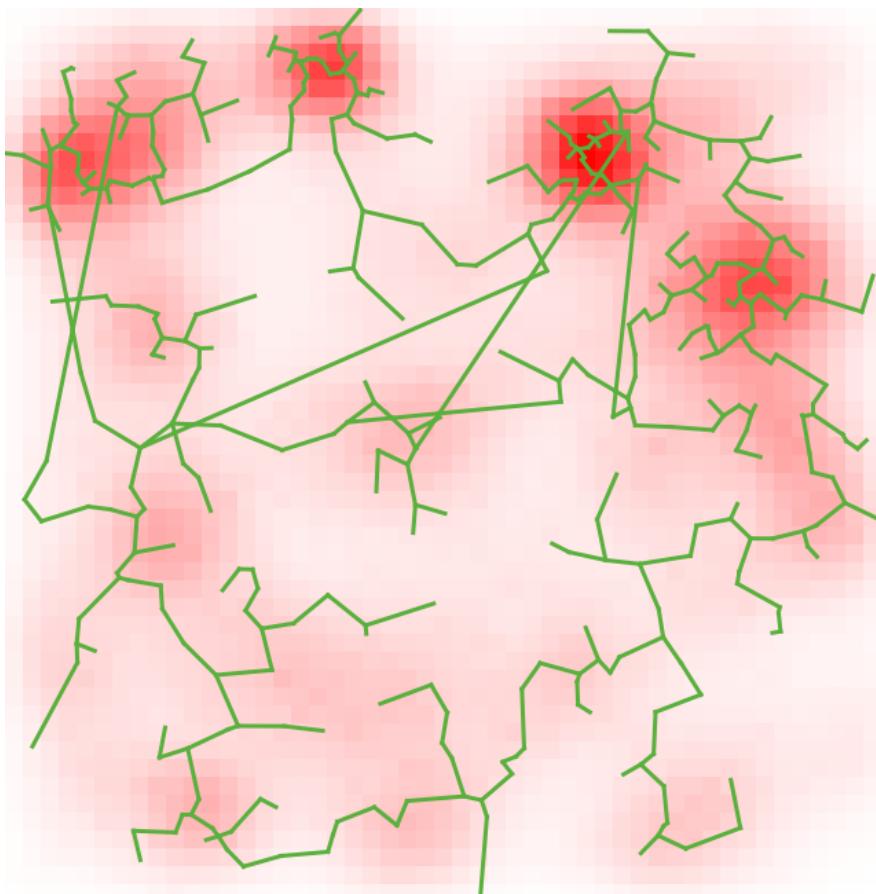


TOWARDS MODELS COUPLING URBAN GROWTH AND TRANSPORTATION NETWORK GROWTH

JUSTE RAIMBAULT



PhD Thesis First Year Preliminary Memoire

Under the supervision of Arnaud Banos and Florent Le Néchet

UMR CNRS 8504 Géographie-cités
and UMR-T 9403 LVMT

Université Paris VII

March 2016 – version 1.1

Juste Raimbault : *Towards Models Coupling Urban Growth and Transportation Network Growth*, PhD Thesis First Year Preliminary Memoire, © March 2016

ABSTRACT

READING NOTES

This provisory Memoire must be read as a work in progress, as it details progresses after one year of Doctorate. Many parts are given at the state of project, and not omitted as playing a role in the current research questioning. Its purpose is to set up a plan and examine the achieved work and corresponding directions, but also to share research ideas at this important step of one year.

PUBLICATIONS

Les travaux suivants contiennent une grande partie du contenu de cette thèse :

PUBLICATIONS

Publications

Antelope, C., Hubatsch, L., Raimbault, J., and Serna, J. M. (2016). An interdisciplinary approach to morphogenesis. *Forthcoming in Proceedings of Santa Fe Institute CSSS 2016*.

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

Raimbault, J. (2016). Investigating the Empirical Existence of Static User Equilibrium, *forthcoming in EWGT 2016 proceedings, Transportation Research Procedia*. arxiv :1608.05266

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

Communications

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*.

TABLE DES MATIÈRES

0.1	Interdisciplinarity	4
I	FOUNDATIONS	11
1	INTERACTIONS BETWEEN NETWORKS AND TERRITORIES	13
1.1	Territories and Networks	14
1.2	Modeling Interactions	20
1.3	Observation Flottante	23
1.4	Research Question	24
2	METHODOLOGICAL DEVELOPMENTS	25
2.1	Reproducibility	26
2.2	An unified framework for stochastic models of urban growth	30
2.3	Sensitivity of Urban Scaling Laws to Spatial Extent . .	33
2.4	Statistical Control on Initial Conditions by Synthetic Data Generation	35
2.5	Spatio-temporal Correlations	37
2.6	Generation of Correlated Synthetic Data	40
3	A DISCREPANCY-BASED FRAMEWORK TO COMPARE ROBUSTNESS BETWEEN MULTI-ATTRIBUTE EVALUATIONS	43
3.1	Introduction	43
3.2	Framework Description	46
3.3	Results	50
3.4	Discussion	55
4	QUANTITATIVE EPISTEMOLOGY	59
4.1	Algorithmic Systematic Review	60
4.2	Indirect bibliometrics through Complex Network analysis	66
4.3	Towards modeling purpose and context automatic extraction	71
II	MATERIALS	73
5	INVESTIGATING THE EMPIRICAL EXISTENCE OF STATIC USER EQUILIBRIUM	75
5.1	Introduction	75
5.2	Data collection	77
5.3	Methods and Results	78
5.4	Discussion	83
5.5	Conclusion	86
6	EMPIRICAL ANALYSIS : INSIGHTS FROM STYLIZED FACTS	89
6.1	Static correlations of urban form and network shape .	90
6.2	Disentangling co-evolutions from causal relations . .	96
6.3	Real Estate Trajectories	100

6.4	South-African historical events as instruments	102
7	MODELING	103
7.1	A simple model of urban growth	104
7.2	Correlated generation of territorial configurations . .	111
7.3	Network Growth Models	121
8	TOWARDS MORE COMPLEX MODELS	123
8.1	The Lutecia Model	123
III SYNTHESIS		131
9	A ROADMAP FOR AN OPERATIONAL FAMILY OF MODELS OF COEVOLUTION	133
9.1	Objectives	133
9.2	Case Studies	133
9.3	Roadmap	134
IV OPENING		135
10	THEORETICAL FRAMEWORK	137
10.1	Geographical Theoretical Context	138
10.2	A theoretical Framework for the Study of Socio-technical Systems	142
V APPENDIX		153
11	GENERATION OF CORRELATED SYNTHETIC DATA	155
12	AN INTERDISCIPLINARY APPROACH TO MORPHOGENESIS	163
13	TECHNICAL DEVELOPMENTS	165
13.1	Derivations for Urban Growth Models	165
13.2	Sensitivity of Urban Scaling	167
14	ARCHITECTURE AND SOURCES FOR ALGORITHMS AND MODELS OF SIMULATION	171
14.1	Algorithmic Systematic Review	171
14.2	Indirect Bibliometrics	172
14.3	Density Urban Growth	172
14.4	Correlated data generation	173
14.5	Lutecia Model	173
14.6	Network analysis	174
15	TOOLS AND WORKFLOW FOR AN OPEN REPRODUCIBLE RESEARCH	175
15.1	NetLogo documentation generator	175
15.2	git as a reproducibility tool	175
15.3	git-data	175
15.4	Towards a git-compatible figures metadata handler . .	176
15.5	TorPool	176

TABLE DES FIGURES

FIGURE 1	Reproducibility and visualization	28
FIGURE 2	53
FIGURE 3	54
FIGURE 4	Systematic review algorithm workflow	63
FIGURE 5	Convergence and sensitivity analysis of systematic review algorithm	64
FIGURE 6	Heterogeneous Bibliographical Data Collection	67
FIGURE 7	Properties of the citation network	69
FIGURE 8	Semantic network of concepts in quantitative geography	70
FIGURE 9	79
FIGURE 10	80
FIGURE 11	81
FIGURE 12	82
FIGURE 13	84
FIGURE 14	Empirical Distribution of Morphological Indicators	91
FIGURE 15	Geographical Distribution of Morphologies . .	92
FIGURE 16	Clustering Analysis of Morphologies	93
FIGURE 17	Typology of Real Estate trajectories	101
FIGURE 18	Generated Density urban shapes	106
FIGURE 19	LHS exploration of density model	107
FIGURE 20	PSE exploration	108
FIGURE 21	Precise calibration of the model	109
FIGURE 22	Exploration of feasible space for correlations between urban morphology and network structure	115
FIGURE 23	Examples of generated coupled configurations	116
FIGURE 24	Biological Network Growth	121
FIGURE 25	Examples of final configurations	130
FIGURE 26	Validation of network exploration heuristic . .	130
FIGURE 27	Example of the multi-scalar structure of the signal	157
FIGURE 28	Effective correlations obtained on synthetic data	160
FIGURE 29	Performance of a predictive model as a function of simulated correlations	161
FIGURE 30	Synthetic density distribution	169

LISTE DES TABLEAUX

TABLE 1	52
TABLE 2	65
Stationary lexical proximities	

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, seront toujours extrêmement complexes, 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. Yuangsi 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, petite échelle de temps, 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’aggregation 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 des faiblesses structurelles du réseau parisien de transports en commun [gleyze2005vulnerabilite] ; le 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 corres-

pondant à un corpus de connaissances scientifiques dans des disciplines diverses comme la géographie, l'économie urbains, les transports. Cette anecdote amusante 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 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.

ON GENERAL POSITIONING

De la position générale

L'ambition de cette thèse est de ne pas avoir d'ambition. 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.

SCIENTIFIC CONTEXT : COMPLEXITY HAS COME OF AGE

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, ce qui implique une mutation épistémologique pour abandonner le réductionnisme strict qui a échoué dans la majorité de ses tentatives de synthèse [anderson1972more]. Arthur a rappelé récemment [arthur2015complexity] qu'une mutation des méthodes et paradigmes en était également un enjeu, de par la place grandissante prise par les approches computationnelles qui remplacent les résolutions purement analytiques généralement limité 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 (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 sociaux¹. Elles sont au contraire moins connues et acceptées en sciences "dures" comme la physique : LAUGHLIN développe dans [laughlin2006different] une vision de la discipline à la même position de "frontière des connaissances" que d'autre champs pouvant paraître moins matures. La plupart des connaissances actuelles concerne des structures classiques simples, alors qu'un grand nombre de système présentent des propriétés *d'auto-organisation*, au sens où les lois macroscopiques ne sont pas suffisantes pour inférer les propriétés macroscopiques du système à moins que son évolution soit entièrement simulée (plus précisément cette vision peut être prise comme une définition de l'émergence sur laquelle nous reviendrons par la suite, or des propriétés auto-organisées sont par nature émergentes). Cela correspond au premier cauchemar du Démon de Laplace développé dans [deffuant2015visions].

A la croisée de positionnements épistémologiques, de méthodes et de champs d'application, les *Sciences de la complexité* se concentrent sur l'importance de l'émergence et de l'auto-organisation dans la plupart des phénomènes réel, ce qui les place plus proche de la frontière des connaissances que ce que l'on peut penser pour des disciplines classiques (LAUGHLIN, op. cit.). Ces concepts ne sont pas récents et avaient déjà été mis en valeur par ANDERSON [anderson1972more]. On peut aussi interpréter la Cybernétique comme un précurseur des Sciences de la Complexité en la lisant comme un pont entre technologie et sciences cognitives [wiener1948cybernetics]. Plus tard, la Synergétique [haken1980synergetics] a posé les bases d'approches théoriques des phénomènes collectifs en physique. Les causes possibles de

¹ 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 [prigogine1997end].

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 [[varenne2010simulations](#)]. Elles peuvent aussi être à chercher auprès de progrès en épistémologie : introduction de la notion de perspectivisme [[giere2010scientific](#)], reflexions plus fine autour de la nature des modèles [[varenne2013modeliser](#)]². Les potentialités théoriques et empiriques de telles approches jouent nécessairement un rôle dans leur succès³, comme le confirme les domaines très variés d'application (voir [[newman2011complex](#)] pour une revue très générale), comme par exemple la Science de Réseaux [[barabasi2002linked](#)] ; les Neurosciences [[koch1999complexity](#)] ; les Sciences Sociales ; la Géographie [[manson2001simplifying](#)][[pumain1997pour](#)] ; la Finance avec les approches éconophysiques [[stanley1999econophysics](#)] ; l'Ecologie [[grimm2005pattern](#)]. La Feuille de Route des Systèmes Complexes [[2009arXiv0907.2221B](#)] propose une double lecture des travaux en Complexité : une approche horizontale faisant la connexion entre champs d'étude par des questions transversales sur les fondations théoriques de la complexité et des faits stylisés empiriques communs, et une approche verticale, dans le but de construire des disciplines intégrées et les modèles multi-scalaires hétérogènes correspondants. L'interdisciplinarité est ainsi cruciale pour notre contexte scientifique.

0.1 INTERDISCIPLINARITY

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, mais en *Science des Systèmes Complexes* qui se réclame disciplines propre comme le propose PAUL BOURGINE. Ce n'est pas sans risques d'être lu avec méfiance voir défiance par les tenants des disciplines classiques, comme des exemples récents de malentendus ou conflits ont récemment illustré [[dupuy2015sciences](#)]. Le positionnement de BATTY lorsqu'il propose *Une Nouvelle Science des Villes* [[batty2013new](#)] (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.

L'évolution scientifique des sciences de la complexité, qui est vue par certains comme une révolution [[colander2003complexity](#)], ou même

² dans ce cadre, les progrès scientifiques et épistémologiques ne peuvent pas être dissociés et peuvent être vus comme étant en co-évolution

³ même si l'adoption de nouvelles pratiques scientifiques est souvent largement biaisé par l'imitation et le manque d'originalité [[dirk1999measure](#)], ou de façon plus ambiguë, par des stratégies de positionnement puisque le combat pour les fonds est un obstacle croissant à une recherche saine [[bollen2014funding](#)].

comme *un nouveau type de science*, pourrait affronter des difficultés intrinsèques dues aux comportements et a-priori des chercheurs en tant qu'être humains. Plus précisément, le besoin d'interdisciplinarité qui fait la force des Sciences de la Complexité pourrait devenir une de ses grandes faiblesses, puisque la structure fortement en silo de la science peut avoir des impacts négatifs sur les initiatives impliquant des disciplines variées. Nous n'évoquons pas les problèmes de sur-publication, quantification, compétition, qui sont plus liés à des questions de Science Ouverte et de son éthique, tout aussi de grande importance mais d'une autre nature. Cette barrière qui nous hante et que nous pourrions ne pas surmonter, a pour plus évident symptôme des *divergences culturelles disciplinaires*, et les conflits d'opinion en résultant. Ce drame du malentendu scientifique est d'autant plus grave qu'il peut en effet détruire totalement certains progrès en interprétant comme une falsification des travaux qui traitent une question toute différente. L'exemple récent d'un travail sur les inégalités liées aux hauts revenus présenté dans [[aghion2015innovation](#)], et dont les conclusions ont été commentées comme s'opposant aux thèses de Piketty dans [[piketty2013capital](#)], 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. D'où des conclusions divergentes sur le rôles des capitaux personnels dans une économie. Mais des *point de vus* 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 comment le couplage est opéré) et être une avancée scientifique. Cette expérience de pensée illustre les potentialités et la nécessité de l'interdisciplinarité. Dans une autre veine assez similaire, [[2017arXiv170105627H](#)] ré-analyse des données biologiques d'une expérience de 1943 qui prétendait confirmer l'hypothèse des processus d'évolution Darwiniens par rapport aux processus Lamarckiens, et montrent que les conclusions ne tiennent plus dans le contexte actuel d'analyse de données (avances énormes sur la théorie et les possibilités de traitement) et scientifique (avec d'autre nombreuses preuves de nos jours des processus Darwiniens) : c'est un bon exemple de malentendu sur le contexte, et comment 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.

PHYSICS REINVENTS GEOGRAPHY. Comme déjà mentionné, DUPUY et BENGUIGUI soulignent dans [dupuy2015sciences] le fait que les sciences urbaines ont récemment connu des conflits ouverts entre les tenants classiques des disciplines et des nouveaux arrivants, en particulier les physiciens. La disponibilité de grands jeux de données d'un nouveau type (réseaux sociaux, données des nouvelles technologies de la communication) ont attiré leur attention 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 impliquant que la connaissance existante n'est pas scientifique de par sa nature plus qualitative, ils n'ont aucunement révélé de connaissance nouvelle sur les systèmes urbains : pour citer quelques exemples, [barthelemy2013self] conclut que Paris a subit une transition pendant la période d'Haussman et ses opérations de planification globale, qui sont des faits naturellement connus depuis longtemps en Histoire Urbaine et Géographie Urbaine. [chen2009urban] redécouvre que le modèle gravitaire est amélioré par l'introduction de décalages dans les interactions et dérive analytiquement l'expression d'une force d'interaction entre les villes, sans aucun cadre théorique ni thématique. De tels exemples peuvent être multipliés, confirmant l'inconfort courant entre physiciens et géographes. Des bénéfices significatifs pourraient résulter d'une intégration raisonnée des disciplines [o2015physicists] mais la route semble être bien longue encore.

ECONOMIC GEOGRAPHY OR GEOGRAPHICAL ECONOMICS ? Des conflits similaires se rencontrent en économie : comme décrit par [marchionni2004geographical], la discipline de l'économie géographique, traditionnellement proche de la géographie, a fortement critiqué un nouveau courant de pensée nommé *économie géographisée*, dont le but est la spatialisation des techniques économiques classiques. Chacune n'ont pas les mêmes desseins et buts, et le conflit apparaît comme un malentendu complet vu d'un oeil extérieur.

AGENT-BASED MODELING IN ECONOMY Des conflits disciplinaires peuvent aussi se manifester sous la forme d'un rejet de méthodes nouvelles par les courants dominants. Suivant FARMER [farmer2009economy], l'échec opérationnel de la plupart des approches économiques classiques pourrait être compensé par un usage plus systématique de la modélisation et simulation basées agent. L'absence de cadre analy-

tique qui est naturelle pour l'étude de la plupart des systèmes complexes adaptatifs semble rebuter la plupart des économistes.

FINANCE En finance quantitative coexistent divers champs de recherche ayant très peu d'interactions entre eux. On peut considérer deux exemples. D'une part, les statistiques et l'économétrie sont extrêmement avancées en mathématiques théoriques, utilisant par exemple des méthodes de calcul stochastique et de théorie des probabilités pour obtenir des estimateurs très raffinés de paramètres pour un modèle donné (voir par exemple [[barndorff2011multivariate](#)]). D'autre part, l'éconophysique a pour but d'étudier des faits stylisés empiriques et inférer les lois correspondantes pour tenter d'expliquer les phénomènes liés à la complexité des marchés financiers [[stanley1999econophysics](#)], 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. En effet, le couplage d'approches hétérogènes à différents niveaux et échelles sera une clé de voute de cette thèse, la moelle épinière de la philosophie sous-jacente et une composante de la théorie qu'on construira.

COMPLEXITY IN GEOGRAPHY

Paradigmes de la Complexité en Géographie

Pour revenir à notre anecdote introductive, nous nous concentrerons sur l'étude d'un objet thématique qui sera les systèmes territoriaux. Plus généralement, il s'agit par commencer de brosser une revue du rôle de la complexité en géographie. Les géographes sont familiers avec la complexité depuis un certain temps, puisque l'étude des interactions spatiales est l'un de ses objets de prédilection. La variété de champs en géographie (géomorphologie, géographie physique, géographie environnementale, géographie 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 [[pumain2003approche](#)] une histoire subjective de l'émergence des paradigmes de la complexité en géographie.

La cybernétique a produit des théories des systèmes comme celle utilisée par Forrester. 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 [sanderson1992systeme] qui témoigne de l'application des concepts de la synergétique aux dynamiques des systèmes urbains. Enfin, les paradigmes actuels des systèmes complexes se sont introduits par plusieurs entrées. Par exemple, la nature fractale de la forme urbaine a été introduite par [batty1994fractal] et a eu de nombreuses applications jusqu'à des développements plus récents [keersmaecker2003using]. BATTY a aussi introduit les automates cellulaires en modélisation urbaine et propose une synthèse jointe avec les modèles basés agents et les fractales dans [batty2007cities]. 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 [sanderson1997sipop]) rentre dans le cadre théorique de [pumain1997pour]), cette théorie vise à comprendre les systèmes de villes comme des systèmes d'agents adaptatifs en co-évolution, aux interactions multiples, avec différents aspects mis en valeur comme l'importance de la diffusion des innovations. La série des modèles Simpop [pumain2012multi] a été conçue pour tester différentes hypothèses de la théorie. Par exemple, des processus sous-jacent différents ont été mis en évidence pour les systèmes de ville en Europe et aux Etats-unis [bretagnolle2010comparer]. A d'autres échelles de temps et dans d'autres contextes, le modèle SimpopLocal [schmitt2014modelisation] a pour but d'étudier les conditions pour l'émergence de systèmes urbains hiérarchiques à partir d'établissements disparates. Un modèle minimal (au sens de paramètres nécessaires et suffisants) a été isolé grâce à l'utilisation de calcul intensif via le logiciel d'exploration de modèles OpenMole [schmitt2014half], ce qui était un résultat impossible à atteindre de manière analytique pour un tel type de modèle complexe. Les progrès techniques d'OpenMole [reuillon2013openmole] ont été menés simultanément avec les avances théoriques et empiriques. Les avancées épistémologiques ont également été cruciales dans ce cadre, comme REY le développe dans [rey2015plateforme] et de nouveaux concepts comme la modélisation incrémentale [cottineau2015incremental] ont été découverts, avec de puissantes applications concrètes : [cottineau2014evolution] l'applique sur le système de villes soviétique et isole les processus socio-économiques dominants, par un test systématique des hypothèses thématiques et des fonctions d'implémentation. Des directions pour le développement de telles pratiques de Modélisation et Simulation en géographie quantitative ont récemment été introduits par BANOS dans [banos2013pour]. Il conclut par neuf principes⁴, parmi lesquels on peut citer l'importance de l'exploration intensive des mo-

⁴ Je me rappelle RENÉ DOURSAT insister pour la recherche du dernier commandement de BANOS

dèles computationnels et l'importance du couplage de modèles hétérogènes, qui sont avec d'autre principes tel la reproductibilité au centre de l'étude des systèmes complexes géographiques selon le point de vue décrit précédemment. Nous nous positionons dans l'héritage de cette ligne de recherche, travaillant de manière conjointe sur les aspects théoriques, empiriques, épistémologiques et de modélisation.

RESEARCH QUESTION

Question de Recherche

La question de recherche et les objets précis sont délibérément flous pour l'instant, puisque nous postulons que la construction d'une problématique ne peut être dissociée de la production d'une théorie correspondante. De manière réciproque, il n'y a aucun sens à poser des questions sorties de nulle part, sur des objets qui ont été seulement partiellement ou brièvement définis. Notre question préliminaire pour entrer dans le sujet, qu'on peut obtenir à partir de cas concrets comme l'anecdote introductory ou la revue de littérature préliminaire, est la suivante :

Comment définir les systèmes territoriaux, et les échelles et ontologies associées, dans une théorie cohérente, innovante et informative sur les processus sous-jacents ?

Il s'agit bien sûr d'une fausse question à ce stade, mais qui est toujours utile pour diriger la compréhension globale et le lecteur soucieux d'une démarche linéaire classique.

En effet, une caractéristique fondamentale des systèmes territoriaux est leur nature spatio-temporelle, qui est contenue dans leur dynamiques spatio-temporelles. La notion de *processus* au sens de [hypergeo] capture de plus les relations causales entre composantes de ces dynamiques, et est ainsi une approche intéressante pour une compréhension voire explication de ces systèmes. L'échelle doit être comprise ici au sens opérationnel (caractéristiques physiques) end l'*ontologie* comme les objets réels étudiés⁵. Notre question peut être vue grossièrement comme la recherche de théories et modèles qui révèlent des processus impliqués dans des systèmes complexes contenant aux moins des établissements humains, ce dernier point étant crucial pour la construction d'une problématique convergente plutôt que de se

⁵ 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 [livet2010], 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.

perdre dans des propositions irréalistes et non constructives qui pourrait aller de comprendre tout du cerveau (qui peut être vu comme une brique élémentaire des systèmes territoriaux qui émergent des interactions sociales) à l'écosphère qui inclut aussi les systèmes territoriaux.

CONTENTS

Contenu

This provisory Memoire is organized the following way. A first part with four chapters sets the thematic, theoretical and methodological background. The study of geographical systems implies, because of their complexity, a subtle combination of Theoretical constructions and Empirical Analysis, either in an inductive reasoning or in a didactic constitution of knowledge. The first part aims to approach our subject from the theoretical and methodological point of view, and rather as a *necessary foundation* shall be understood as a body of knowledge *coevolving* with Empirical and Modeling Parts. A linear reading is not necessarily the best way to deeply perceive the implications of theory on empirical and modeling experiments and reciprocally. Some methodological developments are necessary but explicit reference will be done when it will be the case. A first chapter starts from the provisory research question given above and frames from a thematic point of view geographical objects and processes to be studied, resulting in precise research questions. The scene is set up for the construction of our theoretical background in a second chapter, that consists in a geographical theory for territorial systems on the one hand and in an epistemological theory of socio-technical systems modeling that frames our approach at a meta-level. We then develop methodological considerations on diverse questions implied by theory and required for modeling. Finally, a chapter of quantitative epistemology finishes to pave the way for modeling directions, unveiling literature gaps precisely linked to our question. A second part develops results obtained from empirical analysis and modeling experiments, along with on-going and planned projects in these fields. It first present empirical analysis aimed at identifying stylized facts. Toy-models of urban growth are then proposed, followed by an example and propositions for more complex models. The third part constructs our research objective for the remaining part of our project and sets a corresponding roadmap. Appendices contain non-digest important parts of our work such as models implementation architecture and details and specific tools developed for a reproducible research workflow.

Première partie

FOUNDATIONS

This part set up foundations, constructing our research precise subject and questions from a thematic point of view, completed with a theoretical construction for framing at thematic and epistemological levels. We also provide methodological digressions, and a quantitative epistemological analysis completing the manual state of the art.

INTERACTIONS BETWEEN NETWORKS AND TERRITORIES

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.

- DENIS DIDEROT [diderot1965entretien]

Cette analogie est idéale pour introduire les notions de causalité et de processus dans les systèmes territoriaux. En voulant traiter naïvement des questions similaires à notre question de recherche préliminaire, certains ont qualifiés les causalités au sein de systèmes complexes comme un problème “de poule et œuf” : si un effet semble causer l'autre et réciproquement, comment est-il possible d'isoler les processus correspondants ? Cette vision est souvent présente dans les approches réductionnistes qui ne postulent pas une complexité intrinsèque au sein des systèmes étudiés. L'idée suggérée par DIDEROT est celle de *co-evolution* qui est un phénomène central dans les dynamiques évolutionnaires des Systèmes Complexes Adaptatifs comme HOLLAND élabore dans [holland2012signals]. Il fait le lien entre la notion d'émergence (ignorée dans les approches réductionnistes), en particulier l'émergence de structures à une plus grande échelle par les interactions entre agents à une échelle donnée, en général concrétisée par un système de limites, 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 leur interrelations et environnements générés conditionnés par le système de limites. Nous explorerons ces idées pour le cas des systèmes territoriaux par la suite.

Ce chapitre introductif est destiné à poser le cadre thématique, le contexte géographique sur lesquels les développements suivants se baseront. 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 4 tandis que le second sera traité plus tôt dans le chapitre 10), mais plutôt comme une construction narrative ayant pour but d'introduire nos objets et positions d'étude, afin de construire naturellement des questions de recherche précises.

1.1 TERRITORIES AND NETWORKS

Réseaux et Territoires

1.1.1 *Territories and Networks : There and Back Again*

HUMAN TERRITORIES Une entrée possible dans l'ensemble des objets géographiques que nous proposons d'étudier est la notion de territoire. En Ecologie, un territoire correspond à l'étendue spatiale occupée par un groupe d'agent ou plus généralement un écosystème. Les *Territoires Humains* sont extrêmement plus complexes de par l'importance de leur représentations sémiotiques, qui jouent un rôle significatifs dans l'émergence des constructions sociétales. Selon RAFFESTIN dans [raffestin1988reperes], la *Territorialité Humaine* est "la conjonction d'un processus territorial avec un processus informationnel", ce qui implique que l'occupation physique et l'exploitation de l'espace par les sociétés humaines n'est pas dissociable des représentations (cognitives et matérielles) de ces processus territoriaux, qui influent en retour 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 systèmes territoriaux, par exemple à travers la propagation d'informations et de représentations, par des processus politiques, ou encore par la correspondance effective entre territoire vécu et territoire perçu. 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), elle insiste sur leur rôle comme lieu de pouvoir et de création de richesse au travers des échanges. Mais la 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. En se restreignant à ce sous-système, il y a toutefois compatibilité entre la théorie de territoires de RAFFESTIN et la théorie évolutive des villes de PUMAIN [pumain2010theorie], 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. Les villes sont ainsi des agents compétitifs qui co-évoluent (au sens donné précédemment). 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. En effet, ce qui est considéré comme une ressource (naturelle ou artificielle) dépendra de la société

(par exemple de ses pratiques et de ses capacité technologiques). Un aspect central des établissements humains qui a une longue tradition d'étude en géographie, et qui est directement relié à la notion de territoire, est celui des *réseaux*. Nous allons voir comment le passage de l'un à l'autre est inévitable et leur définition indissociable.

A TERRITORIAL THEORY OF NETWORKS Nous paraphrasons DUPUY dans [dupuy1987vers] lorsqu'il propose des éléments pour une "théorie territoriale des réseaux" basée sur le cas concret d'un réseau de transport urbain. Cette théorie présente les *réseaux réels* (i.e. les réseaux concrets, incluant les réseaux de transport) comme 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 de "projets transactionnels" qui peuvent être compris comme des interactions potentielles entre les éléments du système territorial (agents et/ou ressources). 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. En tout temps des interactions potentielles ont existé¹ Le réseau d'interaction potentiel est concrétisé quand l'offre s'adapte à la demande, 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*. Un tel processus est loin d'être immédiat, et conduit à de forts effets de non-stationnarité et de dépendance au chemin : l'extension d'un réseau existant dépendra de la configuration précédente, et selon les échelles de temps impliquées, la logique et même la nature des opérateurs peut avoir évolué. RAFFESTIN souligne dans sa préface de [offner1996reseaux] qu'une théorie géographique articulant espaces, réseaux et territoires n'a jamais été formulée de manière 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 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). Selon le régime dans lequel le système considéré se trouve, le rôle respectif du réseau peut être radicalement différent. Selon DURANTON [duranton1999distance], les villes pré-industrielles étaient limitées en croissance de par les limitations des réseaux de transport. Les progrès technologiques ont

¹ même quand le nomadisme devait encore être la règle, des réseaux d'interactions potentielles dynamiques dans l'espace ont du exister, mais devaient avoir moins de chance de se matérialiser en des routes matérielles.

permis de les surmonter et à mené à la 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. Cette approche territoriale des réseaux semble naturelle en géographie, puisque les réseaux sont étudiés conjointement avec des objets géographiques auxquels est associée une théorie, en opposition à la science des réseaux qui étudie brutalement les réseaux spatiaux avec peu de fond thématique [**ducruet2014spatial**].

NETWORKS SHAPING TERRITORIES ? 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. Dans le cas des *réseaux techniques*, une autre désignation des réseaux réels donnée dans [**offner1996reseaux**], de nombreux exemples de tels rétroactions peuvent être mis en évidence : 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 est loin d'évident d'attribuer des mutations territoriales à une évolution du réseau and réciproquement la matérialisation d'un réseau à des dynamiques territoriales précises. Revenir à la citation de Diderot 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, mais comme des composantes fortement couplées 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 [**bonnafous1974detection**, **bonnafous1974methodologies**]. Il aura fallu un certain temps pour un positionnement critique sur l'usage non raisonné et decontextualisé de ces méthodes par les planificateurs et les politiques généralement pour justifier technocratiquement des projets de transports. Cela a été fait en premier par OFFNER dans [**offner1993effets**]. Récemment un édition spéciale du même journal sur ce débat [**espacegeo2014effets**] a rappelé d'une part que les mauvaises interprétations et les mauvais usages étaient encore largement présent aujourd'hui dans les milieux opérationnels de la planification comme [**crozet:halshs-01094554**] confirme, et d'autre part qu'il faudrait encore une certaine quantité de progrès scientifique pour comprendre en profondeur les relations entre réseaux et territoires. PUMAIN souligne que des travaux récents ont

révélé des effets systématiques sur de très longues échelles temporelles (comme e.g. le travail de BRETAGNOLLE sur l'évolution des chemins de fer, qui montre une sorte d'effet structurel sur la nécessité de connexion au réseau des villes, afin de rester actives, mais qui n'est ni suffisant ni totalement causal). A un niveau macroscopique des motifs typiques d'interaction émergent, mais les trajectoires microscopiques du systèmes sont essentiellement chaotiques : la compréhension des dynamiques couplées dépend fortement de l'échelle considérée. A une petite échelle il est peu raisonnable de vouloir montrer des comportement systématiques, comme le rappelle OFFNER. Par exemple, sur des territoires de montagne français comparables, [berne2008ouverture] montre que les réactions à un même contexte d'évolution du réseau de transport peut mener à des réactions territoriales très diverses, certains trouvant de forts bénéfices par la nouvelle connectivité, d'autres au contraire devenant plus fermés. Ces retroactions potentielles des réseaux sur les territoires n'agit pas nécessairement sur des composantes concrètes : CLAVAL montre dans [claval1987reseaux] que les réseaux de transport et de communication contribuent à la représentation collective d'un territoire en agissant sur un sentiment d'appartenance.

TERRITORIAL SYSTEMS Ce voyage des territoires aux réseaux, et retour, nous permet d'esquisser une définition préliminaire d'un système territorial sur laquelle se basera les considérations théoriques suivantes. Comme nous avons mis en exergue le rôle des réseaux, la définition se doit de les prendre en compte.

Définition provisoire. *Un Système Territorial est un territoire humain auquel peuvent être associés à la fois un réseau d'interactions et un réseau réel. Les réseaux réels 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.

