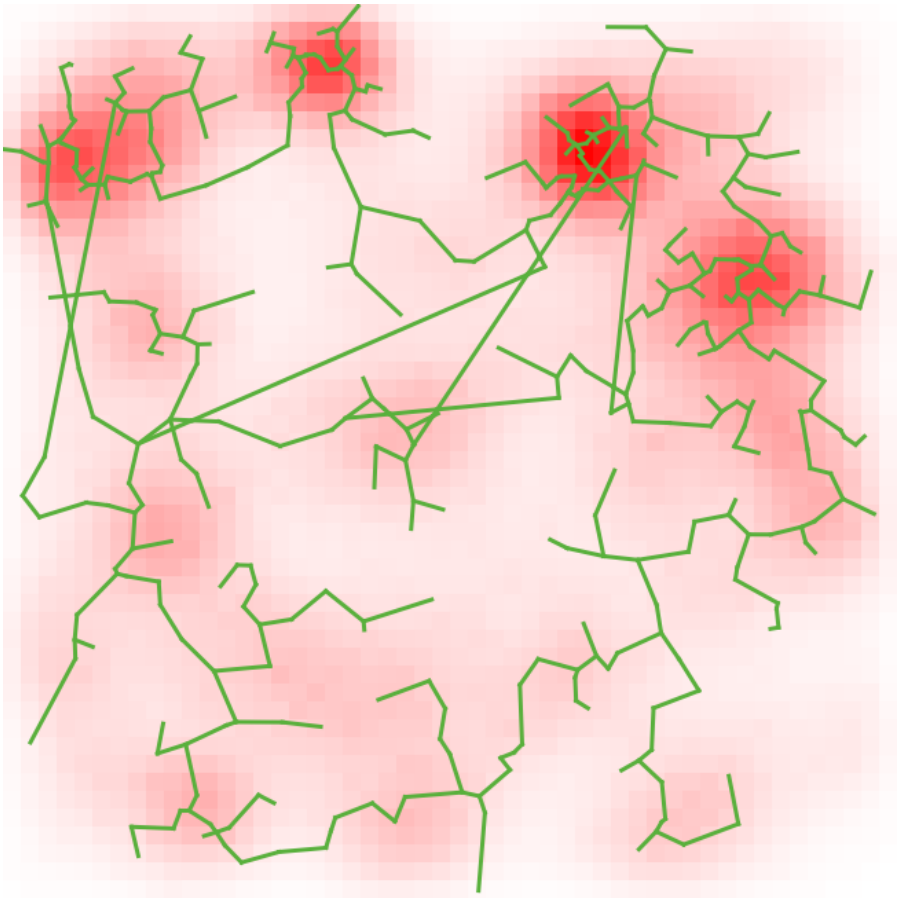
TOWARDS MODELS COUPLING URBAN GROWTH AND TRANSPORTATION NETWORK GROWTH

JUSTE RAIMBAULT



PhD Thesis First Year Preliminary Memoire

Under the supervision of Arnaud Banos and Florent Le Nechet

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

Université Paris VII

March 2016 – version 1.0

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

Our relation to our environment and people changes at time scales we generally expect larger. As a witness, we include this preliminary dedication, for comparison purposes with the final version.

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

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

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 Nechet, F. & Raimbault, J. (2015). Modeling the emergence of metropolitan transport authorithy 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

TABLE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION

It's when you shuffle the anthill that you get a touch of all its complexity.
- 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 écolutions 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 concentrons à présent sur l'originalité du point de vue que nous allons prendre.

SCIENTIFIC CONTEXT : COMPLEXITY HAS COME OF AGE

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 approaches.

Ces considérations sont bien connus 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 sociaux[1]. 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 sense ou les lois macroscopiques ne sont pas suffisantes pour inférer les propriétés macroscopiques du

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**].

systèmes 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. Reasons for the recent growth of works claiming a CS approach may be various. The explosion of computing power is surely one because of the central role of numerical simulations [**varenne2010simulations**]. They could also be the related epistemological progresses : apparition of the notion of perspectivism [**giere2010scientific**], finer reflexions around the notion of model [**varenne2013modeliser**][2]. The theoretical and empirical potentialities of such approaches play surely a role in their success[3], as confirmed in various domains of application (see [**newman2011complex**] for a general survey), as for example Network Science [**barabasi2002linked**] ; Neuroscience [**koch1999complexity**] ; Social Sciences ; Geography [**manson2001simplifying**][**pumain19** Finance with the rising importance of econophysics [**stanley1999econophysics**] ; Ecology [**grimm2005pattern**]. The Complex Systems Roadmap [**2009arXiv0907.2221B**] proposed a double lecture of studies on Complex Systems : an horizontal approach connecting fields of study with transversal questions on theoretical foundations of complexity and empirical common stylized facts, and a vertical conceptions of disciplines, with the aim to construct integrated disciplines and corresponding multi-scale heterogeneous models. Interdisciplinarity is thus central in our scientific background.

2 In that frame scientific and epistemological progress can not be dissociated and can be seen as coevolving

3 Although the adoption of new scientific practices may be strongly biased by imitation and lack of originality [**dirk1999measure**], or more ambivalent, by marketing strategies as the fight for funds is becoming a huge obstacle for research [**bollen2014funding**].

INTERDISCIPLINARITY

We must further insist on the role of interdisciplinarity in the positions taken here. This is not a thesis in Geography nor in Complex Adaptive Systems Modeling, but in *Complex Systems Science* that we claim as a proper discipline following PAUL BOURGINE. It will naturally be seen with defiance by scholars of various concerned disciplines, as recent examples of misunderstandings and conflicts have illustrated [**dupuy2015sciences**]. The positioning of BATTY proposing *A new Science of Cities* [**batty2013new**] (that he subtly presents as *The* new science of cities) is directed towards an integration of disciplines and methods into a science defined by its object of study, cities. Its theoretical and epistemological weaknesses (no theoretical constructions of studied geographical objects on the one hand, approximative contextualization of complexity) combined with an overall impression of *pot-pourri* of forgotten works (space syntax, land-use models), unfortunately avoid us to use it as we will use geographical theories (e.g. evolutive urban theory) in an appropriated epistemological complexity context. Yet our reading of this work may be the result of a misunderstanding due to different cultural backgrounds.

The scientific evolution of complexity that some see as a revolution [**colander2003complexity**], or even as *a new kind of science* [**wolfram2002new**], could indeed face intrinsic difficulties due to behaviors and a-priori of researchers as human beings. More precisely, the need of interdisciplinarity that makes the strength of Complexity Science may be one of its greatest weaknesses, since the highly partitioned structure of science organization has sometimes negative impacts on works involving different disciplines. We do not tackle the issue of overpublication, competition, indexes, which is more linked to a question of open science and its ethics, also of high importance but of an other nature. That barrier we are dealing with and we might struggle to triumph of, is the impact of certains *cultural disciplinary differences* and out-coming conflicts on views and approaches. The drama of scientific misunderstandings is that they can indeed annihilate progresses by interpreting as a falsification some work that answers to a totally different question. The example of a recent work on top-income inequalities given in [**aghion2015innovation**], which conclusions are presented as opposed from the one obtained by Piketty [**piketty2013capital**], follows such a scheme. Whereas Piketty focused on constructing long-time clean databases for income data and showed empirically a recent acceleration of income inequalities, his simple model aiming to link this stylized fact with the accumulation of capital has been criticized as oversimplified. On the other hand, Bergeaud *et al.* prove by a model of innovation economics that *under certain assumptions* income gaps may be beneficial to innovation and consequently a general utility. Thus diverging conclusions about

the role of personal capitals in the economy. But diverging *views* or *interpretations* does not mean a scientific incompatibility, and one could imagine try to gather both approaches in an unified framework and model, yielding possibly similar or different interpretations. This integrated approach has chances to contain more information (depending on how coupling is done) and to be a further advance in Science. We shall now briefly develop other examples to give an overview how conflicts between disciplines can be damaging.

PHYSICS REINVENTS GEOGRAPHY. As already mentioned, DUPUY and BENGUIGUI points out in [**dupuy2015sciences**] the fact that urban sciences have recently known open conflicts between old tenants of the disciplines and new arrivants, especially physicists. The availability of large datasets of new type of data (social networks, ICT data) have drawn their attention towards the study of objects traditionally studied by human science, as analytical and computational methods of statistical physics became applicable. Although these studies are generally presented as the construction of a scientific approach to cities, implying that existing knowledge was not scientific because of their more qualitative aspect, they have not unveil specifically novel knowledge on urban systems : to give some examples, [**barthelemy2013self**] concludes that Paris has followed a transition during Haussman period and that the evolution of a city is the combination of local transformations and global planning operations, what are facts known for a long time in urban history and urban geography. [**chen2009urban**] rediscovers that the gravity model can be improved by adding lags in interactions and theoretically derives the expression of the force of interaction between cities, without any thematic theoretical background. Examples could be multiplied, confirming the current discomfort in communication between physicists and urban geographers. Significant benefices could results from a wise integration of disciplines [**02015physicists**] but the road seems still long.

ECONOMIC GEOGRAPHY OR GEOGRAPHICAL ECONOMICS ? Similar conflict occurred in economics : as [**marchionni2004geographical**] describes, the discipline of economic geography, traditionally close from geography, heavily criticized a new stream of thought named *geographical economics*, which purposes is spatialization of mainstream economic techniques. Both do not have the same purposes and aims, and the conflict appears as a total misunderstanding for an external observer.

AGENT-BASED MODELING IN ECONOMY Disciplinary conflicts may also manifest themselves as the reject of novel methods by mainstream currents. Following [**farmer2009economy**], the operational failure of most classic economic approaches could be compensated by

a broader use of agent-based modeling and simulation practices. The lack of analytical framework that is natural in the study of complex adaptive systems seems to be rebutting for most of economists.

FINANCE In Quantitative Finance coexist various stream of research with a very few interactions. Let consider two examples. On the one hand, Statistics are highly advanced in theoretical mathematics, using stochastic calculus and probabilities to obtain very refined estimators of parameters for a given model (see e.g. [**barndorff2011multivariate**]). On the other hand, Econophysics aims to study empirical stylized facts and infer empirical laws to explain complexity-related phenomena in financial systems [**stanley1999econophysics**], such as e.g. cascades leading to market crashes, fractal properties of asset signals, complex structure of correlation networks. Both have their advantages in a particular context and each would benefit from increased interactions between the fields.

These diverse examples illustrate how interdisciplinarity is both crucial and difficult to achieve. We will try to follow that narrow path in our work, borrowing ideas, theories and methods from various disciplines, aiming for the construction of an integrated knowledge. Indeed, coupling heterogeneous approaches at different levels and scales will be a cornerstone of our thesis, skeleton of the underlying philosophy and building brick of the theory we will propose.

COMPLEXITY IN GEOGRAPHY

Coming back to our introducing anecdote, we will focus on our thematic object of study that are territorial systems. More generally, we propose an overview of the role of complexity in geography. Geographers are familiar with complexity for a long time, as the study of spatial interactions is one of its purposes. The variety of fields in geography (geomorphology, physical geography, environmental geography, human geography, health geography to give a few) has certainly been important in the subtlety of the geographical thinking, that considers heterogeneous and multi-scalar processes.

