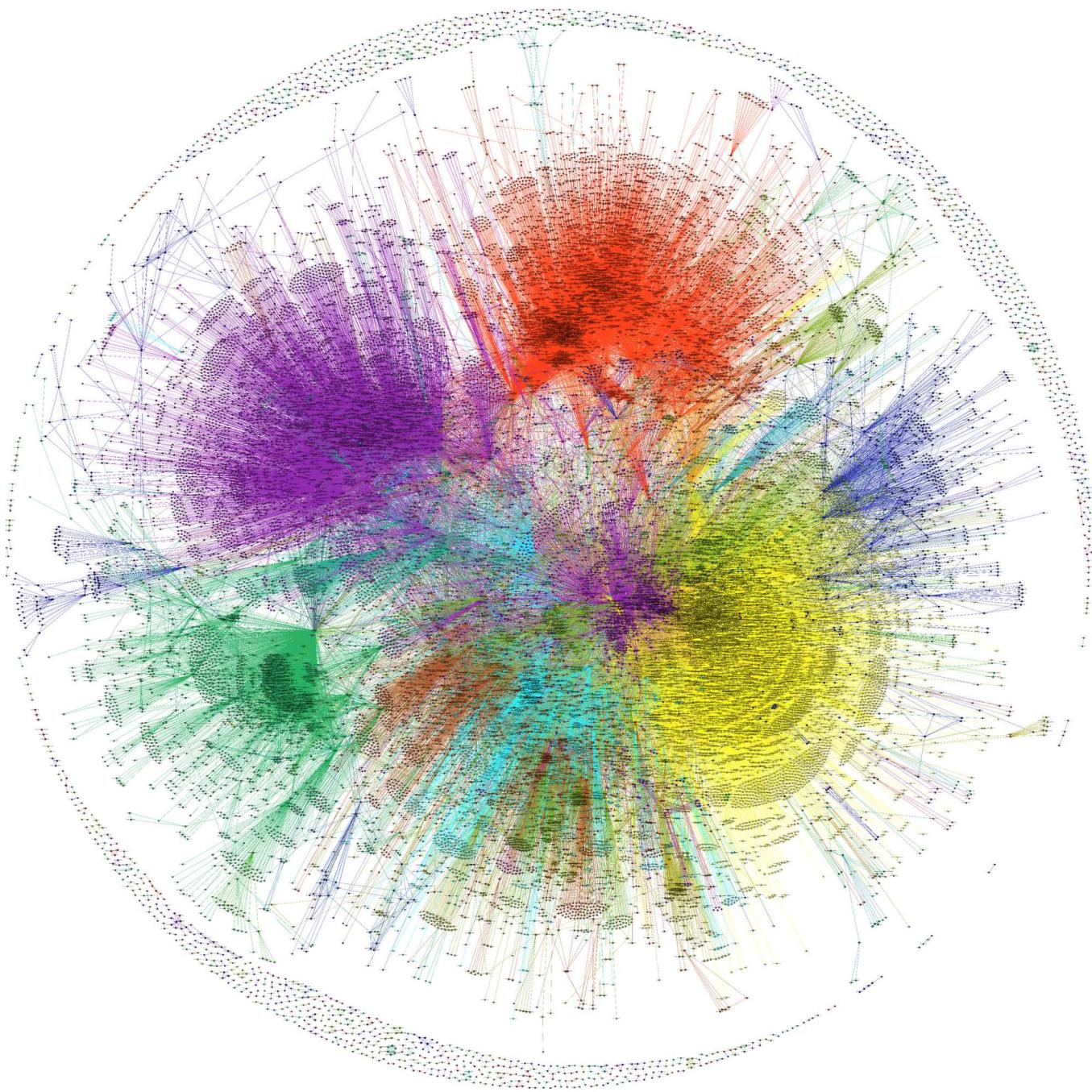


FIGURE 82: An example of semantic network visualization. We show the network obtained for the window 2000-2004, with parameters $\theta_c = 0.06$ and $\theta_w = \theta_w^{(0)} \cdot N_p = 4.5e^{-5} \cdot 9.1e^5$. The corresponding file in a vector format (.svg), that can be zoomed and explored, is available as ??.