1.1.2 *Transportation Networks*

THE PARTICULARITY OF TRANSPORTATION NETWORKS Déjà évoqués dans le cas des effets structurants des réseaux, les réseaux de transports jouent un rôle déterminant dans l'évolution des territoires. 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 ac-

tivités économiques), les réseaux de transport conditionnent d'autres types de réseaux (logistique, échanges commerciaux, interactions sociales concrètes pour donner quelques exemples) and semblent dominer dans les motifs d'évolution territoriale, en particulier dans nos sociétés contemporaines qui sont devenues dépendantes des réseaux de transport [bavoux2005geographie]. Le développement du réseau français à grande vitesse est une illustration pertinente 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 de lignes totalement nouvelles a été présenté comme central pour le développement [zem bri 1997 fondements]. Le manque d'intégration de ces nouveaux réseaux avec l'existant et avec les territoires locaux est à présent observé comme une faiblesse structurelle et des impacts négatifs sur certains territoires ont été prouvés [zem bri 2008 contribution]. Une revue faite dans [bazin 2011 grande] 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 comment les réseaux de transport peuvent avoir des effets à la fois directs et indirects sur les dynamiques territoriales. 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 mène pas au effets initialement désirés [es 119]. Des projets récents comme [l2012ville] ont tenté d'implémenter des idées similaires, mais il manque pour l'instant de recul pour juger de leur succès à produire un territoire effectivement intégré. 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 pour toutes ces raisons évoquées ici.

DECONSTRUCTING ACCESSIBILITY La notion d'accessibilité surgit rapidement lorsqu'on s'intéresse aux réseaux de transport. 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 spatiale² [bavoux2005geographie]. Cet objet est souvent utilisé comme un outil de planification ou comme une variable explicative de localisation des agents par exemple. 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 ter-

² et souvent généralisée comme une *accessibilité fonctionnelle*, par exemple les emplois accessibles aux actifs d'un lieu. Les potentiels d'interaction spatiaux s'exprimant dans les lois de gravité peuvent aussi être compris de cette façon.

ritoriales. La mystification de la notion de *mobilité* a été montrée par COMMENGES dans [commenges:tel-00923682], qui révèle que la majorité des débats sur la modélisation de la mobilité et les notions correspondantes était majoritairement construites de manière ad-hoc par les administrateurs de transports issus du *Corps des Ponts* qui importaient brutalement les outils et méthodes des Etats-Unis sans adaptation ni reflexion adaptée au contexte français. L'accessibilité pourrait 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* [confMangin], cette nouvelle infrastructure de transport métropolitaine

SCALES AND HIERARCHIES

INTERACTIONS BETWEEN TRANSPORTATION NETWORKS AND TERRITORY At this state of progress, we have naturally identified a research subject that seems to take a significant place in the complexity of territorial systems, that is the study of interactions between transportation networks and territories. In the frame of our preliminary definition of a territorial system, this question can be reformulated as the study of networked territorial systems with an emphasize on the role of transportation networks in system evolution processes.

1.2 MODELING INTERACTIONS

Modéliser les Interactions

1.2.1 *Modeling in Quantitative Geography*

La modélisation en Géographie Théorique et Quantitative (TQG), et plus généralement en Sciences Sociales, a une longue histoire dont nous ne pourrons que brosser un bref portrait ici. CUYALA procède dans [cuyala2014analyse] 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. 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 [rey2015plateforme] avec une attention particulière pour la notion de validation de modèles. L'utilisation de ressources de calcul pour la simulation de modèles est antérieur à l'introduction des paradigmes de la complexité, remontant à HÄGERSTRAND et FORRESTER, pionniers 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 modèles 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, voir [schmitt2014half] pour un exemple des possibilités offertes en terme de calibration et de validation de modèle, réduisant le temps de calcul nécessaire de 30 ans à une semaine), qui est également accompagnée par une évolution des pratiques [banos2013pour] et techniques [10.1371/journal.pone.0138212] de modélisation. La modélisation, et en particulier les modèles de simulation, est vue par beaucoup comme une brique fondamentale de la connaissance : [livet2010] rappelle la combinaison des domaines empirique, conceptuel (théorique) et de la modélisation, avec des rétroactions constructives entre chaque. Une modèle peut être un outil d'exploration pour tester des hypothèses, un outil empirique pour valider une théorie sur des jeux de données, un outil explicatif pour révéler des causalités et ainsi des processus internes au système, un outil constructif pour construire itérativement une théorie conjointement avec celle des modèles associés. Ce sont des exemples de fonctions parmi d'autres : Varenne

donne dans [varenne2010simulations] une classification raffinée des diverses fonctions d'un modèle. Nous considérons la modélisation comme un instrument fondamental de connaissance des processus au sein de systèmes complexes adaptatifs, et précisons encore notre question de recherche, qui s'intéressera aux *modèles impliquant des interactions réseaux et territoires*.

1.2.2 Modeling Territories and Networks

Land-Use Transportation Interaction Models

Une partie importante de la littérature proposant des modélisations des interactions entre réseaux et territoires se trouve dans le domaine de la planification urbaine, avec les *modèles d'interaction entre usage du sol et transport (LUTI)*. Ces travaux peuvent être difficiles à cerner car liés à différentes disciplines. Par exemple, du point de vue de l'Economie Urbaine, les propositions de modèle intégrés existent depuis un certain temps [putman1975urban]. La variété des modèles existants a conduit à des comparaisons opérationnelles [paulley1991overview, wegener1991one]. Plus récemment, les avantages respectifs des approches statiques et dynamiques a été étudié par [kryvobokov2013comparison].

Dans tous les cas, ce type de modèle opère généralement à des échelles temporelles et spatiales relativement faibles. [wegener2004land] 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. Le positionnement théorique est plutôt proche des disciplines de l'Economie, de la Planification et de la Sociologie, et relativement de nos raisonnements géographiques qui se veulent de comprendre également des processus sur le temps long. Pas moins de dix-sept modèles sont comparés et classifiés, parmi lesquels aucun n'inclut une évolution endogène du réseau de transport sur les échelles de temps relativement petites des simulations. Une revue complémentaire est faite par [chang2006models], élargissant le contexte avec l'inclusion de classes plus générales de modèles, comme des modèles d'interactions spatiales (parmi lesquels l'attribution du traffic et les modèles à quatre temps), les modèles de planification basés sur la recherche opérationnelle (optimisation des localisations), les modèles microscopiques d'utilité aléatoire, et les modèles de marché foncier. Toutes ces techniques opèrent également à une petite échelle et considèrent au plus l'évolution de l'usage du sol. [iacono2008models] couvre un horizon similaire avec une emphase supplémentaire sur les modèles à automates cellulaires d'évolution d'usage du sol et les modèles basés agent. Les modèles LUTI sont toujours largement étudiés et appliqués, comme par exemple [delons:hal-00319087] qui est utilisé pour la région métropolitaine parisienne. La courte portée 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.

Network Growth

Hybrid Modeling

Models of simulation implementing a coupled dynamic between urban growth and transportation network growth are relatively rare, and always rather poor from a theoretical and thematic point of view. A generalization of the geometrical local optimization model described before was developed in [barthelemy2009co]. As for the road growth model of which it is an extension, no thematic nor theoretical justification of local mechanisms is provided, and the model is furthermore not explored and no geographical knowledge can be drawn from it. [levinson2007co] adopts a more interesting economic approach, similar to a four step model (gravity-based origin-destination flows generation, stochastic user equilibrium traffic assignment) including travel cost and congestion, coupled with a road investment module simulating toll revenues for constructing agents, and a land-use evolution module updating actives and employments through discrete choice modeling. The experiments showed that co-evolving network and land uses lead to positive feedbacks reinforcing hierarchy, but are far from satisfying for two reasons : first network topology does not really evolve as only capacities and flows change within the network, what means that more complex mechanisms on longer time scales are not taken into account, and secondly the conclusions are very limited as model behavior is not known since sensitivity analysis is done on few one-dimensional spaces : exhaustive mechanisms stay thus unrevealed as only particular cases are described in the sensitivity analysis. From an other point of view, [levinson2005paving] is also presented as a model of co-evolution, but corresponds more to coupled statistical analysis as it relies on a Markov-chain predictive model. [rui2011urban] gives a model in which coupling between land-use and network growth is done in a weak paradigm, land-use and accessibility having no feedback on network topology evolution. [achibet2014model] describes a co-evolution model at a very small scale (scale of the building), in which evolution of both network and buildings are ruled by a same agent (influenced differently by network topology and population density) what implies a too strong simplification of underlying processes. Finally, a simple hybrid model explored and applied to a toy planning example in [raimbault2014hybrid], relies on urban activities accessibility mechanisms for settlement growth with a network adapting to urban shape. The rules for network growth are too simple to capture processes we are interested in, but the model produces at a small scale a broad range of urban shapes reproducing typical patterns of human settlements.

Urban Systems Modeling

An approach closer to our current questioning is the one of integrated modeling of system of cities. In the continuity of Simpop models for city systems modeling, SCHMITT described in [schmitt2014modelisation] the SimpopNet model which aim was precisely to integrate co-evolution processes in system of cities on long time scales, typically rules for hierarchical network development as a function of cities dynamics coupled with city dynamics depending on network topology. Unfortunately the model was not explored nor further studied, and furthermore stayed at a toy-level. COTTINEAU proposed transportation network endogenous growth as the last building bricks of her Marius productions but it stayed at a conceptual construction stage. We shall position more in that stream of research in this thesis.

1.2.3 *Sketch of a Modelography*

An ongoing work is the production of a synthesis of this overview, from a modular modeling point of view, combined with a purpose and scale classification. Already mentioned, modular modeling consists in the integration of heterogeneous processes and implementation of processes in order to extract the set of mechanisms giving the best fit to empirical data [cottineau2015incremental]. We can thus classify models described here according to their building bricks in terms of processes implemented and thus identify possible coupling potentialities. This work is a preliminary step for the analysis in quantitative epistemology developed in chapter 4.

1.3 ON QUALITATIVE RESEARCH : AN EXPERIMENT IN FLOATING OBSERVATION

De la Recherche Qualitative : une Experience en Observation Flottante

Si le diable est dans les détails, les systèmes de transport entre autres sont l'allégorie de cette adage. Ce que certains appellent détail contient la majorité de l'information pour d'autres. Logiquement enfermés dans une bulle scientifique, malgré toutes les volontés développées en introduction, on tâchera de rester conscient de la nature et la portée de la connaissance produite ici. Ce que nous pourrions appeler détail, lors de l'étude de l'accessibilité d'un réseau de transport par exemple, tel des impressions ressenties par les usagers ou les relations sociales induites par les situations découlant des dynamiques du systèmes, seront le centre du questionnement pour un anthropologue ou sociologue. Une telle connaissance, qui trouverait certainement une place dans nos problématiques, est hors de notre portée de par l'absence de *terrain* de longue durée.

1.4 RESEARCH QUESTION

Question de Recherche

To close this thematic touring introducing chapter, we can state a general research question that frames our further theoretical constructions and first modeling attempts. It is roughly the same as the problematic given at the end of previous section, but adding the insight of modeling as the approach to understand these complex systems.

General research Question. *To what extent a modeling approach to territorial systems as networked human territories can help disentangling complexly involved processes ?*

This question will be refined by theoretical developments in the next chapter and experiments in the followings.

2

METHODOLOGICAL DEVELOPMENTS

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

- MARC BARTHÉLÉMY

Such a shocking phrase was pronounced during the introduction of a *Network* course for students of Complex System Science. Besides the fact that the spirit of CSS 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 [[2009arXiv0907.2221B](#)], it reveals how methodological considerations shape the perceptions of disciplines. From a background in Physics, "rigorous" implies the use of tools and methods judged more rigorous (analytical derivations, large datasets statistics, etc.). But what is rigorous for someone will not be for an other discipline¹, depending on the purpose of each piece of research (perspectivism [[giere2010scientific](#)] 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.

We first propose a kind of essay insisting on the importance of reproducibility in science. More than a guideline, it is a way to practice science that a necessary condition for its rigor. Any non-reproducible work is not scientific. We then derive technical results on models of urban growth and on the sensitivity of scaling laws, that are both recurrent themes in the modeling of complex urban systems. We then introduce a method in the context of systematic model exploration and model behavior. We finally work on a link between static and dynamic correlations in a geographical system. This chapter is rather heterocline as sections may correspond to a particular technical need at a point in the thesis, to global methodological directions, or global research directions.

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 envelope".

2.1 REPRODUCIBILITY

Reproducibilité

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

As scientific reproducibility is an essential requirement for any study, its practice seems to be increasing [[stodden2010scientific](#)] and technical means to achieve it are always more developed (as e.g. ways to make data openly available, or to be transparent on the research process such as git [[ram2013git](#)], or to integrate document creation and data analysis such as knitr [[xie2013knitr](#)]), at least in the field of numerical modeling and simulation. However, the devil is indeed in the details and obstacles judged at first sight as minor become rapidly a burden for reproducing and using results obtained in some previous researches. We describe two cases studies where models of simulation are apparently highly reproducible but unveil as puzzles on which research-time balance is significantly under zero, in the sense that trying to exploit their results may cost more time than developing from scratch similar models.

2.1.1 *On the Need to Explicit the Model*

A current myth is that providing entire source code and data will be a sufficient condition for reproducibility. It will work if the objective is to produce exactly same plots or statistical analysis, assuming that code provided is the one which was indeed used to produce the given results. It is however not the nature of reproducibility. First, results must be as much implementation-independent as possible for clear robustness purposes. Then, in relation with the precedent point, one of the purposes of reproducibility is the reuse of methods or results as basis or modules for further research (what includes implementation in another language or adaptation of the method), in the sense that reproducibility is not replicability as it must be adaptable [[drummond2009replicability](#)].

Our first case study fits exactly that scheme, as it was undoubtedly aimed to be shared with and used by the community since it is a model of simulation provided with the Agent-Based simulation platform NetLogo [[wilensky1999netlogo](#)]. The model is also available online [[de2007netlogo](#)] and is presented as a tool to simulate socio-economic dynamics of low-income residents in a city based on a synthetic urban environment, generated to be close in stylized facts from

the real town of Tijuana, Mexico. Beside providing the source code, the model appears to be poorly documented in the literature or in comments and description of the implementation. Comments made thereafter are based on the study of the urban morphogenesis part of the model (setup for the “residential dynamics” component) as it is our global context of study [**raimbault2014vers**]. In the frame of that study, source code was modified and commented, which last version is available on the repository of the project².

RIGOROUS FORMALIZATION An obvious part of model construction is its rigorous formalization in a formal framework distinct from source code. There is of course no universal language to formulate it [**banos2013pour**], and many possibilities are offered by various fields (e.g. UML, DEVS, pure mathematical formulation). No paper nor documentation is provided with the model, apart from the embedded NetLogo documentation since it only thematically describes in natural language the ideas behind each step without developing more and provides information about role of different elements of the interface.

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

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

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

² at <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Reproduction/UrbanSuite>



FIGURE 1 : Example of simple improvement in visualization that can help understanding mechanisms implied in the model. *Left* : example of original output; *Middle* : visualization of main roads (in red) and underlying patches attribution, suggesting possible implementation bias in the use of discretized trace of roads to track their positions; *Right* : Visualization of land values using a more readable color gradient. This step confirms the hypothesis, through the form of value distribution, that the morphogenesis step is an unnecessary detour to generate a random field for which simple diffusion method should provide similar results, as detailed in the paragraph on implementation.

EXPECTED MODEL BEHAVIOR Whatever the definition, a model can not be reduced to its formulation and/or implementation, as expected model behavior or model usage can be viewed as being part of the model itself. In the frame of GIERE's perspectivism [giere2010scientific], the definition of model includes the purpose of use but also the agent who aims to use it. Therefore a minimal explication of model behavior and exploration of parameter roles is highly advised to decrease chances of misuses or misinterpretations of it. It includes simple runtime charts that are immediate on the NetLogo platform, but also indicators computations to evaluate outputs of the model. It can also be improved visualizations during runtime and model exploration, such as showed in Fig. 1.

2.1.2 *On the Need of Exactitude in Model Implementation*

Possible divergences between model description in a paper and the effectively implemented processes may have grave consequences on the reproducibility of science. The road network growth model given in [barthelemy2008modeling] is one example that we are currently investigating. A strict implementation of model mechanisms provide slightly different results than the one presented in the paper, and as source code is not provided we need to test different hypotheses on possible mechanisms added by the programmer (that seems to be a connexion rule to intersections under a certain distance threshold). Lessons that could be possibly drawn from this examples are

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

ral specifications) in order to identify possible implementation biaises

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

The last point, if first not provided, may ensure a limited risk of scientific falsification as it may be more complicated to fake false exploration results than to effectively explore the model. A joint project currently done is the writing of a false modeling paper in the spirit of [zilsel2015canular], in which opposite results to the effective results of a given model are provided, without providing model implementation. A first bunch of test is the acceptance of a clearly non-reproducible paper in diverse journals, possibly with a control on textual elements (using or not “buzz-words” associated to the journal, etc.). Depending on results, a second experiment may be tested with providing open source code for model implementation but still with false results, to verify if reviewers effectively try to reproduce results when they pretend to want the code (in reasonable computational power limits of course, HPC being not currently broadly available in Humanities).

2.1.3 *Perspectives*

Again, reproducibility and transparency is a non-negotiable feature of contemporaneous science, along with Open practices and Open Access. Too much examples (see a very recent one in experimental economics [camerer2016evaluating]) show in various disciplines the lack of reproducibility of experiments, that is a falsification of previous results or a result in itself. Falsification is a costly practice, and even if necessary [chavaliarias2005nobel], could be made more efficient through more transparency and direct reproducibility, increase therein the global workflow of science. We develop in parallel of this thesis various tools aimed to ease reproducibility, for which an overview is given in appendix 15.

2.2 AN UNIFIED FRAMEWORK FOR STOCHASTIC MODELS OF URBAN GROWTH

Un cadre unifié pour les modèles stochastiques de croissance urbaine

Urban growth modeling fall in the case of tentatives to find self-consistent rules reproducing dynamics of an urban system, and thus in our logic of system morphogenesis. We examine here methodological issues linked to different frameworks of urban growth.

2.2.1 *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 [[simon1955class](#)] (later generalized as e.g. [[haran1973modified](#)]) and Gibrat models. Many examples can be given across disciplines. [[benguigui2007dynamic](#)] give an equation-based dynamical model, whereas [[gabaix1999zipf](#)] solves a stationary model. [[Gabaix20042341](#)] reviews urban growth approaches in economics. A model adapted from evolutive urban theory is solved in [[favarro2011gibrat](#)] 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 [[bretagnolle2002time](#)]. 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.

2.2.2 *Framework*

PRESENTATION What we propose as a framework can be understood as a meta-model in the sense of [[cottineau2015incremental](#)], 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. 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 [[cottineau2015incremental](#)]). It constructs a sort of model classification or modelography.

We are still at the stage of different derivations of links between models that are presented hereafter.

2.2.3 *Derivations*

Generalization of Preferential Attachment

[yamasaki2006preferential] give a generalization of the classical Preferential Attachment Network Growth model, as a birth and death model with evolving entities. More precisely, network units gain and lose population (equivalent to links connexions) at fixed probabilities, and new unit can be created at a fixed rate.

Link between Gibrat and Preferential Attachment Models

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 $\sigma_i(t) = \mathbb{E}[R_i(t)^2]$. 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 $m > 0$. 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.

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 13.

Link between Simon and Preferential Attachment

A rewriting of Simon model yields a particular case of the generalized preferential attachment, in particular by vanishing death probability.

Link between Favaro-Pumain and Gibrat

[favaro2011gibrat] generalizes Gibrat models with innovation propagation dynamics, being therefore a generalization of that model. Theoretically, a process-based model equivalent to the Favaro-pumain should then fill the missing case in model classification at the corresponding discretization. Simpop models do not fill that case as they stay at the scale of city systems, as for Marius models [cottineau2014evolution]. These must also have their counterparts in discrete microscopic formulation.

Link between Bettencourt-West and Pumain

We are considering to study Bettencourt-West model for urban scaling laws [**bettencourt2008large**] as entering the stochastic urban growth framework as stationary component of a random growth model, but investigation are still ongoing.

Other Models

[**gabaix1999zipf**] 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.

2.3 SENSITIVITY OF URBAN SCALING LAWS TO SPATIAL EXTENT

Sensibilité des Lois d’Echelle Urbaines à l’Etendue Spatiale

At the center of evolutive urban theory are hierarchy and associated scaling laws. We sketch here a methodological investigation on the sensitivity of scaling laws to city definition.

2.3.1 *Introduction*

Scaling laws have been shown to be universal of urban systems at many scales and for many indicators. Recent studies question however the consistence of scaling exponents determination, as their value can vary significantly depending on thresholds used to define urban entities on which quantities are integrated, even crossing the qualitative border of linear scaling, from infra-linear to supra-linear scaling. We use a simple theoretical model of spatial distribution of densities and urban functions to show analytically that such behavior can be derived as a consequence of the type of spatial distribution and the method used. Numerical simulation confirm the theoretical results and reveals that results are reasonably independent of spatial kernel used to distribute density.

Scaling laws for urban systems, starting from the well-known rank-size Zipf’s law for city size distribution [[gabaix1999zipf](#)], have been shown to be a recurrent feature of urban systems, at many scales and for many types of indicators. They reside in the empirical constata-
tion that indicators computed on elements of an urban system, that can be cities for system of cities, but also smaller entities at a smaller scale, do fit relatively well a power-law distribution as a function of entity size, i.e. that for entity i with population P_i , we have for an integrated quantity A_i , the relation $A_i \simeq A_0 \cdot \left(\frac{P_i}{P_0}\right)^\alpha$. Scaling exponent α can be smaller or greater than 1, leading to infra- or supra-linear effects. Various thematic interpretation of this phenomena have been proposed, typically under the form of processes analysis. The economic literature has produced abundant work on the subject (see [[Gabaix20042341](#)] for a review), but that are generally weakly spatial, thus of poor interest to our approach that deals precisely with spatial organization. Simple economic rules such as energetic equilibria can lead to simple power-laws [[bettencourt2008large](#)] but are difficult to fit empirically. A interesting proposition by Pumain is that they are intrinsically due to the evolutionary character of city systems, where complex emergent interaction between cities generate such global distributions [[pumain2006evolutionary](#)]. Although a tempting parallel can be done with self-organizing biological systems, Pumain insists on the fact that the ergodicity assumption for such systems is not reasonable in the case of geographical systems and that the analogy cannot be exploited [[pumain2012urban](#)]. Other

explanations have been proposed at other scales, such as the urban growth model at the mesoscopic scale (city scale) given in [[2014arXiv1401.8200L](#)] that shows that the congestion within transportation networks may be one reason for city shapes and corresponding scaling laws. Note that “classic” urban growth models such as Gibrat’s model do provide first order approximation of scaling systems, but that interactions between agents have to be incorporated into the model to obtain better fit on real data, such as the Favaro-Pumain model for innovation cycles propagation proposed in [[favaro2011gibrat](#)], that generalize a Gibrat model and provide better fits on data for French cities.

However, the blind application of scaling exponents computations was recently pointed as misleading in most cases [[louf2014scaling](#)], confirmed by empirical works such as [[2013arXiv1301.1674A](#)] that showed the variability of computed exponents to the parameters defining urban areas, such as density thresholds. An ongoing work by Cottineau & *al.* presented at [[cottineau2015scaling](#)], studies empirically for French Cities the influence of 3 parameters playing a role in city definition, that are a density threshold θ to delimitate boundaries of an urban area, a number of commuters threshold θ_c that is the proportion of commuters going to core area over which the unity is considered belonging to the area, and a cut-off parameter P_c under which entities are not taken into account for the linear regression providing the scaling exponent. Remarkable results are that exponents can significantly vary and move from infra-linear to supra-linear when threshold varies. A systematic exploration of parameter space produces phase diagrams of exponents for various quantities. One question raising immediately is how these variation can be explained by the features of spatial distribution of variables. Do they result from intrinsic mechanisms present in the system or can they be explained more simply by the fact that the system is particularly spatialized? We prove on a toy analytical model that even simple distributions can lead to such significant variations in the exponents, along one dimension of parameters (density threshold), directing the response towards the second explanation.

The derivations in the simple case of exponential mixture density, are done in Appendix 13.

2.4 STATISTICAL CONTROL ON INITIAL CONDITIONS BY SYNTHETIC DATA GENERATION

Contrôle statistique pour les conditions initiales par génération de données synthétiques

2.4.1 Context

When evaluating data-driven models, or even more simple partially data-driven models involving simplified parametrization, an unavoidable issue is the lack of control on “underlying system parameters” (what is a ill-defined notion but should be seen in our sense as parameters governing system dynamics). Indeed, a statistics extracted from running the model on enough different datasets can become strongly biased by the presence of confounding in the underlying real data, as it is impossible to know if result is due to processes the model tries to translate or to a hidden structure common to all data.

We formalize briefly a proposition of method that would allow to add controls on meta-parameters, in the sense of parameters driving the represented system at a higher temporal and spatial scale, for a model of simulation. We make the hypothesis that such method is valid under constraints of disjunction for scales and/or ontologies between the model of simulation and the domain of meta-parameters.

2.4.2 Description

An advanced knowledge of the behavior of computational models on their parameter space is a necessary condition for deductions of thematic conclusions or their practical application [banos2013pour]. But the choice of varying parameters is always subjective, as some may be fixed by a real-world parametrization, or other may be interpreted as arbitrarily fixed initial conditions. It raises methodological and epistemological issues for the sensitivity analysis, as the scope of the model may become ill-defined.

Let consider the concrete example of the Schelling Segregation model [schelling1971dynamic]. One of its crucial features on which the literature has been rather controversial is the influence of the spatial structure of the container on which agents evolve. The thematic aim of the project developed in [cottineau2015revisiting] is to clarify this point through a systematic model exploration. A methodological contribution is the construction of a framework allowing the analysis of the sensitivity of models to *meta-parameters*, i.e. to parameters considered as fixed initial conditions (e.g. the spatial structure for the Schelling model), or to parameters of another model generating an initial configuration yielding thus a *simple coupling* between models (serial coupling). The benefits of such an approach are various

but include for example the knowledge of model behavior in an extended frame, the possibility of statistical control when regressing model outputs, a finer exploration of model derivatives than with a naive approach. Some remarks can be made on the approach :

- What knowledge are brought by adding the upstream model, rather than for example in the Schelling case exploring a large set of initial geometries ?

→ *to obtain a sufficiently large set of initial configuration, one quickly needs a model to generate them; in that case a quasi-random generation followed by a filtering on morphological constraint will be a morphogenesis model, which parameters are the ones of the generation and the filtering methods. Furthermore, as detailed further, the determination of the derivative of the downstream model is made possible by the coupling and knowledge of the upstream model.*
- Statistical noise is added by coupling models

→ *Repetitions needed for convergence are indeed larger as the final expectance has to be determined by repeating on the first times the second model; but it is exactly the same as exploring directly many configuration, to obtain statistical robustness in that case one must repeat on similar configurations.*
- Complexity is added by coupling models

→ *In the sense of Varenne [varenne2010framework] , coupling is simple and no complexity is thus added.*

2.4.3 Formal Description

One has the composition of the derivative along the meta-parameter

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

→ *the sensitivity of the downstream model (Schelling) can be determined by studying the serial coupling and the upstream model; thematic knowledge : sensitivity to an implicit meta-parameter; and computational gain : generation of controlled differentiates in the “initial space” is quasi impossible.*

The question of stochasticity in simply coupled models causes no additional issue as $E[X] = E[E[X|Y]]$. It naturally multiplies the number of repetition needed for convergence what is the expected behavior.

2.5 LINKING DYNAMIC AND STATIC SPATIO-TEMPORAL CORRELATIONS UNDER SIMPLIFIED ASSUMPTIONS

Lien entre correlation spatio-temporelles statiques et dynamiques sous hypothèses simplifiées

Space and Time are both crucial for the study of geographical systems when aiming to understand *processes* (by definition dynamical [**hypergeo**]) evolving in a *spatial structure* in the sense of [**dollfus1975some**]. Space is more than coordinates for elements of the system, but a dimension in itself that drives interactions and thus system properties. Reading geographical systems from the point of view of *spatio-temporal processes* emphasizes the fact that *space actually matters*. Space and time are closely linked in such processes, and depending on the underlying mechanisms, one can expect to extract useful information from one on the other : in certain cases that we will investigate in this part, it is for example possible to learn about dynamics from static information. Spatio-temporal correlations approaches, linked to spatio-temporal dynamics, are present in very broad fields such as dynamical image processing (including video compression) [**chalidabhongse1997fast**, **hansen2004accelerated**, **ke2007spatio**], target tracking [**belouchrani1997direction**, **vuran2004spatio**], climate science [**cressie1999classes**], Earth sciences [**ma2002spatio**], city systems dynamics [**hernando2015memory**, **pigozzi1980interurban**], among others.

The capture of neighborhood effects in statistical models is a wisely used practice in spatial statistics, as the technique of Geographically Weighted Regression illustrates [**brunsdon1998geographically**]. A possible interpretation among many definitions of spatial autocorrelation [**griffith1992spatial**] yields that by estimating a plausible characteristic distance for spatial correlations or auto-correlations, 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 [**griffith1980towards**]. We propose now to study possible links between spatial and temporal correlations, using spatio-temporal covariance structure to infer information on dynamical processes.

2.5.1 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

$$C_t(\vec{x}_0) = \text{Var}[\vec{Y}[\vec{x}_0, \cdot]]$$

³ note that the formal link between models of spatial autocorrelation (see e.g. [**griffith2012advanced**]) is not clear and should be further investigated

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, reviewing simple cases in which a formal link can be established.

2.5.2 Wave Equation

In the case of propagating waves, there is an immediate link. Let assume that a wave equation is verified by “deterministic” parts of components

$$c^2 \cdot \partial_t^2 \bar{Y}_i = \Delta \bar{Y}_i \quad (1)$$

with $Y_i = \bar{Y}_i + \varepsilon_i$. If errors are uncorrelated and processes are stationary, we have then directly

$$\mathbf{C}_t [\partial_t^2 Y_i, \partial_t^2 Y_j] = \frac{1}{c^2} \cdot \mathbf{C}_x [\Delta Y_i, \Delta Y_j] \quad (2)$$

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.

2.5.3 Fokker-Planck Equation

An other interesting approach may be 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)

$$\partial_t P(x_i, t) = -d \cdot \partial_x P(x_i, t) + \frac{\sigma^2}{2} \partial_x^2 P(x_i, t) \quad (3)$$

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.

2.5.4 Master Equation

In the case of a master equation on probabilities of discrete states of the system

$$\partial_t \vec{P} = \mathbb{W} \vec{P} \quad (4)$$

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.

2.5.5 Consistent spatio-temporal sampling

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 , $|\rho_{\|\Delta\vec{x}\| < l_\rho} [Y_i(\vec{x} + \Delta\vec{x}, t), Y_i(\vec{x}, t)]| > 0$.
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. 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.

2.6 GENERATION OF CORRELATED SYNTHETIC DATA

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.

2.6.1 *Context*

L'utilisation de données synthétiques, au sens de populations statistiques d'individus générées aléatoirement sous la contrainte de reproduire certaines caractéristiques du système étudié, est une pratique méthodologique largement répandue dans de nombreuses disciplines, et particulièrement pour des problématiques liées aux systèmes complexes, telles que par exemple l'évaluation thérapeutique [abadie2010synthetic], l'étude des systèmes territoriaux [moeckel2003creating, pritchard2009advances], l'apprentissage statistique [bolon2013review] ou la bio-informatique [van2006syntren]. Il peut s'agir d'une désagrégation par création d'une population au niveau microscopique présentant des caractéristiques macroscopiques données, ou bien de la création de nouvelles populations au même niveau d'agrégation qu'un échantillon donné avec un critère de ressemblance aux données réelles. Le niveau de ce critère peut dépendre des applications attendues et peut par exemple aller de la fidélité des distributions statistiques pour un certain nombre d'indicateurs à des contraintes plus faibles de valeurs pour des indicateurs agrégés, c'est à dire l'existence de motifs macroscopiques similaires.

Dans le cas de systèmes chaotiques ou présentant de fortes caractéristiques d'émergence, une contrainte microscopique n'implique pas nécessairement le respect des motifs macroscopiques, et arriver à les reproduire est justement un des enjeux des pratiques de modélisation et simulation en sciences de la complexité. La donnée, qu'elle soit simulée, mesurée ou hybride est au cœur de l'étude des systèmes complexes de par la maturation de nouvelles approches computационnelles [arthur2015complexity], il est donc essentiel d'étudier des procédures d'extraction d'information des données (fouille de données) et de simulation d'une information similaire (génération de données synthétiques).

Si le premier ordre est de manière générale bien maîtrisé, il n'est pas systématique ni aisément de contrôler le second ordre, c'est à dire les structures de covariance entre les variables générées, même si des exemples spécifiques existent, comme dans [ye2011investigation] où la sensibilité des sorties de modèles de choix discrets à la forme des distributions des variables aléatoires ainsi qu'à leur structures de dépendance. Il est également possible d'interpréter les modèles de génération de réseaux complexes [newman2003structure] comme la création d'une structure d'interdépendance au sein d'un système, représentée par la topologie des liens. Nous proposons ici une méthode générique prenant en compte l'interdépendance lors de la génération de données synthétiques, sous la forme de correlations.

L'ensemble des méthodologies mentionnées en introduction sont trop variées pour être résumées par un même formalisme. Nous proposons ici une formulation générique ne dépendant pas du domaine d'application, ciblée sur le contrôle de la structure de correlation des données synthétiques.

2.6.2 *Formalization*

Soit un processus stochastique multidimensionnel \vec{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 $\mathbf{X} = (X_{i,j})$, de générer une population statistique $\tilde{\mathbf{X}} = \tilde{X}_{i,j}$ 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 , $\|f(\mathbf{X}) - f(\tilde{\mathbf{X}})\| < \varepsilon$
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 corrélation. 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 12. 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 corrélations 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.

A DISCREPANCY-BASED FRAMEWORK TO COMPARE ROBUSTNESS BETWEEN MULTI-ATTRIBUTE EVALUATIONS

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.

3.1 INTRODUCTION

Introduction

3.1.1 General Context

Les problèmes multi-objectifs sont organiquement liés à la complexité des systèmes sous-jacents. En effet, que ce soit dans le champ des *Systèmes Complexes Industriels*, dans le sens de systèmes conçus par ingénierie, où la construction de Systèmes de Systèmes (SoS) par couplage et intégration induit souvent des objectifs contradictoires [marler2004survey], ou dans le champ des *Systèmes Complexes Naturels*, au sens de systèmes non désignés, physiques, biologiques ou sociaux, qui présentent des propriétés d'émergence et d'auto-organisation, pour lesquels les objectifs peuvent e.g. être le résultat de l'interaction d'agents hétérogènes (voir [newman2011complex] pour une revue étendue des types de systèmes concernés par cette approche), l'optimisation multi-objectifs peut être explicitement introduite pour étudier ou désigner le système, mais régit généralement déjà implicitement les mécanismes

internes du système. Le cas des Systèmes Complexes Sociaux-techniques est particulièrement intéressant puisque selon Haken [haken2003face], ils peuvent être vus comme des systèmes hybrides embarquant des agents sociaux dans des “artefacts techniques” (parfois jusqu’à un niveau inattendu, créant ce que PICON décrit comme *cyborgs* [picon2013smart]), et cumulent ainsi la potentialité d’être à l’origine de problèmes multi-objectifs¹. La notion récente d’*éco-quartier* [souami2012ecoquartiers] est un exemple typique pour lequel la durabilité implique des objectifs contradictoires. L’exemple des systèmes de transport, dont la conception a glissé durant la seconde moitié du 20ème siècle d’analyses coût-bénéfices à la price de décision multi-critères, est également typique de tels systèmes [bavoux2005geographie]. Les systèmes géographiques sont à présent bien étudiés d’un tel point de vue, en particulier grâce à l’intégration des cadres multi-objectifs au sein des Systèmes d’Information Géographiques [carver1991integrating]. Comme dans le cas microscopique des éco-quartiers, la planification et le design urbains mésoscopiques et macroscopiques peuvent être rendus durables grâce aux évaluations par indicateurs [jegou2012evaluation].