PUMAIN recalls in [**pumain2003approche**] a subjective history of the emergence of complexity paradigms in geography. Cybernetics yielded system theories as the one developed by Forrester. Later the shift to self-organized criticality and self-organisation concepts in physics conducted to corresponding developments in geography, as [**sanders1992systeme**] witnesses the application of the concepts of synergetics for the dynamics of an urban system. Finally, Complex Systems paradigms as we currently know them appeared from various points of view. For example, the fractal nature of urban shape was introduced in [**batty1994fractal**] and had numerous application including more recent developments [**keersmaecker2003us**
BATTY also introduced cellular automata in urban modeling and pro-

posed a joint synthesis with agent-based modeling and fractals in [**batty2007cities**]. An other incursion of complexity in geography was for the case of urban systems through the evolutive urban theory of Pumain. In close relation with modeling from the beginning (the first Simpop model described in [**sanders1997simpop**] enters the theoretical framework of [**pumain1997pour**]), this theory aims to understand system of cities as systems of co-evolving adaptive agents, interacting in many ways, with particular features emphasized such as the diffusion of innovation. The series of Simpop models [**pumain2012multi**] focused in testing various assumptions of the theory. For example, different underlying mechanisms were revealed for european city systems and city system of the united states [**bretagnolle2010comparer**]. At other time scales and in other contexts, the SimpopLocal model [**schmitt2014modelisation**] aimed to investigate the conditions for the emergence of hierarchical urban systems from disparate settlements. A minimal model (in the sense of sufficient and necessary parameter) has been isolated thanks to the use of intensive computation with the model exploration software OpenMole [**schmitt2014half**], what was a result analytically not derivable for this kind of complex model. The technical progresses of OpenMole [**reuillon20130penmole**] were done simultaneously with theoretical and empirical advancements. Epistemological advances were also essential to this framework, as Rey develops in [**rey2015plateforme**], and novel concepts such as incremental modeling [**cottineau2015incremental**] were found, with powerful concrete applications : [**cottineau2014evolution**] implemented it on the soviet city system and isolated dominating socio-economic processes, by systematic testing of thematic assumptions and implementation functions. Directions for the development of such Modeling and Simulation practices in quantitative geography were recently introduced by Banos in [**banos2013pour**]. He concludes with nine principles[4], from which we can cite the importance of intensive exploration of computational models and the importance of heterogeneous model coupling, that are among other principles such as reproducibility at the center of the study of complex geographical systems from the point of view described just before. Positioning in the legacy of this line of research, we will conjointly work in the theoretical, empirical, epistemological and modeling domains.

research question

Research question and precise objects are deliberately fuzzy for now, as we postulate that the construction of a problematic can not be dissociated from the production of a corresponding theory. Reciprocally, it makes no sense to ask questions out of the blue, on objects that

4 I remember René Doursat insisting on the search of the last commandement of Banos

have been only partially or rapidly defined. Our preliminary question to enter the subject, that we can obtain from concrete cases such as our introducing anecdote or from preliminary literature review, is the following :

Is it possible to produce a definition of territorial systems, and corresponding scales and ontologies, that would yield a natural, consistent and informational view on processes ?

Indeed, a necessary characteristic of territorial systems is their spatio-temporal nature, that is contained in spatio-temporal dynamics. The notion of *process* in the sense of [**hypergeo**] captures furthermore causal relationships in these dynamics, and is thus an interesting approach for an understanding of such systems. *Scale* must be understood here in the operational sense (physical characteristic) and *ontology* as real-world studied objects[5]. Our question may be roughly viewed as a search for theories and models that would unveil some processes involved in complex systems containing at least human settlements, the last requirement being crucial for a convergent problematic construction rather than ending in non-realistic and non-constructive propositions to understand everything between the brain (that can be seen as one building brick of territorial systems as they emerge from human social constructions) and the ecosphere that includes territorial systems.

CONTENTS

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

5 this use of ontology here naturally biaises our research towards modeling paradigms as it is close from the notion of ontology used in [**livet2010**], but we take the position (largely developed further) to understand any scientific construction as *models*, making the frontier between theory and modeling less relevant than in standard views. Any theory has to make choices on described objects, relations and processes, and therefore contains an ontology in that sense.

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

THEMATIC, THEORETICAL AND METHODOLOGICAL 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.

1

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 originairement 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 on 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 grand échelle par les interactions entre agents à une échelle donnée, en général concrétisée par un systèmes 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 2), 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

1.1.1 *Territories and Networks : There and Back Again*

HUMAN TERRITORIES The notion of territory can be taken as a basis to explore the scope of geographical objects we will study. In Ecology, a territory corresponds to a spatial extent occupied by a group of agents or more generally an ecosystem. *Human Territories* are far more complex in the sense of semiotic representations of these that are a central part in the emergence of societies. For RAFFESTIN in [**raffestin1988reperes**], the so-called *Human Territoriality* is the "conjonction of a territorial process with an informational process", what means that physical occupation and exploitation of space by human societies is not dissociable from the representations (cognitive and material) of these territorial processes, driving in return its further evolutions. In other words, as soon as social constructions are assumed in the constitution of human settlements, concrete and abstract social structures will play a role in the evolution of the territorial system, through e.g. propagation of information and representations, political processes, conjonction or disjonction between lived and perceived territory. Although this approach does not explicitly give the condition for the emergence of a seminal system of aggregated settlements (i.e. the emergence of cities), it insists on the role of these that become places of power and of creation of wealth through exchange. But the city has no existence without its hinterland and the territorial system can not be summarized by its cities as a system of cities. There is however compatibility on this subsystem between RAFFESTIN approach to territories and PUMAIN's evolutive theory of urban systems [**pumain2010theorie**], in which cities are viewed as an auto-organized complex dynamical systems, and act as mediators of social changes : for example, cycles of innovation occur within cities and propagate between them. Cities are thus competitive agents that co-evolve (in the sense given before). The territorial system can be understood as a spatially organized social structure, including its concrete and abstract artifacts. A imaginary free-of-man spatial extent with potential ressources will not be a territory if not inhabited, imagined, lived, and exploited, even if the same ressources would be part of the corresponding habited territorial system. Indeed, what is considered as a ressource (natural or artificial) will depend on the corresponding society (e.g. of its practices and technological potentialities). A crucial aspect of human settlements that were studied in geography for a long time, and that relate with the previous notion of territory, are *networks*. Let see how we can switch from one to the other and how their definition may be indissociable.

A TERRITORIAL THEORY OF NETWORKS We paraphrase DUPUY in [**dupuy1987vers**] when he proposes elements for "a territorial theory of networks" based on the concrete case of Urban Transportation Networks. This theory sees *real networks* (i.e. concrete networks, including transportation networks) as the materialization of *virtual networks*. More precisely, a territory is characterized by strong spatio-temporal discontinuities induced by the non-uniform distribution of agents and ressources. These discontinuities naturally induce a network of "transactional projects" that can be understood as potential interactions between elements of the territorial system (agents and/or ressources). For example today, people need to access the ressource of employments, economic exchanges operate between specialized production territories. At any time period, potential interactions existed[1]. The potential interaction network is concretized as offer adapts to demand, and results of the combination of economic and geographical constraints with demand patterns, in a non-linear way through agents designed as *operators*. This process is not immediate, leading to strong non-stationarity and path-dependance effects : the extension of an existing network will depend on previous configuration, and depending on involved time scales, the logic and even the nature of operators may have evolved. RAFFESTIN points out in his preface of [**offner1996reseaux**] that a geographical theory articulating space, network and territories had never been consistently formulated. It appears to still be the case today, but the theory developed just before is a good candidate, even if it stays at a conceptual level. The presence of a human territory necessarily imply the presence of abstract interaction networks and concrete networks used for transportation of people and ressources (including communication networks as information is a crucial ressource). Depending on regime in which the considered system is, the respective role of different networks may be radically different. Following DURANTON in [**duranton1999distance**], pre-industrial cities were limited in growth because of limitations of transportation networks. Technological progresses have lead to the end of these limitations and the preponderance of land markets in shaping cities (and thus a role of transportation network as shaping prices through accessibility), and recently to the rising importance of telecommunication networks that induce a "tyranny of proximity" as physical presence is not replaceable by virtual communication. This territorial approach to networks seems natural in geography, since networks are studied conjointly with geographical objects with an underlying theory, in opposition to network science that studies brutally spatial networks with few thematic background [**ducruet2014spatial**].

1 even when nomadism was still the rule, spatially dynamic networks of potential interactions necessarily existed, but should have less chance to materialize into concrete routes.

NETWORKS SHAPING TERRITORIES ? However networks are not only a material manifestation of territorial processes, but play their part in these processes as they evolution may shape territories in return. In the case of *technical networks*, an other designation of real networks given in [**offner1996reseaux**], many examples of such feedbacks can be found : the interconnectivity of transportation networks allows multi-scalar mobility patterns, thus shaping the lived territory. At a smaller scale, changes in accessibility may result in an adaptation of a functional urban space. Here emerges again an intrinsic difficulty : it is far from evident to attribute territorial mutations to some network evolutions and reciprocally materialization of a network to precise territorial dynamics. Coming back to Diderot should help, in the sense that one must not consider network nor territories as independent systems that would have causal relationships but as strongly coupled components of a larger system. The confusion on possible simple causal relationships has fed a scientific debate that is still active. Methodologies to identify so-called *structural effects* of transportation networks were proposed by planners in the seventies [**bonnafous1974detection**, **bonnafous1974methodologies**]. It took some time for a critical positioning on unreasoned and decontextualized use of these methods by planners and politics generally to technocratically justify transportation projects, that was first done by OFFNER in [**offner1993effets**]. Recently the special issue [**espacegeo2014effets**] on that debate recalled that on the one hand misconceptions and misuses were still greatly present in operational and planning milieus as [**crozet:halshs-01094554**] confirmed, and on the other hand that a lot of scientific progresses still need to be made to understand relations between networks and territories as PUMAIN highlights that recent works gave evidence of systematic effects on very long time scales (as e.g. the work of BRETAGNOLLE on railway evolution, that shows a kind of structural effect in the necessity of connectivity to the network for cities to "stay in the game", but that is not fully causal as not sufficient). At a macroscopic level typical patterns of interaction emerge, but microscopic trajectories of the system are essentially chaotic : the understanding of coupled dynamics strongly depends on the scale considered. At a small scale it seems indeed impossible to show systematic behavior, as OFFNER pointed out. For example, on comparable French mountain territories, [**berne2008ouverture**] shows that reactions to a same context of evolution of the transportation network can lead to very different reactions of territories, some finding a huge benefit in the new connectivity, whereas others become more closed. These potential retroactions of networks on territories does not necessarily act on concrete components : CLAVAL shows in [**claval1987reseaux**] that transportation and communication networks contribute to the collective representation of territories by acting on territorial belonging feeling.

TERRITORIAL SYSTEMS This detour from territories, to networks and back again, allows us to give a preliminary definition of a territorial system that will be the basis of our following theoretical considerations. As we emphasized the role of networks, the definition takes it into account.

Preliminary Definition. *A territorial system is a human territory to which both interaction and real networks can be associated. Real networks are a component of the system, involved in evolution processes, through multiples feedbacks with other components at various spatial and temporal scales.*

This reading of territorial systems is conditional to the existence of networks and may discard some human territories, but it is a deliberate choice that we justify by previous considerations, and that drives our subject towards the study of interactions between networks and territories.

1.1.2 *Transportation Networks*

THE PARTICULARITY OF TRANSPORTATION NETWORKS Already evoked in relation to the question of structural effects of networks, transportation networks play a determining role in the evolution of territories. Although other types of networks are also strongly involved in the evolution of territorial systems (see e.g. the discussions of impacts of communication networks on economic activities), transportation networks shape many other networks (logistics, commercial exchanges, social concrete interactions to give a few) and are prominent in territorial evolution patterns, especially in our recent societies that has become dependent of transportation networks [**bavoux2005geographie**]. The development of French High Speed Rail network is a good illustration of the impact of transportation networks on territorial development policies. Presented as a new era of railway transportation, a top-down planning of totally novel lines was introduced as central for developments [**zembri1997fondements**]. The lack of integration of these new networks with existing ones and with local territories is now observed as a structural weakness and negative impacts on some territories have been shown [**zembri2008contribution**]. A review done in [**bazin2011grande**] confirms that no general conclusions on local effects of High Speed lines connection can be drawn although it keeps a strong place in imaginaries. These are examples of how transportation networks have both direct and indirect impacts on territorial dynamics. Integrated planning, in the sense of a joint planning of transportation infrastructures and urban development, considers the network as a determining component of the territorial system. Parisian *Villes Nouvelles* are such a case, that witnesses of the complexity of such planning actions that generally do not lead to the

desired effect [**es119**]. Recent projects as [**l2012ville**] have try to implement similar ideas but we have now not enough temporal scope to judge their success in effectively producing an integrated territory. Transportation networks are anyway at the center of these approaches of urban territories. We will focus in our work on transportation networks for the various reasons given here.