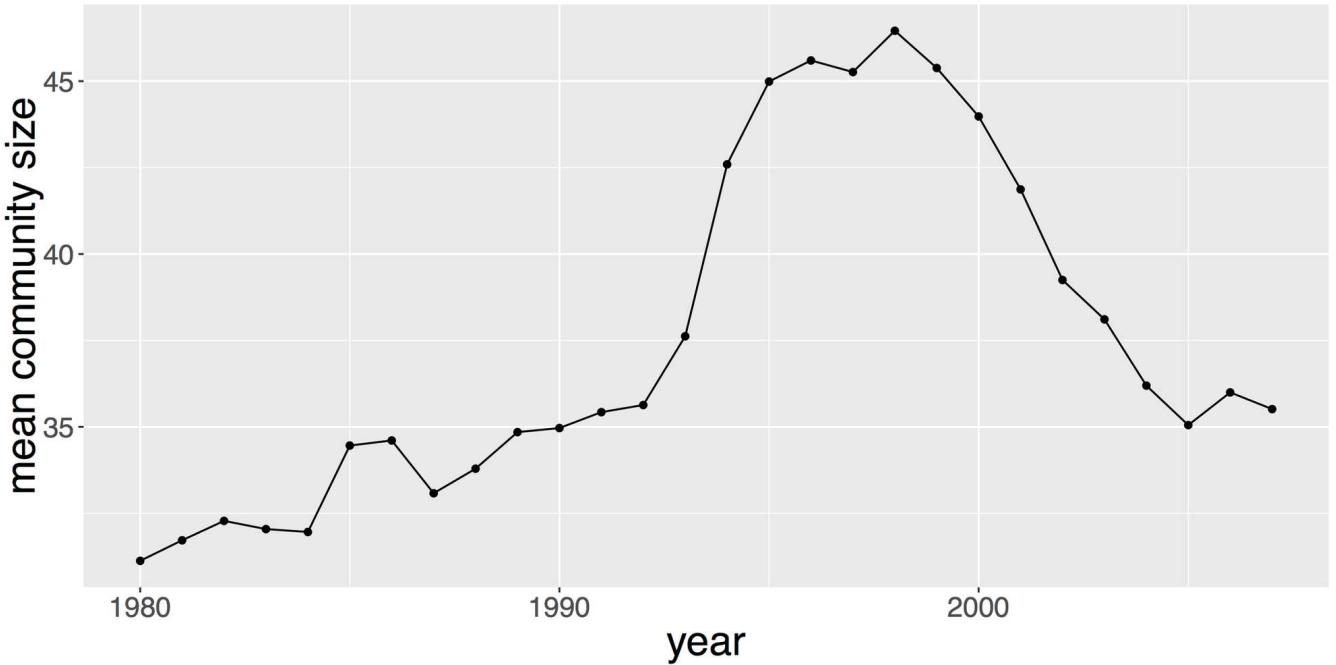


FIGURE 83: This figure plots the average number of keywords by semantic class for each time window $[t - 4; t]$ from $t = 1980$ to $t = 2007$.

arc preventing or extinguishing devices"). As our goal is to release a comprehensive database in which each patent is associated with a set of semantic classes, it is necessary to give an insight on what these classes represent by associating a short description or a title as in [TSENG, LIN et LIN, 2007]. In our case, such description is taken as a subset of keywords taken from $K(k, t)$. For the vast majority of semantic classes that have less than 5 keywords, we decide to keep all of these keywords as a description. For the remaining classes which feature around 50 keywords on average, we rely on the topological properties of the semantic network. [YANG et al., 2000] suggest to retain only the most frequently used terms in $K(k, t)$. Another possibility is to select 5 keywords based on their network centrality with the idea that very central keywords are the best candidates to describe the overall idea captured by a community. For example, the largest semantic class in 2003-2007 is characterized by the keywords : Support Packet; Tree Network; Network Wide; Voic Stream; Code Symbol Reader.

SIZE OF TECHNOLOGICAL AND SEMANTIC CLASSES We consider a specific window of observations (for example 2000-2004), and we define Z the number of patents which appeared during that time window. For each patent $i = 1, \dots, Z$ we associate a vector of proba-

bility where each component $p_{ij}^{(sem)} \in [0, 1]$, with $j = 1, \dots, N^{(sem)}$ and where

$$\sum_{j=1}^{N^{(sem)}} p_{ij}^{(sem)} = 1$$

(when there is no room for confusion, we drop the subscript t in $N_t^{(sem)}$). On average across all time windows, a patent is associated to 1.8 semantic classes with a positive probability. Next we define the size of a semantic class as

$$S_j^{(sem)} = \sum_{i=1}^Z p_{ij}^{(sem)}.$$

Correspondingly, we aim to provide a consistent definition for technological classes. For that purpose, we follow the so-called “fractional count” method, which was introduced by the USPTO and consists in dividing equally the patents between all the classes they belong to. Formally, we define the number of technological classes as $N^{(tec)}$ (which is not time dependent contrary to the semantic case) and for $j = 1, \dots, N^{(tec)}$ the corresponding matrix of probability is defined as

$$p_{ij}^{(tec)} = \frac{B_{ij}}{\sum_{k=1}^{N^{(tec)}} B_{ik}},$$

where B_{ij} equals 1 if the i th patent belongs to the j th technological class and 0 if not. When there is no room for confusion, we will drop the exponent part and write only p_{ij} when referring to either the technological or semantic matrix. Empirically, we find that both classes exhibit a similar hierarchical structure in the sense of a power-law type of distribution of class sizes as shown in Fig. ???. This feature is important, it suggests that a classification based on the text content of patents has some separating power in the sense that it does not divide up all the patents in one or two communities.

Potential Refinements of the Method

Our semantic classification method could be refined by combining it with other techniques such as Latent Dirichlet Allocation which is a widely used topic detection method (e.g. [BLEI, NG et JORDAN, 2003]), already used on patent data as in [KAPLAN et VAKILI, 2015] where it provides a measure of idea novelty and the counter-intuitive stylized facts that breakthrough invention are likely to come out of local search in a field rather than distant technological recombination. Using this approach should first help further evaluate the robustness of our qualitative conclusions (external validation). Also, depending on the level of orthogonality with our classification, it can potentially