Un aspect crucial de l’évaluation est une certaine notion de sa fiabilité, que nous nommerons ici *robustesse*. Les méthodes statistiques incluent naturellement cette notion puisque la construction et l’estimation de modèles statistiques donne divers indicateurs de la consistance des résultats [launer2014robustness]. Le premier exemple venant à l’esprit est l’application de la loi des grands nombres pour obtenir la *p-valeur* d’une estimation de modèle, qui peut être interprété comme une mesure de confiance en les valeurs estimées. D’autre part, les intervalles de confiance et le *beta-power* sont d’autres indicateurs importants de robustesse statistique. L’inférence bayésienne fournit également des mesures de robustesse quand la distribution des paramètres est estimée de manière séquentielle. Concernant les optimisations multi-objectifs, en particulier par des algorithmes heuristiques (comme par exemple les algorithmes génétiques, ou les solveurs de recherche opérationnelle), la notion de robustesse d’une solution consiste plus en la stabilité de la solution dans l’espace des phases du système dynamique correspondant. Des progrès récents ont été faits vers une formulation unifiée de la robustesse pour les problèmes d’optimisation multi-objectifs, comme dans [deb2006introducing] où les fronts de Pareto robustes sont définis comme des solutions insensibles aux petites perturbations. Dans [1688537], la notion de degré de robustesse est introduite, formalisée comme une sorte de continuité des autres solutions dans des voisinages successifs d’une solution.

¹ Nous désignons ici par *Evaluation Multi-objectifs* toutes les pratiques incluant le calcul de multiples indicateurs d’un système (il peut s’agir d’optimisation multi-objectif pour un design de système, une évaluation multi-objectif d’un système existant, une évaluation multi-attributs ; notre cadre particulier correspondra au dernier cas).

Cependant, il n'existe pas de méthode générique qui permettrait une évaluation de la robustesse de façon indépendante au modèle, i.e. qui serait extraite de la structure des données et des indicateurs mais ne dépendrait pas de la méthode utilisée. Un avantage serait par exemple une estimation *a priori* de la robustesse potentielle d'une évaluation et de décider ainsi si elle vaut la peine d'être faite. Nous proposons un cadre répondant à cette contrainte dans le cas particulier des évaluations multi-attributs, i.e. quand le problème est rendu unidimensionnel par agrégation des objectifs. Il est basé sur les données et non sur les modèles, au sens où l'estimation de la robustesse ne dépendra pas de la manière dont les indicateurs sont calculés, tant qu'ils respectent certaines hypothèses détaillées par la suite.

3.1.2 *Proposed Approach*

OBJECTIVES AS SPATIAL INTEGRALS Nous supposons que les objectifs peuvent être exprimés comme intégrales spatiales, ce qui devrait s'appliquer à tout système territorial, et nos cas d'application sont des systèmes urbains. Ce n'est pas si restrictif en terme d'indicateurs possibles si l'on utilise les bonnes variables et noyaux intégrés : de façon analogue à la méthode de Regression Géographique Pondérée [**brunsdon1998geographically**], toute variable spatiale peut être intégrée contre des noyaux réguliers de taille variable et le résultat 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.

LINEARLY AGGREGATED OBJECTIVES 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, [**dobbie2013robustness**] a proposé de comparer la robustesse des différentes techniques d'agrégation par une analyse de sensibilité, effectuée par simulations de Monte-Carlo pour produire des données synthétiques, ce qui permet d'obtenir la distribution des biais pour les différentes techniques, certaines étant signi-

ficativement 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.

3.2 FRAMEWORK DESCRIPTION

Description du Cadre

3.2.1 *Intuitive Description*

Nous décrivons à présent le cadre proposé pour permettre théoriquement de comparer la robustesse d'évaluation de deux systèmes urbains différents. Ce cadre est une généralisation d'une méthode empirique proposée dans [ecodistrictReport] pour accompagner une étude dans un autre contexte effectuant une comparaison du sens et de la pertinence des indicateurs dans un contexte de durabilité. Intuitivement, la base empirique se base sur les principes suivants :

- Les systèmes urbains peuvent être vus selon l'information disponible, i.e. les données brutes décrivant le système. Dans une approche basée sur les données, celles-ci sont la base de notre cadre et la robustesse sera déterminée par leur structure.
- A partir des données sont capturés des indicateurs (fonctions objectifs). Nous supposons qu'un choix d'indicateurs est une intention particulière de traduire des aspects particuliers du système, i.e. de capturer une réalisation d'un "fait urbain" au sens de MANGIN [mangin1999projet] - une sorte de fait stylisé en terme de processus et de mécanismes, ayant différentes réalisations sur des systèmes distincts dans l'espace, dépendant de chaque contexte géographique précis.
- Etant donné plusieurs systèmes et indicateurs associés, un espace commun peut être construit pour les comparer. Dans cet espace, les données représentent plus ou moins bien le système réel, c'est à dire qu'elles sont imprécises en fonction de l'échelle initiale, de la précision effective des données. Nous proposons de capturer exactement ces différents aspects au travers de la notion de discrépance d'un nuage de points, qui est un outil ma-

thématique provenant des théories d'échantillonnage, permettant d'exprimer la façon dont un jeu de données rempli l'espace dans lequel il s'insère [**dick2010digital**].

Synthétisant ces contraintes, nous proposons une notion de *Robustesse* d'une évaluation qui capture à la fois, en combinant la fiabilité des données à l'importance relative des indicateurs,

1. *Données manquantes* : une évaluation se basant sur des jeux de données plus raffinés sera naturellement plus robuste.
2. *Importance des indicateurs* : les indicateurs avec plus d'importance relative pèsent plus dans la robustesse totale.

3.2.2 Formal Description

INDICATORS Soit $(S_i)_{1 \leq i \leq N}$ un nombre fini de systèmes territoriaux géographiquement disjoints, que nous supposons décrits par les données brutes et des indicateurs intermédiaires, donnés par $S_i = (X_i, Y_i) \in \mathcal{X}_i \times \mathcal{Y}_i$ avec $\mathcal{X}_i = \prod_k \mathcal{X}_{i,k}$ tel que chaque sous-espace contient des matrices réelles : $\mathcal{X}_{i,k} = \mathbb{R}^{n_{i,k}^X p_{i,k}^X}$ (de la même façon pour \mathcal{Y}_i). Nous définissons également une fonction d'indice ontologique $I_X(i, k)$ (resp. $I_Y(i, k)$) prenant des valeurs entières qui coïncident si et seulement si les deux variables ont même ontologie au sens de [**livet2010**], c'est à dire qu'elles sont supposées représenter le même objet réel. On distingue les "données brutes" X_i à partir desquelles les indicateurs sont calculés généralement par des fonctions déterministes explicites, des "indicateurs intermédiaires" Y_i qui sont déjà intégrés et peuvent être par exemple les sorties de modèles élaborés simulant certains aspects du système urbain. Nous définissons l'espace caractéristique du "fait urbain" par

$$(\mathcal{X}, \mathcal{Y}) \underset{\text{def}}{=} \left(\prod \tilde{\mathcal{X}}_c \right) \times \left(\prod \tilde{\mathcal{Y}}_c \right) = \left(\prod_{\mathcal{X}_{i,k} \in \mathcal{D}_X} \mathbb{R}^{p_{i,k}^X} \right) \times \left(\prod_{\mathcal{Y}_{i,k} \in \mathcal{D}_Y} \mathbb{R}^{p_{i,k}^Y} \right) \quad (5)$$

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. $q_c = \int_{(\tilde{X}_c, \tilde{Y}_c)} h$ est une fonction décrivant le "fait urbain" (l'indicateur en lui-même)

Des exemples typiques de noyaux peuvent être :

- Une moyenne des lignes de $X_{i,c}$ est calculée par $h(x) = x \cdot f_{i,c}(x)$ où $f_{i,c}$ est la densité de la distribution de la variable sous-jacente.
- Un taux d'éléments du jeu de données respectant une condition donnée C , $h(x) = f_{i,c}(x) \chi_{C(x)}$.
- Pour des variables déjà agrégées Y , une distribution de Dirac permet de les exprimer également comme des intégrales de noyaux.

AGGREGATION La détermination des poids est en fait le point crucial des processus de prise de décision multi-attributs, et de nombreuses méthodes sont disponibles (voir [[wang2009review](#)]) pour une revue dans le cas particulier de la gestion de l'énergie durable). Définissons les poids pour l'agrégation linéaire. Nous supposons les indicateurs normalisés, i.e. $q_c \in [0, 1]$, pour une construction plus simple des poids relatifs. Pour i, c et $h_c \in H_c$ donnés, le poids $w_{i,c}$ est simplement constitué par l'importance relative de l'indicateur $w_{i,c}^L = \frac{\hat{q}_{i,c}}{\sum_c \hat{q}_{i,c}}$ où $\hat{q}_{i,c}$ est un estimateur de q_c pour les données $X_{i,c}$ (i.e. la valeur calculée effectivement). On peut noter que cette étape n'est pas contrainte et que cela peut être étendu à tout ensemble d'attribution de poids, en prenant par exemple $\tilde{w}_{i,c} = w_{i,c} \cdot w'_{i,c}$ si w' sont les poids fixés par le preneur de décisions. Nous nous concentrerons sur l'influence relative des attributs et pour cela choisissons cette forme simple pour les poids.

ROBUSTNESS ESTIMATION La scène est à présent prête pour permettre d'estimer la robustesse d'une évaluation faite par la fonction d'agrégation. Pour cela, nous appliquons un théorème d'approximation d'intégrale similaire au méthodes introduites dans [[varet2010developpement](#)], puisque la forme intégrée des indicateurs permet justement de bénéficier de tels résultats théoriquement puissant. Soit $X_{i,c} = (\vec{X}_{i,c,l})_{1 \leq l \leq n_{i,c}}$ et $D_{i,c} = \text{Disc}_{\tilde{X}_c, L^2}(X_{i,c})$ le discrépance du jeu de données² [[niederreiter1972discrepancy](#)]. Avec $h \in H_c$, on a la borne supérieure sur l'erreur d'approximation de l'intégrale

² 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.

$$\left\| \int h_c - \frac{1}{n_{i,c}} \sum_l h_c(\vec{X}_{i,c,l}) \right\| \leq K \cdot \|h_c\| \cdot D_{i,c}$$

où K est une constante indépendante des points de données et des fonctions objectifs. Cela donne directement

$$\left\| \int \sum w_{i,c} h_c - \frac{1}{n_{i,c}} \sum_l w_{i,c} h_c(\vec{X}_{i,c,l}) \right\| \leq K \sum_c |w_{i,c}| \|h_c\| \cdot D_{i,c}$$

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

$$R_{i,i'} = \frac{\sum_c w_{i,c} \cdot D_{i,c}}{\sum_c w_{i',c} \cdot D_{i',c}} \quad (6)$$

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.

3.3 RESULTS

Résultats

IMPLEMENTATION Le pré-traitement des données géographiques est fait via QGIS [[qgis2011quantum](#)] pour des raisons de performances. L'implémentation du cœur est faite en R [[team2000r](#)] pour la flexibilité de la gestion des données et du traitement statistique. De plus, le package DiceDesign [[franco20092](#)] conçu pour les expériences numériques et l'échantillonnage, permet un calcul efficient et direct des discordances. Enfin, tout aussi important, l'ensemble du code source est disponible de manière ouverte sur le dépôt git du projet³ pour permettre la reproductibilité et la réutilisation [[ram2013git](#)].

3.3.1 *Implementation on Synthetic Data*

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.

DATA COLLECTION Le cas virtuel se base sur des données géographiques réelles, en particulier pour les arrondissements parisiens. Nous utilisons les données disponibles par le projet OpenStreetMap [[bennett2010openstre](#)] 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.

VIRTUAL CASES 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 don-

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

nées 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 voiture à bâtiments proportionnel au taux de motorisation attendu α_m 0.4 à Paris), d_w distance des individus à leur travail (généré à partir du bâtiment vers un point aléatoire distribué uniformément dans l'étendue spatiale du jeu de données), et d_{max} le diamètre de l'aire de Paris, $\bar{d}_w = 1 - \frac{1}{|b \in A(a)|} \cdot \sum_{b \in A(a)} n_{cars}(b) \cdot \frac{d_w}{d_{max}}$
- Le complémentaire des flots moyens de voitures des rues dans la zone, approché par, avec $\varphi(s)$ flot relatif dans le segment de rue s , généré par le minimum entre 1 et une distribution log-normale ajustée pour avoir 95% de masse plus petite que 1, ce qui mimique la distribution hiérarchique de l'utilisation des rues (qui correspond à la centralité de chemin), et $l(s)$ longueur du segment, $\bar{\varphi} = 1 - \frac{1}{|s \in A(a)|} \cdot \sum_{s \in A(a)} \varphi(s) \cdot \frac{l(s)}{\max(l(s))}$
- Longueur relative de rues piétonnes \bar{p} , calculé via une dummy variable aléatoire uniforme ajustée pour obtenir une proportion fixée de segments piédestre.

Comme les données synthétiques sont stochastiques, les simulations sont lancées pour chaque quartier $N = 50$ fois, ce qui était un compromis raisonnable entre convergence statistique et temps nécessaire au calcul. La table 1 montre les résultats (moyennes et déviations standard) des valeurs des indicateurs et le calcul du ratio de robustesse. Les déviations standard obtenues confirment que ce nombre de simulations donnent des résultats constants. Les indicateurs obtenus en fixant un ratio fixe montre peu de variabilité, ce qui peut être une limite de cette approche jouet. On obtient toutefois le résultat intéressant que la majorité des arrondissements donne des évaluations plus robustes que le 1er arrondissement, ce qui était attendu par la taille et la fonction de ce quartier : il s'agit en effet d'un petit quartier avec de grands bâtiments administratifs, ce qui implique moins d'éléments spatiaux et pour cela une évaluation moins robuste selon la définition qu'on en a donnée.

3.3.2 Application to a Real Case : Metropolitan Segregation

Le premier exemple avait pour but de montrer les potentialités de la méthode mais était purement synthétique, ne pouvant pour cela

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 ± 0.0000
2 th	0.723225 ± 0.032539	0.844350 ± 0.036085	0.209467 ± 0.058675	1.002098 ± 0.0399
3 th	0.713716 ± 0.044789	0.797313 ± 0.057480	0.185541 ± 0.065089	0.999341 ± 0.0488
4 th	0.712394 ± 0.042897	0.861635 ± 0.030859	0.201236 ± 0.044395	0.973045 ± 0.0369
5 th	0.715557 ± 0.026328	0.894675 ± 0.020730	0.209965 ± 0.050093	0.963466 ± 0.0407
6 th	0.733249 ± 0.026890	0.875613 ± 0.029169	0.206690 ± 0.054850	0.990676 ± 0.0316
7 th	0.719775 ± 0.029072	0.891861 ± 0.026695	0.209265 ± 0.041337	0.966103 ± 0.0371
8 th	0.713602 ± 0.034423	0.931776 ± 0.015356	0.208923 ± 0.036814	0.973975 ± 0.0338
9 th	0.712441 ± 0.027587	0.910817 ± 0.015915	0.202283 ± 0.049044	0.971889 ± 0.0353
10 th	0.713072 ± 0.028918	0.881710 ± 0.021668	0.210118 ± 0.040435	0.991036 ± 0.0389
11 th	0.682905 ± 0.034225	0.875217 ± 0.019678	0.203195 ± 0.047049	0.949828 ± 0.0351
12 th	0.646328 ± 0.039668	0.920086 ± 0.019238	0.198986 ± 0.023012	0.960192 ± 0.0348
13 th	0.697512 ± 0.025461	0.890253 ± 0.022778	0.201406 ± 0.030348	0.960534 ± 0.0337
14 th	0.703224 ± 0.019900	0.902898 ± 0.019830	0.205575 ± 0.038635	0.932755 ± 0.0336
15 th	0.692050 ± 0.027536	0.891654 ± 0.018239	0.200860 ± 0.024085	0.929006 ± 0.0316
16 th	0.654609 ± 0.028141	0.928181 ± 0.013477	0.202355 ± 0.017180	0.963143 ± 0.0332
17 th	0.683020 ± 0.025644	0.890392 ± 0.023586	0.198464 ± 0.033714	0.941025 ± 0.0349
18 th	0.699170 ± 0.025487	0.911382 ± 0.027290	0.188802 ± 0.036537	0.950874 ± 0.0286
19 th	0.655108 ± 0.031857	0.884214 ± 0.027816	0.209234 ± 0.032466	0.962966 ± 0.0341
20 th	0.637446 ± 0.032562	0.873755 ± 0.036792	0.196807 ± 0.026001	0.952410 ± 0.0387

TABLE 1

fournir pas de conclusion concrete ni d'implications pour la gouvernance. Nous proposons maintenant de l'appliquer à des données réelles dans le cas de la ségrégation métropolitaine.

DATA 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'INSEE⁴. Les données sont associées à l'étendue géographique des unités statistiques, permettant le calcul d'indicateurs d'analyse spatiale.

⁴ <http://www.insee.fr>

INDICATORS Nous utilisons ici trois indicateurs de ségrégation intégrés sur une zone géographique. Supposons la zone divisée en unités couvrantes S_i pour $1 \leq i \leq N$ avec pour centroïdes (x_i, y_i) . Chaque unité a des caractéristiques de population P_i et de revenu médian X_i . On définit des poids spatiaux utilisés pour quantifier l'intensité des interactions géographiques entre unités i, j , avec d_{ij} distance euclidienne entre centroïdes : $w_{ij} = \frac{P_i P_j}{(\sum_k P_k)^2} \cdot \frac{1}{d_{ij}}$ si $i \neq j$ et $w_{ii} = 0$. Les indicateurs normalisés sont les suivants

- Indice d'autocorrelation spatiale de Moran, défini comme la covariance pondérée normalisée du revenu médian par $\rho = \frac{N}{\sum_{ij} w_{ij}} \cdot \frac{\sum_{ij} w_{ij} (X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}$
- Indice de dissimilarité (proche du Moran mais intégrant les dissimilarités locales plutôt que les corrélations), donné par $d = \frac{1}{\sum_{ij} w_{ij}} \sum_{ij} w_{ij} |X_i - \tilde{X}_i|$ avec $\tilde{X}_i = \frac{X_i - \min(X_k)}{\max(X_k) - \min(X_k)}$
- Le complémentaire de l'entropie de la distribution des revenus, qui est une façon de capturer des inégalités globales $\varepsilon = 1 + \frac{1}{\log(N)} \sum_i \frac{X_i}{\sum_k X_k} \cdot \log \left(\frac{X_i}{\sum_k X_k} \right)$

De nombreuses mesures de ségrégation avec différentes signification à différentes échelles existent, comme par exemple à l'échelle d'une unité spatiale élémentaire par comparaison de la distribution de revenus empirique avec un modèle nul [louf2015patterns]. Le choix est ici arbitraire, afin d'illustrer la méthode avec un nombre raisonnable de dimensions.

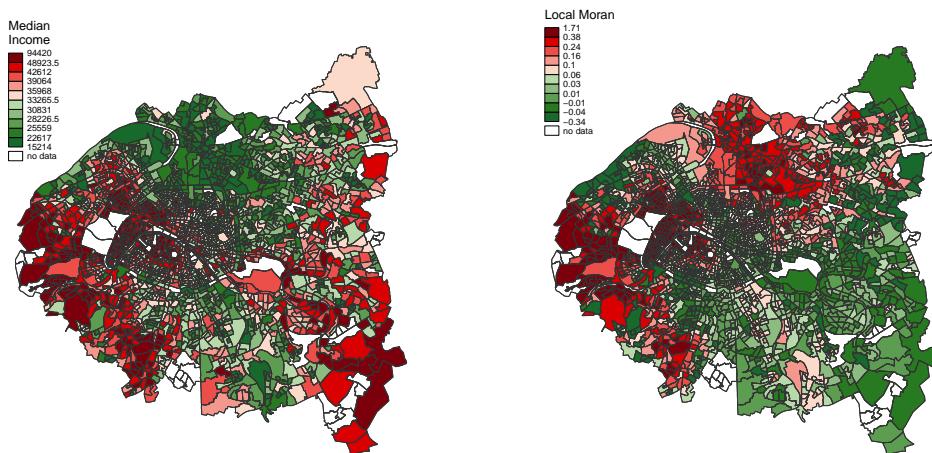


FIGURE 2

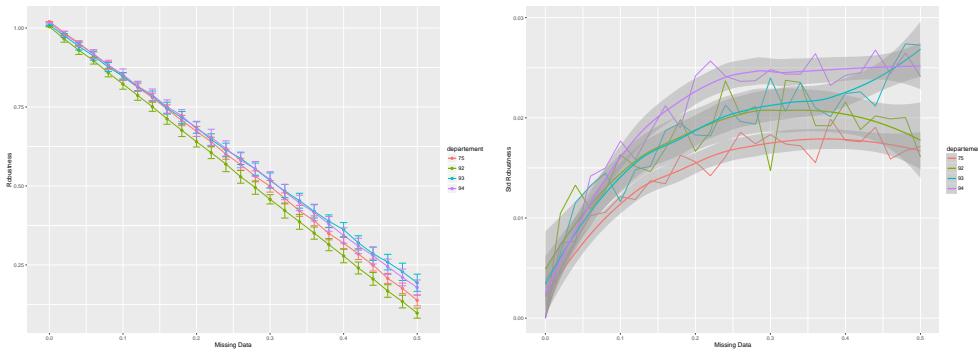


FIGURE 3

RESULTS La méthode est appliquée avec ces indicateurs à la zone du Grand Paris, constitué de 4 département qui sont des niveaux administratifs intermédiaires. La création récente d'un nouveau système de gouvernance métropolitaine [gilli2009paris] met en évidence des interrogations sur sa pertinence, notamment sur ses capacités d'atténuer les inégalités spatiales. On peut voir en Fig. ?? les cartes de la distribution spatiale du revenu médian et de l'index local d'autocorrelation spatiale correspondant. La dichotomie bien connue entre est et ouest est retrouvée ainsi que la disparité des quartiers intra-muros, comme cela été présenté par diverses études, comme [guerois2009dynamique] à travers l'analyse des dynamiques des transactions immobilières. Notre cadre d'étude est ensuite appliqué à une question concrète ayant des implications pour la prise de décision : *dans quelle mesure une évaluation de la ségrégation au sein de différents territoires est sensible aux données manquantes ?* Pour cela, on procède à des simulations de Monte-Carlo (75 répétitions) pour lesquelles une proportion fixe de données est supprimée aléatoirement, et l'indice de robustesse correspondant est évalué avec les indicateurs normalisés. Les simulations sont faites sur chaque département de façon indépendante, à chaque fois pour une robustesse relative à l'évaluation du Grand Paris complet. Les résultats sont présentés en Fig. ???. Toutes les zones ont une robustesse légèrement meilleure que la référence, ce qui pourrait être expliqué par une homogénéité locale et donc des indices de ségrégation plus fiables. Les implications pour la prise de décision qui peuvent être par exemple tirées sont des comparaisons directes entre les zones : une perte de 30% de l'information sur le 93 correspond à une perte de seulement 25% pour le 92. La première zone étant déjà défavorisée socio-économiquement, l'inégalité est augmentée par cette qualité moindre de l'information statistique. L'étude des déviations standard suggère des études plus approfondies comme différents régimes de réponse à la suppression de données semblent exister.

3.4 DISCUSSION

Discussion

3.4.1 *Applicability to Real situations*

IMPLICATIONS FOR DECISION-MAKING L’application de notre méthode à des situations concrètes de prise de décision peu être pensée de différentes manières. Tout d’abord dans le cas d’un processus multi-attributs à but comparatif, comme la détermination d’un corridor pour une nouvelle infrastructure de transport, l’identification des territoires sur lesquels l’évaluation pourrait être biaisée (i.e. avec une mauvaise robustesse relative) devrait permettre une attention particulière pour ceux-ci, et l’adaptation des jeux de données ou la révision des points en conséquence. Dans tous les cas le processus total devrait être plus fiable. Une autre possibilité ressemble 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.

INTEGRATION WITHIN EXISTING FRAMEWORKS L’applicabilité de la méthode à des cas réels dépendra directement de son intégration potentielle dans des environnements existants. Au delà des difficultés techniques qui apparaissent nécessairement en essayant de coupler ou d’intégrer des implémentations existantes, des obstacles plus théoriques pourraient émerger, comme des formulations floues des fonctions ou des types de données, la cohérence des bases de données, etc. De tels cadres multi-critères sont nombreux. Un développement possible serait l’intégration dans un cadre open-source, comme par exemple celui décrit dans [[tivadar2014oasis](#)] qui calcule divers indices de ségrégation urbaine, comme on l’a déjà illustré pour l’application à la ségrégation métropolitaine.

AVAILABILITY OF RAW DATA 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 res-

treint). 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.

3.4.2 *Validity of Theoretical Assumptions*

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 [kwan1998space] 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 [banos2013HDR], car cela impliquerait qu'il pourrait exister un langage universel pour la modélisation, malgré qu'il ne puisse retranscrire certains systèmes, ayant pour conséquences des conclusions errantes à cause d'une rupture d'ontologie dans le cas d'une formulation sur-contrainte.

3.4.3 *Framework Generality*

Nous soutenons qu'un des avantages fondamentaux de notre cadre est sa généralité et sa flexibilité, puisque la robustesse des évaluations est obtenue seulement par la structure des données si l'on relaxe les hypothèses sur les valeurs des poids. Des approfondissement pourraient inclure une formulation plus générale, en supprimant par exemple l'hypothèse d'agrégation linéaire. Des fonctions d'agrégation non-linéaires demanderaient toutefois de vérifier certaines propriétés regardant les inégalités intégrales. Par exemple, des résultats similaires pourraient être obtenus en s'orientant vers des inégalités intégrales pour fonctions Lipschitziennes, comme les résultats en une dimension de [dragomir1999ostrowski].

CONCLUSION

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.

4

QUANTITATIVE EPISTEMOLOGY

The Social Construction of What ?
- IAN HACKING [[hacking1999social](#)]

Under this provocative book title by HACKING are implied complex mechanisms in the production of scientific knowledge. Animated debates on constructivism would be due to different metaphysical conceptions that are by essence not provable. As we have already evoked with perspectivism, scientific enterprises may have different purposes and be difficultly transferable to other contexts as we intent to do in our broad thematic vision developed in chapter 1.

A corollary of theoretical background proposed in chapter 10 is the need of an understanding of involved disciplines themselves to be able to build integrated heterogeneous models. The potentialities of couplings and integrations are greatly determined by existing approaches and corresponding gaps. This implies an advanced epistemological study in each field, that we propose to tackle in a systematic and quantitative way. This deliberate choice may shadow elaborated epistemological considerations but fits our purpose of preliminary investigations for the construction of models, as it may reveal investigation directions.

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

4.1 ALGORITHMIC SYSTEMATIC REVIEW

Revue Systématique Algorithmique

Une étude bibliographique étendue suggère une rareté 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 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 procédons à une analyse de sensibilité. Nous l'appliquons ensuite à des requêtes représentatives de notre question spécifique, pour lesquelles les résultats tendent à confirmer l'hypothèse d'isolation des disciplines.

4.1.1 *In search of models of co-evolution*

Les réseaux de transport et l'usage du sol urbain sont connus pour être des composantes fortement couplées des systèmes urbains à différentes échelles [[bretagnolle2009organization](#)]. Une approche commune est de les considérer comme étant en co-évolution, tout en évitant les interprétations trompeuses comme le mythe des effets structurants des infrastructures de transport [[offner1993effets](#)]. Une question qui se présente rapidement est l'existence de modèles endogénisant 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. Nous proposons dans cette section, après un état de l'art rapide de la littérature existante, de développer cette approche en formalisant l'algorithme, dont les résultats sont ensuite présentés et discutés.

4.1.2 *Modeling Interactions between Urban Growth and Network Growth : An Overview*

Land-Use Transportation Interaction Models.

Une large classe de modèle développés essentiellement dans des objectifs de planification, les modèles d'interaction entre transport de usage du sol, sont un premier type pouvant rentrer dans notre problématique. Voir les diverses revues [[chang2006models](#)], [[iacono2008models](#)] et [[wegerer2004land](#)] pour avoir un aperçu de l'hétérogénéité des approches incluses, qui existent depuis plus de 30 ans. Des versions récentes avec divers raffinements sont toujours développés aujourd'hui, comme [[delons:hal-00319087](#)] qui inclut le marché immobilier pour la

région parisienne. Différents aspects du même système peuvent être traduits par divers modèles (comme e.g. [[wegener1991one](#)]), et le trafic, les dynamiques résidentielles et d'emploi, l'évolution de l'usage du sol en découlant, influencée aussi par un réseau de transport statique, sont généralement pris en compte.

Network Growth Approaches

A l'opposé de nombreux travaux ont pris la logique inverse, i.e. essayent de reproduire la croissance du réseau étant donné des hypothèses sur l'environnement urbain, comme résumé dans [[zhang2007economics](#)].

Dans [[xie2009modeling](#)], les travaux économiques empiriques sont positionnés parmi les autres approches de la croissance des réseaux, comme des travaux de physiciens proposant des modèles de croissance géométrique locale [[barthelemy2008modeling](#)]. L'analogie avec la biologie a également déjà été faite, permettant de reproduire les propriétés typiques de robustesse des réseaux de transport [[tero2010rules](#)].

Hybrid Approaches

Peu de travaux couplant croissance urbaine et croissance du réseau sont disponibles dans la littérature. [[barthelemy2009co](#)] couple l'évolution de la densité avec la croissance du réseau dans un modèle jouet. Dans [[raimbault2014hybrid](#)], un automate cellulaire simple couplé à un réseau évolutif reproduit les faits stylisés des Etablissements Humains décrits par Le Corbusier. A une plus petite échelle, [[achibet2014model](#)] propose un modèle de co-évolution entre routes et bâtiments, en suivant des règles géométriques. Ces approches restent cependant limitées et rares.

4.1.3 Bibliometric Analysis

La revue de littérature est une étape préliminaire cruciale à toute entreprise scientifique, et sa qualité et son étendue peut avoir un impact conséquent sur le résultat final. Des techniques de revue systématique ont été développées, des revues qualitatives aux meta-analyses quantitatives qui permettent de produire des nouveaux résultats par combinaison d'études existantes [[rucker2012network](#)]. Passer sous silence certaines références peut même être considéré comme une erreur scientifique dans le contexte de l'émergence des systèmes d'information [[lissacksubliminal](#)]. 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. 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:tel-00923682](#)]

pour le concept de mobilité, pour lequel un “petit monde d’acteurs” relativement fermé 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 [2013arXiv1310.8220N] ou les réseaux de co-auteurs [2014arXiv1402.7268S], nous proposons d’utiliser un paradigme moins exploré, basé sur l’analyse textuelle, introduit par [chavaliarias2013phylomemetic], qui obtient une cartographie dynamique des disciplines scientifiques en se basant sur leur contenu sémantique. 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 of the Algorithm

Soit A un alphabet, A^* les mots correspondants et $T = \cup_{k \in \mathbb{N}} A^{*k}$ 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 à l’itération n sera noté $\mathcal{C} \subset T^3$. Nous supposons l’existence d’un ensemble de mots-clés \mathcal{K}_n , les mots-clés initiaux étant \mathcal{K}_0 . Une itération procède de la manière suivante :

1. Un corpus intermédiaire brut \mathcal{R}_n est obtenu par une requête à un catalogue auquel on fourni les mots-clés précédents \mathcal{K}_{n-1} .
2. Le corpus total est actualisé par $\mathcal{C}_n = \mathcal{C}_{n-1} \cup \mathcal{R}_n$.
3. Les nouveaux mot-clés \mathcal{K}_n sont extraits du corpus par Traitement du Langage Naturel (NLP), étant donné un paramètre N_k fixant le nombre de mot-clés.

L’algorithme termine quand la taille du corpus devient stable ou quand un nombre maximal d’itérations défini par l’utilisateur est atteint. La figure 4 montre le processus général.

Results

IMPLEMENTATION De par l’hétérogénéité des opérations requises par l’algorithme (organisation des références, requêtes au catalogue,

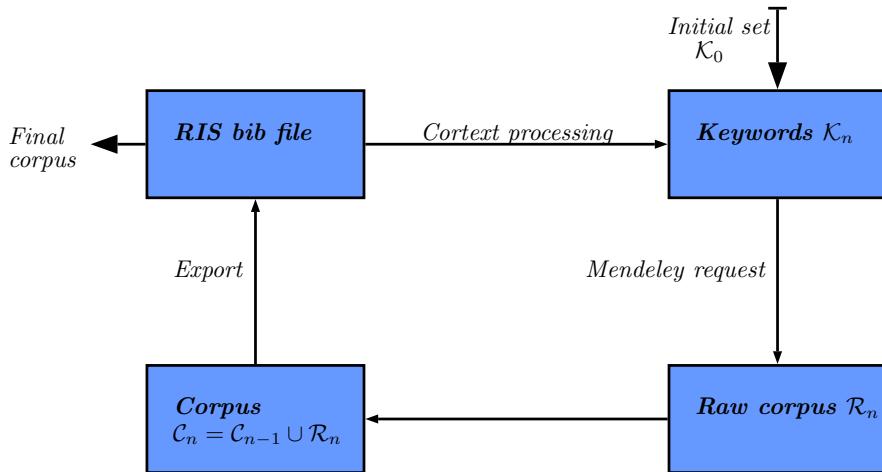


FIGURE 4 : Global workflow of the algorithm, including implementation details : catalog request is done through Mendeley API; final state of corpuses are RIS files.

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] qui permet un accès ouvert à une base de données conséquente. L'extraction des mots-clés est effectuée par techniques d'Analyse Textuelle (NLP) selon le processus donné dans [chavalarias2013phylometic], via un script Python qui utilise [bird2006nltk].

CONVERGENCE AND SENSITIVITY ANALYSIS 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 5, 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.