DECONSTRUCTING ACCESSIBILITY The notion of accessibility comes rapidly when considering transportation networks. Based on the possibility to access a place through a transportation network (including transportation speed, difficulty of travel), it is generally described as a potential of spatial interaction[2] [**bavoux2005geographie**]. This object is often used as a planning tool or as an explicative variable of agents localisation for example. We must warn here on the potential dangers of its unconditional use. More precisely, it may be a construction that misses a consistent part of territorial dynamics. The mystification of the notion of *mobility* was shown by COMMENGES in [**commenges:tel-00923682**], which proved than most of debates on modeling mobility and corresponding notions were mostly made-of by transportation administrators of *Corps des Ponts* who roughly imported ideas from the United States without adaptation and reflexion fit to the totally different French context. Accessibility may be such a social construct and have no theoretical root since it is mostly a modeling and planning tool. Recent debates on the planification of *Grand Paris Express* [**confMangin**], a totally novel metropolitan transportation infrastructure planned to be built in the next twenty years, have revealed the opposition between a vision of accessibility as a right for disadvantaged territories against accessibility as a driver of economic development for already dynamic areas, both being difficultly compatible since corresponding to very different transportation corridors. Such operational issues confirm the complexity of the role of transportation networks in the dynamics of territorial systems, and we shall give in our work elements of response to a definition of accessibility that would integrate intrinsic territorial dynamics.

SCALES AND HIERARCHIES An incontournable aspect of transportation networks that we will need to take into account in further developments is hierarchy. Transportation networks are by essence hierarchical, depending on scales they are embedded in. [**10.1371/journal.pone.0102007**] showed empirical scaling properties for public transportation networks for a consequent number of metropolitan areas across the world, and scaling laws reveal the presence of hierarchy within a system, as for size hierarchy for system of cities expressed by Zipf's law [**nitsch2005zipf**]

2 and often generalized as *functional accessibility*, for example employments accessible for actives at a location. Spatial interaction potentials ruling gravity law can also been understood this way.

or other urban scaling laws[**2013arXiv1301.1674A**, **2015arXiv151000902B**]. Transportation network topology has been shown to exhibit such scaling also for the distribution of its local measures such as centrality [**samaniego2008cities**]. Hierarchy seems to play a particular role on interaction processes, as BRETAGNOLLE [**bretagnolle:tel-00459720**] highlighted an increasing correlation between urban hierarchy and network hierarchy for French railway network, marker of positive feedbacks between urban rank and network centralities. Different regimes in space and times were identified : for French railway network evolution e.g., a first phase of adaptation of the network to the existing urban configuration was followed by a phase of co-evolution i.e. in the sense that causal relations become difficult to identify. Railway evolution in the United States followed a different pattern, without hierarchical diffusion, shaping locally urban growth. It emphasizes the presence of path-dependance for trajectories of urban systems : the presence in France of a previous city system and network (postal roads) strongly shaped railway development, whereas its absence in the US lead to completely different dynamics. An open question is if generic processes underlie both evolutions, each being different realizations with different initial conditions and different meta-parameters (different *regimes* in the sense of settlement systems transitions introduced in the current ANR Research project TransMonDyn, as a transition can be understood as a change of stationarity for meta-parameters of a general dynamic). In terms of dynamical systems formulation, it is equivalent to ask if dynamics of attractors (long time scale components) obey similar equations as the position and nature of attractors for a stochastic dynamical system give its current regime, in particular if it is in a divergent state (positive local Liapounov exponent) or is converging towards stable mechanisms [**sanders1992systeme**]. To answer this question together with a disentangling of co-evolution processes for that regime, [**bretagnolle:tel-00459720**] proposes modeling as a constructive element of answer. We will see in next section how modeling can bring knowledge about territorial processes.

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

1.2.1 *Modeling in Quantitative Geography*

Modeling in Theoretical and Quantitative Geography (TQG), and more generally in Social Science, has a long history on which we can not go further than a general context. CUYALA does in [**cuyala2014analyse**] an analysis of the spatio-temporal development of French speaking TQG movement and underlines the emergence of the discipline as the combination between quantitative analysis (e.g. spatial analysis or modeling and simulation practices) and theoretical constructions, an integration of both allowing the construction of theories from empirical facts that yield theoretical hypothesis to be tested on empirical data. These approach were born under the influence of the *new geography* in Anglo-saxon countries and Sweden. A broad history of the genesis of models of simulation in geography is done by REY in [**rey2015plateforme**] with a particular emphasis on the notion of validation of models. The use of computation for simulation of models is anterior to the introduction of paradigms of complexity, coming back to HÄGERSTRAND and FORRESTER, pioneers of spatial economic models inspired by Cybernetics. With the increase of computational possibilities epistemological transformations have also occurred, with the apparition of explicative models as experimental tools. REY compares the dynamism of seventies when computation centers were opened to geographers to the democratization of High Performance Computing (transparent grid computing, see [**schmitt2014half**] for an exemple of the possibilities offered in terms of model validation and calibration, decreasing the computational time from 30 years to one week), that is also accompanied by an evolution of modeling practices [**banos2013pour**] and techniques [**10.1371/journal.pone.0138212**]. Modeling (in particular computational models of simulation) is seen by many as a fundamental building brick of knowledge : [**livet2010**] recalls the combination of empirical, conceptual (theoretical) and modeling domains with constructive feedbacks between each. A model can be an exploration tool to test assumptions, an empirical tool to validate a theory against datasets, an explicative tool to reveal causalities (and thus internal processes of a system), a constructive tool to iteratively build a theory with an iterative construction of an associated model. These are example among others : VARENNE proposes in [**varenne2010simulations**] a refined classifications of diverse functions of a model. We will consider modeling as a fundamental instrument of knowledge on processes within complex adaptive systems, as already evoked, and restraining again our question, will focus on *models involving interactions between transportation networks and territories*.

1.2.2 *Modeling Territories and Networks*

Concerning our precise question of interactions between transportation networks and territories, we propose an overview of existing approaches. Following [**bretagnolle2002time**], the "*thoughts of specialists in planning aimed to give definitions of city systems, since 1830, are closely linked to the historical transformations of communication networks*". It implies that ontologies and corresponding models addressed by geographers and planners are closely linked to their current historical preoccupations, thus necessarily limited in scope and purpose. In a perspectivist vision of science [**giere2010scientific**] such boundaries are the essence of the scientific entreprise, and as we will argue in chapter 2 their combination and coupling in the case of models is a source of knowledge.

Land-Use Transportation Interaction Models

A subsequent bunch of literature in modeling interaction between networks and territories can be found in the field of planning, with the so-called *Land-use Transportation Interaction Models*. These works are difficult to be precisely bounded as they may be influenced by various disciplines. For example, from the point of view of Urban Economics, propositions for synthesizing models have existed for a relatively long term [**putman1975urban**]. The variety of possible models has lead to operational comparisons [**paulley19910verview**, **wegener19910ne**]. More recently, the respective advantages of static and dynamic modeling was investigated in [**kryvobokov2013comparison**]. Generally these type of models operate at relatively small temporal and spatial scales. [**wegener2004land**] reviewed state of the art in empirical and modeling studies on interactions between land-use and transportation. It is positioned in economic, planning and sociological theoretical contexts, and is relatively far from our geographical approach aiming to understand long-time processes. Seventeen models are compared and classified, none of which implements actually network endogenous evolution on the relatively small time scales of simulation. A complementary review done in [**chang2006models**] broadens the scope with inclusion of more general classes of models, such as spatial interaction models (including traffic assignment and four steps models), operational research planning models (optimal localisations), micro-based random utility models, and urban market models. These techniques operate also at small scales and consider at most land-use evolution. [**iacono2008models**] covers a similar scope with a further emphasis on cellular automata models of land-use change and agent-based models. These type of models are still largely developed and used today, as for example [**delons:hal-00319087**] which is used for Parisian metropolitan region. The short-term range of application and their operational character makes them useful for

planning, what is far from our preoccupation to obtain explicative models for geographical processes.

Network Growth

Network growth can be used to design modeling entreprises that aim to endogenously explain growth of transportation networks, generally from a bottom-up point of view, i.e. by exhibiting local rules that would allow to reproduce network growth over long time scales (generally the road network). Economists have proposed such models : [**zhang2007economics**] reviews transportation economics literature on network growth within an endogenous growth theory [**aghion1998endogenous**], recalling the three main features studied by economists on that subject that are road pricing, infrastructure investment and ownership regime, and describes an analytical model combining the three. [**xie2009modeling**] develops a broad review on network growth modeling extending to other fields : transportation geography early developed empirical-based models but which did concentrate on topology reproduction rather than on mechanisms according to [**xie2009modeling**] ; statistical models on case studies provide mitigated conclusions on causal relations between offer and demand ; economists have studied infrastructure provision from both microscopic and macroscopic point of views, generally non-spatial ; network science has provided toy-models of network growth based on structural and topological rules rather on mechanism-based rules. An other approach not mentioned that we will develop further is biologically inspired network design. We first give some example of economic-based and geometrical-based network growth modeling attempts. [**yerra2005emergence**] shows through a reinforcement economic model including investment rule based on traffic assignment that local rules are enough to make hierarchy of roads emerge for a fixed land-use. A very similar model in [**louf2013emergence**] with simpler cost-benefits obtains the same conclusion. Whereas these models based on processes focus on reproducing macroscopic patterns of networks (typically scaling), geometrical optimization models aim to ressemble topologically real networks. [**barthelemy2008modeling**] proposes a model based on local energy optimization but it stays very abstract and unvalidated. The morphogenesis model given in [**courtat2011mathematics**] based on local potential and connectivity rules, even if not calibrated, seems to reproduce more reasonably real street patterns. Very close work is done in [**rui2013exploring**]. Other tentatives [**de2007netlogo**, **yamins2003growing**] are closer to procedural modeling [**lechner2004procedural**, **watson2008procedural**] and therefore not of interest in our purpose as they can difficultly be used as explicative models. Finally, an interesting and original approach to network growth are biological networks. These belong to the field of morphogenetic engineering pioneered by DOURSAT that aim to design artificial complex system inspired from natural complex systems and in which a control of emer-

ging properties is possible [**doursat2012morphogenetic**]. *Physarum Machines*, that are models of a self-organized mould (slime mould) have been shown to provide efficient bottom-up solution to computationally heavy problems such as routing problems [**tero2006physarum**] or NP-complete navigation problems such as the Travelling Salesman Problem [**zhu2013amoeba**]. It has been shown to produce networks with Pareto-efficient cost-robustness properties [**tero2010rules**], relatively close in shape to real networks (under certain conditions, see [**adamatzky2010road**]). This type of models can be of interest for us since auto-reinforcement mechanisms based on flows are analog to mechanisms of link reinforcement in transportation economics.