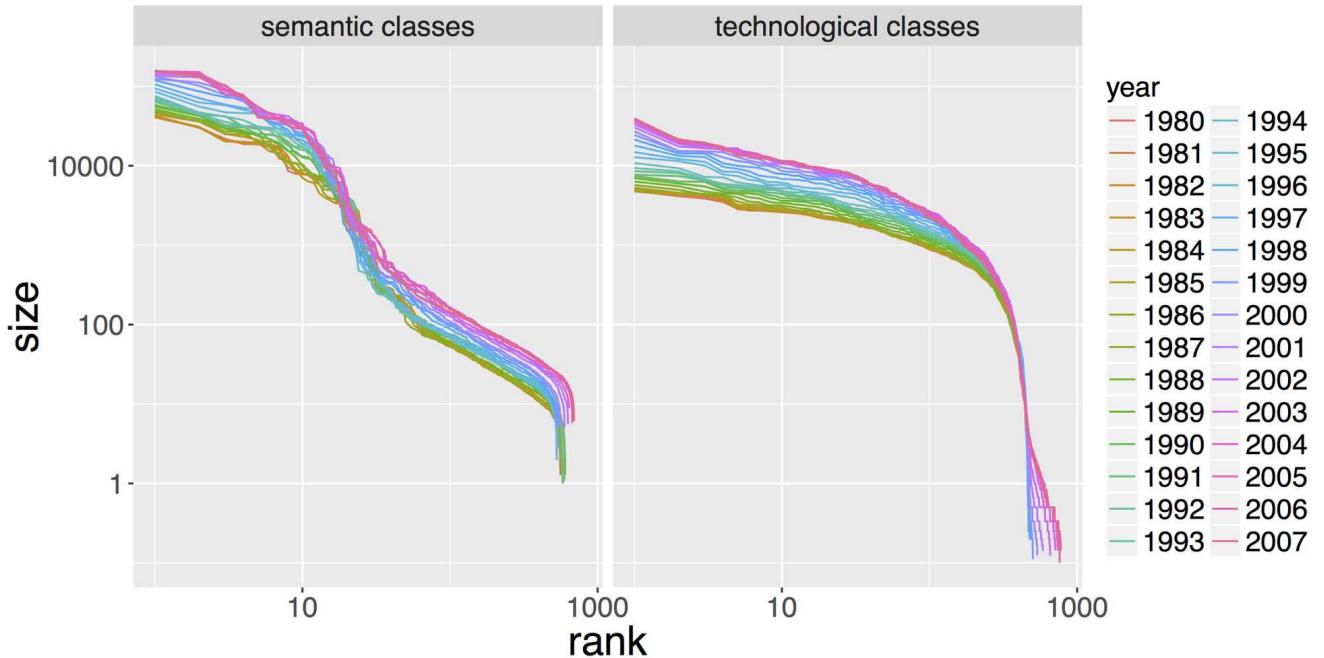


FIGURE 84: Sizes of classes. Yearly from $t = 1980$ to $t = 2007$, we plot the size of semantic classes (left-side) and technological classes (right-side) for the corresponding time window $[t - 4, t]$, from the biggest to the smallest. The formal definition of size can be found in Section . Each color corresponds to one specific year. Yearly semantic classes and technological classes present a similar hierarchical structure which confirms the comparability of the two classifications. This feature is crucial for the statistical analysis in Section . Over time, curves are translated and levels of hierarchy stays roughly constant.

bring an additional feature to characterize patents, in the spirit of multi-modeling techniques where neighbor models are combined to take advantage of each point of view on a system.

Our use of network analysis can also be extended using newly developed techniques of hyper-network analysis. Indeed, patents and keywords can for example be nodes of a bipartite network, or patents be links of an hyper-network, in the sense of multiple layers with different classification links and citation links. The combination of citation network modeling by Stochastic Block Modeling with topic modeling was studied for scientific papers by [ZHU et al., 2013b], outperforming previous link prediction algorithms. [IACOVACCI, Wu et BIANCONI, 2015] provide a method to compare macroscopic structures of the different layers in a multilayer network that could be applied as a refinement of the overlap, modularity and statistical modeling studied in this paper. Furthermore, it has recently been shown that measures of multilayer network projections induce a significant loss of information compared to the generalized corresponding measure [DE DOMENICO et al., 2015], which confirms the relevance of such development that we left for further research.

An other potential research development would be to further exploit the temporal structure of our dataset. Indeed, large progress have recently been made in complex network analysis of time-series data (see [GAO, SMALL et KURTHS, 2017] for a review). For example, [GAO et al., 2015] develops a method to construct multiscale network from time series, which could in our case be a solution to identify structures in patents trajectories at different levels, and be an alternative to the single scale modularity analysis we use.

Results

In this section, we present some key features of our resulting semantic classification showing both complementary and differences with the technological classification. We first present several measures derived from this semantic classification at the patent level : Diversity, Originality, Generality (Section) and Overlapping (Section). We then show that the two classifications show highly different topological measures and strong statistical evidence that they feature a different model (Sections and).

Patent Level Measures

Given a classification system (technological or semantic classes), and the associated probabilities p_{ij} for each patent i to belong to class j (that were defined in Section), one can define a patent-level diversity measure as one minus the Herfindhal concentration index on p_{ij} by

$$D_i^{(z)} = 1 - \sum_{j=1}^{N^{(z)}} p_{ij}^2, \text{ with } z \in \{\text{tec, sem}\}.$$

We show in Fig. ?? the distribution over time of semantic and technological diversity with the corresponding mean time-series. This is carried with two different settings, namely including/not including patents with zero diversity (i.e. single class patents). We call other patents “complicated patents” in the following. First of all, the presence of mass in small probabilities for semantic but not technological diversity confirms that the semantic classification contains patent spread over a larger number of classes. More interestingly, a general decrease of diversity for complicated patents, both for semantic and technological classification systems, can be interpreted as an increase in invention specialization. This is a well-known stylized fact as documented in [ARCHIBUGI et PIANTA, 1992]. Furthermore, a qualitative regime shift on semantic classification occurs around 1996. This can be seen whether or not we include patents with zero diversity. The diversity of complicated patents stabilizes after a constant decrease, and the overall diversity begins to strongly decrease. This means that on the one hand the number of single class patents begins to increase and on the other hand complicated patents do not change in diversity. It can be interpreted as a change in the regime of specialization, the new regime being caused by more single-class patents.

More commonly used in the literature are the measures of originality and generality. These measures follow the same idea than the above-defined diversity in quantifying the diversity of classes (whether technological or semantic) associated with a patent. But instead of looking at the patent’s classes, they consider the classes of the patents that are cited or citing. Formally, the originality O_i and the generality G_i of a patent i are defined as

$$O_i^{(z)} = 1 - \sum_{j=1}^{N^{(z)}} \left(\frac{\sum_{i' \in I_i} p_{i'j}}{\sum_{k=1}^{N^{(z)}} \sum_{i' \in I_i} p_{i'k}} \right)^2 \quad \text{and} \quad G_i^{(z)} = 1 - \sum_{j=1}^{N^{(z)}} \left(\frac{\sum_{i' \in \tilde{I}_i} p_{i'j}}{\sum_{k=1}^{N^{(z)}} \sum_{i' \in \tilde{I}_i} p_{i'k}} \right)^2,$$

where $z \in \{\text{tec, sem}\}$, I_i denotes the set of patents that are cited by the i th patent within a five year window (i.e. if the i th patent appears at year t , then we consider patents on $[t - T_0, t]$) when considering the originality and \tilde{I}_i the set of patents that cite patent i after less than five years (i.e. we consider patents on $[t, t + T_0]$) in the case of generality. Note that the measure of generality is forward looking in the sense that $G_i^{(z)}$ used information that will only be available 5 years after patent applications. Both measures are lower on average