Lorsque l'algorithme a été partiellement validé, on peut l'appliquer à notre question. 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"). Nous prenons l'hypothèse la plus faible pour le paramètre $N_k = 100$, au sens

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

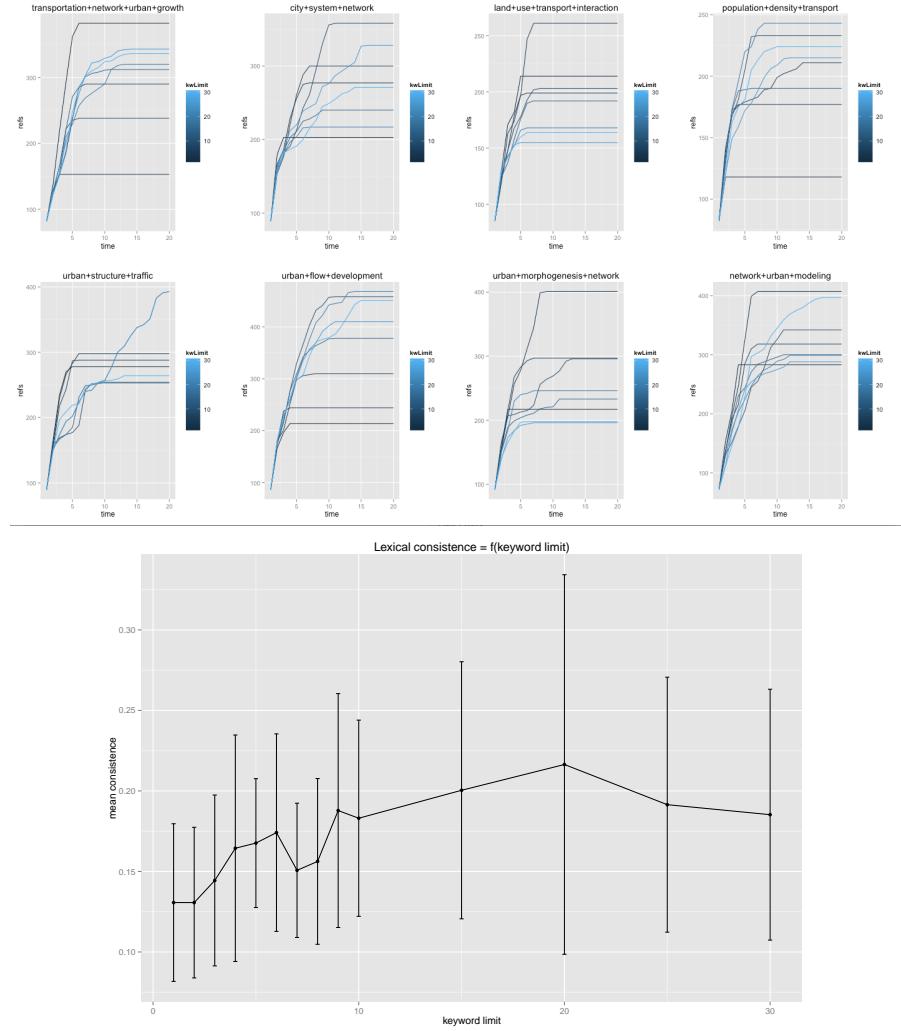


FIGURE 5 : Convergence and sensitivity analysis. Left : Plots of number of references as a function of iteration, for various queries linked to our theme (see further), for various values of N_k (from 2 to 30). We obtain a rapid convergence for most cases, around 10 iterations needed. Final number of references appears to be very sensitive to keyword number depending on queries, what seems logical since encountered landscape should strongly vary depending on terms. Right : Mean lexical consistence and standard error bars for various queries, as a function of keyword number. Lexical consistence is defined though co-occurrences of keywords by, with N final number of keywords, f final step, and $c(i)$ co-occurrences in references, $k = \frac{2}{N(N-1)} \cdot \sum_{i,j \in \mathcal{K}_f} |c(i) - c(j)|$. The stability confirms the consistence of final corpuses.

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

TABLE 2 : Symmetric matrix of lexical proximities between final corpuses, defined as the sum of overall final keywords co-occurrences between corpuses, normalized by number of final keywords (100). We obtain very low values, confirming that corpuses are significantly far. Size of final corpuses is given as W.

où les domaines atteints devraient être moins restreints. 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 2 montre les valeurs de la proximité lexicale relative, qui est significativement basse, confirmant notre hypothèse.

Les développements possibles incluent la construction de réseaux de citation via un accès automatique à Google Scholar 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 a priori de modèles qui simulent la coévolution 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. Nous avons proposé 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. 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:tel-00923682](#)], dans laquelle des questionnaires avec des acteurs historiques fournissent des informations extrêmement pertinentes.

4.2 INDIRECT BIBLIOMETRICS THROUGH COMPLEX NETWORK ANALYSIS

Bibliométrie Indirecte par Analyse de Réseaux Complexes

4.2.1 *Context*

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

Previous works in quantitative epistemology using various types of networks have shown interesting potentialities. For the citation network, a good predicting power for citation patterns is for example obtained in [2013arXiv1310.8220N]. Co-authorship networks can also be used for predictive models [2014arXiv1402.7268S]. A multilayer network approach was recently proposed in [2016arXiv160106075O], using bipartites networks of papers and scholars, in order to produce measures of interdisciplinarity. Disciplines can be stratified into layers to reveal communities between them and therein collaboration patterns [2015arXiv150601280B]. Keyword networks are used in other fields such as economics of technology [choi2014patent, shibata2008detecting].

4.2.2 *Application to a scientific journal*

Presentation

We briefly describe here an ongoing study that implemented the ideas given above for the particular case of a scientific journal for which bibliographical data is difficult to obtain, that is *cybergeo*, an electronic journal in theoretical and quantitative geography, that is concerned with open science issues such as peer-review ethics transparency [10.1371/journal.pone.0147913]. Our approach combine semantic communities analysis (as done in [2016arXiv160208451P] but with keyword extraction; [2015arXiv151003797G] analyses semantic networks of political debates) with citation network to extract e.g. interdisciplinarity measures.

Implementation

The general architecture for data collection is presented in Fig. 6. Citation data is collected from *Google Scholar*, that is the only source

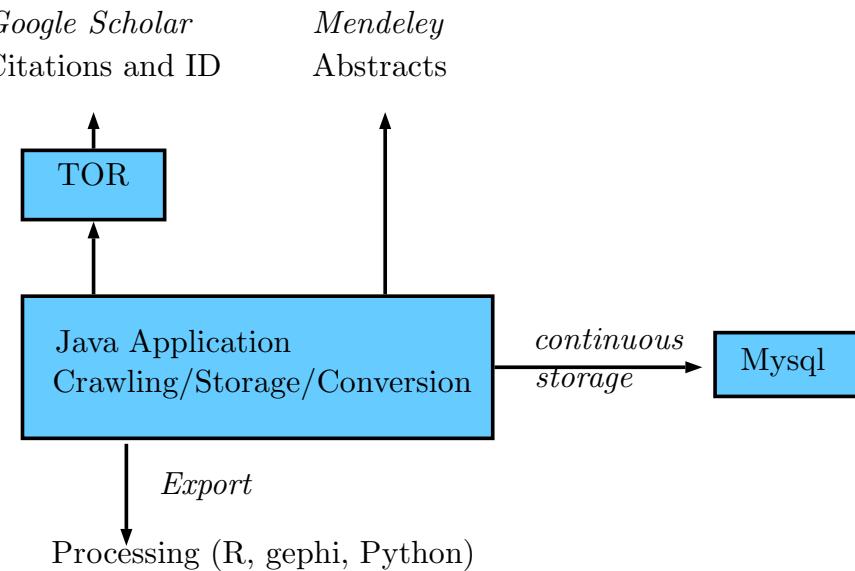


FIGURE 6 : Heterogeneous Bibliographical Data Collection. Architecture of the application for content (semantic data), metadata and citation data collection.

for incoming citations [**noruzi2005google**] in our case as the journal is not referenced in other databases. We are aware of the possible biases using this single source [**bohannon2014scientific**]², but these critics are more directed towards search results than citation counts.

Text processing is done the same way as in previous section, expect that a particular treatment is done to language detection using *stopwords* and a specific tagger TreeTagger is used for other languages than english [**schmid1994probabilistic**].

Results

We show in figures 7 and 8 preliminary results on citation and semantic network. We are able by the reconstruction of the citation network at depth ± 1 from the original 1000 references of the journal to retrieve around $45 \cdot 10^6$ references, on which $2.1 \cdot 10^6$ are retrieved with abstract text allowing semantic analysis. We retrieve by community detection in the semantic network typical geographical disciplines, such as :

- Hydrology : water, basin, river, capac
- Traffic : traffic, road, vehicl
- Biogeography : habitat, soil, veget, ecosystem
- Political Science : polit, cultur, societi, debat
- Economy : market, economi, privat, competit, industri

² or see <http://iscpif.fr/blog/2016/02/the-strange-arithmetic-of-google-scholars/>

- Transportation : *transport, travel*
- Teledetection : *cluster, imag, classif, satellit*
- Education : *educ, age, student, school*
- Health : *diseas, infect*
- GIS : *gi, geograph inform system*
- Social geography : *neighborhood, resid*

Distribution of keywords within reconstructed disciplines provides an article-level interdisciplinarity, and we can construct various measures at the journal level. Combination of citation and semantic layers in the hyper-network provide second order interdisciplinarity measures. The construction of null models for comparison and the collection of currently missing data (journals for other papers) are currently ongoing so these results are not presented here.

4.2.3 *Application*

We will try to reconstruct the same way disciplines around our thematic, and by for example identifying bridge articles (nodes with high centrality or vulnerability) identify crucial thematic elements and research directions.

An other application will be the reflexivity of our thesis : we attend to proceed to similar analysis on our proper bibliography (and its evolution, available via git history), to understand our patterns of knowledge, possible gaps or unveil unexpected developments.

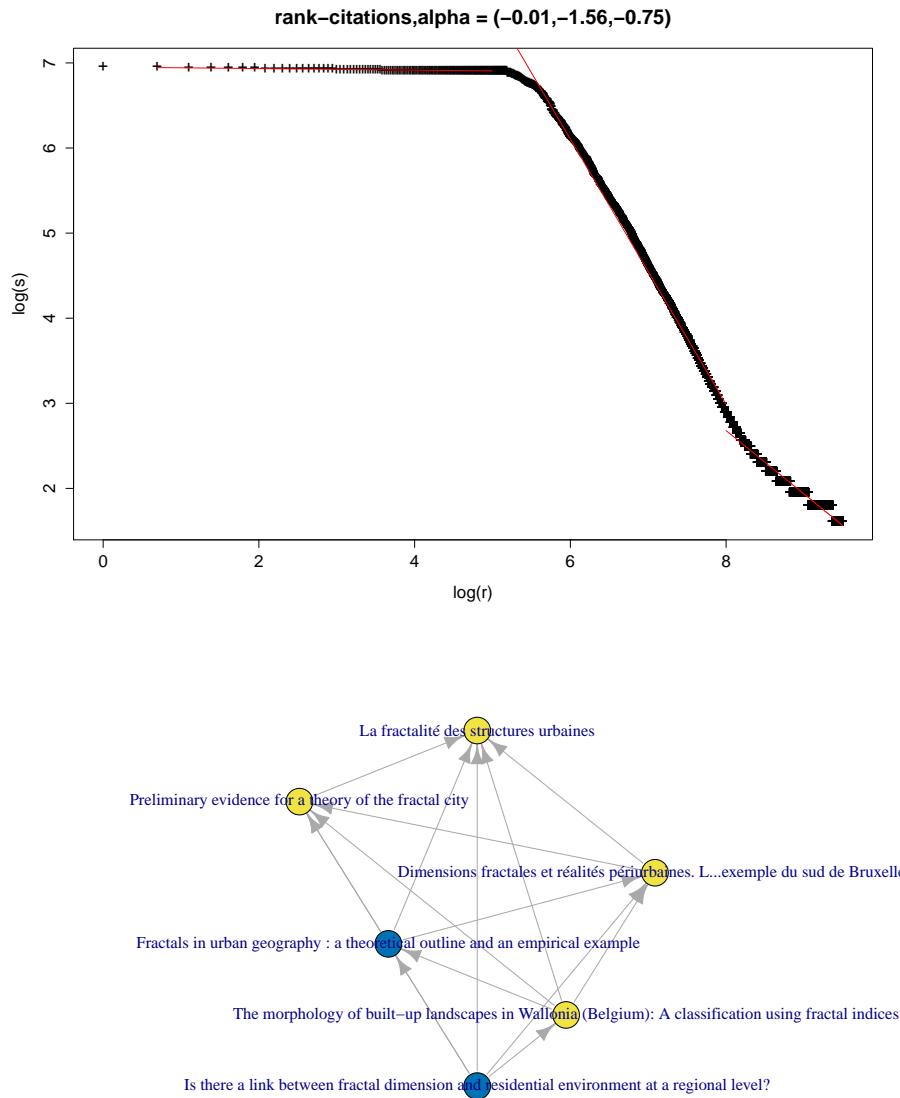


FIGURE 7 : Properties of the citation network. Top : Rank-size plot of in-degrees ; three superposing successive regimes must correspond to different literature types or practices across disciplines. Bottom : example of a maximal clique in the citation network, paper of *cybergeo* being in blue.

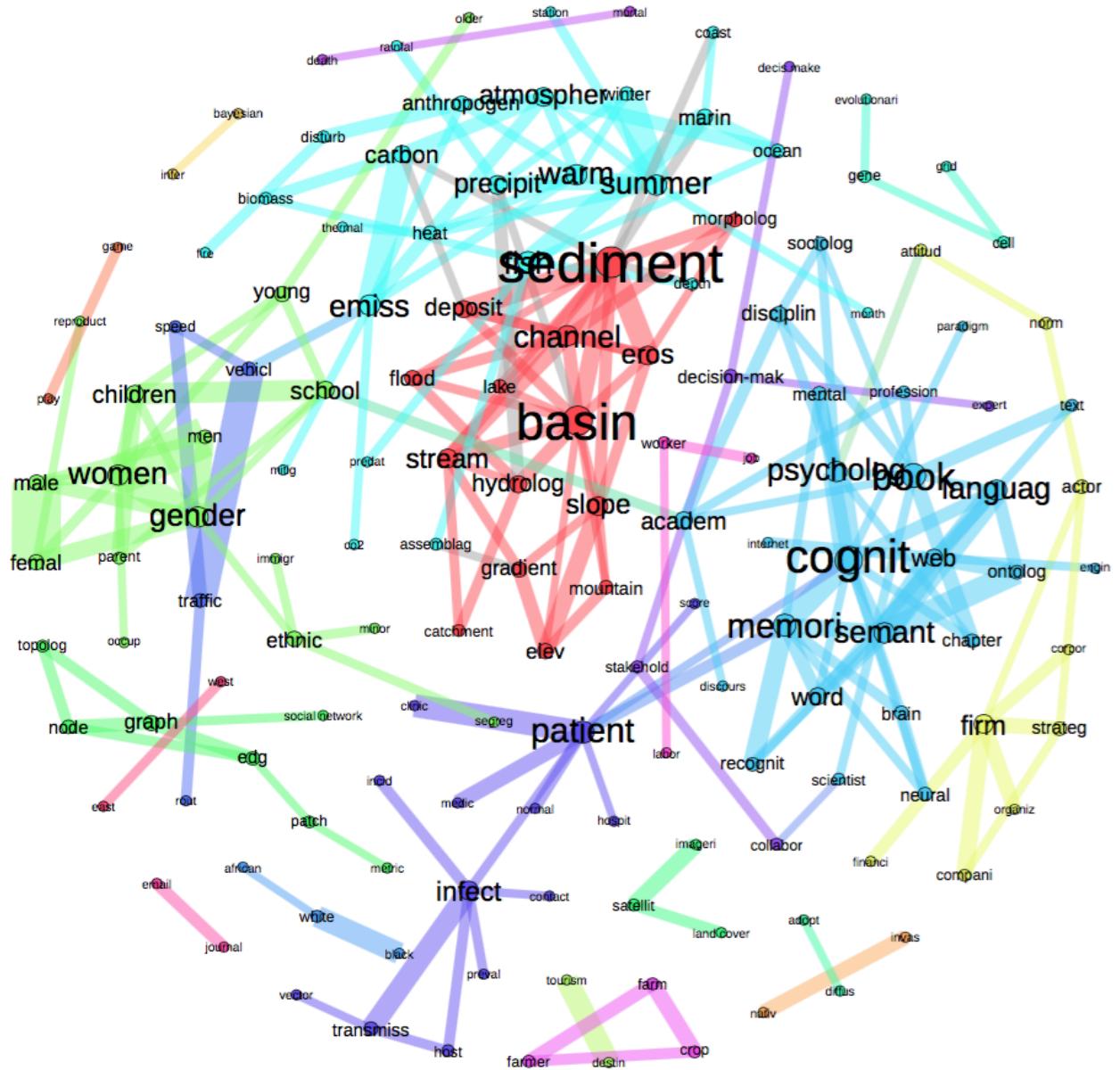


FIGURE 8 : Semantic network of concepts in quantitative geography. Corpus consists of around $2 \cdot 10^5$ abstracts of publications at a topological distance shorter than 2 from the journal *cybergeo* in the citation network. Relevance of keywords were estimated with a bootstrap method, and semantic network is constructed by co-occurrences of keywords (cut at larger degrees, 10% here to delete hubs such as *model* or *space* and efficiently reveal communities).

4.3 TOWARDS MODELING PURPOSE AND CONTEXT AUTOMATIC EXTRACTION

Vers une modélisation des thèmes et une extraction automatique du contexte

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

Depending on extended results of the two previous sections and on thematic requirements (huge need of knowledge on precise models structure, that may appear when trying to construct more specialized operational models), this project may be conducted with more or less investment.

Deuxième partie

MATERIALS

This part aims at producing knowledge from the empirical analysis of case studies and from first modeling experiments. Explicit testing of hypothesis drawn from the theory is not achieved yet as these are preliminary steps for a reasoned insight into empirical and modeling domains.

INVESTIGATING THE EMPIRICAL EXISTENCE OF STATIC USER EQUILIBRIUM

L'Equilibre Utilisateur Statique est un cadre puissant pour l'étude théorique du trafic. Malgré l'hypothèse restreignant 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.

5.1 INTRODUCTION

Introduction

La modélisation du trafic a été largement étudiée depuis les travaux séminaux de Wardrop ([wardrop1952road]) : les enjeux économiques et techniques justifient entre autre le besoin d'une compréhension fine des mécanismes régissant les flots de trafic à différentes échelles. Différentes approches aux objectifs différents co-existent aujourd’hui, parmi lesquels on trouve par exemple les modèles dynamiques de micro-simulation, généralement opposés aux techniques de basant sur l'équilibre. Tandis que la validité des modèles microscopiques a été étudiée de façon conséquente et leur application souvent questionnée, la littérature est relativement pauvre en études empiriques assurant l'hypothèse d'équilibre stationnaire du cadre de l'Equilibre Utilisateur Statique (EUS). De nombreux développements plus réalistes on été documentés dans la littérature, tels l'Equilibre Utilisateur Dynamique Stochastique (EUDS) (voir pour une description par exemple [han2003dynamic]). A un niveau inter-

médiaire entre les cadres statiques et stochastiques se trouve l'Equilibre Utilisateur Stochastique Restreint, pour lequel les choix d'itinéraire des utilisateurs sont contraints à un ensemble d'alternatives réalistes ([\[rasmussen2015stochastic\]](#)). D'autres extensions prenant en compte le comportement de l'utilisateur via des modèles de choix ont été proposé plus récemment, comme [\[zhang2013dynamic\]](#) qui inclut à la fois l'influence de la 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 [\[mahmassani1987boundedly\]](#). Dans ce cadre, l'utilisateur est satisfait si son utilité tombe dans un intervalle et l'équilibre est achevé lorsque chaque utilisateur est satisfait. Les dynamiques résultantes sont plus complexes comme révélé par l'existence d'équilibres multiples, et permet de rendre compte de faits stylisés spécifiques comme des évolutions irréversibles du réseau comme développé par [\[guo2011bounded\]](#). D'autres modèles d'attribution de trafic inspirés d'autres domaines ont également été plus récemment proposés : dans [\[puzis2013augmented\]](#), une définition étendue de la centralité de chemin qui combine linéairement la centralité des flots non-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 sur l'Equilibre Utilisateur Statique. La région parisienne utilise par exemple un modèle statique (MODUS) pour gérer et planifier le trafic. [\[leurent2014user\]](#) introduit un modèle statique de flots qui inclut les recherches locales et le choix du parking : il est légitime de s'interroger, en particulier à de si faibles échelles, si la stationnarité de la distribution des flots est une réalité. Une example d'exploration empirique des hypothèses classiques est donné par [\[zhu2010people\]](#), pour lequel les choix d'itinéraires révélés sont étudiés. Les conclusions questionnent le "premier principe de Wardrop" qui implique que les utilisateurs choisissent parmi un ensemble d'alternatives parfaitement connu. Dans le même esprit, nous étudions l'existence possible de l'équilibre en pratique. Plus précisément, l'EUS suppose une distribution stationnaire des flots sur l'ensemble du réseau. Cette hypothèse reste valable dans le cas d'une stationnarité locale, tant que l'échelle temporelle d'évolution des paramètres est considérablement plus grande que les échelles typiques de voyage. Le second cas qui est plus plausible et de plus compatible avec les cadres théoriques dynamiques est testé ici.

La suite de ce travail s'organise ainsi : la procédure de collection de données ainsi que le jeu de données sont décrits ; nous présentons en-

suite 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.

5.2 DATA COLLECTION

Collecte des données

5.2.1 *Dataset Construction*

Nous proposons de travailler sur l'étude de cas de la région métropolitaine de Paris. Un jeu de données ouvert a été construit, comprenant les liens autoroutiers dans la région, par collecte des données publiques en temps réel des temps de parcours (disponible sur www.sytadin.fr). Comme rappelé par [**bouteiller2013open**], la disponibilité de jeux de données ouverts pour les transports est loin d'être la règle, et nous contribuons ainsi à une ouverture par la construction de notre jeu de données. La procédure de collecte de données consiste en les points suivants, exécutés toutes les deux minutes par un script python :

- récupération de la page web brute donnant les informations de trafic
- parsing du code html afin de récupérer les identifiants des liens de trafic et les temps de parcours correspondants
- insertion des liens dans une base sqlite avec le temps courant.

Le script automatisé de collection des données continue d'enrichir la base au fur et à mesure du temps, permettant des développements futurs de ce travail sur un jeu de données plus large, et une réutilisation potentielle pour des travaux scientifiques ou opérationnels. La dernière version du jeu de données au format sqlite est disponible en ligne sous une Licence *Creative Commons*¹.

5.2.2 *Data Summary*

Une granularité de deux minutes a été obtenue pour une période de trois mois (de février 2016 à avril 2016 inclus). La granularité spatiale est en moyenne de 10km, les temps de trajet étant fournis pour les liens majeurs. Le jeu de données contient 101 liens. La variable brute utilisée est le temps de trajet effectif, à partir duquel il est possible

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

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 par inversion d'un fonction BPR 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.

5.3 METHODS AND RESULTS

Méthodes and Résultats

5.3.1 *Visualization of spatio-temporal congestion patterns*

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 [grimm2005pattern], nous nous intéressons au motifs macroscopiques à des échelles temporelles et spatiales données : d'une manière équivalente aux faits stylisés qui sont dans cette approches extraits d'un système avant de tenter de le modéliser, nous devons explorer les données de manière interactive dans le temps et l'espace afin d'identifier des motifs pertinents et les échelles associées. Une application web interactive a ainsi été implémentée pour explorer les données, à l'aide des packages R `shiny` et `leaflet`². 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 9 présente une capture d'écran de l'interface. La conclusion majeure de l'exploration interactive des données est qu'une grande hétérogénéité spatiale et temporelle est la règle. Le motif temporel le plus récurrent, la périodicité journalière des heures de pointe, est perturbée pour une proportion non négligeable de jours. En première approximation, les heures creuses peuvent être approchées par une distribution localement stationnaire des flots, tandis que les heures de pointe sont trop courtes pour pouvoir impliquer la validation de l'hypothèse d'équilibre. Concernant l'espace, aucun motif spatial particulier n'émerge clairement. Cela signifie que dans le cas d'une validité de l'équilibre utilisateur statique, les méta-paramètres régissant son établissement doivent varier à des échelles temporelles plus courtes qu'un

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

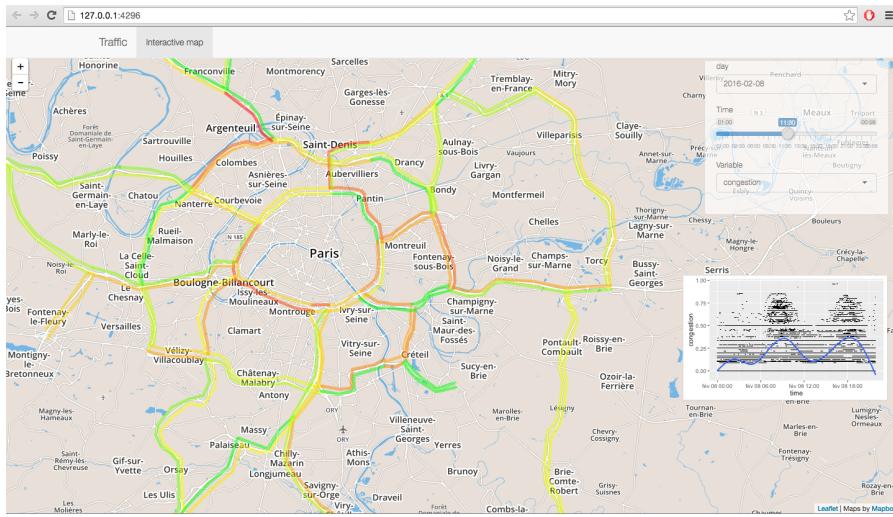


FIGURE 9

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.

5.3.2 Spatio-temporal Variability of Travel Path

A la suite de l'exploration interactive des données, nous proposons de quantifier la variabilité spatiale des motifs de congestion pour valider ou invalider l'intuition que si l'équilibre existe par rapport au temps, il est fortement dépendant de l'espace et localisé. La variabilité spatio-temporelle des plus courts chemins de trajet est une première façon d'étudier la stationnarité des flots d'un point de vue de théorie des jeux. En effet, l'Equilibre Utilisateur Statique est la distribution stationnaire des flots sous laquelle aucun utilisateur ne peut augmenter son temps de trajet en changeant son itinéraire. Une forte 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 10 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

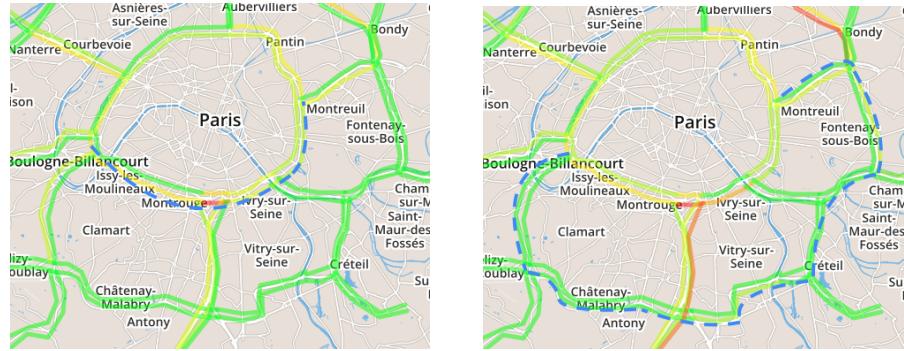


FIGURE 10

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.

5.3.3 *Stability of Network measures*

La variabilité des trajectoires potentielles observée dans la section précédente peu être confirmée par l'étude de la variabilité des propriétés du réseau. En particulier, les mesures topologiques de réseau capturent les motifs globaux dans un réseau de transport. Les mesures de centralité et de connectivité des noeuds sont des indicateurs classiques pour la description des réseaux de transport comme rappelé par [bavoux2005geographie]. La littérature en transports a développé des mesures de réseau élaborées et opérationnelles, comme des mesures de robustesse pour identifier les liens critiques et mesurer la résilience globale du réseau aux perturbations (un exemple parmi d'autres est l'indice de *Robustesse du Réseau Effective* introduit dans [sullivan2010identifying]).

Plus précisément, nous étudions la centralité de chemin du réseau de transport, défini pour un noeud comme le nombre de plus courts chemins passant par celui-ci, i.e. par l'équation

$$b_i = \frac{1}{N(N-1)} \cdot \sum_{o \neq d \in V} \mathbb{1}_{i \in p(o \rightarrow d)} \quad (7)$$

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.

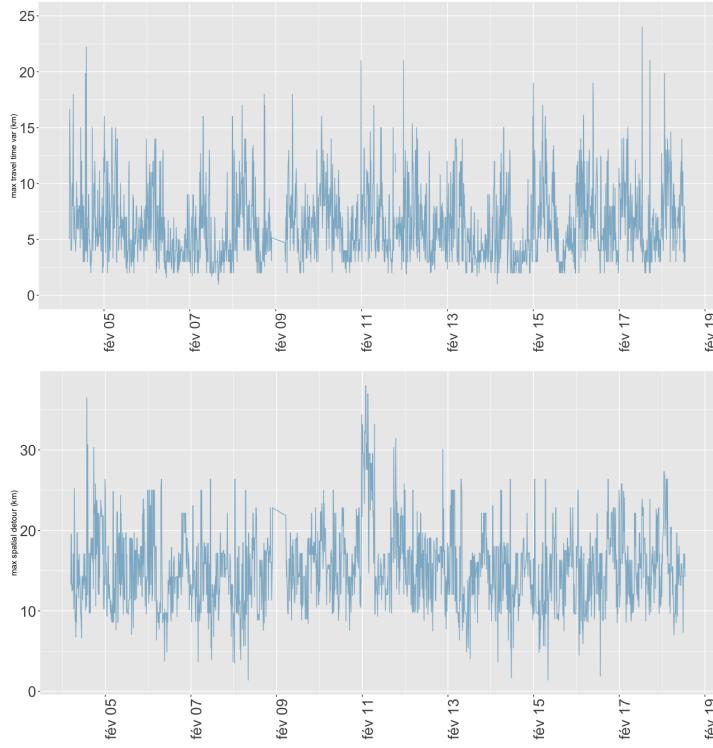


FIGURE 11

Nous montrons en Figure 4 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

$$\Delta b(t) = \frac{|\max_i(b_i(t + \Delta t)) - \max_i(b_i(t))|}{\max_i(b_i(t))} \quad (8)$$

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 12) implique dans le cas d'une variation négative, qu'au moins cette proportion de trajectoires potentielles ont changé et que la potentielle congestion locale a décrue de la même proportion. Dans le cas d'une variation positive, un seul noeud a capturé au moins 20% des trajets. Sous l'hypothèse (qu'on ne tente pas de vérifier ici et qu'on peut également supposer non vérifiée comme montré par [zhu2010people]), mais que l'on utilise comme un outil pour donner une intuition sur la signification concrète de la variabilité de la centralité) que les utilisateurs choisissent rationnellement le plus court chemin, et supposant que la majorité des trajets est réalisées, une telle variation de la centralité implique une variation similaire dans les flots effectifs, conduisant

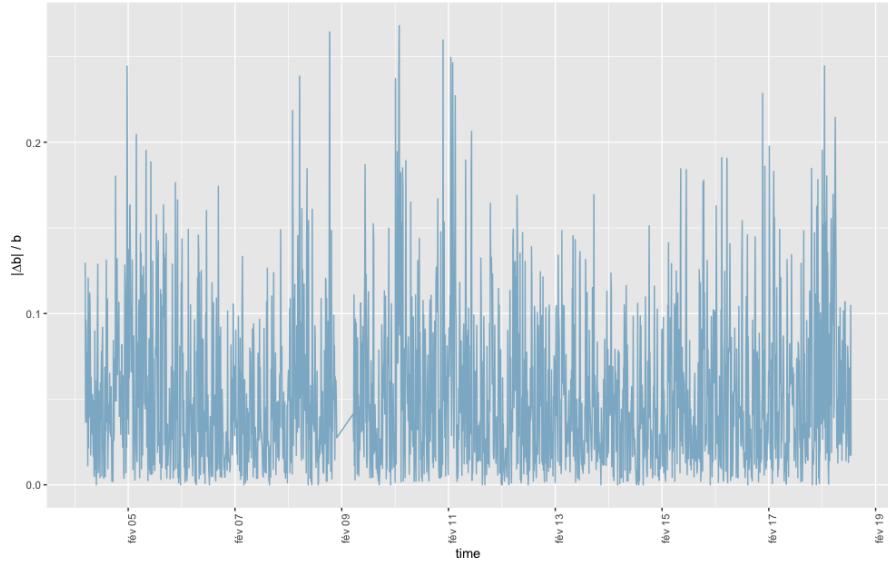


FIGURE 12

à la conclusion qu'ils ne peuvent être stationnaires ni dans le temps (au moins sur une échelle plus grande que Δt) ni dans l'espace.

5.3.4 Spatial heterogeneity of equilibrium

Afin d'obtenir un point de vue différent sur la variabilité spatiale des motifs de congestion, nous proposons d'utiliser un indice d'auto-corrélation spatiale, l'indice de Moran (défini par exemple dans [\[tsai2005quantifying\]](#)). 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 trafic localisée. A un point donnée de l'espace, l'auto-corrélation locale pour la variable c est calculée par

$$\rho_i = \frac{1}{K} \cdot \sum_{i \neq j} w_{ij} \cdot (c_i - \bar{c})(c_j - \bar{c}) \quad (9)$$

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 13 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.