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 sim-

plification 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 RESEARCH QUESTION

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

THEORETICAL FRAMEWORK

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

Theory is a key element of any scientific construction, especially in Human Sciences in which object definition and questioning are more open but also determining for research directions. We develop in this chapter a self-consistent theoretical background. It naturally emerges from thematic considerations of previous chapter, empirical explorations done in chapter 5 and modeling experiments conducted in chapter 6, as a linear structure of knowledge is not appropriate to translate the type of scientific entreprise we are conducting, typically in the spirit of Sanders in [**livet2010**] for which the simultaneous conjonction of empirical, conceptual and modeling domains is necessary for the emergence of knowledge. This theoretical construction is however presented to be understood independently, and is used as a structuring skeleton for the rest of the thesis.

We propose first to construct the *geographical theory* that will pose the studied objects and their meaning in the real world (their ontology), with their interrelations. This yields precise assumptions that will be sought to be confirmed or proven false in the following. Staying at a thematic level appears however to be not enough to obtain general guidelines on the type of methodologies and the approaches to use. More precisely, even if some theories imply an more natural use of some tools[1], at the subtler level of contextualization in the sense of the approach taken to implement the theory (as models or empirical analysis), the freedom of choice may mislead into unappropriated techniques or questionings (see [**raimbault2016cautious**] on the example of incautious use of big data and computation). We develop therefore in a second section a theoretical framework at a meta-level, aiming to give a vision and framing for modeling socio-technical systems.

1 to give a rough example, a theory emphasizing the complexity of relations between agents in a system will conduct generally to use agent-based modeling and simulation tools, whereas a theory based on macroscopic equilibrium will favorise the use of exact mathematical derivations.

2.1 GEOGRAPHICAL THEORETICAL CONTEXT

2.1.1 *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 in 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 interrelations 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" interpretations of scaling laws developed by physicists incompatible with evo-

lutive urban theory [**pumain2012urban**]. We will interpret territorial systems following that idea of complex adaptive systems.

Urban Morphogenesis

The idea of morphogenesis was introduced by TURING in [**turing1952chemical**] when trying to isolate simple chemical rules that could lead to the emergence of the embryo and its form. The morphogenesis of a system consists in self-consistent evolution rules that produce the emergence of its successives states, i.e. the precise definition of self-organization. Progresses towards the understanding of embryo morphogenesis (in particular the isolation of processes producing the differentiation of cells from an unique cell) has been made only recently with the use of Complexity Approaches in integrative biology [**delile2016chapitre**]. In the case of urban systems, the idea of urban morphogenesis, i.e. of self-consistent mechanisms that would produce the urban form, is more used in the field of architecture and urban design [**hachi2013master**] (as ALEXANDER generative grammar "Pattern Language" e.g.), in relation with theories of Urban Form [**moudon1997urban**]. This idea can be pushed into very small scales such as the building [**whitehand1999urban**] but we will use it more at a mesoscopic scale, in terms of land-use changes within an intermediate scale territorial system, in the same ontologies as Urban morphogenesis modeling literature (for example [**bonin2012modele**] describes a model of urban morphogenesis with qualitative differentiation, whereas [**makse1998modeling**] give a model of urban growth based on a mono-centric population distribution perturbed with correlated noises). The notion of morphogenesis will be important in our theory in link with modularity and scale. Modularity of a complex system consists in its decomposition into relatively independent sub-modules, and modular decomposition of a system can be seen as a way to disentangle non-intrinsic correlations [**2015arXiv150904386K**] (think of a block diagonalisation of a first order dynamical system). The isolation of a subsystem yields a corresponding characteristic scale. Isolating possible morphogenesis processes imply a controlled isolation (controlled boundary conditions e.g.) of the considered system, corresponding to a modularity level and thus a scale. When self-consistent processes are not enough to explain the evolution of the system (with reasonable action on boundary conditions), a change of scale is necessary, caused by an underlying phase transition in modularity. The example of metropolitan growth is a good example : complexity of interactions within the metropolitan region will grow with size and diversity of functions leading to a change in scale necessary to understand processes. The emergence of an international airport will strongly influence local development, what corresponds to the significant integration within a larger system. The characteristic scales and processes for which these change occur will be precise questions to be investigated through

modeling. It is interesting to remark that a territorial subsystem in which morphogenesis has a sense can be seen as an *autopoietic system* in the extended sense of BOURGINE in [**bourgine2004autopoiesis**], as a network of auto-reproducing processes[2] regulating their boundary conditions, what emphasizes boundaries on which we will last insist.

Co-evolution

Our last pillar is a clarification of the notion of *co-evolution*, on which HOLLAND shed light through an approach of complex adaptive systems by a theory of CAS as signal processing agents operating thanks to their boundaries [**holland2012signals**]. In this theory, complex adaptive systems form aggregates at diverse hierarchical levels, that correspond to different level of self-organization, and boundaries are vertically and horizontally intricate in a complex way. That approach introduces the notion of *niche* as a relatively independent subsystem in which ressources circulate (the same way as network communities) : numerous illustrations are given such as economical niches or ecological niches. Agents within a niche are said to be *co-evolving*. Co-evolution thus means strong interdependences (implying circular causal processes) and a certain independence regarding the exterior of the niche. The notion is naturally flexible as it will depend on ontologies, resolution, thresholds etc. considered to define the system. This concept is easily transmissible to the evolutive urban theory and converges with the notion of co-evolution described by PUMAIN : co-evolving agents in a system of cities consist in a niche with its flows, signals and boundaries and thus co-evolving entities in the sense of HOLLAND. This notion will be important for us in the definition of territorial subsystems and their coupling.

2.1.2 *Synthesis : an theory of co-evolutive networked territorial systems*

We synthesize our pillars as a short self-consistent geographical theory of territorial systems in which networks play a central role in the co-evolution of components of the system. See the foundation subsection for definitions and references. The formulation is intended to be minimalistic.

Definition 1 - *Territorial System.* *A territorial system is a set of networked human territories, i.e. human territories in and between which real networks exist.*

At this step complexity and dynamical evolutive characters of territorial systems are implied but not an explicit part of the theory. We

2 which are however not cognitive, making this auto-organized systems fortunately not alive in the sense of autopoietic and cognitive systems

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 $\mathrm{Cov}\left[dX_j, dX_{j'}\right] \simeq 0$ 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 ans 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 Systems[3]. 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.

3 Using a representation theorem [**knight1975predictive**], any discrete stationary process is a *Hidden Markov Model*. Given the definition of a causal state as $\mathbb{P}[future|A] = \mathbb{P}[future|B]$, 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.

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

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.

2.2.1 *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*[4] 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 systems[5], 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 Science[6] 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 analytical and "exact" approaches to computational and evidence-

4 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**].

5 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

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

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 environment[7], 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.

7 geographical systems in the sense of [**dollfus1975some**] 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 objects[8].
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.

2.2.2 *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* $O = (O_j)_j$.

8 *as already explained before, this positioning along with the importance of structure may be related to Ontic Structural Realism [***frigg2011everything***] in further 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 structure[9]. 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 forms[10] :

- *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*

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

10 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* $\mathcal{P}(\mathcal{O})$ *:*

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

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* $\mathcal{P}(\mathcal{O})$ *and edges of two types :* $E_W = \{(O, O') | O' \preccurlyeq O\}$ *and* $E_N = \{(O, O') | O' \Subset O\}$
2. *Nodes are reduced*[11] *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.*

11 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 $O \to O' \to \ldots \to O$ 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.

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 $\mathbb{T}$ 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 operator[12] $\|\cdot\|_S$ and that the state variable has a certain level of differentiability, the *thematic scale* if defined as $\|(d^k\vec{X}(t))_k\|_S$

2.2.3 *Application*

The particular case of geographical systems

In [**dollfus1975some**] 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.

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

2.2.4 *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".

2.2.5 *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 5.

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 3.

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 7.

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.

3

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 discipline[1], depending on the purpose of each piece of research (perspectivism [**giere2010scientific**] poses the *model*, that includes methods, as the articulating core of research entreprises). 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 heteroclite 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 enveloppe".

3.1 REPRODUCIBILITY

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

3.1.1 *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 project[2].

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`

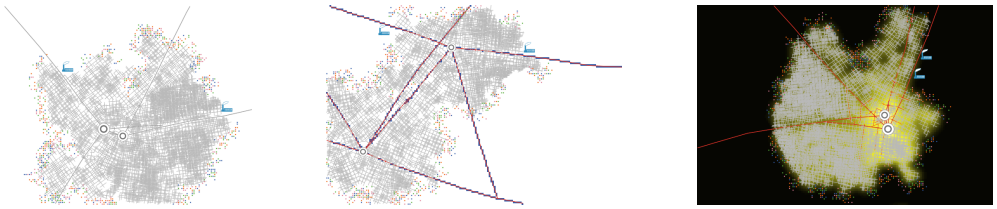


FIGURE 1 : Example of simple improvement in visualization that can help understanding mechanisms implied in the model. *Left :* example of original output ; *Middle :* visualization of main roads (in red) and underlying patches attribution, suggesting possible implementation bias in the use of discretized 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.

3.1.2 *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).

3.1.3 *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 [**chavalarias2005nobel**], 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 13.

3.2 AN UNIFIED FRAMEWORK FOR STOCHASTIC MODELS OF URBAN GROWTH

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

3.2.1 *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 [**favaro2011gibrat**] 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.

3.2.2 *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.

3.2.3 *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

Let consider a strictly positive growth Gibrat model given by $P_i(t) = R_i(t) \cdot P_i(t-1)$ with $R_i(t) > 1$, $\mu_i(t) = \mathbb{E}[R_i(t)]$ and $\sigma_i(t) = \mathbb{E}\left[R_i(t)^2\right]$. On the other hand, we take a simple preferential attachment, with fixed attachment probability $\lambda \in [0,1]$ and new arrivants number $m > 0$. We derive that Gibrat model can be statistically equivalent to a limit of the preferential attachment model, assuming that the moment-generating function of $R_i(t)$ exists. Classical distributions that could be used in that case, e.g. log-normal distribution, are entirely defined by two first moments, making this assumption reasonable.

Lemma 1 *The limit of a Preferential Attachment model when $\lambda \ll 1$ is a linear-growth Gibrat model, with limit parameters $\mu_i(t) = 1 + \frac{\lambda}{m \cdot (t-1)}$.*

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

$$\{P_i(t) = k\} = \bigcup_{\delta=0}^{\infty} \left(\{P_i(t-1) = k - \delta\} \cap \{P_i \leftarrow P_i + 1\}^{\delta} \right)$$

where the second event corresponds to city i being increased δ times between $t-1$ and t (note that events are empty for $\delta \geqslant k$). 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 $q = \lambda \cdot \frac{\bar{P}_i(t-1)}{P_0 + m \cdot (t-1)}$

$$\begin{aligned}
\bar{P}_i(t) &= \sum_{k=0}^{\infty}\sum_{\delta=0}^{\infty} k\cdot\left(\lambda\cdot\frac{\bar{P}_i(t-1)}{P_0+m\cdot(t-1)}\right)^{\delta}\cdot\mathbb{P}[P_i(t-1)=k-\delta] \\
&= \sum_{\delta'=0}^{\infty}\sum_{k'=0}^{\infty}\left(k'+\delta'\right)\cdot q^{\delta'}\cdot\mathbb{P}\left[P_i(t-1)=k'\right] \\
&= \sum_{\delta'=0}^{\infty} q^{\delta'}\cdot\left(\delta'+\bar{P}_i(t-1)\right) = \frac{q}{(1-q)^2}+\frac{\bar{P}_i(t-1)}{(1-q)} \\
&= \frac{\bar{P}_i(t-1)}{1-q}\left[1+\frac{1}{\bar{P}_i(t-1)}\frac{q}{(1-q)}\right]
\end{aligned}$$

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

$$\mathbb{E}\left[P_i(t)^2\right] = \mathbb{E}\left[R_i(t)^2\right]\cdot\mathbb{E}\left[P_i(t-1)^2\right]$$

and

$$\mathbb{E}\left[P_i(t)^2\right] = \sum_{k=0}^{+\infty} k^2\,\mathbb{P}[P_i(t)=k]$$

We obtain the same way

$$\begin{aligned}
\mathbb{E}\left[P_i(t)^2\right] &= \sum_{\delta'=0}^{\infty}\sum_{k'=0}^{\infty}\left(k'+\delta'\right)^2\cdot q^{\delta'}\cdot\mathbb{P}\left[P_i(t-1)=k'\right] \\
&= \sum_{\delta'=0}^{\infty} q^{\delta'}\cdot\left(\mathbb{E}\left[P_i(t-1)^2\right]+2\delta'\bar{P}_i(t-1)+\delta'^2\right) \\
&= \frac{\mathbb{E}\left[P_i(t-1)^2\right]}{1-q}+\frac{2q\bar{P}_i(t-1)}{(1-q)^2}+\frac{q(q+1)}{(1-q)^3} \\
&= \frac{\mathbb{E}\left[P_i(t-1)^2\right]}{1-q}\left[1+\frac{q}{\mathbb{E}[P_i(t-1)^2]}\left(\frac{2\bar{P}_i(t-1)}{1-q}+\frac{(1+q)}{(1-q)^2}\right)\right]
\end{aligned}$$

We have therefore an equivalence between the Gibrat model as a continuous formulation of a Preferential Attachment (or Simon model) in a certain limit. ∎

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.