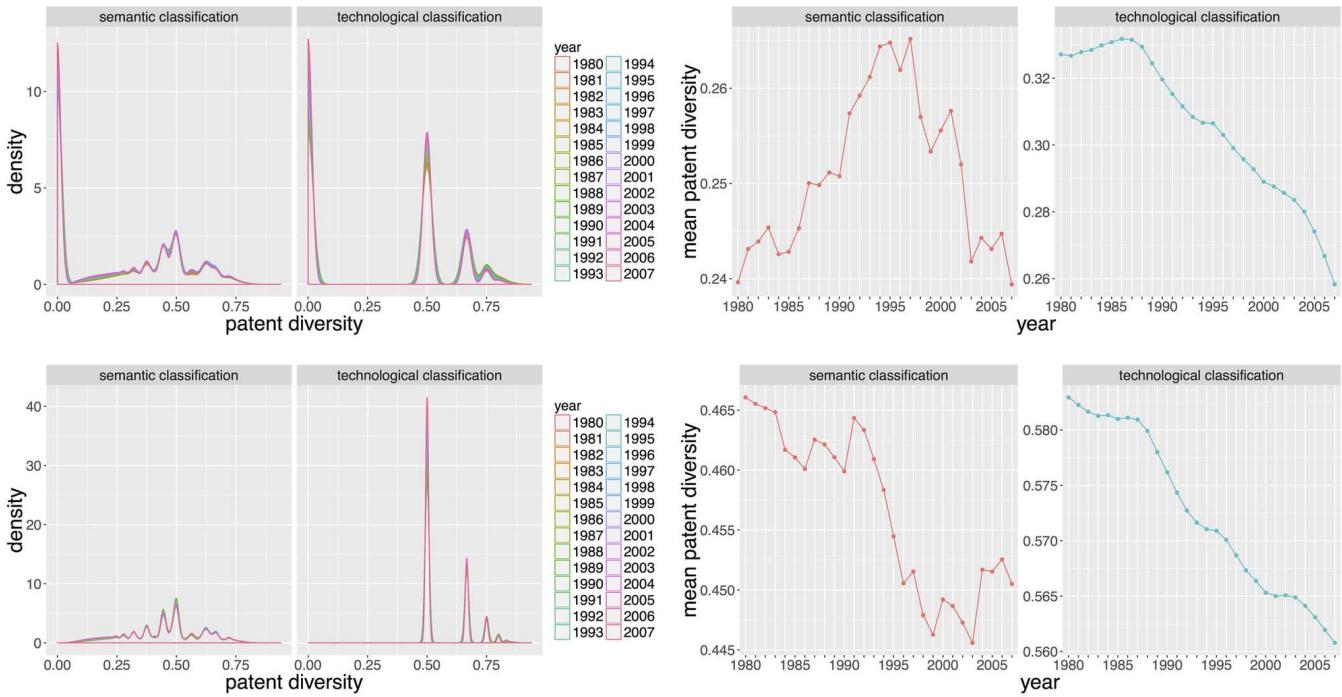


FIGURE 85: Patent level diversities. Distributions of diversities (Left column) and corresponding mean time-series (Right column) for $t = 1980$ to $t = 2007$ (with the corresponding time window $[t - 4, t]$). The first row includes all classified patents, whereas the second row includes only patents with more than one class (i.e. patents with diversity greater than 0).

based on semantic classification than on technological classification.
Fig. ?? plots the mean value of $O_i^{(sem)}$, $O_i^{(tec)}$, $G_i^{(sem)}$ and $G_i^{(tec)}$.

Classes overlaps

A proximity measure between two classes can be defined by their overlap in terms of patents. Such measures could for example be used to construct a metrics between semantic classes. Intuitively, highly overlapping classes are very close in terms of technological content and one can use them to measure distance between two firms in terms of technology as done in [BLOOM, SCHANKERMAN et REENEN, 2013]. Formally, recalling the definition of (p_{ij}) as the probability for the i th patent to belong to the j th class and N_p as the number of patents it writes

$$\text{Overlap}_{jk} = \frac{1}{N_p} \cdot \sum_{i=1}^{N_p} p_{ij} p_{ik}. \quad (30)$$

The overlap is normalized by patent count to account for the effect of corpus size : by convention, we assume the overlap to be maximal when there is only one class in the corpus. A corresponding relative

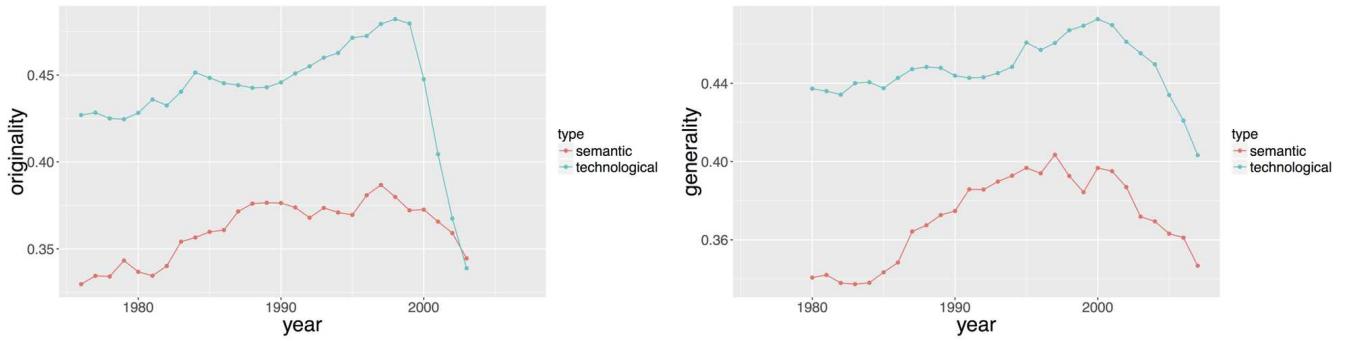


FIGURE 86: Patent level originality (left hand side) and generality (right hand side) for $t = 1980$ to $t = 2007$ (with the corresponding time window $[t - 4, t]$) as defined in subsection .

overlap is computed as a set similarity measure in the number of patents common to two classes A and B, given by $\sigma(A, B) = 2 \cdot \frac{|A \cap B|}{|A| + |B|}$.