5.4 DISCUSSION

Discussion

5.4.1 *Theoretical and practical implications of empirical conclusions*

Nous prétendons 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 taille et variés, et pour des objectifs d'application différents (prédition du traffic, reproduction de faits stylisés, etc.), à l'image des revues systématiques qui sont la règle en évaluation thérapeutique par exemple. Cela implique cependant des pratiques de partage des données et des modèles plus larges que celles existant couramment. La connaissance précise des potentialités d'application d'un cadre donné peut induire des développements inattendus comme l'intégration dans des modèles plus larges. L'exemple des études des interaction entre Transport et Usage du Sol (modèles *LUTI*) est une bonne illustration d'un cas où le EUS peut toujours être utilisé avec des motivations plus larges que la modélisation du traffic. [kryvobokov2013comparison] décrit deux modèles *LUTI*, dont l'un inclut deux équilibres pour les

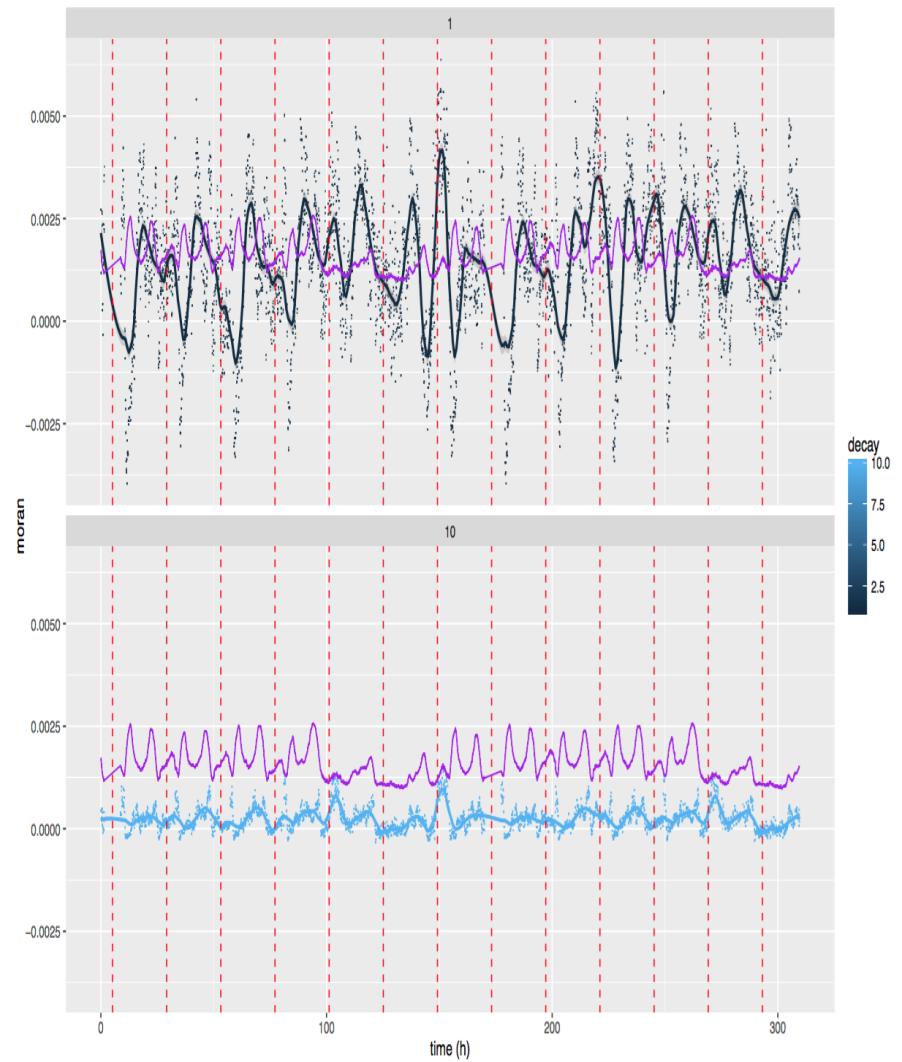


FIGURE 13

modèles de transport à quatre temps et pour l'évolution de l'usage du sol (localisation des ménages et emplois), l'autre étant dynamique. La conclusion est que chaque modèle a ses avantages au regard de l'objectif poursuivi, et que le modèle statique peut être utilisé pour comparer des politiques sur le temps long, tandis que le modèle dynamique fournit de l'information plus précise à de plus petites échelles temporelles. Dans le premier cas, un module de transport plus compliqué aurait été plus difficile à inclure, ce qui est un avantage du EUS dans ce cas.

Concernant les applications pratiques, il semble naturel que les modèles statiques ne devraient pas être utilisés pour la prédiction et la gestion du trafic sur de petites échelles temporelles (semaine ou jour) et que des efforts doivent être faits pour implémenter des modèles plus réalistes. Cependant, l'utilisation des modèles par la communautés des ingénieurs et des planificateurs n'est pas directement reliée aux enjeux académiques et à l'état de l'art dans le domaine. Dans le cas particulier de la France et des modèles de mobilité, [[commenges2013invention](#)] a montré que les ingénieurs allaient jusqu'au point de construire des problèmes inexistant et d'implémenter les modèles correspondants qu'ils avaient importé d'un contexte géographique totalement différent (la planification aux Etats-Unis). L'utilisation d'un cadre ou d'un type de modèle a des raisons historiques qui peuvent être difficiles à surmonter.

5.4.2 *Towards explanatory interpretations of non-stationarity*

Une hypothèse qu'on peut formuler concernant l'origine de la non-stationnarité des flots dans le réseau, au regard de l'exploration des données et des analyses quantitatives, est que le réseau est au moins la moitié du temps fortement congestionné et dans un état critique. Les heures creuses sont les plus grandes fenêtres temporelles potentielles de stationnarité spatiale et temporelle, mais couvre moins de la moitié du temps. Comme déjà interprété dans le comportement de l'indicateur d'auto-corrélation, un comportement chaotique pourrait être à l'origine d'une telle variabilité lors des heures congestionnées. A la manière d'un fluide supercritique qui condense sous une perturbation externe infinitésimale, l'état d'un lien peut qualitativement changer par un petit incident, produisant une perturbation du réseau qui se propage et peut même s'amplifier. L'effet direct des événements du trafic (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.

5.4.3 *Possible developments*

Le travail futur pourra être planifié dans la direction d'une étude raffinée 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 [goldhirsch1987stability]. La valeur des exposants de Liapounov fournit l'échelle de temps sur laquelle le système instable s'éloigne de l'équilibre. Leur comparaison avec la durée des heures de pointe et le temps de trajet moyen, sur différentes zones spatiales et différentes échelles, devrait fournir plus d'information sur une possible validité de l'hypothèse de stationnarité locale. Cette technique a déjà été introduite à une autre échelle dans les études de transport, comme e.g. [tordeux2016jam] qui étudie la stabilité des modèles de régulation de vitesse à l'échelle microscopique pour éviter l'émergence de congestion.

D'autres directions de recherche peuvent consister en le test des autres hypothèses du EUS (comme le choix rationnel du plus court chemin, qui serait cependant difficile à tester à un tel niveau d'agrégation, impliquant l'utilisation de modèles de simulation calibrés et cross-validés sur le jeu de données pour comparer différentes hypothèses, sans toutefois nécessairement une validation ou invalidation directe de l'hypothèse), ou le calcul empirique des paramètres dans les cadres d'Equilibre Utilisateur Stochastique ou Dynamique.

5.5 CONCLUSION

Nous avons décrit une étude empirique ayant pour but une étude simple, mais selon notre point de vue nécessaire, de l'existence de l'équilibre utilisateur statique, plus précisément de sa stationnarité dans le temps et l'espace pour un réseau routier métropolitain principal. Un jeu de données de congestion du trafic est construite par collection de données, pour le réseau du Grand Paris sur 3 mois avec une granularité temporelle de 2 minutes. L'exploration interactive du jeu de données via une application web permettant la visualisation spatio-temporelle aide à guider les analyses quantitatives. La variabilité spatio-temporelle des plus courts chemins et de la topologie du réseau, en particulier la centralité de chemin, révèle que l'hypothèse de stationnarité ne tient généralement pas, ce qui est confirmé par l'étude de l'auto-corrélation spatiale de la congestion du réseau. Nous suggérons que nos résultats soulignent un besoin général de

plus grandes connexions entre les études théoriques et empiriques, puisque cette étude permet de chasser les incompréhensions théoriques sur l'Equilibre Utilisateur Statique, et guider le choix d'application potentielles.

6

EMPIRICAL ANALYSIS : INSIGHTS FROM STYLIZED FACTS

*Mais ce n'est pas une question
d'âge, de chiffres et de stats
Moi je te parle surtout de rage, de kif
et d'espoir*

- YOUSSEOPHA , Esperance de Vie

As this quote suggests, a purely quantitative view of the world makes no sense without qualitative counterbalancing. More precisely, we argue that the *cliché* of an opposition between quantitative and qualitative analysis is an illusion. No distinct boundary exists between both. We propose to call quantitative any process involving computation by a Turing machine, whereas the qualitative will be for us the modeling design process and its interpretations. Therefore both are necessarily closely interlaced in any of our approaches. In particular concerning the construction and the validation or refutation of our theory, empirical analysis on real case studies, implying the extraction and qualification of stylized facts, follows that schema.

We propose in this chapter various empirical analysis on different objects at different scales. A first section begins the examination of static spatial correlations between morphological measures of population density and road network measures on Europe at a 500m resolution. Applying last section of the methodological chapter should provide information on typical spatial scales of interaction between these indicators of territory and network and on dynamical correlations between these. These computation furthermore provide empirical measures on which one model will be calibrated. We then describe a roadmap for statistical analysis on dynamical data of interactions for Bassin Parisien in the last fifty years. An other project using Real Estate transaction data for Parisian Metropolitan Region aim at seeking early warning of network breakdowns. We finally describe potential analyses on South African historical data.

6.1 STATIC CORRELATIONS OF URBAN FORM AND NETWORK SHAPE

Corrélations Statiques entre Forme Urbaine et Forme de Réseau

Spatio-temporal processes implying diffusion or propagation phenomena generally have a specific structure of correlation. In particular, as derived in section 2.5, a static computation of correlation between different instances of a system may under certain conditions provide information on dynamical correlations implied.

6.1.1 Morphological Measures of European Population Density

Context

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 (voir e.g. l'ontologie des modèles Simpop [**pumain2012multi**] ou de leur successeur Marius [**cottineau2014evolution**]). 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.

Empirical Analysis

We study systematically morphological indicators for constant size areas covering European Community. The choice of fixed size areas can be questioned regarding definition of a territorial system, that can be otherwise understood as a consistent spatial entity at a given scale and along certain criteria : *Human territories* as defined by Raffestin (op. cit.) or more generally functionally autonomous spaces¹. Here we choose the mesoscopic scale of a metropolitan center ($\simeq 50\text{km}$) for comparability purposes and because greater scale are no more relevant regarding urban form, whereas smaller scales must contain too much noise.

Data is the European population density grid [**eurostat**] and indicators computation is implemented in parallel using R with Fast convolution raster functions. We show in next figures computed values of morphological indicators (see [**le2015forme**] for a precise formulation

¹ for example, a tentative of definition of a *Parisian* territory would present many facets. From the subjective territory point of view, intra-muros Parisians consider a strict boundary at *Boulevard Peripherique*, whereas close and even further suburbs will be seen as Parisians from the Province. The functional territory of *Metropolitain* extends slightly further than the administrative boundary. Governance perimeters are currently mutating with the Metropolitan governance project. Complementary perceptions of the territory can thus be multiplied.

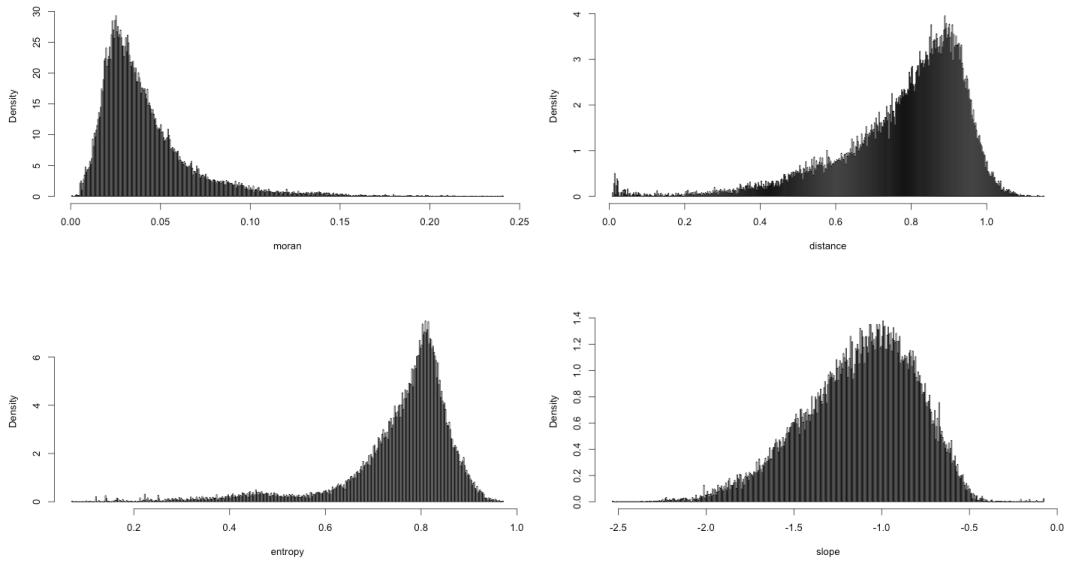


FIGURE 14 : Empirical Distribution of Morphological Indicators

of indicators that are Moran index, average distance, entropy and hierarchy).

Further developments

In [10.1371/journal.pone.0107042] density grids for other countries across the world (ex. China) are provided² so we may repeat our analysis to other regions for comparison purposes.

6.1.2 Network Measures

We consider network aggregated indicators as a way to characterize transportation network properties on a given territory, the same way morphological indicators yielded information on urban structure. We propose to compute some simple indicators on same extents as for morphology, to be able to explore relations between these static measures. Static network analysis has been extensively documented in the literature, see [louf2014typology] for a cross-sectional study of cities or [2015arXiv151201268L] for exploration of new measures for the road network.

Data preprocessing

We work in a first time on road network, which structure is finely conditioned to territorial configuration of population densities. Furthermore, data for present day road network is available through

² available at <http://www.worldpop.org.uk/>

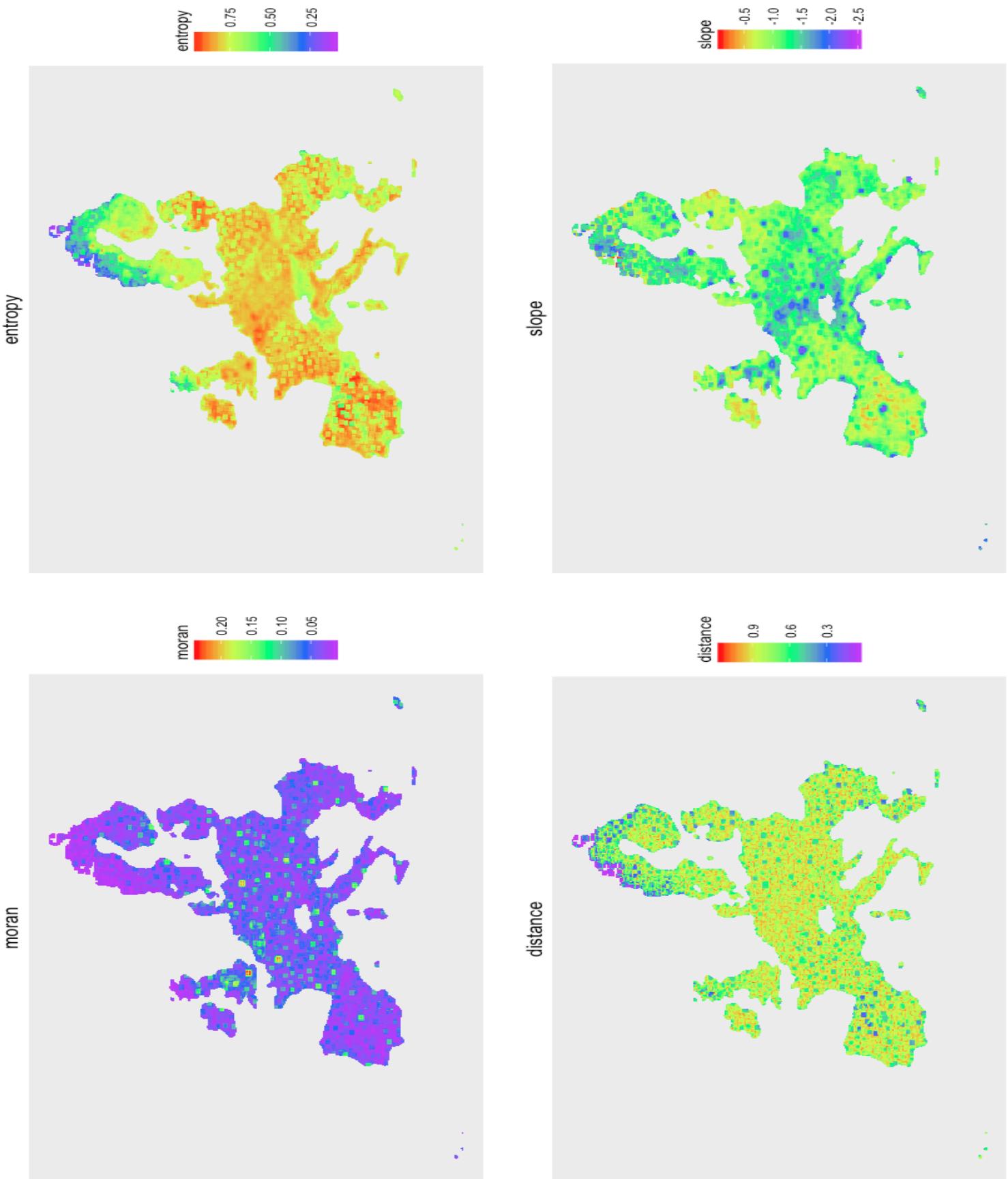
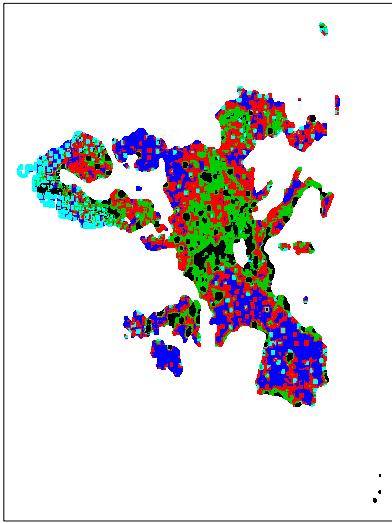
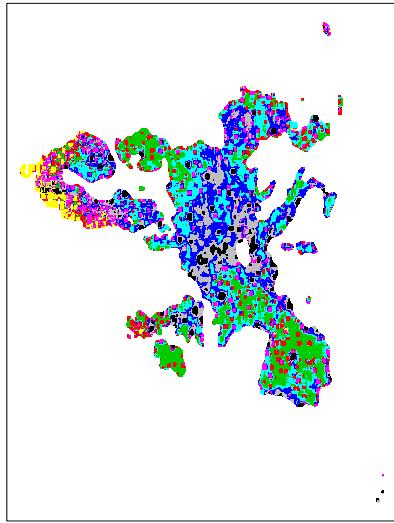


FIGURE 15 : Geographical Distribution of Morphologies : value of indicators across Europe.

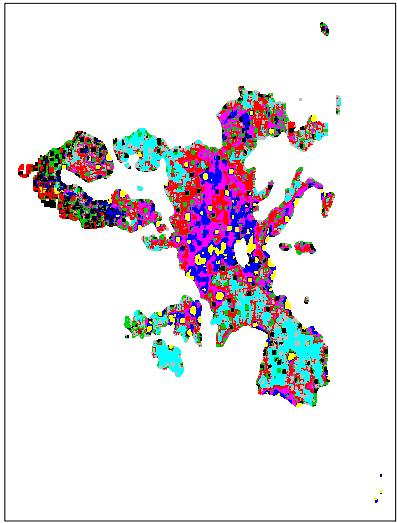
k=5 ; withinProp=0.258568287232286



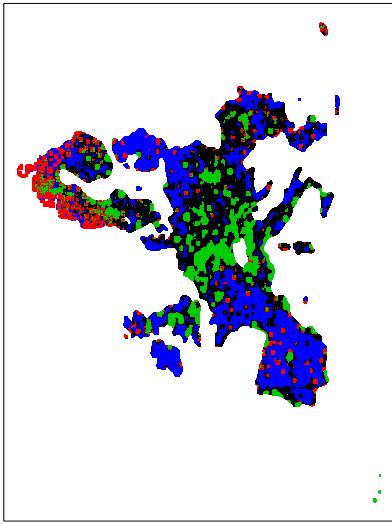
k=8 ; withinProp=0.179550720275406



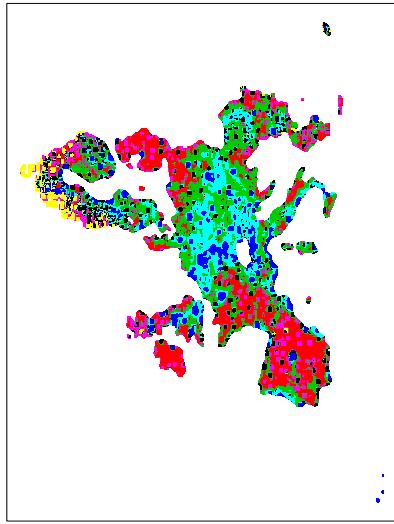
k=11 ; withinProp=0.148765308222633



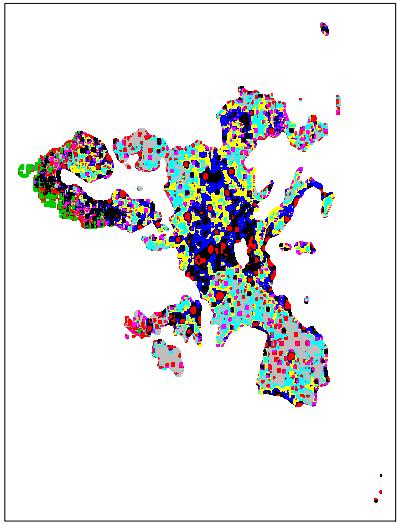
k=4 ; withinProp=0.304934256837235



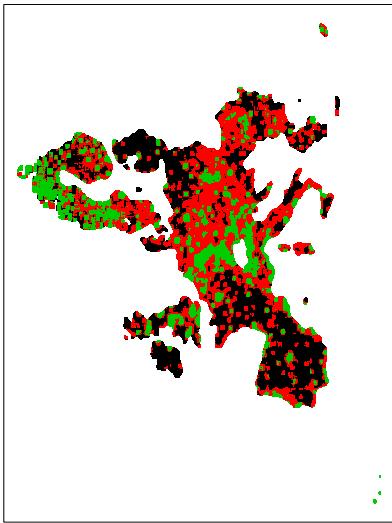
k=7 ; withinProp=0.20159558807077



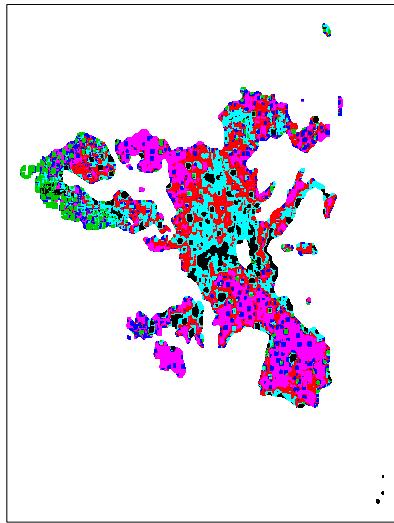
k=10 ; withinProp=0.156221620124614



k=3 ; withinProp=0.379727175801079



k=6 ; withinProp=0.224023913068682



k=9 ; withinProp=0.167033521857729

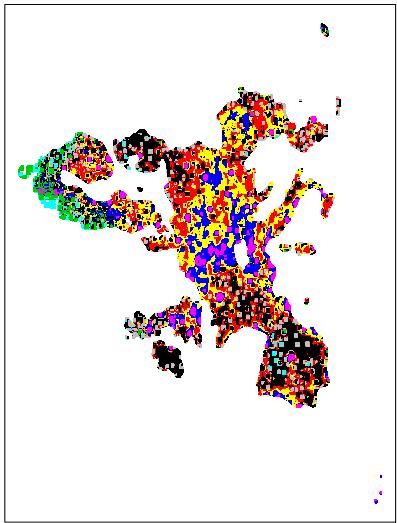


FIGURE 16 : Clustering Analysis of Morphologies. We present the results of an average k-means for different values of

the OpenStreetMap project [[openstreetmap](#)]. Its quality was investigated for different countries such as England [[haklay2010good](#)] and France [[girres2010quality](#)]. It was found to be of a quality equivalent to official surveys for the primary road network.

SIMPLIFICATION ALGORITHM For a given dataset corresponding to a subset of the overall road network, it is necessary to simplify network structure by spatial aggregation as initial data presents very detailed features and thus a very large numbers of nodes ($\simeq 10^{10}$ for Europe dataset). Such a level of precision is not needed in our study since density data is already aggregated at 500m resolution. It is possible to drastically reduce network size by spatial aggregation of nodes and link replacements. 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.

IMPLEMENTATION 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.

SENSITIVITY TO SIMPLIFICATION PARAMETERS Sensitivity of indicators to raster resolution and to degree simplification algorithm must still be tested to ensure the relevance of data preprocessing.

Indicators

Network macroscopic structure is summarized by the following set of indicators, after the simplifications and reductions done in the previous step. Assuming network given by $N = (V, E)$, nodes spatial positions $\vec{x}(V)$ and edges *effective distances* $d(E)$ taking into account impedances and real distances (to include basically network hierarchy), we have indicators :

- connectivity
- degree distribution
- centrality, taken as normalized mean *betweenness-centrality*
- average path length
- network diameter
- mean network speed

These indicators are used to capture a rough picture of the structure. Refined work at smaller scales (intra-urban road network) and with more elaborated measures that allow to differentiate more precisely local form, was recently done by Lagesse in [2015arXiv151201268L].

Results

Les indicateurs de réseau ont été calculés sur des zones similaires aux indicateurs de forme urbaine,

6.2 DISENTANGLING CO-EVOLUTIONS FROM CAUSAL RELATIONS : A CASE STUDY ON *bassin parisien*

Isoler la Co-évolution des Relations causales

Spatial statistics studies on dynamical relations between network and territories are relatively rare. [levinson2008density] does so on London metropolitan area and identifies causalities using lagged variables, but does not disentangle relations in the sense of coupled statistical models that would isolate endogenous effects from coupling effects.

6.2.1 *Context Formalization*

We assume a dynamic transportation network $n(\vec{x}, t)$ within a dynamic territorial landscape $\vec{T}(\vec{x}, t)$, which 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

6.2.2 *On Accessibility*

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 [weibull1976axiomatic], 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 ; 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).

6.2.3 Data

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.

6.2.4 Statistical Tests

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

- On raw data :
 - Multivariate models

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

- Autocorrelated univariate models

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

- Autocorrelated multivariate models

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

- Geographically Weighted Regression [brunsdon1998geographically]

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

- Granger causality tests : [xie2009streetcars] 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?

6.2.5 Méthode Générique

Description

Nous décrivons ici une méthode générique, basée sur un test similaire à la causalité de Granger [], pour tenter d'identifier des relations causales dans des systèmes spatiaux. Soit $X_j(\vec{x}, t)$ des processus aléatoires spatiaux unidimensionnels. Une réalisation d'un sous-système territorial est donnée par des ensembles de trajectoires pour chaque processus $x_{i,j,t}$. On suppose l'existence de fonctions de correspondance $\Phi_{j1,j2}$ permettant de faire correspondre les réalisations de chaque composantes à un index unique (dans le cas le plus simple, on associera les variables sur les mêmes patches). Si $\text{argmax}_{\tau} \hat{\rho} [x_{j1}, x_{j2}]$ est clairement défini, son signe donnera alors le sens de la causalité entre les composantes $j1$ et $j2$.

Données Synthétiques

CASE STUDY Cette méthode doit dans un premier temps être testée et partiellement validée, ce que nous proposons de faire sur des données synthétiques, approche dont l'utilisation est documentée et illustrée au chapitre ???. [raimbault2014hybrid] est un modèle simple de morphogénèse urbaine (modèle RBD) faisant un candidat intéressant pour notre test. En effet, les variables explicatives de la croissance urbaine, les processus d'extension du réseau et le couplage entre densité urbaine et réseau sont assez élémentaires. Cependant, hormis dans des cas extrêmes (distance au centre détermine valeur foncière uniquement, le réseau dépendra de manière causale de la densité, ou distance au réseau seule, la causalité devrait être 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 explorons une grille de l'espace des paramètres du modèle RBD. Pour chaque valeur des paramètres, nous procérons à $N =$ répétitions.

6.3 EARLY WARNINGS OF NETWORK BREAKDOWNS : SOCIO-ECONOMIC AND REAL ESTATE TRAJECTORIES

Trajectoires de Marchés Immobiliers

6.3.1 *Context*

Des aspects très variés des territoires sont concernés par l'interaction avec les réseaux. Dans nos études précédentes, aucun aspect socio-économique des populations habitant le territoire ni des valeurs économiques pour le foncier et l'immobilier n'ont été considérés. Il s'agit cependant d'éléments cruciaux des dynamiques territoriales et sont étudiés de manière intensive dans des champs comme l'analyse territoriale ou l'économie urbaine : par exemple, [homocianu:tel-00359302] étudie les choix résidentiels des ménages pour comprendre les interactions entre usage du sol et transport. Nous proposons ici d'utiliser une base de données de transactions immobilières pour la région parisienne sur les 20 dernières années, avec une granularité temporelle de 2 ans et coordonnées spatiales exactes. [guerois2009dynamique] l'utilise pour établir une typologie des dynamiques spatiales du marché immobilier parisien.

6.3.2 *Preliminary Results*

We show in Fig. 17 typologies of temporal transactional profiles for total stocks. Temporal dynamics show different reactions of local territories to the 2008 crisis, in particular a strong differentiation between urban and rural areas. More precise classification into urban territories are still to be investigated when the analysis will be pushed further.

6.3.3 *A strategy to investigate early warnings of network breakdowns*

The span of the end of this database coincides with planification phases of the Grand Paris Express that we already mentioned. We aim to seek for early warnings of potential station implantation, in correspondance with different stages of the project, in order to verify if intrinsic territorial dynamics were already present or if the announcement of a new station induced a local phase transition.

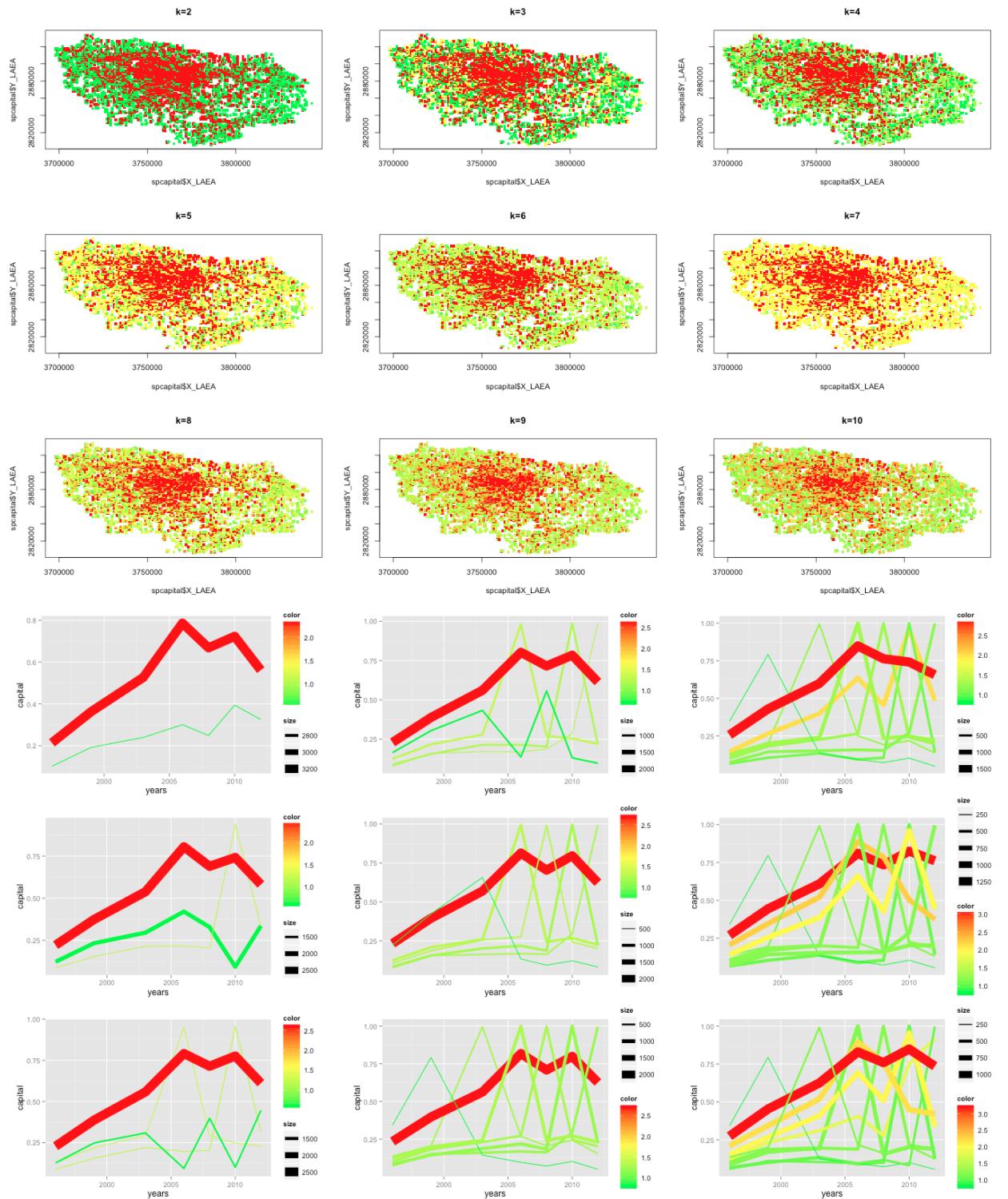


FIGURE 17 : Typology of Real Estate trajectories. Locations were categorized using averaged k-means on time-series. We show maps and time series for value of k from 2 to 10.

6.4 SOUTH-AFRICAN HISTORICAL EVENTS AS INSTRUMENTS TO UNDERSTAND NETWORK-TERRITORY RELATIONS

Relations Réseaux-territoires en Afrique du Sud

6.4.1 *Context*

BAFFI studied in her thesis project [**baffi2016thesis**] qualitatively the role of South African railways in segregations and integration processes, aims to use an extensive database of railway growth and population dynamics in cities on the last 100 years produced during the thesis. In particular, she showed qualitatively that dynamics between territories and networks profoundly changed at the end of the apartheid, transforming a tool of sordid planned segregation (network shaped was optimized to minimize unwanted accessibility) into an integration tool thanks to recent changes in network topology patterns.

6.4.2 *Objectives*

We can use first the particular shape of that network to control on local and global topology effects (but this is quite equivalent as controlling on accessibility), and in a second time the historical events as statistic instruments, assuming that territorial dynamics and network dynamics responded differently to these. We expect to learn from these project informations on interactions at long time scale and large spatial scale, in a very particular context of constrained growth.

6.4.3 *Possible developments*

The method of instruments in statistics [**angrist1996identification**] is used to identify causal relationships between variables, in a different way than Granger causality test for example. Trying to identify causalities between network dynamics and territorial dynamics is of crucial importance to test our theoretical assumption on the existence of co-evolution.

Do or do not. There is no try.