3.3 ANALYTICAL SENSITIVITY OF URBAN SCALING LAWS TO SPATIAL EXTENT

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

3.3.1 *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 constatation that indicators computed on elements of an urban system, that can be cities for system of cities, but also smaller entities at a smaller scale, do fit relatively well a power-law distribution as a function of entity size, i.e. that for entity i with population P_i, we have for an integrated quantity A_i, the relation $A_i \simeq A_0 \cdot \left(\frac{P_i}{P_0}\right)^{\alpha}$. Scaling exponent α can be smaller or greater than 1, leading to infra- or 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. Remarquable 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 propose to prove by the tractation of 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.

3.3.2 *Formalization*

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

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

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 : $\ln\left[d_i^0 r_i^2\right] = K' - \alpha \ln i$.

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 $A(D) = \mathbb{E}\left[\iint_D a(\vec{x})d\vec{x}\right]$ represents the expected quantity of a in area D. We make the assumption that $a \in \{0; 1\}$ ("counting" indicator) and that its law is given by $\mathbb{P}[a(\vec{x}) = 1] = f(d(\vec{x}))$. Following the empirical work done in **[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 $P_i(\theta) = \iint_{D(\vec{x}_i,\theta)} d$.

3.3.3 *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

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

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 $\alpha(\theta) = \frac{\mathrm{Cov}[(\ln A_i(\theta))_i, (\ln P_i(\theta))_i]}{\mathrm{Var}[(\ln P_i(\theta))_i]}$. As we work on city boundaries, threshold is expected to be significantly smaller than center density, i.e. $\theta/d_i^0 \ll 1$. We can develop the expression in the first order of θ/d_i^0 and use the global scaling law for city sizes, what gives $\ln A_i(\theta) \simeq K_A - \alpha \ln i + (\beta - 1)\ln d_i^0 + \beta \ln \frac{\theta}{d_i^0}\left(\frac{\theta}{d_i^0}\right)^\beta$ and $\ln P_i(\theta) = K_P - \alpha \ln i + \ln\left[\frac{\theta}{d_i^0}\right]\frac{\theta}{d_i^0}$. Developing the covariance and variance gives finally an expression of the scaling exponent as a function of θ, where $k_j, k_j{}'$ are constants obtained in the development :

$$\alpha(\theta) = \frac{k_0 + k_1\theta + k_2\theta^\beta + k_3\theta^{\beta+1} + k_4\theta\ln\theta + k_5\theta^\beta\ln\theta + k_6\theta^\beta(\ln\theta)^2 + k_7\theta^{\beta+1}(\ln\theta)^2 + k_8\theta^{\beta+1}\ln\theta}{k_0' + k_1'\ln\theta + k_2'\theta\ln\theta + k_3'\theta^2 + k_4'\theta^2\ln\theta + k_5'\theta^2(\ln\theta)^2} \quad (2)$$

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.

3.3.4 *Numerical Simulations*

IMPLEMENTATION We implement empirically the density model given in section 3.3.2. 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 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 code[3]. We show in figure 2 example of synthetic density distributions on which the numerical study is

3 available at `https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Scaling`

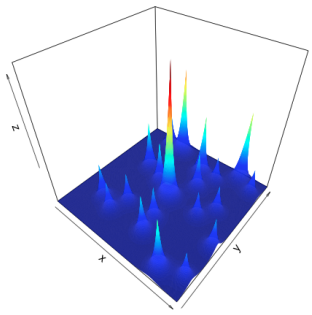


FIGURE 2 : Example of a synthetic density distribution obtained with the exponential mixture, with a grid of size 400×400 and parameters $N = 20$, $r_0 = 10$, $P_{max} = 200$, $\alpha = 0.5$, $\theta_C = 0.01$.

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
- Radiuses 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.

3.4 STATISTICAL CONTROL ON INITIAL CONDITIONS BY SYNTHETIC DATA GENERATION

3.4.1 *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 disjonction for scales and/or ontologies between the model of simulation and the domain of meta-parameters.

3.4.2 *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 configurationyielding 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.*

3.4.3 *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 $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$. It naturally multiplies the number of repetition needed for convergence what is the expected behavior.

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

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 interactions[4]. 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.

3.5.1 *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

$$\mathbf{C}_t(\vec{x}_0) = \mathrm{Var}\left[\vec{Y}[\vec{x}_0, \cdot]\right]$$

and spatial covariance structure at fixed time by

$$\mathbf{C}_x(t) = \mathrm{Var}\left[\vec{Y}[\cdot, t]\right]$$

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

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.

3.5.2 *Wave Equation*

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

$$c^2 \cdot \partial_t^2 \bar{Y}_i = \Delta \bar{Y}_i \tag{3}$$

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

$$\mathbf{C}_t \left[\partial_t^2 Y_i, \partial_t^2 Y_j\right] = \frac{1}{c^2} \cdot \mathbf{C}_x \left[\Delta Y_i, \Delta Y_j\right] \tag{4}$$

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.

3.5.3 *Fokker-Planck Equation*

An other interesting approach may when the process verifies a Fokker-Planck equation on probabilities of the state of the system when it is given by its position (diffusion of particles in that case)

$$\partial_t P(x_i, t) = -d \cdot \partial_x P(x_i, t) + \frac{\sigma^2}{2} \partial_x^2 P(x_i, t) \tag{5}$$

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.

3.5.4 *Master Equation*

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

$$\partial_t \vec{P} = \mathbb{W} \vec{P} \tag{6}$$

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.

3.5.5 *Consistent spatio-temporal sampling*

In a more empirical way, we propose to not assume any contraint 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, $\left|\rho_{\|\Delta\vec{x}\|<l_\rho}[Y_i(\vec{x}+\Delta\vec{x}, t), Y_i(\vec{x}, t)]\right| > 0$.
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.

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 entreprises 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 2 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

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

4.1.1 *In search of models of co-evolution*

Transportation networks and urban land-use are known to be strongly coupled components of urban systems at different scales [**bretagnolle2009organizatio**
One common approach is to consider them as co-evolving, avoiding misleading interpretations such as the myth of structural effect of transportation infrastructures [**offner1993effets**]. A question rapidly arising is the existence of models endogeneizing this co-evolution, i.e. taking into account simultaneous urban and network growth. We try to answer it using an algorithmic systematic review. We propose in this section, after a brief state of the art of existing literature, to present such an approach by formalizing the algorithm, which results are then presented and discussed.

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

Land-Use Transportation Interaction Models.

A wide class of models that have been developed essentially for planning purposes, which are the so-called Land-use Transportation Interaction Models, is a first type answering our research question. See recent reviews [**chang2006models**], [**iacono2008models**] and [**wegener2004land**] to get an idea of the heterogeneity of included approaches, that exist for more than 30 years. Recent models with diverse refinements are still developed today, such as [**delons:hal-00319087**] which includes housing market for Paris area. Diverse aspects of the same system can be translated into many models (as e. g.[**wegener1991one**]), and traffic, residential and employment dynamics, resulting land-use evolution, influenced also by a static transportation network, are generally taken into account.

Network Growth Approaches

On the contrary, many economic literature has done the opposite of previous models, i.e. trying to reproduce network growth given assumptions on the urban landscape, as reviewed in [**zhang2007economics**]. In [**xie2009modeling**], economic empirical studies are positioned within other network growth approaches, such as work by physicists proposing model of geometrical network growth [**barthelemy2008modeling**]. Analogy with biological networks was also done, reproducing typical robustness properties of transportation networks [**ter02010rules**].

Hybrid Approaches

Fewer approaches coupling urban growth and network growth can be found in the literature. [**barthelemy2009co**] couples density evolution with network growth in a toy model. In [**raimbault2014hybrid**], a simple Cellular Automaton coupled with an evolutive network reproduces stylized facts of human settlements described by Le Corbusier. At a smaller scale, [**achibet2014model**] proposes a model of co-evolution between roads and buildings, following geometrical rules. These approaches stay however limited and rare.

4.1.3 *Bibliometric Analysis*

Literature review is a crucial preliminary step for any scientific work and its quality and extent may have a dramatic impact on research quality. Systematic review techniques have been developed, from qualitative review to quantitative meta-analyses allowing to produce new results by combining existing studies [**rucker2012network**]. Ignoring some references can even be considered as a scientific mistake in the context of emerging information systems [**lissacksubliminal**]. We aim to take advantage of such techniques to tackle our issue. Indeed, observing the form of the bibliography obtained in previous section raises some hypothesis. It is clear that all components are present for co-evolutive models to exist but different concerns and objectives seem to stop it. As it was shown by [**commenges:tel-00923682**] for the concept of mobility, for which a "small world of actors" relatively closed invented a notion ad hoc, using models without accurate knowledge of a more general scientific context, we could be in an analog case for the type of models we are interested in. Restricted interactions between scientific fields working on the same objects but with different purposes, backgrounds and at different scales, could be at the origin of the relative absence of co-evolving models. While most of studies in bibliometrics rely on citation networks [**2013arXiv1310.8220N**] or co-autorship networks [**2014arXiv1402.7268S**], we propose to use a less explored paradigm based on text-mining introduced in [**chavalarias2013phylomemetic**], that obtain a dynamic mapping of scientific disciplines based on their

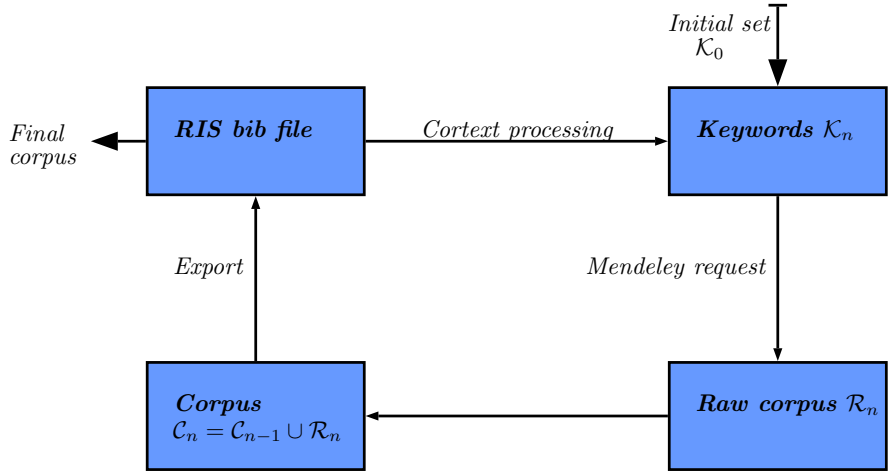


FIGURE 3 : Global workflow of the algorithm, including implementation details : catalog request is done through Mendeley API ; final state of corpuses are RIS files.

semantic content. For our question, it has a particular interest, as we want to understand content structure of researches on the subject. We propose to apply an algorithmic method described in the following. The algorithm proceeds by iterations to obtain a stabilized corpus from initial keywords, reconstructing scientific semantic landscape around a particular subject.