INTRACLASSIFICATION OVERLAPS The study of distributions of overlaps inside each classification, i.e. between technological classes and between semantic classes separately, reveals the structural difference between the two classification methods, suggesting their complementary nature. Their evolution in time can furthermore give insights into trends of specialization. We show in Fig. ?? distributions and mean time-series of overlaps for the two classifications. The technological classification globally always follow a decreasing trend, corresponding to more and more isolated classes, i.e. specialized inventions, confirming the stylized fact obtained in previous subsection. For semantic classes, the dynamic is somehow more intriguing and supports the story of a qualitative regime shift suggested before. Although globally decreasing as technological overlap, normalized (resp. relative) mean overlap exhibits a peak (clearer for normalized overlap) culminating in 1996 (resp. 1999). Looking at normalized overlaps, classification structure was somewhat stable until 1990, then strongly increased to peak in 1996 and then decrease at a similar pace up to now. Technologies began to share more and more until a breakpoint when increasing isolation became the rule again. An evolutionary perspective on technological innovation [ZIMAN, 2003] could shed light on possible interpretations of this regime shift : as species evolve, the fitness landscape first would have been locally favorable to cross-insemination, until each fitness reaches a threshold above which auto-specialization becomes the optimal path. It is very comparable to the establishment of an ecological niche [HOLLAND, 2012], the strong interdependency originating here during the mutual insemination resulting in a highly path-dependent final situation.

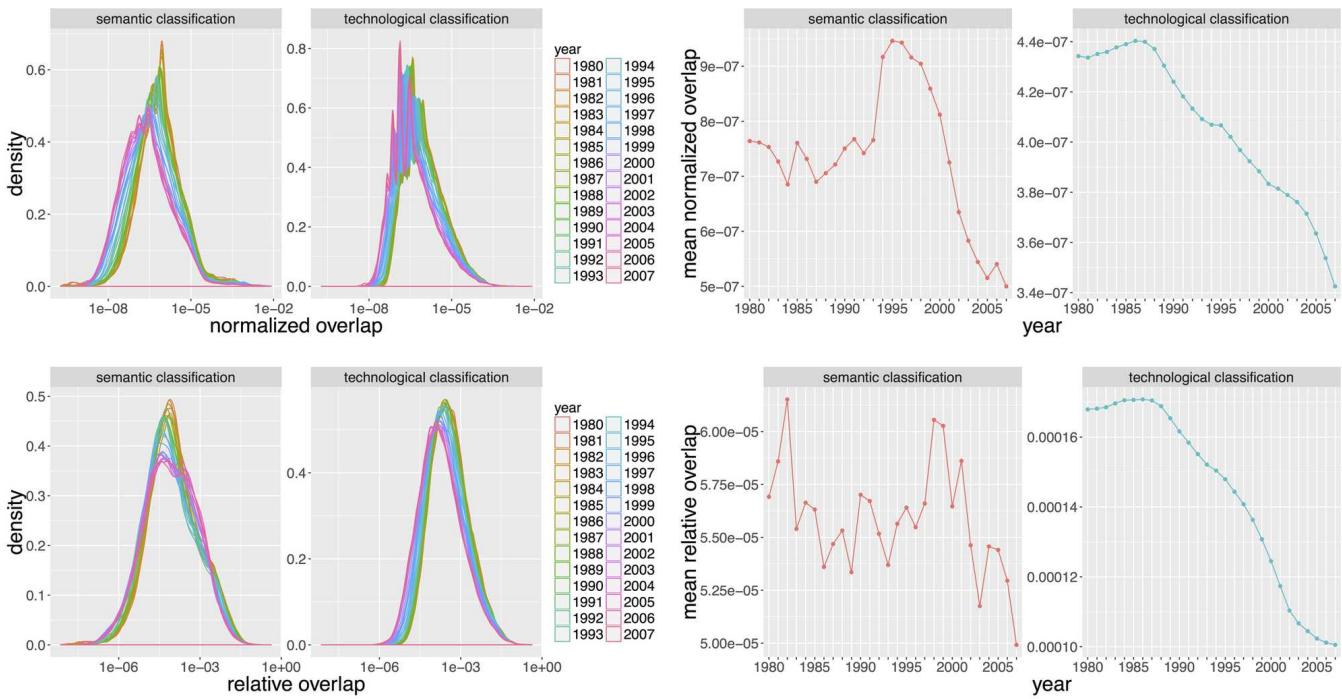


FIGURE 87: Intra-Classification overlaps. (Left column) Distribution of overlaps O_{ij} for all $i \neq j$ (zero values are removed because of the log-scale). Right column) Corresponding mean time-series. (First row) Normalized overlaps. (Second row) Relative overlaps.

INTER-CLASSIFICATION OVERLAPS Overlaps between classifications are defined as in (), but with j standing for the j th technological class and k for the k th semantic class : p_{ij} are technological probabilities and p_{ik} semantic probabilities. They describe the relative correspondence between the two classifications and are a good indicator to spot relative changes, as shown in Fig. ???. Mean inter-classification overlap clearly exhibits two linear trends, the first one being constant from 1980 to 1996, followed by a constant decrease. Although difficult to interpret directly, this stylized fact clearly unveils a change in the *nature* of inventions, or at least in the relation between content of inventions and technological classification. As the tipping point is at the same time as the ones observed in the previous section and since the two statistics are different, it is unlikely that this is a mere coincidence. Thus, these observations could be markers of a hidden underlying structural changes in processes.

Citation Modularity

An exogenous source of information on relevance of classifications is the citation network described in Section ???. The correspondence between citation links and classes should provide a measure of accuracy

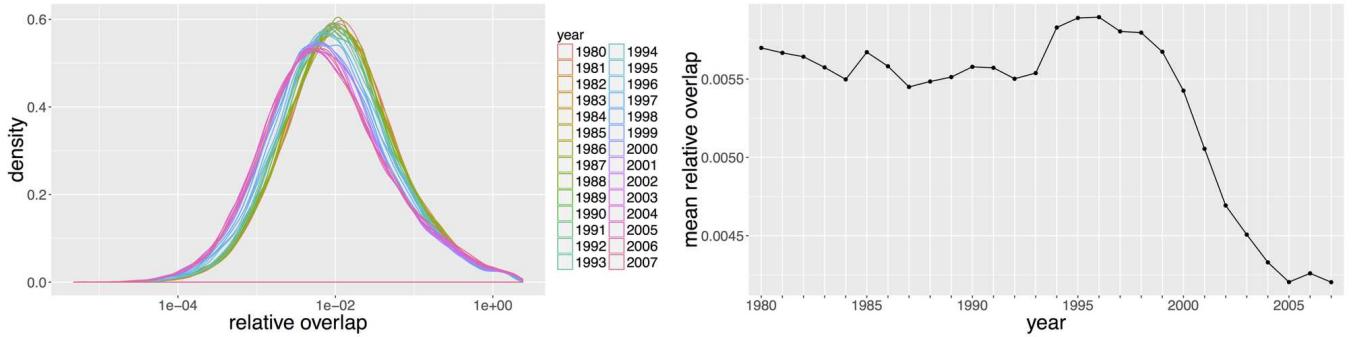


FIGURE 88: Distribution of relative overlaps between classifications. (Left) Distribution of overlaps at all time steps; (Right) Corresponding mean time-series. The decreasing trend starting around 1996 confirms a qualitative regime shift in that period.