- YODA

One does not simply *try* to model something. On that point personal experience confirms indeed that point, as I remember as an early Master student giving in to the call of incautious agent-based modeling, naively thinking that integrated models of any aspect of an urban system could be constructed, producing numerous NetLogo code lines to build a gaz factory with unfounded internal processes, an extremely poor external validation and no internal validation. This was a try and therefore a step towards the dark side of models bricolage. The construction of a computational model of simulation is a rigorous exercise that one can not improvise, as much as statistical modeling. Recent progresses in the field [**banos2013pour**] help to that purpose, and modular model construction and validation is one tool useful to avoid becoming lost in shady places.

We propose in this chapter simple modeling experiments, conceived to be preliminaries for more elaborated tests of our theory. We begin with a simple diffusion-aggregation model of urban growth as a relatively small scale. Beginning with simple assumptions does not mean a non-rigorous exploration of the model, that is therefore explored and calibrated on real data. The fact that we reproduce existing urban forms without the use of networks suggest either the total absence of network influence at this scale, or its very strong influence yielding apparent random effects that disappear in average calibration. We propose then to simply couple this model with a network generation heuristic in order to study feasible correlations between morphology and network. The absence of coupled calibration avoids to draw empirical conclusion but the method is satisfying in itself as it permits the generation of synthetic territorial configurations where correlation structure is controlled. We finally describe a project of benchmark of diverse heuristic models for network generation.

7.1 A SIMPLE MODEL OF URBAN GROWTH

Un modèle simple de croissance urbaine

We propose a stochastic model of urban growth that generates spatial distributions of population densities, at an intermediate scale between economic models at the macro scale and land-use evolution models focusing on local relations. Integrating simply the two opposite key processes of aggregation (“preferential attachment”) and diffusion (urban sprawl), we show that we can capture the whole spectrum of existing urban forms in Europe. An extensive exploration and calibration of the proposed model allows determining the region of parameter space corresponding morphologically to observed European urban systems, providing an validated thematic interpretation to model parameters, and furthermore determining the effective dimension of the urban system at this scale regarding morphological objectives.

7.1.1 *Context*

[andersson2002urban] propose a micro-based model of urban growth, with the purpose to replace non-interpretable physical mechanisms with agent mechanisms, including interactions forces and mobility choices. Local correlations are used in [makse1998modeling] to modulate growth patterns to ressemble real configurations. In the same spirit, our model situates at similar scales and can be qualified as a morphogenesis model.

7.1.2 *Model Description*

RATIONALE Our model is an extension of the diffusion-limited aggregation model studied in [batty2006hierarchy]. Indeed, the tension between antagonist aggregation and sprawl mechanisms may be an important process in urban morphogenesis. [fujita1996economics] opposes centrifugal forces with centripetal forces in the equilibrium view of urban spatial systems, what is easily transferable to non-equilibrium systems in the framework of self-organized complexity : a urban structure is a far-from-equilibrium system that has been driven to this point by this opposite forces. The two contradictory processes of urban concentration and urban sprawl are captured by the model, what allows to reproduce with a good precision a large number of existing morphologies. A generalization of the basic model is proposed in [raimbault2016calibration].

SETTINGS The model D proceeds iteratively the following way. An square grid of width N , initially empty, is represented by popula-

tion $(P_i(t))_{1 \leq i \leq N^2}$. At each time step, until total population reaches a fixed parameter P_m ,

- total population is increased of a fixed number N_G (growth rate), following a preferential attachment such that

$$\mathbb{P}[P_i(t+1) = P_i(t) + 1 | P(t+1) = P(t) + 1] = \frac{(P_i(t)/P(t))^\alpha}{\sum (P_i(t)/P(t))^\alpha}$$

- a fraction β of population is diffused to four closest neighbors is operated n_d times

Indicators

Indicators to qualify model outputs are morphological measures of population density, proposed in [le2015forme], that are entropy, hierarchy, spatial auto-correlation, mean distance.

7.1.3 Results

The model was implemented in a first time in NetLogo for exploration purpose, later in scala for performance reasons and easy integration into OpenMole [reuillon2013openmole] for HPC model exploration.

Generation of urban patterns

The model as few parameters but is able to generate a very wide variety of shapes, extending beyond existing forms. In particular, its dynamical nature allows through P_m parameter to choose final regime that can be non-stationarity (generally chaotic shapes), semi-stationarity or total stationarity. Fig. 18 shows examples of generated shapes.

Model Behavior

CONVERGENCE - INTERNAL MODEL VALIDATION Indicators show good convergence property and bimodal statistical distribution for cumulated points in the parameter space confirm the existence of superposed regimes : gaussian distribution gives stationary configurations, whereas inverse log-normal distribution are close to real data shape and correspond to non-stationary regime. For one point and a large number of repetitions, we find that 50 repetitions are enough to obtain a 95% confidence interval smaller than σ around indicator mean.

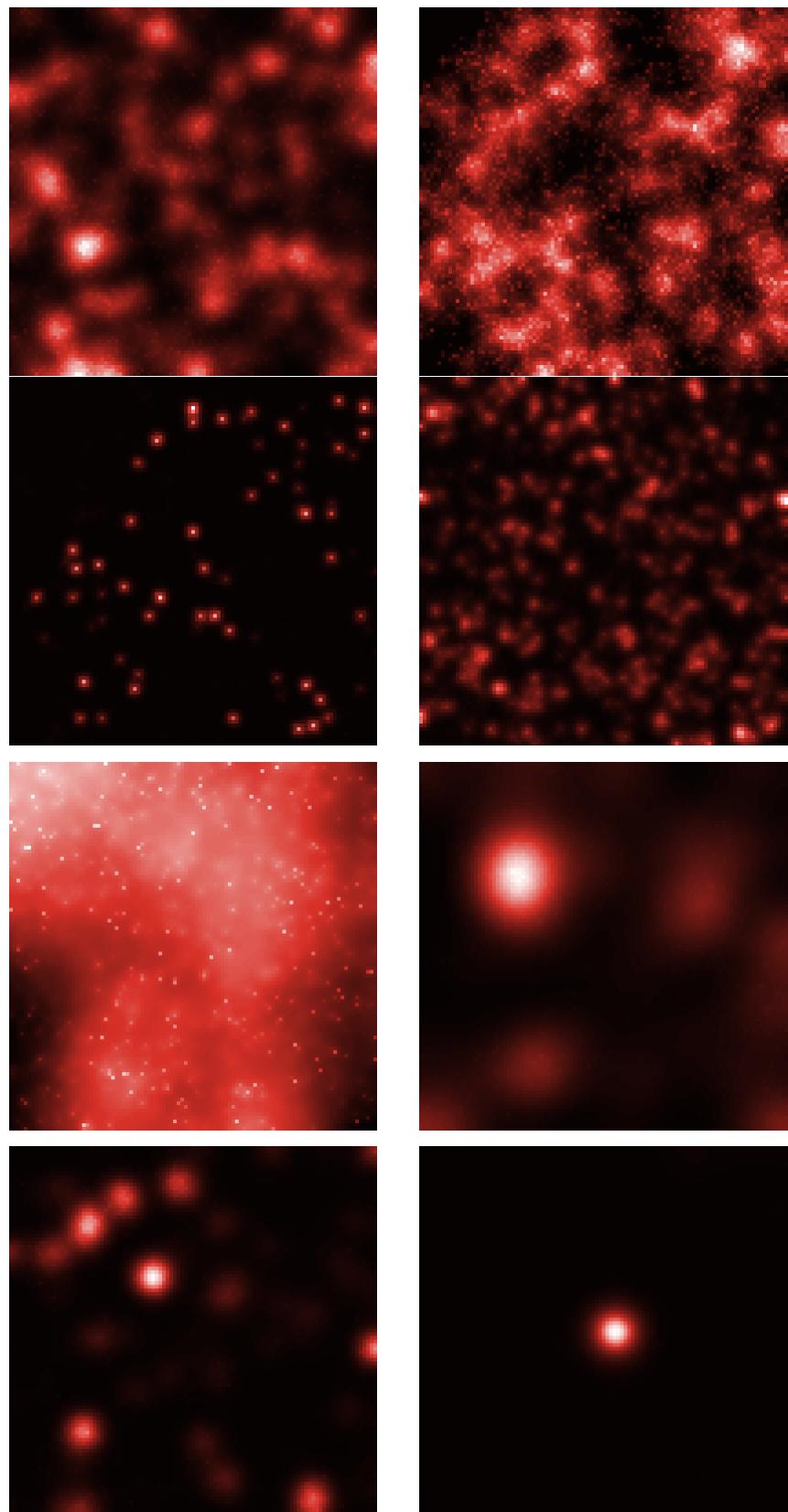


FIGURE 18 : Example of the variety of generated urban shapes

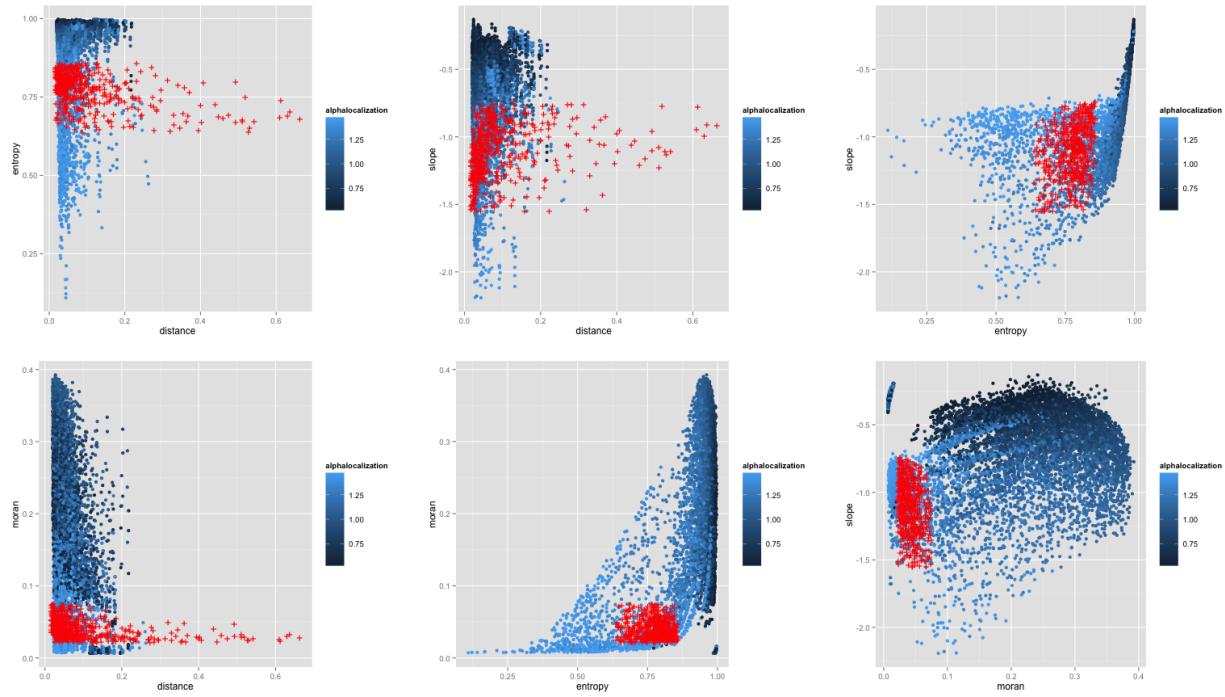


FIGURE 19 : Scatterplots of indicators distribution in an hypercube of the parameter space. We show here the influence of one parameter (localization exponent α). Red points correspond to real data.

EXPLORATION OF PARAMETER SPACE Parameter space is explored using a grid in first experiments, than a Latin Hypercube Sampling exploration. Parameter bounds are $\alpha \in [0.2, 2]$, $\beta \in [0, 0.1]$, $n_d \in \{0, \dots, 4\}$, $N_G \in [500, 3000]$, $P_m \in [2000, 100000]$. Fig.19 shows the result. We also use the parameter space exploration algorithm [[10.1371/journal.pone.0138212](https://doi.org/10.1371/journal.pone.0138212)] implemented in OpenMole, and obtain in Fig. 20 the lower bound in Moran-entropy plan, that unexpectedly exhibit a scaling relationship that we aim to explore further.

STATISTICAL ANALYSIS A statistical analysis (basic models) of indicator behaviors remains to be done and interpreted (one is done conjointly with network in paper corresponding to next section).

Model Calibration

REAL DATA Empirical morphological measures for calibration are the one described in the empirical chapter, i.e. the calibration is done on morphological objectives (entropy, hierarchy, spatial auto-correlation, mean distance) against real values computed on the set of 50km sized grid extracted from european density grid [eurostat].

CALIBRATION PROCESS We use a specific calibration process : a principal component analysis allows to maximize the cumulated dis-

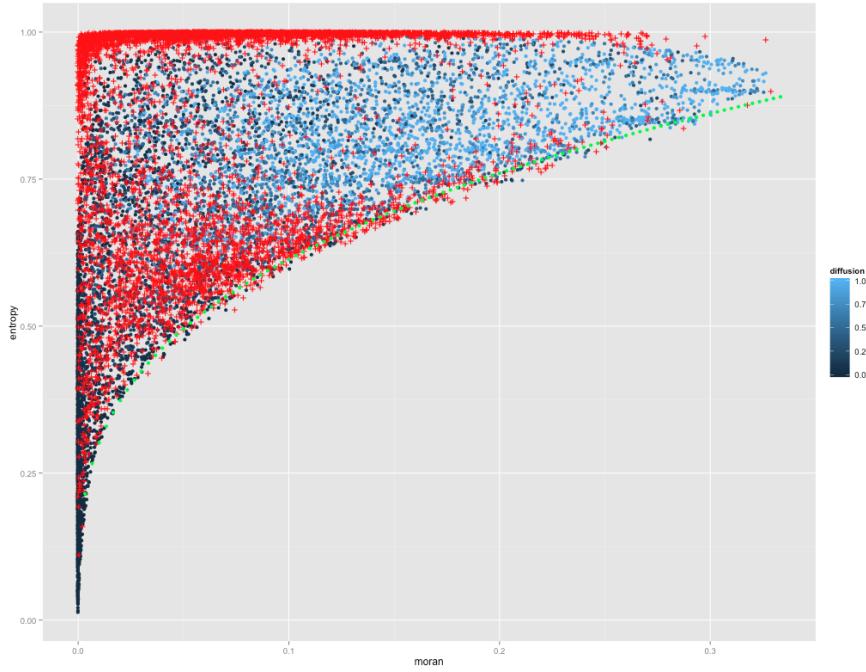


FIGURE 20 : Scatterplots of Moran against Entropy, with blue points obtained with LHS and red with PSE exploration. Lower bound is in green.

tance between generated points and real points. We select then the point cloud that overlaps real points in the (PC_1, PC_2) plan, given a distance threshold. Fig. 21 shows the points we obtain for four different values of the threshold ranging from 10^{-6} to 10^{-3} .

Calibration refinement

We plan in further work to extract the exact parameter space covering all real situations and provide interpretation of its shape (correlations between parameters). Its volume in different directions should give the relative importance of parameters.

7.1.4 Discussion

Thematic interpretation of growth behavior

We still need to interpret the positions of typical shapes within parameter space in order to confirm the thematic interpretation of parameters. Depending on results of calibration refinement, we may obtain necessary and sufficient parameters to explain growth at this scale and a corresponding interpretation.

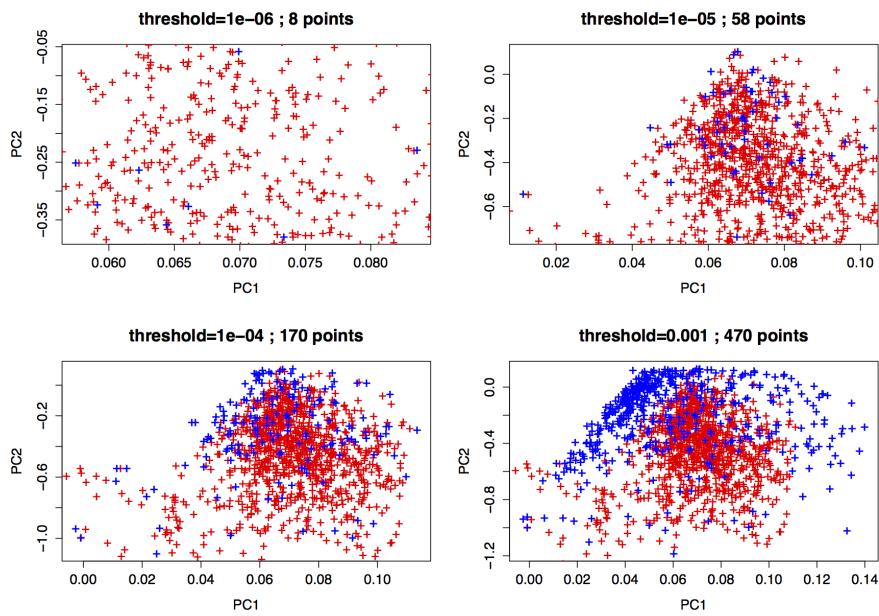


FIGURE 21 : Precise calibration of the model. The principal component analysis is conducted to maximize the spread of the differences between real data and model output, i.e. on the set $\{|R_i - M_j|\}$ where R_i is the set of real points, M_j the set of model outputs. We select then the overlapping cloud at threshold θ , by taking models output closer to real point cloud than θ in the (PC_1, PC_2) plan.

Integration into a multi-scale growth model

It could be possible to couple this model with a Gibrat (or Favaropumain) at Europa scale (macro) (with addition of consistence on migration constraints), where meso growth rates which were exogenous before are top-down determined, and bottom-up feedback is done through local aggregation level, influence importance of each area.

In conclusion, this first modeling step provide an accurately calibrated spatial urban growth model at the mesoscopic scale that can reproduce any European urban pattern in terms of urban form. Further work is needed for an interpretation of parameter influence and the determination of effective independent dimensions of the urban system at this scale. We will use this model for other purposes in the following.

7.2 CORRELATED GENERATION OF TERRITORIAL CONFIGURATIONS

Génération de configurations territoriales corrélées

This section aims to explore the sequential coupling between previous model of density generation and an heuristic of network growth. We explore therein the feasible space of correlations between network measures and morphological measures.

7.2.1 Correlated geographical data of density and network

Context

En géographie, l'utilisation de données synthétiques est plus généralement axée vers l'utilisation de population synthétiques au sein de modèles basés agents (mobilité, modèles *LUTI*) [pritchard2009advances]. 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 [brunsdon1998geographically], 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 [anders1997simpop], 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 [schmitt2014modelisation], 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 [cottineau2015revisiting], méthodologie pouvant être interprétée comme un contrôle par données statistiques spatiales. L'enjeu est de pouvoir alors distinguer effets propres dus à la dynamique intrinsèque du modèle, d'effet particuliers dus à la structure géographique du cas d'application. Celui-ci est crucial pour la validation des conclusions issues des pratiques de modélisation et simulation en géographie quantitative.

Formalization

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 [offner1996reseaux] mais extrêmement complexe et difficile à quantifier [offner1993effets]. Une modélisation dynamique des processus impliqués devrait apporter des connaissances sur ces interactions ([bretagnolle:tel-00459720], p. 162-163). Dans ce cadre, nous développons un couplage *simple* (c'est à dire sans boucle de rétroaction) entre un modèle de morphogenèse urbaine et un modèle de génération de réseau.

DENSITY MODEL Les modèle de densité est celui décrit et exploré dans la section précédente. Nous l'utilisons pour la génération conditionnelle du réseau.

NETWORK MODEL 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.

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 [TeroAl10], une génération par optimisation locale de contraintes géométriques [barthelemy2008modeling] ou un modèle de percolation plus complexe que celui utilisé, peuvent le remplacer. Ainsi, dans le cadre d'une architecture modulaire où le choix entre différentes implémentations d'une brique fonctionnelle peut être vue comme méta-paramètre [cottineau2015incremental], on pourrait choisir la fonction de génération adaptée à un besoin donné (par exemple proximité à des données réelles, contraintes sur les relations entre indicateurs de sortie, variété de formes générées, etc.).

PARAMETER SPACE L'espace des paramètres du modèle couplé¹ est constitué des paramètres de génération de densité $\vec{\alpha}_D = (P_m/N_G, \alpha, \beta, n_d)$ (on s'intéresse pour simplifier au rapport entre population et taux de croissance, i.e. le nombre d'étapes nécessaires pour générer) et des paramètres de génération de réseau $\vec{\alpha}_N = (N_C, k_h, \gamma, r_g, d_0)$. On notera $\vec{\alpha} = (\vec{\alpha}_D, \vec{\alpha}_N)$.

¹ 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.

INDICATORS On quantifie la forme urbaine et la forme du réseau, dans le but de moduler la corrélation entre ces indicateurs. La forme est définie par un vecteur $\vec{M} = (r, \bar{d}, \varepsilon, a)$ donnant auto-corrélation spatiale (indice de Moran), distance moyenne, entropie, hiérarchie (voir [le2015forme] pour une définition précise de ces indicateurs). Les mesures de la forme du réseau $\vec{G} = (\bar{c}, \bar{l}, \bar{s}, \delta)$ sont, avec le réseau noté (V, E) ,

- Centralité moyenne \bar{c} , définie comme la moyenne de la *betweenness-centrality* (normalisée dans $[0, 1]$) sur l'ensemble des liens.
- 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 [banos2012towards], qui correspond à la performance du réseau par rapport au trajet à vol d'oiseau, définie par $\bar{s} = \frac{2}{|V| \cdot (|V|-1)} \sum_{i < j} \frac{d_{ij}}{d_N(i, j)}$.
- Diamètre du réseau $\delta = \max_{i,j} d_N(i, j)$

COVARIANCE AND 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 corrélation associée la corrélation de Pearson estimée de la même façon.

Implementation

Le couplage des modèles génératifs est effectué à la fois au niveau formel et au niveau opérationnel, c'est à dire qu'on fait interagir des implémentations indépendantes. Pour cela, le logiciel OpenMole [reuillon2013openmole] utilisé pour l'exploration intensive, offre le cadre idéal de par son langage modulaire permettant de construire des *workflows* par composition de tâches à loisir et de les brancher sur divers plans d'expérience et sorties. Pour des raisons opérationnelles, le modèle de densité est implémenté en langage *scala* comme un plugin d'OpenMole, tandis que la génération de réseau est implémentée en langage basé-agent NetLogo [wilensky1999netlogo], ce qui facilite l'exploration interactive et construction heuristique interactive. Le code source est disponible pour reproductibilité sur le dépôt du projet².

Results

L'étude du modèle de densité seul est développée dans [raimbault2016calibration]. Il est notamment calibré sur les données de la grille européenne de

² à l'adresse <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Synthetic>

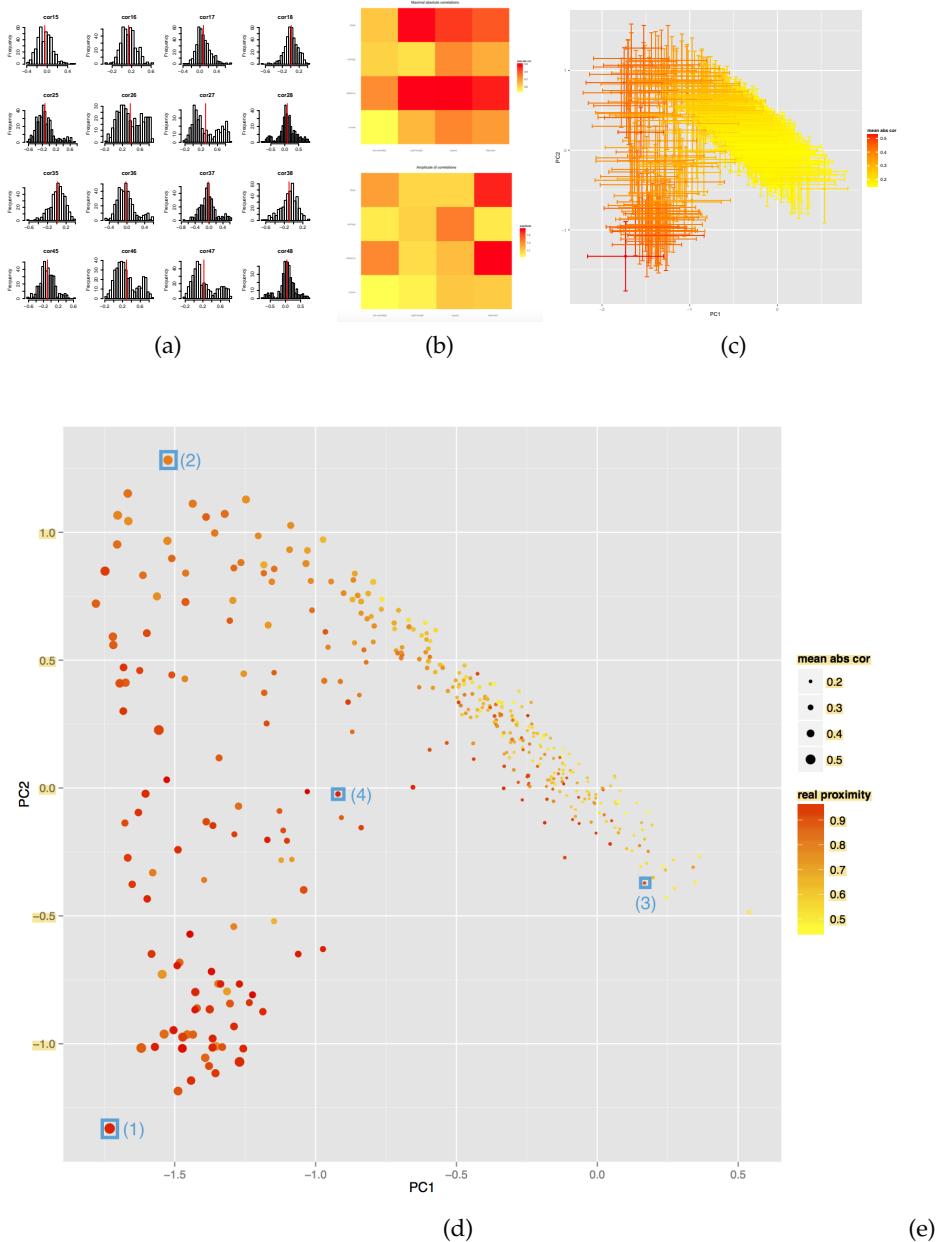


FIGURE 22 : Exploration of feasible space for correlations between urban morphology and network structure | (a) Distribution of crossed-correlations between vectors \vec{M} of morphological indicators (in numbering order Moran index, mean distance, entropy, hierarchy) and \vec{N} of network measures (centrality, mean path length, speed, diameter). (b) Heatmaps for amplitude of correlations, defined as $a_{ij} = \max_k \rho_{ij}^{(k)} - \min_k \rho_{ij}^{(k)}$ and maximal absolute correlation, defined as $c_{ij} = \max_k |\rho_{ij}^{(k)}|$. (c) Projection of correlation matrices in a principal plan obtained by Principal Component Analysis on matrix population (cumulated variances : PC1=38%, PC2=68%). Error bars are initially computed as 95% confidence intervals on each matrix element (by standard Fisher asymptotic method), and upper bounds after transformation are taken in principal plan. Scale color gives mean absolute correlation on full matrices. (d) Representation in the principal plan, scale color giving proximity to real data defined as $1 - \min_r \|\vec{M} - \vec{M}_r\|$ where \vec{M}_r is the set of real morphological measures, point size giving mean absolute correlation.

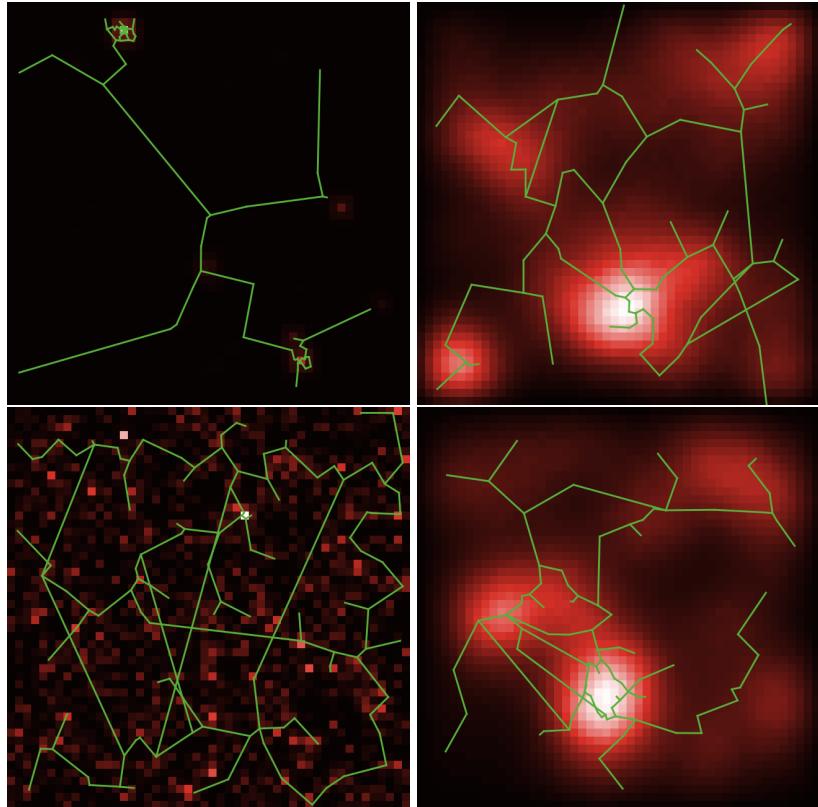


FIGURE 23 : Configurations obtained for parameters giving the four emphasized points in (d), in order from left to right and top to bottom. We recognize polycentric city configurations (2 and 4), diffuse rural settlements (3) and aggregated weak density area (1). See appendix for exhaustive parameter values, indicators and corresponding correlations. For example \bar{d} is highly correlated with \bar{l}, \bar{s} ($\simeq 0.8$) in (1) but not for (3) although both correspond to rural environments; in the urban case we observe also a broad variability : $\rho[\bar{d}, \bar{c}] \simeq 0.34$ for (4) but $\simeq -0.41$ for (2), what is explained by a stronger role of gravitation hierarchy in (2) $\gamma = 3.9, k_h = 0.7$ (for (4), $\gamma = 1.07, k_h = 0.25$), whereas density parameters are similar.

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 $\simeq 80\%$) permet d'isoler un nuage de points de sorties recouvrant assez fidèlement le nuage des points réels, ce qui veut dire que le modèle est capable de reproduire morphologiquement l'ensemble des configurations existantes.

A densité donnée, l'exploration de l'espace des paramètres du modèle de réseau suggèrent une assez bonne flexibilité sur des indicateurs globaux \tilde{G} , ainsi que de bonnes propriétés de convergence. Pour une étude du comportement précis, voir l'appendice donnant les regressions 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 correlations.

Etant donné la grande dimension relative de l'espace des paramètres, une exploration par grille exhaustive est impossible. On utilise un plan d'expérience par criblage (hypercube latin), avec les bornes indiquées ci-dessus pour $\tilde{\alpha}_D$ et pour $\tilde{\alpha}_N$, 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 correlations, une centaine donne des IC (obtenus par méthode de Fisher) de taille moyenne 0.4, on fixe donc $n = 80$ pour l'expérience. La figure 22 donne le détail des résultats de l'exploration. On retiendra les résultats marquants suivants au regard de la génération de données synthétiques corrélées :

- les distributions empiriques des coefficients de correlations entre indicateurs de forme et indicateurs de réseaux ne sont pas simples, pouvant être bimodales (par exemple $\rho_{46} = \rho[r, \bar{l}]$ entre l'index de Moran et le chemin moyen).
- On arrive à générer un assez haut niveau de correlation pour l'ensemble des indicateurs, la correlation absolue maximale variant entre 0.6 et 0.9; l'amplitude varie quant à elle entre 0.9 et 1.6, ce qui permet un large spectre de valeurs. L'espace couvert dans un plan principal a une étendue certaine mais n'est pas uniforme : on ne peut pas moduler à loisir n'importe quel coefficients, ceux-ci étant liés par les processus de génération sous-jacent. Une étude plus fine aux ordres suivants (corrélation des correlations) serait nécessaire pour cerner exactement la latitude dans la génération.

- les points les plus corrélés en moyenne sont également ceux les plus proches des données réelles, ce qui confirme l'intuition d'une forte interdépendance en réalité.
- Des exemples concrets pris sur des points particuliers distants dans le plan principal montre que des configurations de densité proches peuvent présenter des profils de correlations très différents.

Possible developments

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 [[10.1371/journal.pone.0138212](#)].

7.2.2 Discussion

Scientific positioning

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 [[2009arXiv0907.2221B](#)] 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.

Direct applications

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 aisément utilisable les données ouvertes d'OpenStreetMap³ qui sont de qualité raisonnable pour l'Europe, du moins pour la France [girres2010quality], avec 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 "correlation théorique" pour celles-ci, étant donné qu'une correlation empirique n'est en théorie pas calculable puisqu'une seule instance des processus stochastiques est observée. Vu la non-ergodicité des systèmes urbains [pumain2012urban], il y a de fortes chances pour que ces processus soient différents d'une zone géographique à l'autre (ou selon un autre point de vue qu'ils soient dans un autre état des meta-paramètres, dans un autre régime) et que leur interprétation en tant que réalisations d'un même processus stochastique n'ait aucun sens, entraînant l'impossibilité du calcul des covariations. En attribuant un jeu de données synthétiques similaire à une situation donnée, on serait capable de calculer une sorte de *correlation intrinsèque* propre à la situation, qui émerge en fait en réalité des interdépendances temporelles des composantes. Connaitre celle-ci renseigne alors sur ces interdépendances, et donc sur les relations entre réseaux et territoires.

- Comme déjà évoqué, la plupart des modèles de simulation nécessitent un état initial, généré artificiellement à partir du moment où la paramétrisation n'est pas effectuée totalement à partir de données réelles. Une analyse de sensibilité avancée du modèle implique alors un contrôle sur les paramètres de génération du jeu de données synthétique, vu comme méta-paramètre du modèle [cottineau2015revisiting]. Dans le cas d'une analyse statistique des sorties du modèle, on est alors capable d'effectuer un contrôle statistique au second ordre.
- On a étudié des processus stochastiques dans le premier exemple, au sens de séries temporelles aléatoires, alors que le temps ne jouait pas de rôle dans le second. On peut suggérer un couplage fort entre les deux composantes du modèle (ou la construction d'un modèle intégré) et observer les indicateurs et correlations à différents pas de temps de la génération. Dans le cas d'une dynamique, de par les rétroactions, on a nécessairement des effets de propagation et donc l'existence d'interdépendances décalées dans l'espace et le temps [pigazzi1980interurban], étendant le

³ <https://www.openstreetmap.org>

domaine d'étude vers une meilleure compréhension des corrélations dynamiques.