Description of the Algorithm

Let A be an alphabet, A^* corresponding words and $T = \cup_{k \in \mathbb{N}} A^{*k}$ texts of finite length on it. A reference is for the algorithm a record with text fields representing title, abstract and keywords. Set of references at iteration n will be denoted $\mathcal{C} \subset T^3$. We assume the existence of a set of keywords $\mathcal{K}_n$, initial keywords being $\mathcal{K}_0$. An iteration goes as follows :

1. A raw intermediate corpus $\mathcal{R}_n$ is obtained through a catalog request providing previous keywords $\mathcal{K}_{n-1}$.
2. Overall corpus is actualized by $\mathcal{C}_n = \mathcal{C}_{n-1} \cup \mathcal{R}_n$.
3. New keywords $\mathcal{K}_n$ are extracted from corpus through Natural Language Processing treatment, given a parameter N_k fixing the number of keywords.

The algorithm stops when corpus size becomes stable or a user-defined maximal number of iterations has been reached. Fig. 1 shows the global workflow.

Results

IMPLEMENTATION Because of the heterogeneity of operations required by the algorithm (references organisation, catalog requests, text processing), it was found a reasonable choice to implement it in Java. Source code is available on the Github repository of the project1. Catalog request, consisting in retrieving a set of references from a set of keywords, is done using the Mendeley software API [**mendeley**] as it allows an open access to a large database. Keyword extraction is done by Natural Language Processing (NLP) techniques, following the workflow given in [**chavalarias2013phylomemetic**], calling a Python script that uses [**bird2006nltk**].

CONVERGENCE AND SENSITIVITY ANALYSIS A formal proof of algorithm convergence is not possible as it will depend on the empirical unknown structure of request results and keywords extraction. We need thus to study empirically its behavior. Good convergence properties but various sensitivities to Nk were found as presented in Fig. 2. We also studied the internal lexical consistence of final corpuses as a function of keywords number. As expected, small number yields more consistent corpuses, but the variability when increasing stays reasonable.

Once the algorithm is partially validated, we apply it to our question. We start from five different initial requests that were manually extracted from the various domains identified in the manual bibliography (that are "city system network", "land use transport interaction", "network urban modeling", "population density transport", "transportation network urban growth"). We take the weakest assumption on parameter $N_k = 100$, as it should less constrain reached domains. After having constructed corpuses, we study their lexical distances as an indicator to answer our initial question. Large distances would go in the direction of the assumption made in section 2, i.e. that discipline self-centering may be at the origin of the lack of interest for co-evolutive models. We show in Table 1 values of relative lexical proximity, that appear to be significantly low, confirming this assumption.

Further work is planned towards the construction of citation networks through an automatic access to Google Scholar that provides backward citations. The confrontation of inter-cluster coefficients on the citation network for the different corpuses with our lexical consistence results are an essential aspect of a further validation of our results.

The disturbing absence of models simulating the co-evolution of transportation networks and urban land-use, confirmed through a state-of-the-art covering many domain, may be due to the absence of communication between scientific disciplines studying different aspects of that problems. We have proposed an algorithmic method to

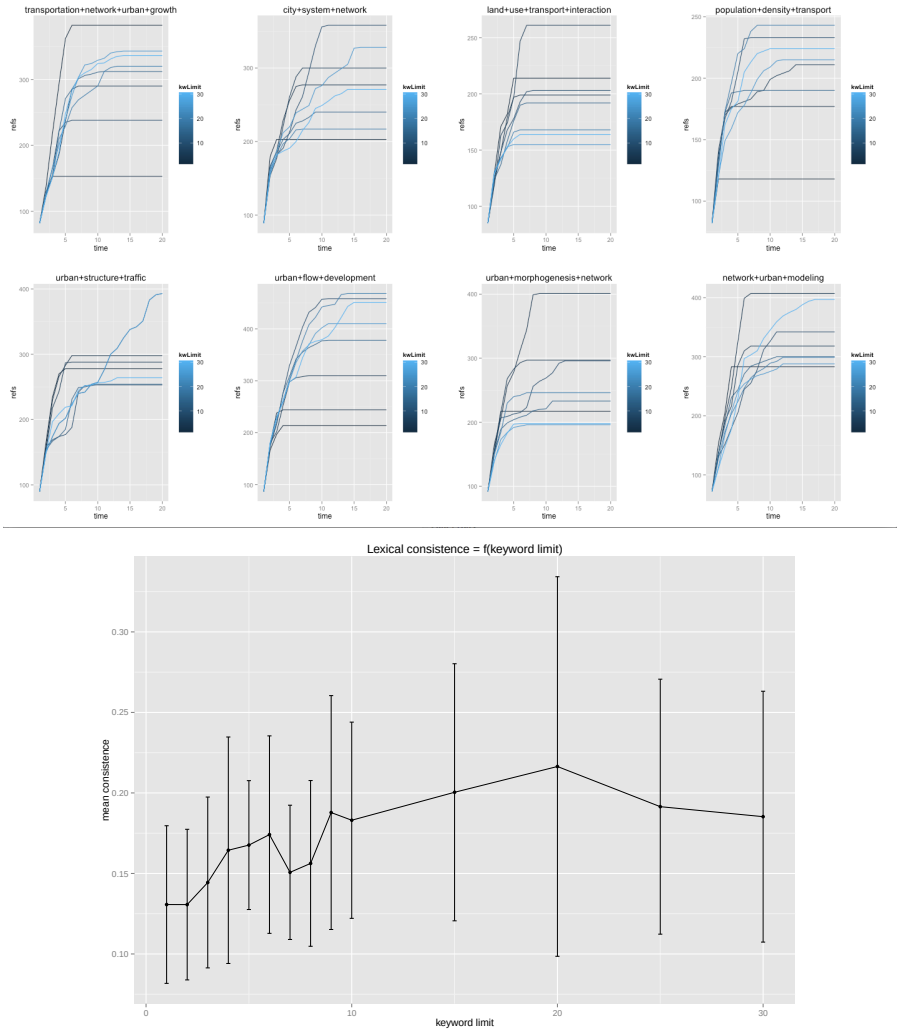


FIGURE 4 : Convergence and sensitivity analysis. Left : Plots of number of references as a function of iteration, for various queries linked to our theme (see further), for various values of N_k (from 2 to 30). We obtain a rapid convergence for most cases, around 10 iterations needed. Final number of 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.

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

TABLE 1 : 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*.

give elements of answers through text-mining-based corpus extraction. First numerical results seem to confirm the assumption. However, such a quantitative analysis should not be considered alone, but rather come as a back-up for qualitative studies that will be the object of further work, such as the one lead in [**commenges:tel-00923682**], in which questionnaires with historical actors of modeling provide highly relevant information.

4.2 REFINING BIBLIOMETRICS THROUGH HYPER-NETWORK ANALYSIS

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. 5. Citation data is collected from `Google Scholar`, that is the only source for incoming citations [**noruzi2005google**] in our case as the journal is not referenced in other databases. We are aware of the possible

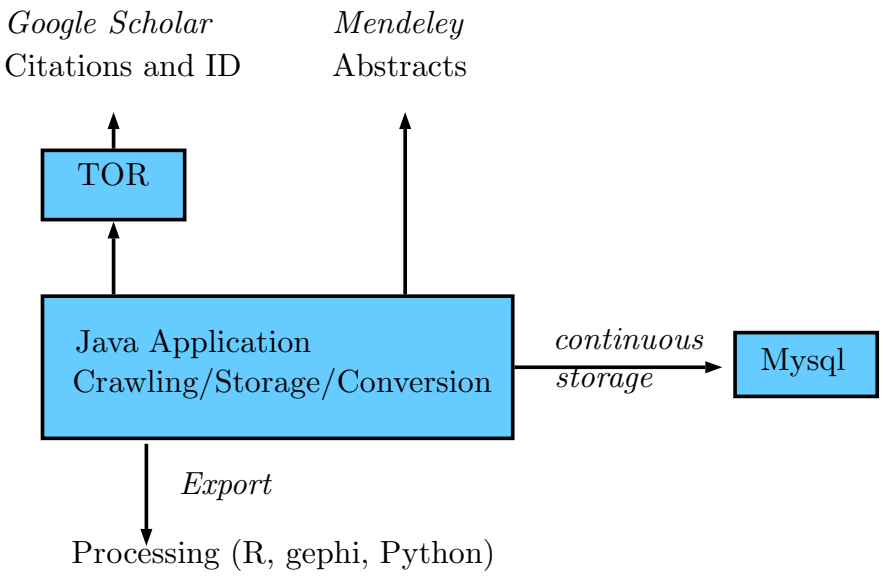


FIGURE 5 : Heterogeneous Bibliographical Data Collection. Architecture of the application for content (semantic data), metadata and citation data collection.

biaises using this single source **[bohannon2014scientific]**[1], 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 *stop-words* and a specific tagger `TreeTagger` is used for other languages than english **[schmid1994probabilistic]**.

Results

We show in figures 6 and 7 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`
- Transportation : `transport, travel`

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

- 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.

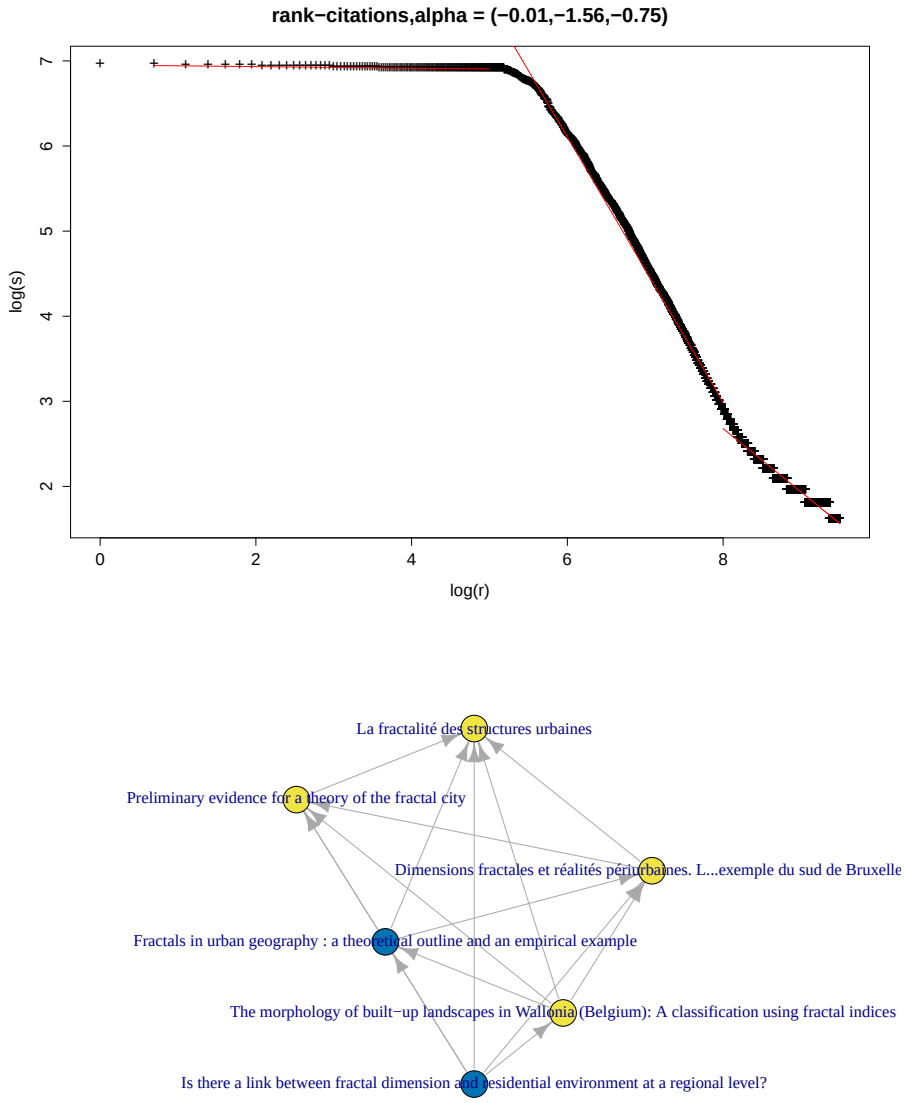


FIGURE 6 : Properties of the citation network. Top : Rank-size plot of in-degrees ; three superposing successive regimes must correspond to different literature types or practices across disciplines. Bottom : example of a maximal clique in the citation network, paper of `cybergeo` being in blue.