of classifications, in the sense of an external validation since it is well-known that citation homophily is expected to be quite high (see, e.g, [ACEMOGLU et KERR, 2016]). This section studies empirically modularities of the citation network regarding the different classifications. To corroborate the obtained results, we propose to look at a more rigorous framework in Section . Modularity is a simple measure of how communities in a network are well clustered (see [CLAUSSET, NEWMAN et MOORE, 2004] for the accurate definition). Although initially designed for single-class classifications, this measure can be extended to the case where nodes can belong to several classes at the same time, in our case with different probabilities as introduced in [NICOSIA et al., 2009]. The simple directed modularity is given in our case by

$$Q_d^{(z)} = \frac{1}{N_P} \sum_{1 \leq i, j \leq N_P} \left[A_{ij} - \frac{k_i^{in} k_j^{out}}{N_P} \right] \delta(c_i, c_j),$$

with A_{ij} the citation adjacency matrix (i.e. $A_{ij} = 1$ if there is a citation from the i th patent to the j th patent, and $A_{ij} = 0$ if not), $k_i^{in} = |I_i|$ (resp. $k_i^{out} = |\tilde{I}_i|$) in-degree (resp. out-degree) of patents (i.e. the number of citations made by the i th patent to others and the number of citations received by the i th patent). Q_d can be defined for each of the two classification systems : $z \in \{\text{tec}, \text{sem}\}$. If $z = \text{tec}$, c_i is defined as the main patent class, which is taken as the first class whereas if $z = \text{sem}$, c_i is the class with the largest probability.

Multi-class modularity in turns is given by

$$Q_{ov}^{(z)} = \frac{1}{N_P} \sum_{c=1}^{N^{(z)}} \sum_{1 \leq i, j \leq N_P} \left[F(p_{ic}, p_{jc}) A_{ij} - \frac{\beta_{i,c}^{out} k_i^{out} \beta_{j,c}^{in} k_j^{in}}{N_P} \right],$$

where

$$\beta_{i,c}^{\text{out}} = \frac{1}{N_P} \sum_j F(p_{ic}, p_{jc}) \text{ and } \beta_{j,c}^{\text{in}} = \frac{1}{N_P} \sum_i F(p_{ic}, p_{jc}).$$

We take $F(p_{ic}, p_{jc}) = p_{ic} \cdot p_{jc}$ as suggested in [NICOSIA et al., 2009]. Modularity is an aggregated measure of how the network deviates from a null model where links would be randomly made according to node degree. In other words it captures the propensity for links to be inside the classes. Overlapping modularity naturally extends simple modularity by taking into account the fact that nodes can belong simultaneously to many classes. We document in Fig. ?? both simple and multi-class modularities over time. For simple modularity, $Q_d^{(\text{tec})}$ is low and stable across the years whereas $Q_d^{(\text{sem})}$ is slightly greater and increasing. These values are however low and suggest that single classes are not sufficient to capture citation homophily. Multi-class modularities tell a different story. First of all, both classification modularities have a clear increasing trend, meaning that they become more and more adequate with citation network. The specializations revealed by both patent level diversities and classes overlap is a candidate explanation for this growing modularities. Secondly, semantic modularity dominates technological modularity by an order of magnitude (e.g. 0.0094 for technological against 0.0853 for semantic in 2007) at each time. This discrepancy has a strong qualitative significance. Our semantic classification fits better the citation network when using multiple classes. As technologies can be seen as a combination of different components as shown by [YOUN et al., 2015], this heterogeneous nature is most likely better taken into account by our multi-class semantic classification.

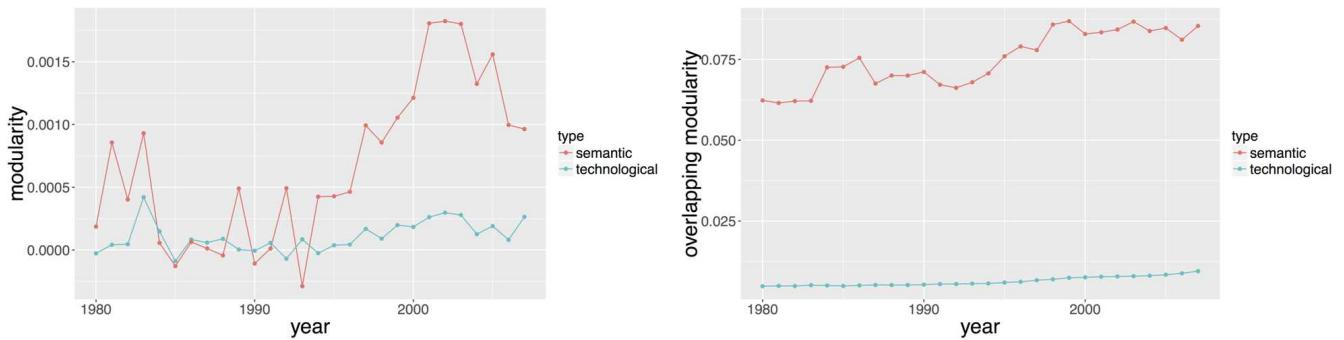


FIGURE 89: Temporal evolution of semantic and technological modularities of the citation network. (Left) Simple directed modularity, computed with patent main classes (main technological class and semantic class with larger probability). (Right) Multi-class modularity, computed following [NICOSIA et al., 2009]