Generalization

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 [chicheportiche2013nested] est une piste de développement intéressante.

7.2.3 Conclusion

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.

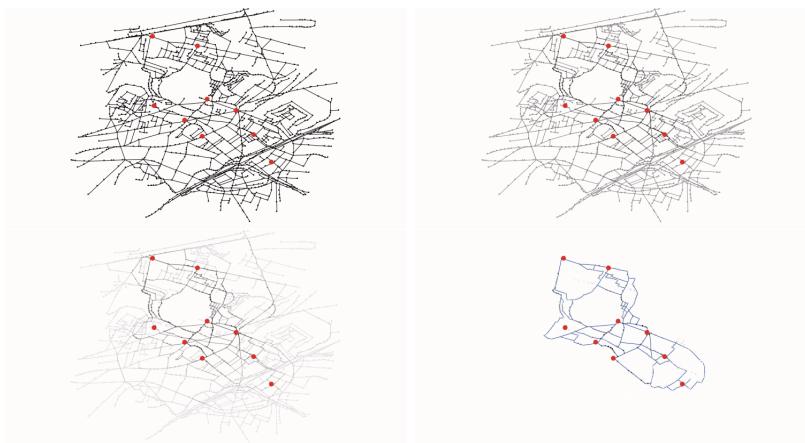


FIGURE 24 : Example of the application of the slime mould network generation model to the computation of an optimal public transportation network design.

7.3 NETWORK GROWTH MODELS : EXPLICATIVE POWER FOR VARIOUS APPROACHES

Modèles de Croissance de Réseau

7.3.1 *Benchmarking Network growth heuristics*

Pour la croissance du réseau en tant que tel, de nombreuses heuristiques existent pour générer un réseaux 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. Un travail futur aura pour but de comparer ces diverses méthodes contre les valeurs réelles des indicateurs pour le réseau de routes européen. La Fig. 24 présente un travail préliminaire présenté dans [raimbault2015labex] qui explore des applications des modèles de croissance de réseau biologique. D'autre part, comme présenté dans la section sur la reproduction, des modèles d'optimisation locale ont également été testés.

7.3.2 *Towards simple models of network morphogenesis*

An interdisciplinary project that was just launched with a Physicist LAGESSE, an Architect HACHI and a Computer Scientist DUGUE aims at finding consistent models of urban street network morphogenesis, regarding urban design particularities, geographical rules and complex network indicators feedbacks. Models of network morphogenesis were already discussed here and the aim of this project is to gain insight from the interdisciplinary vision to explore the potentiality of

such models. In the frame of our thesis, it is logically situated within the morphogenesis theoretical part and network growth modeling heuristics.

8

TOWARDS MORE COMPLEX MODELS

This single section chapter is differentiated from the previous one as it makes a step further towards more complex models. A toy-model introducing governance processes is described. Such exploration logically enters our theoretical framework to try to validate or invalidate the network necessity assumption : if non-linear necessary processes are highlighted and validated against stylized facts, it argues towards the validation of this assumption.

Other targeted projects such as the exploration of an hybrid macro-economic/accessibility-based model to explore transportation companies line implementation strategies are still at the state of ideas and are not described here.

8.1 TAKING GOVERNANCE INTO ACCOUNT IN NETWORK PRODUCTION PROCESSES : THE LUTECIA MODEL

Le Modèle Lutecia

8.1.1 *Thematic Context*

We describe a game-theory based framework, studied in collaboration with LE NECHET which aims to be integrated as behavioral rules for governing agents in a hybrid model introduced in [le2010approche] and formalized then explored in [lenechet2012]. This model couples land-use dynamics with transportation infrastructure evolution and aims to endogeneize transportation infrastructure development at different levels. The framework proposed extends it by allowing cooperation and fusion between governing entities.

As detailed in [lenechet2012], a conceptual city system with local administrative boundaries and corresponding governing agents (mayors), and a global governor (state) is the foundation of the model. A land-use evolution (residences and employments localisations) and transportation (gravity flows) are the first step of an iteration. The transportation infrastructure (road network) is then evolved by constructing a new road. First level of decision (global or local) is chosen randomly according to a fixed probability, and in the case of a local decision, the richest mayor will build the new road. The road is then build optimizing the marginal accessibility for the area corresponding to the builder in charge (all world if global, commun if local).

One thematic aspect lacking in the model and that would be interesting to study is the emergence of larger administrative zones, i.e. the emergence of new levels of governance in polycentric metropolitan areas. The reality is of course not as simple, as bottom-up initiatives such as collaboration between neighbor cities are interlaced with top-down decisions such as e.g. the “Métropole du Grand Paris” which is a new administrative structure for Paris Area decided at the state level [gilli2009paris]. It would be however interesting to test conditions for emergence of governance patterns from the bottom-up in a conceptual way by extending the model and adding interactions and fusion between administrative entities.

The extension shall consist in relaxing the assumption of a single road segment built at each time step and attribute one segment to the N richest mayors. That leads to situation where neighbor towns may want to construct both a new road. As they are likely to communicate with each other, we assume that negotiations take place and that they consider eventually to build in common, in which case they merge after (rough simplifying but stylized assumption). Such negotiations may be interpreted as a game in the sense of Game Theory, which as already been widely applied for modeling in social and political sciences for questions dealing with cognitive interacting agents with individual interests [ordeshook1986game]. Such a framework as already been used in transportation investment studies, as e.g. in [Roumboutsos2008209] where choices of operators (public and privates) to integrate their system in a global consistent commuter system is explored through the notion of Nash equilibrium.

8.1.2 Formalization

The model architecture couples in a complex way a module for land-use evolution with a module for transportation network growth. Sub-modules, detailed in the following, include in particular a governance module that rules processes of network evolution.

Land-use evolution

The following steps are detailed in [lenechet2012] but we recall the big picture :

- Initial distribution of Actives and Employments is done around governance centers at positions \vec{x}_i by

$$A(\vec{x}) = A_{\max} \cdot \exp\left(\frac{\|\vec{x} - \vec{x}_i\|}{r_A}\right); E(\vec{x}) = E_{\max} \cdot \exp\left(\frac{\|\vec{x} - \vec{x}_i\|}{r_E}\right)$$

- For facility patches, employments are added by $E(\vec{x}) = E(\vec{x}) + \frac{k_{\text{ext}} \cdot E_{\max}}{n_{\text{ext}}}$.

- Transportation module : computation of flows ϕ_{ij} are done by solving on p_i, q_j by a fixed point method (Furness algorithm), the system of gravity flows

$$\begin{cases} \phi_{ij} = p_i q_j A_i E_j \exp(-\lambda_{tr} d_{ij}) \\ \sum_k \phi_{kj} = E_j; \sum_k \phi_{ik} = A_i \\ p_i = \frac{1}{\sum_k q_k E_k \exp(-\lambda_{tr} d_{ik})}; q_j = \frac{1}{\sum_k p_k A_k \exp(-\lambda_{tr} d_{kj})} \end{cases}$$

- Trajectories then attributed by effective shortest path, and corresponding congestion c obtained (no Wardrop equilibrium).
- Speed of network is given by a BPR function $v(c) = v_0 (1 - \frac{c}{\kappa})^{\gamma_c}$. Congestion is not used in current studies (infinite capacity κ).
- Land-Use module : we assume that residential/employments relocations are at equilibrium at the time scale of a tick, that corresponds to transportation infrastructure evolution time scale which is much larger [**bretagnolle:tel-00459720**].
- We take a Cobb-douglas function for utilities of actives/employments at a given cell

$$U_i(A) = X_i(A)^{\gamma_A} \cdot F_i(A)^{1-\gamma_A}; F_i(A) = \frac{1}{A_i E_i}$$

$$U_j(E) = X_j(E)^{\gamma_E} \cdot F_j(E)^{1-\gamma_E}; F_j(E) = 1$$

where $X_i(A) = A_i \cdot \sum_j E_j \exp(-\lambda \cdot d_{ij})$ and $X_j(E) = E_j \cdot \sum_i A_i \exp(-\lambda \cdot d_{ij})$.

- Relocations are then done deterministically following a discrete choice model :

$$A_i(t+1) = \sum_i A_i(t) \cdot \frac{\exp(\beta U_i(A))}{\sum_i \exp(\beta U_i(A))}$$

$$E_j(t+1) = \sum_j E_j(t) \cdot \frac{\exp(\beta U_j(E))}{\sum_j \exp(\beta U_j(E))}$$

The default parameter values are taken as follows : $A_{max} = E_{max} = 500; r_A = 1; r_E = 0.8; \gamma_E = 0.9; \gamma_A = 0.65; \beta_l = 1.8; \lambda = 0.005; r_0 = 2$
and

$N_{expl} = 25; I = 0.001; J = 0.0001; \nu = 5; E_{ext}(t_0) = 3E_{max}; t_f = 4$

Effective distances computation

Distance via network are updated in a dynamical programming fashion for efficiency purposes (because of the numerous network updates), the following way :

- Euclidian distance matrix $d(i, j)$ computed analytically

- Network shortest paths between network intersections (rasterized network) updated in a dynamic way (addition of new paths and update/change of old paths if needed when a link is added), correspondance between network patches and closest intersection also updated dynamically; $O(N_{\text{inters}}^3)$
- Weak component clusters and distance between clusters updated; $O(N_{\text{nw}}^2)$
- Network distances between network patches updated, through the heuristic of only minimal connexions between clusters; $O(N_{\text{nw}}^2)$
- Effective distances (taking paces/congestion into account) updated as minimum between euclidian time and

$$\min_{C,C'} d(i, C) + d_{\text{nw}}(p_C(i), p'_C(j)) + d(C', j)$$

, complexity in $O(N_{\text{clusters}}^2 \cdot N^2)$ (Approximated with \min_C only in the implementation, consistent within the interaction ranges ~ 5 patches taken in the model).

Externality

The model allows also to simulate the competition of territory for an external ressource (an airport for example). We implement therefore the option of adding in initial state an area with initial A_{max} employments and that follows an intrinsic growth rule as a geometric law.

Transportation Network growth

The workflow for transportation network development is the following :

- At each time step, N new road segments are built. Choice between local and global is still done through uniform drawing with probability ξ . In the case of local building, roads are attributed successively to mayors with probabilities ξ_i , what means that richer areas may get many roads. It stays consistent with the thematic assumption than each road correspond to the allocation of one public market which are done independently (with N becoming greater, this assumption should be relaxed as attribution of subventions to local areas is of course not proportional to wealth, but we assume that it stays true with small N values).
- Areas building a road without neighbors doing it follow the standard procedure to develop the road network.

- Neighbor areas building a road will enter negotiations. We assume in this first simple version of the model that only bilateral negotiations may occur. Therefore, in the case of clusters with more than two areas, pairing is done at random (uniform drawing) between neighbors until all areas are paired.
- Possible strategies for players (negotiating areas, $i = 1, 2$) are : staying alone (A) and collaborating (C). Strategies are chosen simultaneously (non-cooperative game) as detailed after. For (C, A) and (A, C) couples, the collaborating agent loose its investment and cannot build a road whereas the other continues his business alone. For (A, A) both act as alone, and for (C, C) a common development is done. We denote $Z_i^*(S_1, S_2)$ the optimal infrastructure for area i with $(S_1, S_2) \in \{(A, C), (C, A), (A, A)\}$ which are determined the standard way in each zone separately, and Z_C^* the optimal common infrastructure computed with a 2 segments infrastructure on the union of both areas, which corresponds to the case where both strategies are C. Marginal accessibilities for area i and infrastructure Z is defined as $\Delta X_i(Z) = X_i^Z - X_i$. We introduce the costs of construction which are necessary to build the payoff matrix. They are assumed spatially uniform and noted I for a road segment, whereas a 2 road segment will cost $2 \cdot I - \delta I$ ($\delta I > 0$ cost gain of common technical means, assumed to be equally shared). An interesting generalization would be to divide costs proportional to wealth in the case of a collaboration. The payoff matrix of the game is the following, with κ a normalization constant ("price of accessibility") :

$1 2$	C	A
C	$U_i = \kappa \cdot \Delta X_i(Z_C^*) - I - \frac{\delta I}{2}$	$\begin{cases} U_1 = \kappa \cdot \Delta X_1(Z_1^*) - I \\ U_2 = \kappa \cdot \Delta X_2(Z_2^*) - I - \frac{\delta I}{2} \end{cases}$
A	$\begin{cases} U_1 = \kappa \cdot \Delta X_1(Z_1^*) - I - \frac{\delta I}{2} \\ U_2 = \kappa \cdot \Delta X_2(Z_2^*) - I \end{cases}$	$U_i = \kappa \cdot \Delta X_i(Z_i^*) - I$

We have a typical coordination game for which it is clear that no strategy is dominant for any player. In a probabilistic mixed-strategy case, there always exists a Nash equilibrium that we can easily determine in our case. It is reasonable to make such an assumption since negotiations take generally some time during which agents are able to find the way to optimize rationally their expected utility. If $\mathbb{P}[S_1 = C] = p_1$ and $\mathbb{P}[S_2 = C] = p_2$, we have

$$\begin{aligned} \mathbb{E}[U_1] &= p_1 p_2 U_1(C, C) + p_1 \cdot (1 - p_2) U_1(C, A) + p_2 \cdot (1 - p_1) U_1(A, C) + (1 - p_1)(1 - p_2) U_1(A, A) \\ &= p_1 \cdot \left[p_2 \cdot \left(\kappa \cdot \Delta X_1(Z_C^*) - \frac{\delta I}{2} \right) - \kappa \cdot \Delta X_1(Z_1^*) + I \right] + p_2 \cdot \frac{\delta I}{2} + \kappa \cdot \Delta X_1(Z_1^*) - I \end{aligned}$$

Optimizing the expected utility along p_1 (the variable on which agent 1 has control) imposes the condition on p_2

$$\frac{\partial \mathbb{E}[U_1]}{\partial p_1} = 0 \iff p_2 = \frac{\delta I / 2}{\Delta X_2 Z_C^* - \Delta X_2 Z_2^*}$$

We obtain generally

$$p_i = \frac{J}{\Delta X_i Z_C^* - \Delta X_i Z_i^*}$$

Note that we can directly interpret these expressions, as a player chances to cooperate will decrease with the potential gain of the other player, what is intuitive for a competitive game. It also forces feasibility conditions on I and δI to keep a probability, that are $I \leq \kappa \cdot \min(\Delta X_1(Z_1^*), \Delta X_2(Z_2^*))$ (binary positive cost-benefit conditions) and $I - \delta I > \kappa \cdot \max_i(\Delta X_i(Z_i^*) - \Delta X_i(Z_C^*))$. As soon as accessibility difference stay relatively small, both shall be compatible when $\delta I \ll I$, giving corresponding boundaries for I .

- Agents make choice of strategy following uniform drawings with probability computed above. Corresponding infrastructures are built, and in the case of choices (C, C) , towns merge in a single one with new corresponding variables (employment, actives, etc.).

REMARK FOR THE IMPLEMENTATION To adapt an existing implementation, one just has to add the negotiation stage if conditions are met, using probabilities given above. The accessibility-dimensioned parameters $\alpha = \frac{I}{\kappa}$ and $\delta \alpha = \frac{\delta I}{\kappa}$ should be more simple to deal with.

AN ALTERNATIVE DISCRETE CHOICE “GAME” Using the same payoff matrix with a random utility model allows to obtain also values for probabilities. We have

$$U_i(C) - U_i(NC) = p_i (\Delta X_i Z_C^* - \Delta X_i Z_i^*) - J$$

and therefore p_i verifies the equation that is solved numerically

$$p_i = \frac{1}{1 + \exp \left(-\beta_{DC} \cdot \left(\frac{\Delta X_i Z_C^* - \Delta X_i Z_i^*}{1 + \exp(-\beta_{DC} (p_i \cdot (\Delta X_i Z_C^* - \Delta X_i Z_i^*) - J))} - J \right) \right)}$$

This module is also implemented for comparison purposes.

8.1.3 Results

Implementation

The model was implemented in NetLogo [[wilensky1999netlogo](#)] because of its exploratory and interactive nature. A particular care was taken for the computation of accessibilities and shortest paths, as a dynamic reevaluation of network distance is necessary for each new potential infrastructure, what become rapidly a computational burden. We use thus a dynamical programming shortest path computation, inspired from [[tretyakov2011fast](#)], using distance matrices updates instead of shortest paths full computation at each step. See details in architectural precisions in Appendix ??

Exploration and Validation

We show in Fig. [25](#) and Fig.[26](#) examples of obtained configurations and preliminary validation of governance and network growth heuristic. Internal validation and external validation through stylized facts, and model explorations, including statistical analysis of model behavior, are provisory for now and not presented here (see [[le2015modeling](#)] for preliminary results).

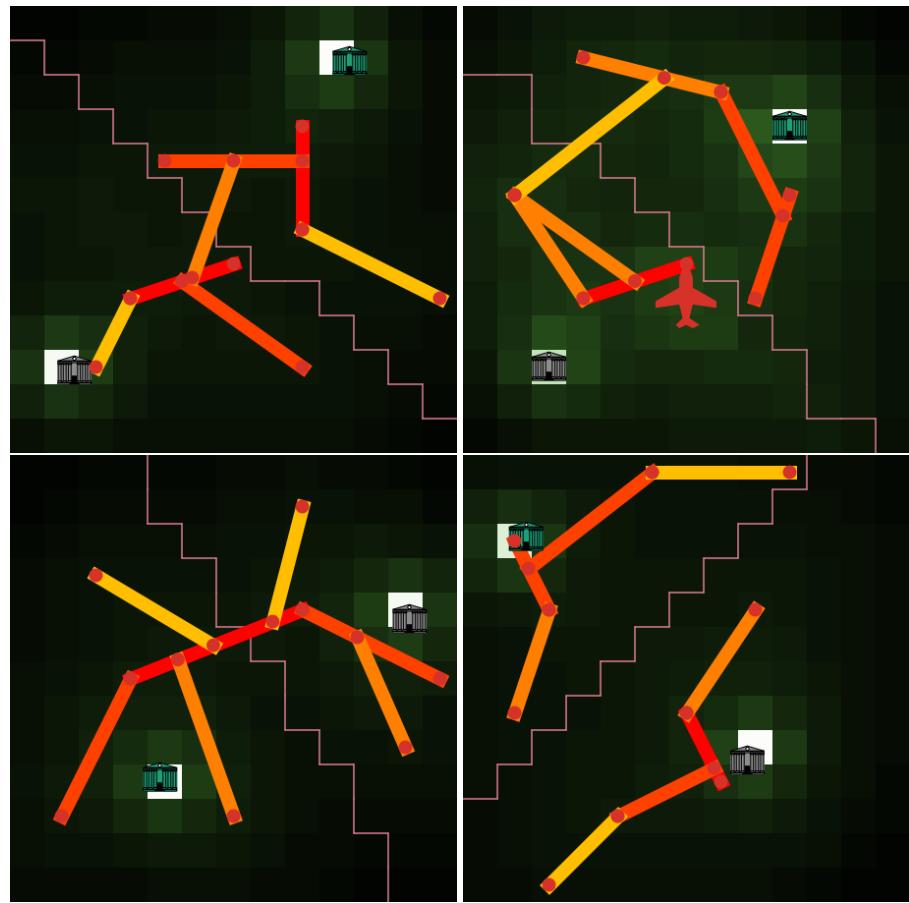


FIGURE 25 : Examples of final configurations, with or without externality, for different values of cooperation parameters.

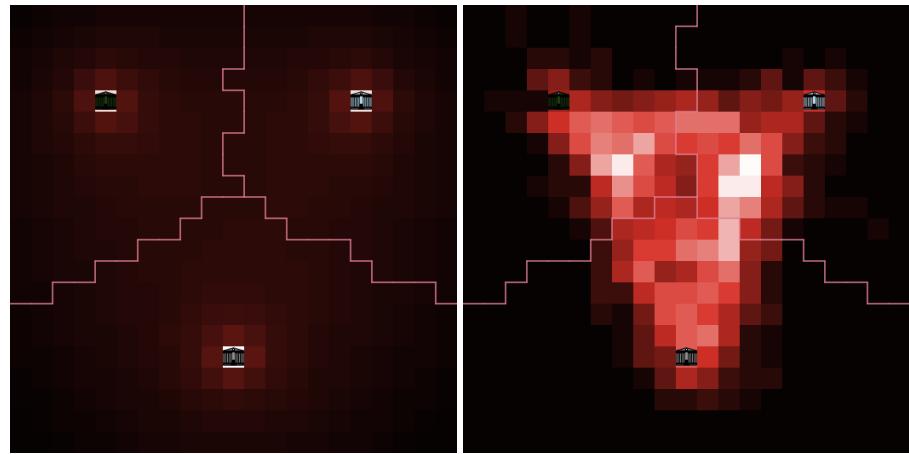


FIGURE 26 : Validation of network exploration heuristic : mean accessibility(left) and network positions on 500 realizations on the same initial configuration. The optimal distribution of network validates network generation heuristic.

Troisième partie

SYNTHESIS

This concluding remark, for now a brief roadmap, is one objective of our thesis as implementation of our theory and thus is expected to become a consequent part. We conclude here this preliminary work by perspectives and roadmap. This part make the synthesis of what was build until now, towards a delicate though robust edifice.

A ROADMAP FOR AN OPERATIONAL FAMILY OF MODELS OF COEVOLUTION

As previously stated, one of our principal aims is the validation of the network necessity assumption, that is the differentiating point with a classic evolutive urban theory. To do so, toy-model exploration and empirical analysis will not be enough as hybrid models are generally necessary to draw effective and well validated conclusions. We briefly give an overview of planned work in the following, that will be the conclusion of this Memoire.

9.1 OBJECTIVES

Objectifs

We expect to product *models of coevolution*, with the emphasis on processes of coevolution, to directly confront the theory. They will be necessary a flexible family because of the variety of scales and concrete cases we can include and we already began to explore in preliminary studies. Processes already studied can serve either as a thematic bases for a reuse as building bricks in a multi-modeling context, or as methodological tools such as synthetic data generator for synthetic control. Finally, we mean by operational models hybrid models, in the sense of semi-parametrized or semi-calibrated on real datasets or on precise stylized facts extracted from these same datasets. This point is a requirement to obtain a thematic feedback on geographical processes and on theory.

9.2 CASE STUDIES

Cas d'étude

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:tel-00459720](#)], that would also provide an interesting case study on other regimes, scales and transportation mode.

9.3 ROADMAP

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 [7](#)), 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.*

Quatrième partie

OPENING

A building is never used the way it was designed, that is a reality which grasping makes the difference between good and excellent architects. The effective functional use give sense to any construction. So goes it for a knowledge edifice. We shall now take a look back on what we constructed and try to take a step back. This part develops first theoretical apparels emerging from the various aspects already tackled. It then proposes to extract fundamental open questions that future research on territorial complex systems will have to tackle in the incoming decades.

10

THEORETICAL FRAMEWORK

Your theory is crazy, but not enough to be true.

- NIELS BOHR

La théorie est un élément essentiel de toute construction scientifique, en particulier en Sciences Humaines pour lesquelles la définition des objets et questions de recherche sont plus ouverts mais aussi plus déterminants des directions de recherche alors prises. Nous développons dans ce chapitre un cadre théorique autonome. Il émerge naturellement des considérations thématiques du chapitre précédent, des explorations empiriques faites dans le chapitre 6 et des expériences de modélisation conduites dans le chapitre 7

Nous proposons d'abord de construire une *Théorie Géographique* qui fixera les objets étudiés et leur nature réelle (leur ontologie), ainsi que leur interrelations. Celle-ci permettra de produire des hypothèses précises qu'on cherchera à confirmer ou infirmer par la suite.

10.1 GEOGRAPHICAL THEORETICAL CONTEXT

Pour une Théorie Géographique

10.1.1 *Foundation**Networked Human Territories*

Our first pillar has already been constructed before in the thematic exploration of the research subject. We rely on the notion of *Human Territory* elaborated by RAFFESTIN as the basis for a definition of territorial systems. It permits to capture complex human geographical systems in their concrete and abstract characteristics and representation. For example, a metropolitan territorial system can be apprehended simply by the functional extent of daily commuting, or by the perceived or lived space of different populations, the choice depending on the precise question asked. Note that this approach to territory is a position and that other (possibly compatible) entries could be taken [murphy2012entente]. The concrete of this pillar is reinforced by the territorial theory of networks of DUPUY, yielding the notion of networked human territory, as a human territory in which a set of potential transactional networks have been realized, which is in accordance with vision of the territory as networked places [champollion:halshs-00999026]. We make therein the assumption that real networks are necessary elements of territorial systems.

Evolutive Urban Theory

The second pillar of our theoretical construction is the Evolutive Urban Theory of PUMAIN, closely linked to the complexity approach we take. This theory was first introduced in [pumain1997pour] which argues for a dynamical vision of city systems, in which self-organization is key. Cities are interdependent evolutive spatial entities whose inter-relations produces the macroscopic behavior at the scale of city system. The city system is also designed as a network of city what emphasizes its view as a complex system. Each city is itself a complex system in the spirit of [berry1964cities], the multi-scale aspect being essential in this theory, since microscopic agents convey system evolution through complex feedbacks between scales. The positioning within Complex System Sciences was later confirmed [pumain2003approche]. It was shown that this theory provide an interpretation for the origin of pervasive scaling laws, resulting from the diffusion of innovation cycles between cities [pumain2006evolutionary]. The aspect of resilience of system of cities, induced by the adaptive character of these complex systems, implies that cities are drivers and adapters of social change [pumain2010theorie]. Finally, path dependance yield non-ergodicity within these systems, making "universal" interpreta-

tions of scaling laws developed by physicists incompatible with evolutive urban theory [pumain2012urban]. We will interpret territorial systems following that idea of complex adaptive systems.

Urban Morphogenesis

Co-evolution

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 traitant des signaux grâce à leur frontières [holland2012signals]. Dans cette théorie, les systèmes complexes adaptatifs forment des agrégats à différents niveaux hiérarchiques

10.1.2 Synthesis : an theory of co-evolutive networked territorial systems

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 minimalistique.

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.*

At this step complexity and dynamical evolutive characters of territorial systems are implied but not an explicit part of the theory. We will assume to simplify a discrete definition of temporal, spatial and ontological scales under modularity and local stationarity assumptions.

Proposition 1 - Discrete scales. *Assuming a discrete modular decomposition of a territorial system, the existence of a discrete set (τ_i, x_i) of temporal and functional scales for the territorial system is equivalent to the local temporal stationarity of a random dynamical system specification of the system.*

Proof (Sketch of). We underlie that any territorial system can be represented by random variables, what is equivalent to have well defined objects and states and use the Transfer Theorem on events of successive states. If $X = (X_j)$ is the modular decomposition, we have necessarily quasi-independence of components in the sense that $\text{Cov}[dX_j, dX_{j'}] \simeq 0$ at any time. General stationarity transitions induce modular transitions that are kept or not depending if they correspond to an effective transition within the subsystem, what provide

temporal scales as characteristic times of sub-dynamics. Functional scales are the corresponding extent in the state space. ■

This proposition induce a discrete representation of system dynamics in time. Note that even in the case of no modular representation, the system as a whole will verify the property. This definition of scales allows to explicitly introduce feedback loops and thus emergence and complexity, making our theory compatible with the evolutive urban theory.

Assumption 1 - Scales and Subsystems intrication. *Complex networks of feedbacks exist both between and inside scales, what impose the existence of weak emergence [bedau2002downward]. Furthermore a horizontal and vertical hierarchical imbrication of boundaries is not the rule.*

Within these complex subsystems intrications we can isolate co-evolving components using morphogenesis. The following proposition is a consequence of the equivalence between the independence of a niche and its morphogenesis. Morphogenesis provides the modular decomposition (local stationarity assumed) needed for the existence of scale, giving minimal vertically (scale) and horizontally (space) independent subsystems.

Proposition 2 - Co-evolution of components. *Morphogenesis processes of a territorial system are an equivalent formulation of the existence of co-evolutive subsystems.*

Finally we make a key assumption putting real networks at the center of co-evolutive dynamics, introducing their necessity to explain dynamical processes of territorial systems.

Assumption 2 - Necessity of Networks. *Network evolution can not be explained only by the dynamics of other territorial components and reciprocally, i.e. co-evolving territorial subsystems include real networks. They can thus be at the origin of regime changes (transition between stationarity regimes) or more dramatic bifurcations in dynamics of the whole territorial system.*

On long time scale, an overall co-evolution has been shown for the french railway network by [bretagnolle:tel-00459720]. At smaller scales it is less evident (debate on structural effects) but we postulate that co-evolution effects are present at any scale. Regional examples may illustrate that : Lyon has not the same dynamical relations with Clermont than with Saint-Etienne and network connectivity has necessarily a role in that (among intrinsic interaction dynamics and distance). At a smaller scale, we think that effects are even less observable, but precisely because of the fact that co-evolution is stronger

and local bifurcations will occur with stronger amplitude and greater frequency than in macroscopic systems where attractors are more stable and stationarity scales greater. We will try to identify bifurcation or phase transitions in toy models, hybrid models and empirical analysis, at different scales, on different case studies and with different ontologies.

One difficulty in our construction is the stationarity assumption. Even if it seems a reasonable assumptions on large scales and has already been observed in empirical data [**sanders1992systeme**], we shall verify it in our empirical studies. Indeed, this question is at the center of current research efforts to apply deep learning techniques to geographical systems : BOURGINE has recently developed a framework to extract patterns of Complex Adaptive Systems¹. The issues are then if the stationarity assumption be tackled through augmentation of system states, and if heterogeneous and asynchronous data be used to bootstrap long time-series necessary for a correct estimation of the neural network. These issue are related to the stationarity assumption for the first and to non-ergodicity for the second.

¹ Using a representation theorem [**knight1975predictive**], any discrete stationary process is a *Hidden Markov Model*. Given the definition of a causal state as $\mathbb{P}[\text{future}|A] = \mathbb{P}[\text{future}|B]$, the partition of system states induced by the corresponding equivalence relations allows to derive a *Recurrent Network* that is enough to determine next state of the system, as it is a *deterministic* function of previous state and hidden states [**shalizi2001computational**] : $(x_{t+1}, s_{t+1}) = F[(x_t, s_t)]$. The estimation of Hidden States and of the Recurrent Function thus captures through deep learning entirely dynamical patterns of the system, i.e. full information on its dynamics and internal processes.

10.2 A THEORETICAL FRAMEWORK FOR THE STUDY OF SOCIO-TECHNICAL SYSTEMS

Un Cadre Théorique pour l'Etude des Systèmes Sociaux-techniques

After having set up the thematic theoretical framework, we develop a more general framework in which the previous can enter. At an epistemological level, it is essential to frame generally our directions of research.

10.2.1 *Introduction*

Scientific Context

The structural misunderstandings between Social Sciences and Humanities on one side, and so-called Exact Sciences on the other side, far from being a generality, seems to have however a significant impact on the structure of scientific knowledge [2015arXiv151103981H]. In particular, the place of theory (and indeed the signification of this term itself) in the elaboration of knowledge has a totally different place, partly because of the different *perceived complexities*² of studied objects : for example, mathematical constructions and by extent theoretical physics are *simple* in the sense that they are mostly entirely analytically solvable, whereas Social Science subjects such as humans or society (to give a *cliché* exemple) are *complex* in the sense of complex systems³, thus a stronger need of a constructed theoretical (generally empirically based) framework to identify and define the objects of research that are necessarily more arbitrary in the framing of their boundaries, relations and processes, because of the multitude of possible viewpoints : Pumain suggests indeed in [pumain2005cumulativite] a new approach to complexity deeply rooted in social sciences that "would be measured by the diversity of disciplines needed to elaborate a notion". These differences in backgrounds are naturally desirable in the spectrum of science, but things can get nasty when playing on "common" terrains, typically complex systems problematics as already detailed, as the exemple of geographical urban systems has recently shown [dupuy2015sciences]. Complex System Science⁴ is presented by some as a "new kind of Science" [wolfram2002new], and would at least be a symptom of a shift in scientific practices, from

² We used the term *perceived* as most of systems studied by physics might be described as simple whereas they are intrinsically complex and indeed not well understood [laughlin2006different].

³ for which no unified definition exists but of which fields of application range broadly from neuroscience to quantitative finance, including e.g. quantitative sociology, quantitative geography, integrative biology, etc. [newman2011complex], and for which study various complementary approaches may be applied, such as Dynamical Systems, Agent-based Modeling, Random Matrix Theory

⁴ that we deliberately call that way although there is a running debate on whether it can be seen as a Science in itself or more as a different way to do Science.

analytical and “exact” approaches to computational and evidence-based approaches [arthur2015complexity], but what is sure is that it brings, together with new methodologies, new scientific fields in the sense of converging interests of various disciplines on transversal questions or of integrated approaches on a particular field [2009arXiv0907.2221B].

Objectives

Within that scientific context, the study of what we will call *Socio-technical Systems*, which we define in a rather broad way as hybrid complex systems including social agents or objects that interact with technical artifacts and a natural environment⁵, lies precisely between social sciences and hard sciences. The example of Urban Systems is the best example, as already before the arrival of approaches claiming to be “more exact” than soft approaches (typically by physicists, see e.g. the rather disturbing introduction of [louf2014scaling]), but also by scientists coming from social sciences such as Batty [batty2013new]), many aspects of urban systems were already in the field of exact sciences, such as urban hydrology, urban climatology or technical aspects of transportation systems, whereas the core of their study relied in social sciences such as geography, urbanism, sociology, economy. Therefore a necessary place of theory in their study : following [livet2010], the study of complex systems in social science is an interaction between empirical analysis, theoretical constructions, and modeling.

We propose in this paper to construct a theory, or rather a theoretical framework, that would ease some aspects of the study of such systems. Many theories already exist in all fields related to this kind of problems, and also at higher levels of abstraction concerning methods such as agent-based modeling e.g., but there is to our knowledge no theoretical framework including all of the following aspects that we consider as being crucial (and that can be understood as an informal basis of our theory) :

1. a precise definition and emphasis on the notion of coupling between subsystems, in particular allowing to qualify or quantify a certain degree of coupling : dependence, interdependence, etc. between components.
2. a precise definition of scale, including timescale and scales for other dimensions.
3. as a consequence of the previous points, a precise definition of what is a system.

⁵ geographical systems in the sense of [dolfus1975some] are the archetype of such systems, but that definition may cover other type of systems such as an extended transportation system, social systems taken with an environmental context, complicated industrial systems taken with users, etc.

4. the inclusion of the notion of emergence in order to capture multi-scale aspects of systems.
5. a central place of ontology in the definition of systems, i.e. of the sense in the real world given to its objects⁶.
6. taking into account heterogeneous aspects of the same system, that could be heterogeneous components but also complementary intersecting views.

The rest of this section is organized as follows : we construct the theory in the following part, staying at an abstract level, and propose a first application to the question of co-evolving subsystems. We then discuss positioning regarding existing theories, and possible developments and concrete applications.