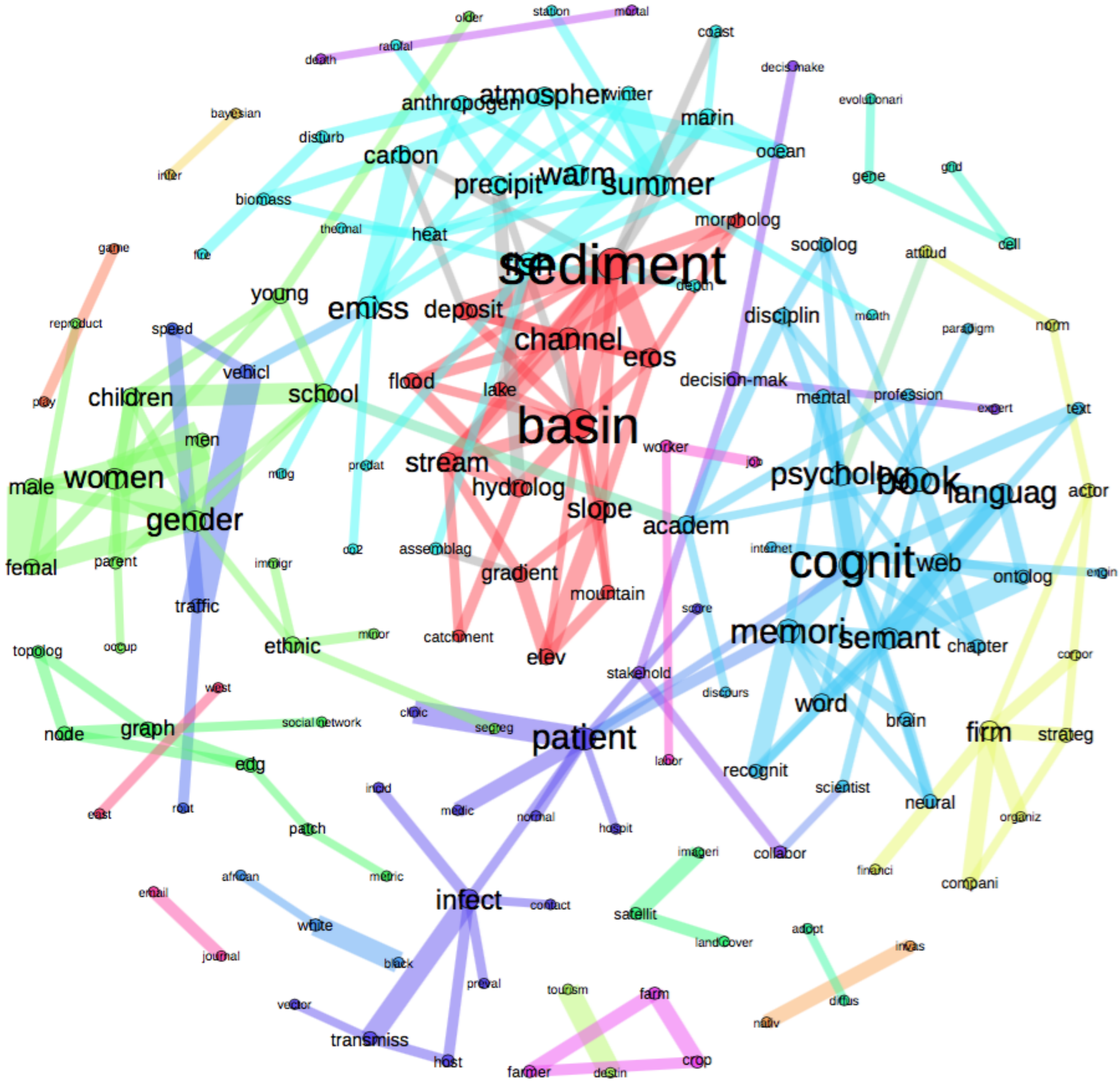


FIGURE 7 : Semantic network of concepts in quantitative geography. Corpus consists of around $2 \cdot 10^5$ abstracts of publications at a topological distance shorter than 2 from the journal `cybergeo` in the citation network. Relevance of keywords were estimated with a bootstrap method, and semantic 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

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

MODELING AND EMPIRICAL ANALYSIS

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.

5

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
- Youssoupha , *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.

5.1 STATIC CORRELATIONS OF URBAN FORM AND NETWORK SHAPE

Spatio-temporal processes implying diffusion or propagation phenomena generally have a specific structure of correlation. In particular, as derived in section 3.5, a static computation of correlation between different instances of a system may under certain conditions provide information on dynamical correlations implied.

5.1.1 *Morphological Measures of European Population Density*

Context

At the macroscopic scale of system of cities, spatialization of the urban system is reasonably captured by cities position, associated with aggregated city variable to represent entirely the system (see e.g. ontologies of Simpop models [**pumain2012multi**] or its successor Marius [**cottineau2014evolution**]). At the mesoscopic scale at which we aim to capture morphological manifestations of interactions between transportation networks and territories, structure of the territorial system can be specified by more refined indicators for the morphological aspect.

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 spaces[1]. Here we choose the mesoscopic scale of a metropolitan center ($\simeq$ 50km) 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 of indicators that are Moran index, average distance, entropy and hierarchy).

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 Periphérique*, 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.

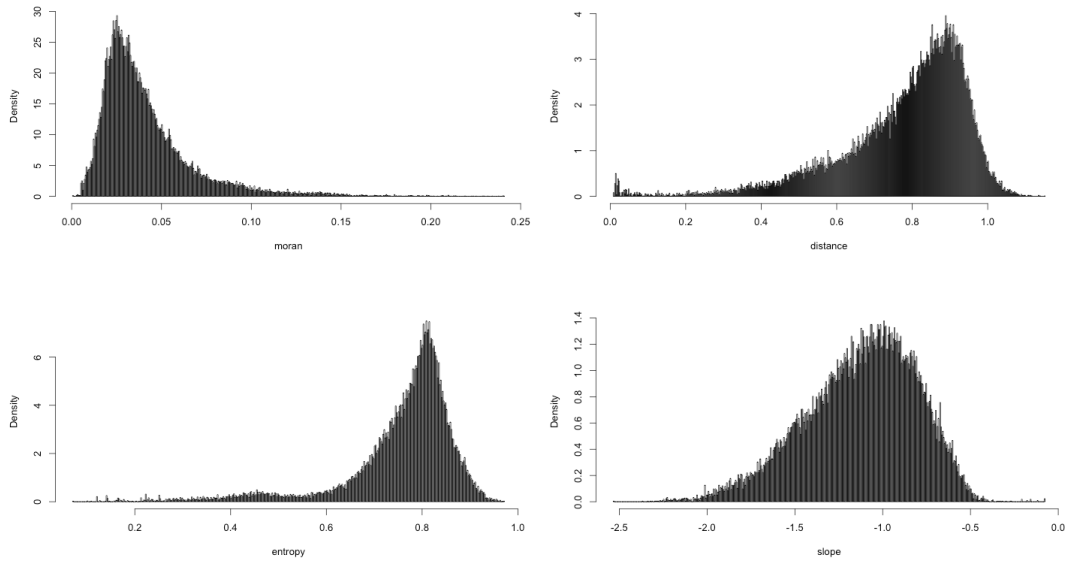


FIGURE 8 : Empirical Distribution of Morphological Indicators

Further developments

In [**10.1371/journal.pone.0107042**] density grids for other countries across the world (ex. China) are provided[2] soo we may repeat our analysis to other regions for comparison purposes.

5.1.2 *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 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.

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

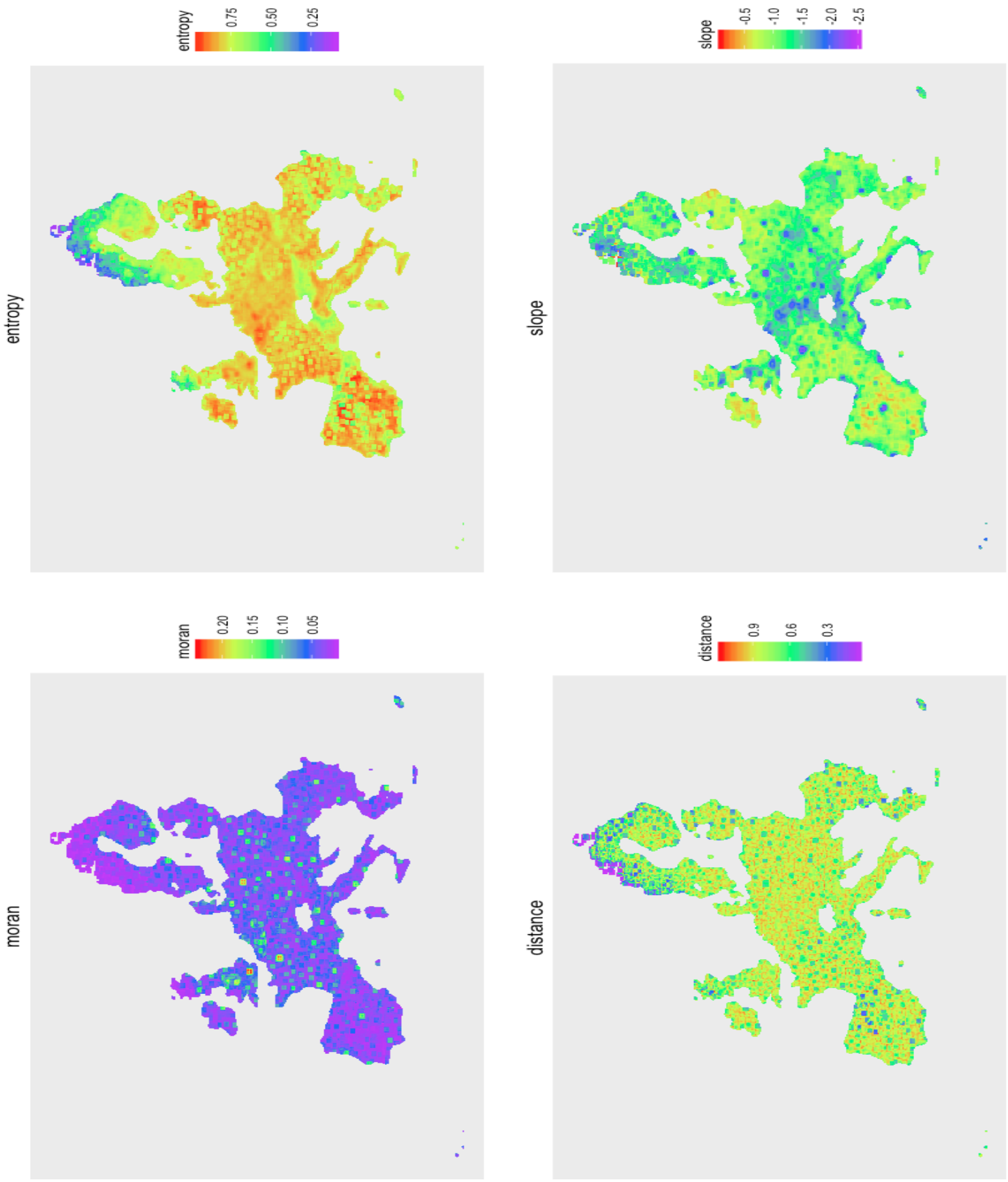


FIGURE 9 : Geographical Distribution of Morphologies : value of indicators across Europe.