Statistical Model

In this section, we develop a statistical model aimed at quantifying performance of both technological and semantic classification systems. In particular, we aim at corroborating findings obtained in Section . The mere difference between this approach and the citation modularity approach lies in the choice of the underlying model, and the according quantities of interest. In addition for the semantic approach, we want to see if when restricting to patents with higher probabilities to belong to a class, we obtain better results. To do that, we choose to look at within class citations proportion (for both technological and semantic approaches). We provide two obvious reasons why we choose this. First, the citations are commonly used as a proxy for performance as mentioned in Section . Second, this choice is “statistically fair” in the sense that both approaches have focused on various goals and not on maximizing directly the within class proportion. Nonetheless, the within class proportion is too sensitive to the distribution of the shape of classes. For example, a dataset where patents for each class account for 10% of the total number of patents will mechanically have a better within class proportion than if each class accounts for only 1%. Consequently, an adequate statistical model, which treats datasets fairly regardless of their distribution in classes, is needed. This effort ressembles to the previous study of citation modularity, but is complementary since the model presented here can be understood as an elementary model of citation network growth. Furthermore, the parameters fitted here can have a direct interpretation as a citation probability.

We need to introduce and recall some notations. We consider a specific window of observations $[t - T_0, t]$, and we define Z the number of patents which appeared during that time window. We let t_1, \dots, t_Z their corresponding appearance date by chronological order, which for simplicity are assumed to be such that $t_1 < \dots < t_Z$. For each patent $i = 1, \dots, Z$ we consider C_i the number of distinctive couples {cited patent, cited patent’s class} made by the i th patent (for instance if the i th patent has only made one citation and that the cited patent is associated with three classes, then $C_i = 3$). Let $z \in \{\text{tec}, \text{sem}\}$, we define $N_i^{(z)}$ the number of patents associated to at least one of the i th classes at time t_{i-1} . For $l = 1, \dots, C_i$ we consider the variables $B_{l,i}$, which equal 1 if the cited patent’s class is also common to the i th patent. We assume that $B_{l,i}$ are independent of each other and conditioned on the past follow Bernoulli variables

$$B\left(\min\left\{1, \frac{N_i^{(z)}}{i-1} + \theta^{(z)}\right\}\right),$$

where the parameter $0 \leq \theta^{(z)} \leq 1$ indicates the propensity for any patent to cite patents of its own technological or semantic class. When $\theta^{(z)} = 0$, the probability of citing patents from its own class is sim-

ply $N_i^{(z)}(i-1)^{-1}$, which corresponds to the observed proportion of patents which belong to at least one of the i th patent's classes. Thus this corresponds to the estimated probability of citing one patent if we assume that the probability of citing any patent $k = 1, \dots, i-1$ is uniformly distributed, which could be a reasonable assumption if classes were assigned randomly and independently from patent abstract contents. Conversely if $\theta^{(z)} = 1$, we are in the case of a model where there are 100% of within class citations. A reasonable choice of $\theta^{(z)}$ lies between those two extreme values. Finally, we assume that the number of distinctive couples C_i are a sequence of independent and identically distributed random variables following the discrete distribution C , and also independent from the other quantities.

We estimate $\theta^{(z)}$ via maximum likelihood, and obtain the corresponding maximum likelihood estimator (MLE) $\hat{\theta}^{(z)}$. The likelihood function, along with the standard deviation expression and details about the test, can be found in ???. The fitted values, standard errors and p-values corresponding to the statistical test $\theta^{(sem)} = \theta^{(tec)}$ (with corresponding alternative hypothesis $\theta^{(sem)} > \theta^{(tec)}$) on non-overlapping blocks from the period 1980-2007 are reported on Table 14. Note that the estimation included patents up until 2010 in the period 2006-2007 and not the patents from 1980 in the period 1980-1985 for homogeneity in size with other periods. This doesn't affect the significativity of the results. Semantic values are reported for four different chosen thresholds $p^- = .04, .06, .08, .1$. It means that we restricted to the couples (i th patent, j th class) such that $p_{ij} \geq p^-$.

The choice of considering non-overlapping blocks (instead of overlapping blocks) is merely statistical. Ultimately, our interest is in the significance of the test over the whole period 1980-2007. Thus, we want to compute a global p-value. This can be done considering the local p-values (by local, we mean for instance computed on the period 2001-2005) assuming independence between them. This assumption is reasonable only if the blocks are non-overlapping. All of this can be found in ???. Finally, note that from a statistical perspective, including overlapping blocks wouldn't yield more information.

The values reported in Table 14 are overwhelmingly against the null hypothesis. The global estimates of $\theta^{(sem)}$ are significantly bigger than the estimate of $\theta^{(tec)}$ for all the considered thresholds. Although the corresponding p-values (which are also very close to 0) are not reported, it is also quite clear that the bigger the threshold, the higher the corresponding $\theta^{(sem)}$ is estimated. This is consistently seen for any period, and significant for the global period. This seems to indicate that when restricting to the couples (patent, class) with high semantic probability, the propensity to cite patents from its own class $\theta^{(sem)}$ is increasing. We believe that this might provide extra information to patent officers when making their choice of citations. Indeed,

they could look first to patents which belong to the same semantic class, especially when patents have high probability semantic values.

Note that the introduced model can be seen as a simple model of citations network growth conditional to a classification, which can be expressed as a stochastic block model (e.g. [DECCELLE et al., 2011], [VALLES-CATALA et al., 2016]). The parameters are estimated computing the corresponding MLE. In view of [NEWMAN, 2016], this can be thought as equivalent to maximizing modularity measures.

Conclusion

The main contribution of this study was twofold. First we have defined how we built a network of patents based on a classification that uses semantic information from abstracts. We have shown that this classification share some similarities with the traditional technological classification, but also have distinct features. Second, we provide researchers with materials resulting from our analysis, which includes : (i) a database linking each patent with its set of semantic classes and the associated probabilities ; (ii) a list of these semantic classes with a description based on the most relevant keywords ; (iii) a list of patent with their topological properties in the semantic network (centrality, frequency, degree, etc.). The availability of this data suggests new avenues for further research. Linking our dataset with existing open ones can lead to various powerful developments. For example, using it together with the disambiguated inventor database provided by [LI et al., 2014] could be a way to study semantic profiles of inventors, or of cities as inventor addresses are provided. The investigation of spatial diffusion of innovation between cities, which is a key component of Pumain's Evolutive Urban Theory [PUMAIN, 2010], would be made possible.