10.2.2 *Construction of the theory*

Perspectives and Ontologies

The starting point of the theory construction is a perspectivist epistemological approach on systems introduced by Giere [[giere2010scientific](#)]. To sum up, it interprets any scientific approach as a perspective, in which someone pursues some objective and uses what is called *a model* to reach it. The model is nothing more than a scientific medium. Varenne developed [[varenne2010framework](#)] model typologies that can be interpreted as a refinement of this theory. Let for now relax this possible precision and use perspectives as proxies of the undefined objects and concepts. Indeed, different views on the same object (being complementary or diverging) have the property to share at least the object in itself, thus the proposition to define objects (and more generally systems) from a set of perspectives on them, that verify some properties that we formalize in the following.

A perspective is defined in our case as a dataflow machine M (that corresponds to the model as medium) in the sense of [[golden2012modeling](#)] that gives a convenient way to represent it and to introduce timescales, to which is associated an ontology O in the sense of [[livet2010](#)], i.e. a set of elements each corresponds to a *thing* (it can be an object, an agent, a process, etc.) in the real world. We include only two aspect (the model and the objects represented) of Giere's theory, making the assumption that purpose and user of the perspective are indeed contained in the ontology.

Definition 2 *A perspective on a system is given by a dataflow machine $M = (i, o, \mathbb{T})$ and an associated ontology O . We assume that the ontology can be decomposed into atomic elements $O = (O_j)_j$.*

⁶ *as already explained before, this positioning along with the importance of structure may be related to Ontic Structural Realism [[frigg2011everything](#)] in further developments.*

The atomic elements of the ontology can be particular elements such as agents or components of the system, but also processes, interactions, states, or concepts for example. The ontology can be seen as the rigorous description of the content of the perspective. The assumption of a dataflow machine implies that possible inputs and outputs can be quantified, what is not necessarily restrictive to quantitative perspectives, as most of qualitative approaches can be translated into discrete variables as long as the set of possibles is known or assumed.

The system is then defined “reversely”, i.e. from a set of perspectives on a system :

Definition 3 *A system is a set of perspectives on a system : $S = (M_i, O_i)_{i \in I}$, where I may be finite or not.*

We denote by $\mathcal{O} = (O_{j,i})_{j,i \in I}$ the set of all elements within ontologies.

Note that at this level of construction, there is not necessarily any structural consistence in what we call a system, as given our broad definition could allow for example to consider as a system a perspective on a car together with a perspective on a system of cities what makes reasonably no sense at all. Further definitions and developments will allow to be closer from classical definition of a system (interacting entities, designed artifacts, etc.). The same way, the definition of a subsystem will be given further. The introduced elements of our approach help to tackle so far points three, five and six of the requirements.

PRECISION ON THE RECURSIVE ASPECT OF THE THEORY One direct consequence of these definitions must be detailed : the fact that they can be applied recursively. Indeed, one could imagine taking as perspective a system in our sense, therefore a set of perspectives on a system, and do that at any order. If ones takes a system in any classical sense, then the first order can be understood as an epistemology of the system, i.e. the study of diverse perspectives on a system. A set of perspectives on related systems may in some conditions be a domain or a field, thus a set of perspectives on various related systems the epistemology of a field. These are more analogies to give the idea behind the recursive character of the theory. It is indeed crucial for the meaning and consistence of the theory because of the following arguments :

- The choice of perspectives in which a system consists is necessarily subjective and therefore understood as a perspective, and a perspective on a system if we are able to build a general ontology.

- We will use relations between ontologies in the following, which construction based on emergence is also subjective and seen as perspectives.

Ontological Graph

We propose then to capture the structure of the system by linking ontologies. This approach could eventually be linked to structural realism epistemological positioning [frigg2011everything] as knowledge of the world is partly contained here in structure of models. Therefore, we choose to emphasize the role of emergence as we believe that it may be one practical minimalist way to capture quite well complex systems structure⁷. We follow on that point the approach of Bedau on different type of emergences, in particular his definition of weak emergence given in [bedau2002downward]. Let recall briefly definitions we will use in the following. Bedau starts from defining emerging properties and then extends it to phenomena, entities, etc. The same way, our framework is not restricted to objects or properties and wrapped thus the generalized definitions into emergence between ontologies. We will apply the notion of emergence under the two following forms⁸ :

- *Nominal emergence* : one ontology O' is included in an other O but the aspect of O that is said to be nominally emergent regarding O' does not depend on O' .
- *Weak emergence* : one part of an ontology O can be derived by aggregation of elements and interactions between elements of an ontology O' .

As developed before, the presence of emergence, and especially weak emergence, will consist in itself in a perspective. It can be conceptual and postulated as an axiom within a thematic theory, but also experimental if clues of weak emergence are effectively measured between objects. In any case, the relation between ontologies must be encoded within an ontology, which was not necessarily introduced in the initial definition of the system.

We make therefore the following assumption for next developments :

Assumption 3 *A system can be partially structured by extending it with an ontology that contains (not necessarily only) relations between elements of ontologies of its perspectives. We name it the coupling ontology and*

⁷ what of course can not been presented as a provable claim as it depends on system definition, etc.

⁸ the third form Bedau recalls, *Strong emergence* will not be used, as we need only to capture dependance and autonomy, and weak emergence is more satisfying in terms of complex systems, as it does not assume “irreducible causal powers” to the greater scale objects. Nominal emergence is used to capture inclusion between ontologies.

assume its existence in the following. We assume furthermore its atomicity, i.e. if O is in relation with O' , then any subsets of O, O' can not be in relation, what is not restrictive as a decomposition into several independent subsets ensures it if it is not the case.

It allows to exhibit emergence relations not only within a perspective itself but also between elements of different perspectives. We define then pre-order relations between subsets of ontologies :

Proposition 3 *The following binary relationships are pre-orders on $\mathcal{P}(O)$:*

- *Emergence (based on Weak Emergence) : $O' \preccurlyeq O$ if and only if O weakly emerges from O' .*
- *Inclusion (based on Nominal Emergence) : $O' \Subset O$ if and only if O nominally emerges from O' .*

Proof With the convention that it can be said that an object emerges from itself, we have reflexivity (if such a convention seems absurd, we can define the relationships as O emerges from O' or $O = O'$). Transitivity is clearly contained in definitions of emergence.

Note that the inclusion relation is more than an inclusion between sets, as it translates an inclusion “inside” the elements of the ontology.

These relations are the basis for the construction of a graph called the *ontological graph* :

Definition 4 *The ontological graph is constructed by induction the following way :*

1. *A graph with vertices elements of $\mathcal{P}(O)$ and edges of two types : $E_W = \{(O, O') | O' \preccurlyeq O\}$ and $E_N = \{(O, O') | O' \Subset O\}$*
2. *Nodes are reduced⁹ by : if $o \in O, O'$ and $(O' \preccurlyeq O$ or $O' \Subset O)$ but not $(O \preccurlyeq O'$ or $O \Subset O')$, then $O' \leftarrow O' \setminus o$*
3. *Nodes with intersecting sets are merged, keeping edges linking merged nodes. This step ensures non-overlapping nodes.*

Minimal Ontological Tree

The topological structure of the graph, that contains in a way the *structure of the system*, can be reduced into a minimal tree that contains hierarchical structure essential to the theory.

We need first to give consistence to the system :

Definition 5 *A consistent part of the ontological graph is a weakly connected component of the graph. We assume for now to work on a consistent part.*

⁹ the reduction procedure aims to delete redundancy, keeping an entity at the higher level it exists.

The notion of consistent system, together with subsystem or nodes timescales that will be defined later, requires to reconstruct perspectives from ontological elements, i.e. the inverse operation of what was done in our deconstruction procedure.

Assumption 4 *There exists $\mathcal{O}' \subset \mathcal{P}(\mathcal{O})$ such that for any $O \subset \mathcal{O}'$, there exists a corresponding dataflow machine M such that the corresponding perspective is consistent with initial elements of the system (i.e. machines on ontology overlaps are equivalent). If $\Phi : M \mapsto O$ is the initial mapping, we denote this extended reciprocal construction by $M' = \Phi^{<-1>}(O)$.*

REMARK. This assumption could eventually be changed into a provable proposition, assuming that the coupling ontology is indeed a coupling perspective, which dataflow machine part is consistent with coupled entities. Therein, the decomposition postulate of [golden2012modeling] should allow to identify basic components corresponding to each element of the ontology, and then construct the new perspective by induction. We find however these assumptions too restrictive, as for example various ontological elements may be modeled by an irreducible machine, as a differential equations with aggregated variables. We prefer to be less restrictive and postulate the existence of the reverse mapping on some sub-ontologies, that should be in practice the ones where couplings can be effectively modeled.

Given this assumption, we can define the consistent system as the reciprocal image of the consistent part of the ontological graph. It ensures system connectivity what is a requirement for tree construction.

Proposition 4 *The tree decomposition of the ontological graph in which nodes contains strongly connected components is unique. The corresponding reduced tree, that corresponds to the ontological graph in which strongly connected components have been merged with edges kept, is called the Minimal Ontological Tree.*

Proof (sketch of) The unicity is obtained as nodes are fixed as strongly connected components. It is trivially a tree decomposition (with no edges) as in a directed graph, strongly connected components do not intersect, thus the consistence of the decomposition.

Any loop $O \rightarrow O' \rightarrow \dots \rightarrow O$ in the ontological graph assumes that all its elements are equivalent in the sense of \preccurlyeq . This equivalence loops should help to define the notion of strong coupling as an application of the theory (see applications).

The Minimal Ontological Tree (MOT) is a tree in the undirected sense but a forest in the directed sense. Its topology contains a sort of system hierarchy. Consistent subsystems are defined from the set \mathcal{B} of branches of the forest, as $(\Phi^{<-1>}(\mathcal{B}), \mathcal{B})$. The timescale of a node, and by extension of a subsystem, is the union of timescales

of corresponding machines. Levels of the tree are defined from root nodes, and the emergence relations between nodes implies a vertical inclusion between timescales.

Action on Data

Scales

Finally, we propose to define scales associated to a system. Following [manson2008does], an epistemological continuum of visions on scale is a consequence of differences between disciplines in the way we developed in the introduction. This proposition is indeed compatible with our framework, as the construction of scales for each level of the ontological tree results in a broad variety of scales.

Let (M, O) a subsystem and T the corresponding timescale. We propose to define the “thematic scale” (for example spatial scale) assuming a representation theorem, i.e. that an aspect (thematic aspect) of the machine can be represented as a dynamic state variable $\vec{X}(t)$. Assuming a scale operator¹⁰ $\|\cdot\|_s$ and that the state variable has a certain level of differentiability, the *thematic scale* if defined as $\|(d^k \vec{X}(t))_k\|_s$

10.2.3 Application

The particular case of geographical systems

In [dolfus1975some] DURAND-DASTÈS proposes a definition of geographical structure and system, structure would be the spatial container for systems viewed as complex open interacting systems (elements with attributes, relations between elements and inputs/outputs with external world). For a given system, its definition is a perspective, completed by structure to have a system in our sense. Depending on the way to define relations, it may be easy to extract ontological structure.

Modularity and co-evolving subsystems

For the example of Urban Systems, urban evolutionary theory enters this framework using our previous thematic theory ? The decomposition into uncorrelated subsystems yields precisely strongly coupled components as co-evolving components. The correlation between subsystems should be positively correlated with topological distance in the tree. If we define elements of a node before merging as *strongly coupled elements*, in the case of dynamic ontologies, it provides a definition of *co-evolution* and co-evolving subsystems equivalent to the thematic definition.

¹⁰ that can be of various nature : extent, probabilistic extent, spectral scales, stationarity scales, etc.

10.2.4 *Discussion*

LINK WITH EXISTING FRAMEWORKS A link with the Cottineau-Chapron framework for multi-modeling [[10.1371/journal.pone.0138212](#)] may be done in the case they add the bibliographical layer, which would correspond to the reconstruction of perspectives. [[reymond2013logique](#)] proposes the notion of “interdisciplinary coupling” what is close to our notion of coupling perspectives. A correspondance with System of Systems approaches (see e.g. [[luzeaux2015formal](#)] for a recent general framework englobing system modeling and system description) may be also possible as our perspectives are constructed as dataflow machines, but with the significant difference that the notion of emergence is central.

CONTRIBUTIONS TO THE STUDY OF COMPLEX SYSTEMS

- We do not claim to provide a theory of systems (beware of cybernetics, systemics etc. that could not model everything), but more a framework to guide research questions (e.g. in our case the direct outcomes will be quantitative epistemology that comes from system construction as perspectives research; empirical to construct robust ontologies for perspectives; targeted thematic to unveil causal relationship/emergence for construction of ontological network; study of coupling as possible processes containing co-evolution; study of scales; etc.). It may be understood as meta-theory which application gives a theory, the thematic theory developed before being a specific implementation to territorial networked systems.
- We Emphasize the notion of socio-technical system, crossing a social complex system approach (ontologies) with a description of technical artifacts (dataflow machines), taking the “best of both worlds”.

10.2.5 *Research Directions*

We can draw from the construction of this theoretical framework a set of research directions, that give a general line on how trying to answer to research questions asked after the thematic theory construction.

1. The perspectivist approach implies a broad understanding of existing perspectives on a system, and of possibility of coupling between them; thus an emphasis on applied epistemology, i.e. **Algorithmic Systematic Review** (exploration of the knowledge space), **Disciplines Mapping**(extraction of its structure) and **Datamining for Content Analysis**(refinement at the atomic level in scientific knowledge) that correspond to the three sections of chapter [4](#).

2. At a finer level of particularization, the knowledge of perspectives means **Knowledge of stylized facts**, i.e. empirical analysis of cases studies. These are the object of chapter [6](#).
3. The emphasis on coupled subsystems at different scales implies a deep understanding of coupling mechanisms, thus the need of methodological and technical developments : **Methods for Statistical Control**, **Methods for Model Exploration**, **Theoretical Study of Coupling**, **Multi-Modeling**, of which some are developed and other proposed in the methodological chapter [2](#).
4. Furthermore, the possibility of hidden elements within the ontology implies the test for causal relations and intermediate processes at the origin of emergence, thus e.g. the exploration of new paradigms such as role of governance within complex models as done in chapter [8](#).
5. Finally, the idea behind system structure contained within the ontological forest is a large set of coupled models for a given system : it means that a proper system definition (i.e. thematic problematization and exploration) and construction should yield to a structured family of models : parallel branches can be different implementations of the same process or various processes trying to explain the emerging ontology ; therefore the final objective of a family of models tackling the thematic question.

Cinquième partie

APPENDIX

GENERATION OF CORRELATED SYNTHETIC DATA

APPLICATION : FINANCIAL TIME-SERIES

Application : Séries temporelles financières

Context

Un premier domaine d'application proposé pour notre méthode est celui des séries temporelles financières, signaux typiques de systèmes complexes hétérogènes et multiscalaires [[mantegna2000introduction](#)] et pour lesquels les corrélations ont fait l'objet d'abondants travaux. Ainsi, l'application de la théorie des matrices aléatoires peut permettre de débruiter, ou du moins d'estimer la part de signal noyée dans le bruit, une matrice de correlations pour un grand nombre d'actifs échantillonnés à faible fréquence (retours journaliers par exemple) [[2009arXiv0910.1205B](#)]. De même, l'analyse de réseaux complexes construits à partir des corrélations, selon des méthodes type arbre couvrant minimal [[2001PhyA..299...16B](#)] ou des extensions raffinées pour cette application précise [[tumminello2005tool](#)], ont permis d'obtenir des résultats prometteurs, tels la reconstruction de la structure économique des secteurs d'activités. A haute fréquence, l'estimation précise de paramètres d'interdépendance dans le cadre d'hypothèses fixées sur la dynamique, fait l'objet d'importants travaux théoriques dans un but de raffinement des modèles et des estimateurs [[barndorff2011multivariate](#)]. Les résultats théoriques doivent alors être testés sur des jeux de données synthétiques, qui permettent de contrôler un certain nombre de paramètres et de s'assurer qu'un effet prédit par la théorie est bien observable *toutes choses égales par ailleurs*. Par exemple, [[potiron2015estimation](#)] dérive une correction du biais de l'estimateur de *Hayashi-Yoshida* qui est un estimateur de la covariance de deux browniens corrélés à haute fréquence dans le cas de temps d'observation asynchrones, par démonstration d'un théorème de la limite centrale pour un modèle généralisé endogénisant les temps d'observations. La confirmation empirique de l'amélioration de l'estimateur est alors obtenue sur un jeu de données synthétiques à un niveau de corrélation fixé.

Formalization

FRAMEWORK Considérons un réseau d'actifs $(X_i(t))_{1 \leq i \leq N}$ échantillonnés à haute fréquence (typiquement 1s). On se place dans un cadre multi-scalaire (utilisé par exemple dans les approches par onde-

lettes [ramsey2002wavelets] ou analyses multifractales du signal [bouchaud2000apparent] pour interpréter les signaux observés comme la superposition de composantes à des multiples échelles temporelles : $X_i = \sum_{\omega} X_i^{\omega}$. On notera $T_i^{\omega} = \sum_{\omega' \leq \omega} X_i^{\omega'}$ le signal filtré à une fréquence ω donnée. Prédire l'évolution d'une composante à une échelle donnée est alors un 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 corrélations respectives entre actifs et il est alors intéressant d'utiliser la méthode pour évaluer celle-ci en fonction de niveaux de corrélation à plusieurs échelles. On assume une dynamique de Black-Scholes [jarrow1999honor] pour les actifs, i.e. $dX = \sigma \cdot dW$ avec W processus de Wiener, ce qui permettra d'obtenir facilement des niveaux de corrélation voulus.

DATA GENERATION 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 $dW_1 \perp\!\!\!\perp dW_1^{\perp\!\!\!\perp}$ (et $\sigma_1 < \sigma_2$ pour fixer les idées, quitte à échanger les actifs), alors $W_2 = \rho_{12}W_1 + \sqrt{1 - \frac{\sigma_1^2}{\sigma_2^2} \cdot \rho_{12}^2} W_1^{\perp\!\!\!\perp}$ est tel que $\rho(dW_1, dW_2) = \rho_{12}$. Les signaux suivants sont construits de la même manière par orthonormalisation de Gram. On isole alors la composante à la fréquence ω_1 voulue par filtrage, c'est à dire $\tilde{X}_i^{\omega_1} = W_i - \mathcal{F}_{\omega_0}[W_i]$ (avec \mathcal{F}_{ω_0} filtre passe-bas à fréquence de coupure ω_0), puis on reconstruit les signaux synthétiques par $\tilde{X}_i = T_i^{\omega_0} + \tilde{X}_i^{\omega_1}$.

Implementation and Results

METHODOLOGY 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 échantillonées à l'ordre de la seconde, consiste dans un premier temps à la détermination du support temporel commun

¹ voir [gell1995quark] 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.

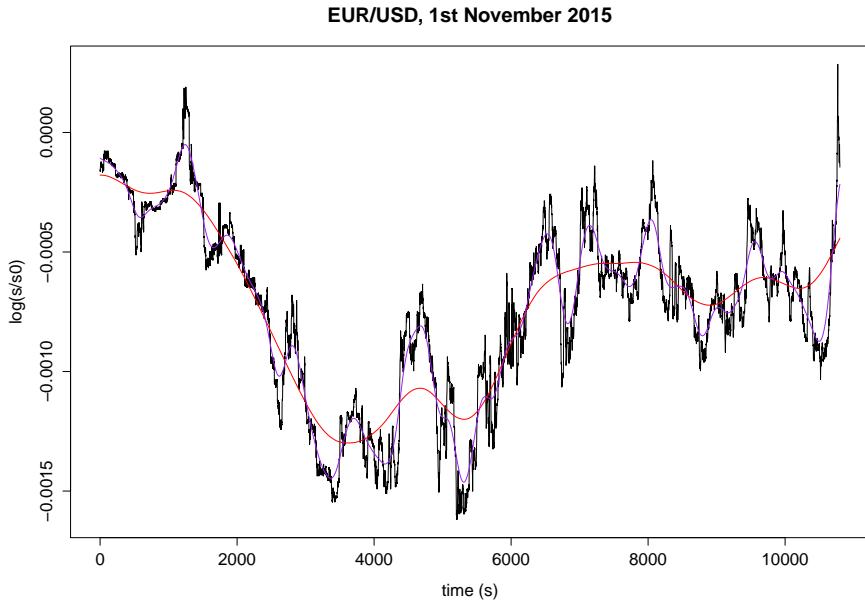


FIGURE 27 : Example of the multi-scalar structure of the signal, basis of the construction of synthetic signals | Log-prices are represented on a time window of around 3h for November 1st 2015 for asset EUR/USD, with 10min (purple) and 30min trends.

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 similaires 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 27 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

$$\rho_e = [\varepsilon_1 \varepsilon_2 \rho_0 + \rho] \cdot \left[1 - \frac{1}{2} (\varepsilon_1^2 + \varepsilon_2^2) \right] \quad (10)$$

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 $\hat{\rho}[X1, X2] = \frac{\hat{C}[X1, X2]}{\sqrt{\hat{V}\text{ar}[X1] \hat{V}\text{ar}[X2]}}$, où $\hat{C}[X1, X2] = \frac{1}{(T-1)} \sum_t X_1(t)X_2(t) - \frac{1}{T \cdot (T-1)} \sum_t X_1(t) \sum_t X_2(t)$ 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 $t + 1$.

IMPLEMENTATION L'implémentation est faite en language R, utilisant en particulier la bibliothèque MTS [[Tsay:2015xy](#)] pour les modèles de séries temporelles. Les données nettoyées et le code source sont disponibles de manière ouverte sur le dépôt `git` du projet⁴.

RESULTS 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 prédite par 10 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 [[potiron2016estimating](#)] qui suppose une dynamique localement paramétrique, mais pour lequel on fixe les métaparamètres de la dynamique. On pourrait imaginer estimer un T_W variable qui s'adapterait pour une meilleure estimation locale, à l'image de l'estimation de paramètres en traitement du signal Bayesien effectuée via augmentation de l'état par les paramètres.

⁴ at <https://github.com/JusteRaimbault/SynthAsset>

1h, 1h30, 2h sont présentés en figure 29. 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

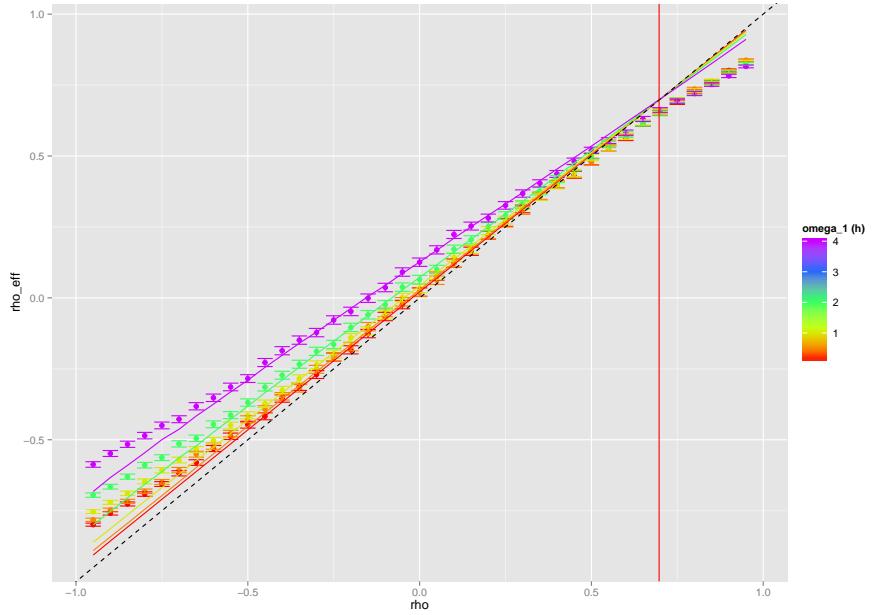


FIGURE 28 : Effective correlations obtained on synthetic data | Dots represent estimated correlations on a synthetic dataset corresponding to 6 months between June and November 2015 (error-bars give 95% confidence intervals obtained with standard Fisher method); scale color gives filtering frequency $\omega_1 = 10\text{min}, 30\text{min}, 1\text{h}, 2\text{h}, 4\text{h}$; solid lines give theoretical values for ρ_e obtained by 10 with estimated volatilities (dotted-line diagonal for reference); vertical red line position is the theoretical value such that $\rho = \rho_e$ with mean values for ε_i on all points. We observe for high absolute correlations values a deviation from corrected values, what should be caused by non-verified independence and centered returns assumptions. Asymmetry is caused by the high value of $\rho_0 \simeq 0.71$.

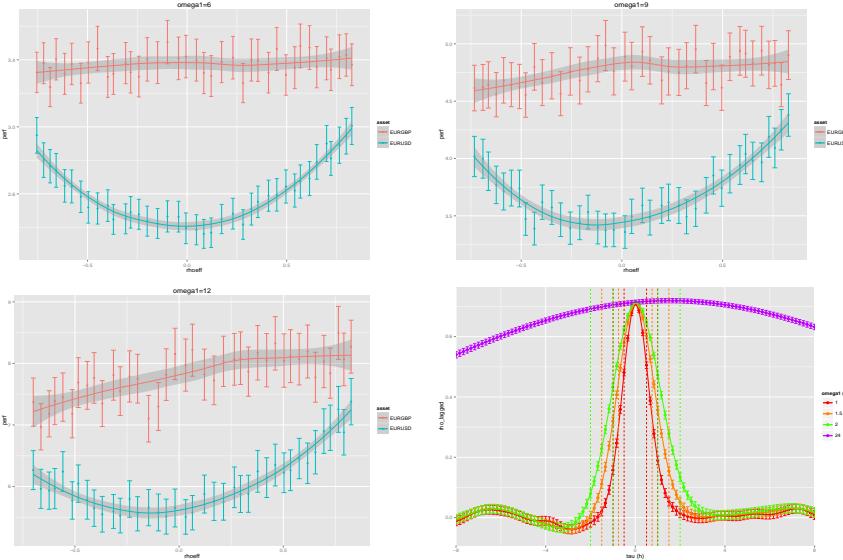


FIGURE 29 : Performance of a predictive model as a function of simulated correlations | From left to right and top to bottom, three first graphs show for each asset the normalized performance of an ARMA model ($p = 2, q = 0$), defined as $\pi = \left(\frac{1}{T} \sum_t (\tilde{X}_i(t) - M_{\omega_1}[\tilde{X}_i](t))^2 \right) / \sigma[\tilde{X}_i]^2$ (95% confidence intervals computed by $\pi = \bar{\pi} \pm (1.96 \cdot \sigma[\pi]) / \sqrt{T}$, local polynomial smoothing to ease reading). It is interesting to note the U-shape for EUR/USD, due to interference between components at different scales. Correlation between simulated noises deteriorates predictive power. The study of *lagged correlations* (here $\rho[\Delta X_{EURUSD}(t), \Delta X_{EURGBP}(t - \tau)]$) on real data clarifies this phenomenon : fourth graph show an asymmetry in curves at any scale compared to zero lag ($\tau = 0$) what leads fundamental components to increase predictive power for the dollar, amelioration then perturbed by correlations between simulated components. Dashed lines show time steps (in equivalent τ units) used by the ARMA at each scale, what allows to read the corresponding lagged correlation on fundamental component.

AN INTERDISCIPLINARY APPROACH TO
MORPHOGENESIS

This Appendix was submitted as an Essay Paper with C. Antelope (U. California), L. Hubatsch (F. Crick Institute) and J.M. Serna (Université Paris 7), as :

Antelope, C., Hubatsch, L., Raimbault, J., and Serna, J. M. (2016). An interdisciplinary approach to morphogenesis. *Forthcoming in Proceedings of Santa Fe Institute CSSS 2016.*

13

TECHNICAL DEVELOPMENTS

C'est hardcore tes calculs.

- ANONYME

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

13.1 DERIVATIONS FOR URBAN GROWTH MODELS

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. ■

13.2 SENSITIVITY OF URBAN SCALING

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 [anas1998urban]. Taking a geographical space as \mathbb{R}^2 , we take for any $\vec{x} \in \mathbb{R}^2$ the density of population as

$$d(\vec{x}) = \sum_{i=1}^N d_i(\vec{x}) = \sum_{i=1}^N d_i^0 \cdot \exp\left(\frac{-\|\vec{x} - \vec{x}_i\|}{r_i}\right) \quad (11)$$

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 [cottineau2015scaling], 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 [keersmaecker2003using] and that it implies necessarily a power-law distribution [chen2010characterizing]. 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 $P_i(\theta) = \iint_{D(\vec{x}_i, \theta)} d$.

13.2.1 Analytical Derivation of Sensitivity

With above notations, let derive the expression of estimated exponent for quantity a as a function of density threshold parameter θ . The quantity computed for a given city i is, thanks to the monocentric assumption and in a spatial range and a range for θ such that $\theta \gg \sum_{j \neq i} d_j(\vec{x})$, allowing to approximate $d(\vec{x}) \simeq d_i(\vec{x})$ on $D(\vec{x}_i, \theta)$, is computed by

$$\begin{aligned} A_i(\theta) &= \lambda \cdot \iint_{D(\vec{x}_i, \theta)} d^\beta = 2\pi\lambda d_i^{0\beta} \int_{r=0}^{r_0 \ln \frac{d_i^0}{\theta}} r \exp\left(-\frac{r\beta}{r_0}\right) dr \\ &= \frac{2\pi d_i^{0\beta} r_0^2}{\beta^2} \left[1 + \beta \ln \frac{\theta}{d_i^0} \left(\frac{\theta}{d_i^0} \right)^\beta - \left(\frac{\theta}{d_i^0} \right)^\beta \right] \end{aligned}$$

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 :

$$\alpha(\theta) = \frac{k_0 + k_1 \theta + k_2 \theta^\beta + k_3 \theta^{\beta+1} + k_4 \theta \ln \theta + k_5 \theta^\beta \ln \theta + k_6 \theta^\beta (\ln \theta)^2 + k_7 \theta^{\beta+1} (\ln \theta)^2}{k_0' + k_1' \ln \theta + k_2' \theta \ln \theta + k_3' \theta^2 + k_4' \theta^2 \ln \theta + k_5' \theta^2 (\ln \theta)^2} \quad (12)$$

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.

13.2.2 Numerical Simulations

IMPLEMENTATION We implement empirically the density model given in section 13.2. 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 . In practice, adapting N to world size allows to respect the monocentric condition. Population are distributed in order to follow the scaling law

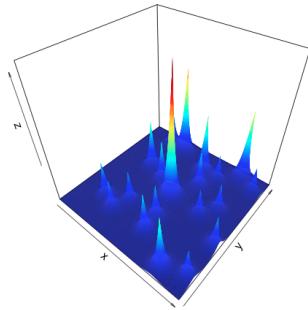


FIGURE 30 : Example of a synthetic density distribution obtained with the exponential mixture, with a grid of size 400×400 and parameters $N = 20$, $r_0 = 10$, $P_{\max} = 200$, $\alpha = 0.5$, $\theta_C = 0.01$.

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:2015fk] and using the package `kernlab` for efficient kernel mixture methods [Karatzoglou:2004uq] are given as comments in source code¹. We show in figure 30 example of synthetic density distributions on which the numerical study is conducted. 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.

RANDOM PERTURBATIONS 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.

KERNEL TYPE 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.

¹ available at <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Scaling>

*You must not be afraid of putting
code in your thesis, code is not dirty*
- ALEXIS DROGOUL

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.

14.1 ALGORITHMIC SYSTEMATIC REVIEW

Revue Systématique Algorithmique

OBJECTIVE Implement systematic literature review algorithm.

LOCATION <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Biblio/AlgoSR/AlgoSRJavaApp>

CHARACTERISTICS

- Language : Java
- Size : 7116

PARTICULARITIES

- 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.

ADDITIONAL SCRIPTS R for result exploration and visualization.

14.2 INDIRECT BIBLIOMETRICS

Bibliométrie Indirecte

OBJECTIVE Hypernetworks analysis of cybergeo journal.

LOCATION <https://github.com/Geographie-cites/cybergeo20/tree/master/HyperNetwork>

<https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Biblio/AlgoSR/AlgoSRJavaApp> for common Java part.

CHARACTERISTICS

- Language : Python, R and Java.
- Size : -

PARTICULARITIES Polyglot

ARCHITECTURE See schema chapter 3.

ADDITIONAL SCRIPTS -

14.3 DENSITY URBAN GROWTH

Croissance Urbaine

OBJECTIVE Simple density urban growth model.

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

CHARACTERISTICS

- Language : NetLogo then scala.
- Size : 4355

PARTICULARITIES Morphological indicators in scala implemented with Fast Fourier transform; with R communication in NetLogo.

ARCHITECTURE Nothing particular.

ADDITIONAL SCRIPTS R for result exploration and morphological analysis.
 oms for model exploration.

14.4 CORRELATED DATA GENERATION

Génération des Données Synthétiques Corrélées

OBJECTIVE Weak coupling of density generation and network generation.

LOCATION https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Synthetic/Network_20151229

CHARACTERISTICS

- Language : NetLogo (network) and scala.
- Size : 3188

PARTICULARITIES Network heuristic easier to implement and explore in netlogo

ARCHITECTURE OpenMole allows coupling between modules through exploration script.

ADDITIONAL SCRIPTS R for result exploration.
 oms for model exploration.

14.5 LUTECIA MODEL

Modèle Lutecia

OBJECTIVE Implementation of Lutecia model, chapter 7.

LOCATION <https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Governance/MetropolSim/Lutecia>

CHARACTERISTICS

- Language : NetLogo
- Size : 4791

PARTICULARITIES Shortest path dynamical programming using matrices.

ARCHITECTURE Pseudo object architecture in agent environment.

ADDITIONAL SCRIPTS R for result exploration.
oms for model exploration.

14.6 NETWORK ANALYSIS

Analyse des Réseaux

OBJECTIVE Simplification of european road network

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

CHARACTERISTICS

- Language : R, Shell, PostgreSQL
- Size : 505

PARTICULARITIES Handling of large size databases imposes sequential processing; use of external program osmosis for conversion from osm data to pgsql.

ARCHITECTURE Shell script lead manoeuvres.

ADDITIONAL SCRIPTS -

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.

15.1 NETLOGO DOCUMENTATION GENERATOR

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>.

15.2 GIT AS A REPRODUCIBILITY TOOL

git comme outil de reproductibilité

The use of git as a reproducibility and transparency tool was emphasized in [ram2013git] (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.

15.3 GIT-DATA

git-data

git-data is a shell based (experimental) git extension, available at <https://github.com/JusteRaimbault/gitdata>, that allows automated backup of large file within a git repository, their transparent integration in ignored files and the creation of symbolic links for a transparent local use.

15.4 TOWARDS A GIT-COMPATIBLE FIGURES METADATA HANDLER

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.

15.5 TORPOOL

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.