A first potential application is to use the patents' topological measures inherited from their relevant keywords. The fact that these measures are backward-looking and immediately available after the publication of the patent information is an important asset. It would for example be very interesting to test their predicting power to assess the quality of an innovation, using the number of forward citations received by a patent, and subsequently the future effect on the firm's market value.

Regarding firm innovative strategy, a second extension could be to study trajectories of firms in the two networks : technological and semantic. Merging these information with data on the market value of firms can give a lot of insight about the more efficient innovative strategies, about the importance of technology convergence or about acquisition of small innovative firms. It will also allow to observe innovation pattern over a firm life cycle and how this differ across technology field.

A third extension would be to use dig further into the history of innovation. USPTO patent data have been digitized from the first patent in July 1790. However, not all of them contain a text that is directly exploitable. We consider that the quality of patent's images is good enough to rely on Optical Character Recognition techniques to retrieve plain text from at least 1920. With such data, we would be able to extend our analysis further back in time and to study how technological progress occurs and combines in time. [AKCIGIT, KERR et NICHOLAS, 2013] conduct a similar work by looking at recombination and apparition of technological subclasses. Using the fact that communities are constructed yearly, one can construct a measure of proximity between two successive classes. This could give clear view on how technologies converged over the year and when others became obsolete and replaced by new methods.

TABLE 14 : Estimated values of $\theta^{(tec)}$ and $\theta^{(sem)}$ and corresponding standard errors obtained from a Maximum Likelihood estimator as presented in section .

Approach	Estimated Value	st. er.	p-value
1980-1985 period			
technological	.664	.008	
semantic $p^- = .04$.741	.047	.053
semantic $p^- = .06$.799	.081	.049
semantic $p^- = .08$.828	.126	.097
semantic $p^- = .10$.834	.166	.153
1986-1990 period			
technological	.634	.007	
semantic $p^- = .04$.703	.022	.001
semantic $p^- = .06$.768	.040	.0004
semantic $p^- = .08$.804	.069	.007
semantic $p^- = .10$.832	.114	.041
1991-1995 period			
technological	.619	.006	
semantic $p^- = .04$.655	.009	.0004
semantic $p^- = .06$.713	.017	9e-08
semantic $p^- = .08$.731	.025	7e-06
semantic $p^- = .10$.750	.037	9e-06
1996-2000 period			
technological	.551	.003	
semantic $p^- = .04$.585	.002	≈ 0
semantic $p^- = .06$.638	.004	≈ 0
semantic $p^- = .08$.660	.006	≈ 0
semantic $p^- = .10$.686	.008	≈ 0
2001-2005 period			
technological	.567	.003	
semantic $p^- = .04$.621	.004	≈ 0
semantic $p^- = .06$.676	.007	≈ 0
semantic $p^- = .08$.701	.010	≈ 0
semantic $p^- = .10$.710	.013	≈ 0
2006-2007 period			
technological	.600	.007	
semantic $p^- = .04$.683	.016	1e-06
semantic $p^- = .06$.732	.025	2e-07
semantic $p^- = .08$.760	.036	6e-06
semantic $p^- = .10$.782	.048	9e-05
1980-2007 global period			
technological	.606	.002	
semantic $p^- = .04$.665	.009	8e-11
semantic $p^- = .06$.721	.017	9e-12
semantic $p^- = .08$.747	.025	9e-09
semantic $p^- = .10$.782	.035	2e-07

D

DONNÉES

when possible, specify data citation (ex. traffic data : Transportatio-nEquilibrium paper) ; try to put all on dataverse ; laius sur dataverse, partage des données etc.

Les données comme domaine de connaissance propre : décrire opération de collecte des données et de construction des jeux.

D.1 DONNÉES DE TRAFFIC DU GRAND PARIS

D.2 PRIX DE L'ESSENCE AUX ETATS-UNIS

D.3 RÉSEAU ROUTIER EUROPÉEN

D.4 INTERVIEWS

E

OUTILS

E.1 SOFTWARES AND PACKAGES

This appendix lists and describes the different open datasets created and used in the thesis.

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*

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 [REY-COYREHOURCQ, 2015]

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 Simple density urban growth model.

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

CARACTÉRISTIQUES

- Language : NetLogo then 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

Open for Discovery
- PLoS

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*

Documentation generation is central for reproducibility as it can automatize implementation description. NetLogo does not provide a documentation generator and we are thus currently writing a Doxygen wrapper for NetLogo code, that basically consists in transforming NetLogo code into Java code and parsing documentation comment blocks. An experimental version is available at <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Doc>.

E.3.2 *git comme outil de reproductibilité*

The use if git as a reproducibility and transparency tool was emphasized in [RAM, 2013] (for various reasons such as exact history tracing, easy cloning, past commit branching). It furthermore can help individual workflow for advantages such as automatic backup, organisation, experiments tracking. We use it actively and develop extensions for it.

E.3.3 *Vers un gestionnaire de métadonnées compatible avec git*

The issue of meta-data for figures is a crucial issue, as it is often difficult to keep a trace of all parameter values that have generated it, along with the corresponding code. Tricks may furthermore happen in script environments such as R or python when variables are accidentally modified without code modification. Keeping an exhaustive trace of the exact dataset, code and history that has generated a precise figure is a necessary condition for exact reproducibility. We are elaborating a git-compatible tool that would automatically handle these metadata, for example by branching and associating the unique commit hash to the figure. To become not an organizational burden nor a repository perturbation, we must still make some experiments. The final idea would be to have under each figure a unique identifier linking to the associated reproducing environment.

E.3.4 *TorPool*

TorPool is a java based Tor wrapper available with an api (currently only java, R version projected) at <https://github.com/JusteRaimbault/TorPool>. It allows among other purposes tricky data retrieval.



QUANTITATIVE ANALYSIS OF THESIS REFLEXIVITY

Quantitative Analysis of Thesis reflexivity

C : faire un graphe des concepts; compare to semantic network of concepts in Gödel Escher Bach.