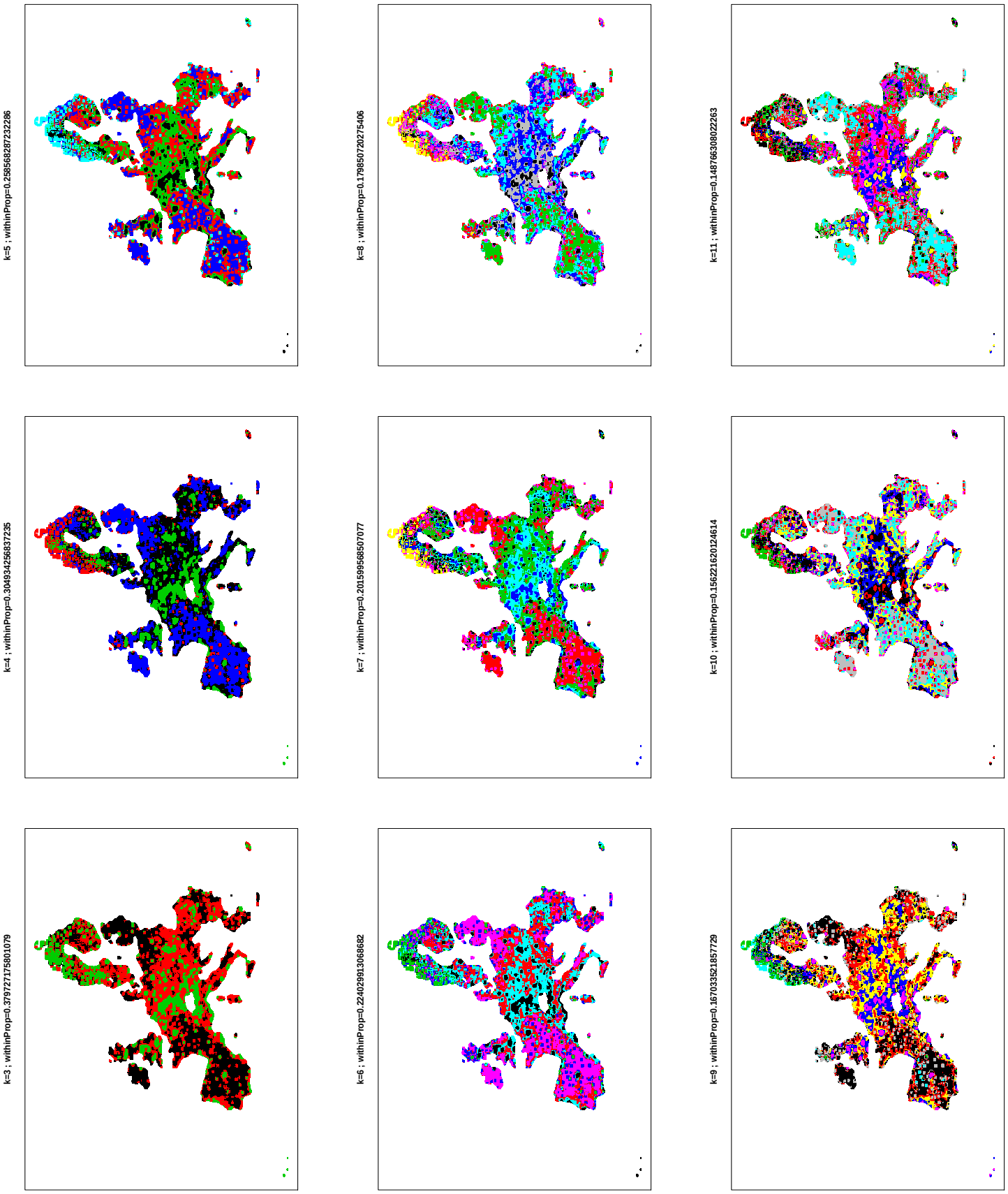


FIGURE 10 : Clustering Analysis of Morphologies. We present the results of an average k-means for different values of

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

Computations of network simplification are still ongoing (local parametrization possible only, estimated effective computation time is around 3 weeks). Computations of network indicators and static correlations must then provide the aforementioned insights into interaction processes.

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

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.

5.2.1 *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 $\mathbf{T} = \left(\vec{T}(\vec{x}_i, t_j^{(T)})\right)_{i,j}$ with $1 \leqslant i \leqslant N$ and $1 \leqslant j \leqslant T$. 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 $\mathbf{N}$

5.2.2 *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

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

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 th 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 $n, \vec{T}$, Helmoltz theorem states that it derives from a potential (can be adapted to non-stationary dynamics with a time-varying potential).

5.2.3 *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.

5.2.4 *Statistical Tests*

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

- On raw data :
  - Multivariate models

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

  - Autocorrelated univariate models

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

  - Autocorrelated multivariate models

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

  - Geographically Weighted Regression **[brunsdon1998geographically]**

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

  - Granger causality tests : **[xie2009streetcars]** use for example Granger causality to link transit with land-use changes.
- On data returns :

– Autoregressive multivariate models

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

– Autoregressive autocorrelated multivariate models : idem with spatial autocorrelation term.

– Synthetic Instrumental Variables : static territory and/or network ?

5.2.5 *Expected results*

We expect from these analyses to test at these spatial and temporal scales, and on a particular metropolitan case study, the assumption on network necessity for the territorial system of functional job commutings.

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

5.3.1 *Context*

Various aspects of territories are concerned by interactions with networks. In previous empirical studies, no socio-economic attributes of populations inhabiting the territory nor economic values for land and real estate was considered. Both are however crucial elements of territorial dynamics and are extensively studied in fields such as territorial analysis or urban economics : for example, [**homocianu:tel-00359302**] studies households residential choices to understand land-use transportation interactions. We propose here to use a database of Real Estate transactions for Parisian region on the last 20 years, with 2 years temporal granularity and exact spatial coordinates. [**gueroisdynamique**] used it to make typologies of spatial dynamics of Parisian real estate. This project is conjointly done with LE CORRE whose strong thematic knowledge on Real Estate financial properties will bring insight for spatial typologies of temporal trajectories.

5.3.2 *Preliminary Results*

We show in Fig. 11 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.

5.3.3 *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.

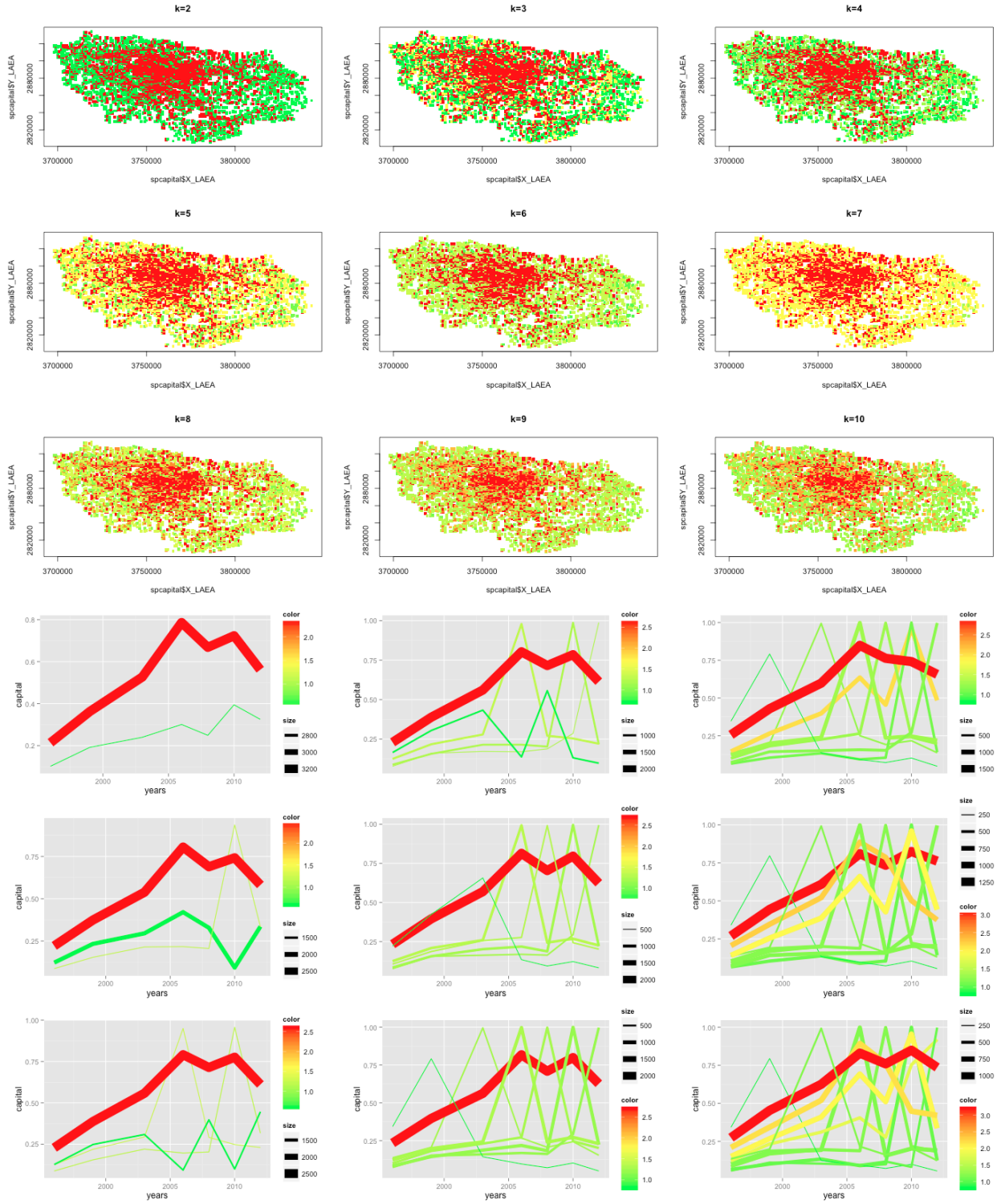


FIGURE 11 : Typology of Real Estate trajectories. Locations were categorized using averaged k-means on time-series. We show maps and time series for value of k from 2 to 10.

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

5.4.1 *Context*

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.

A project planned in collaboration with BAFFI, based on her thesis project [**baffi2016thesis**] that studied 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.

5.4.2 *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

MODELING

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.

6.1 A SIMPLE MODEL OF URBAN GROWTH

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.

6.1.1 *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.

6.1.2 *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 population $(P_i(t))_{1 \leqslant i \leqslant N^2}$. 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.

6.1.3 *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 **[reuillon20130penmole]** 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. 12 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.

EXPLORATION OF PARAMETER SPACE Parameter space is explored using a grid in first experiments, than a Latin Hypercube Sampling exploration. Parameter bounds are $\alpha \in [0.2, 2]$, $\beta \in [0, 0.1]$, $n_d \in \{0, \ldots, 4\}$, $N_G \in [500, 3000]$, $P_m \in [2000, 100000]$. Fig.13 shows the result. We also use the parameter space exploration algorithm **[10.1371/journal.pone.0138212]**

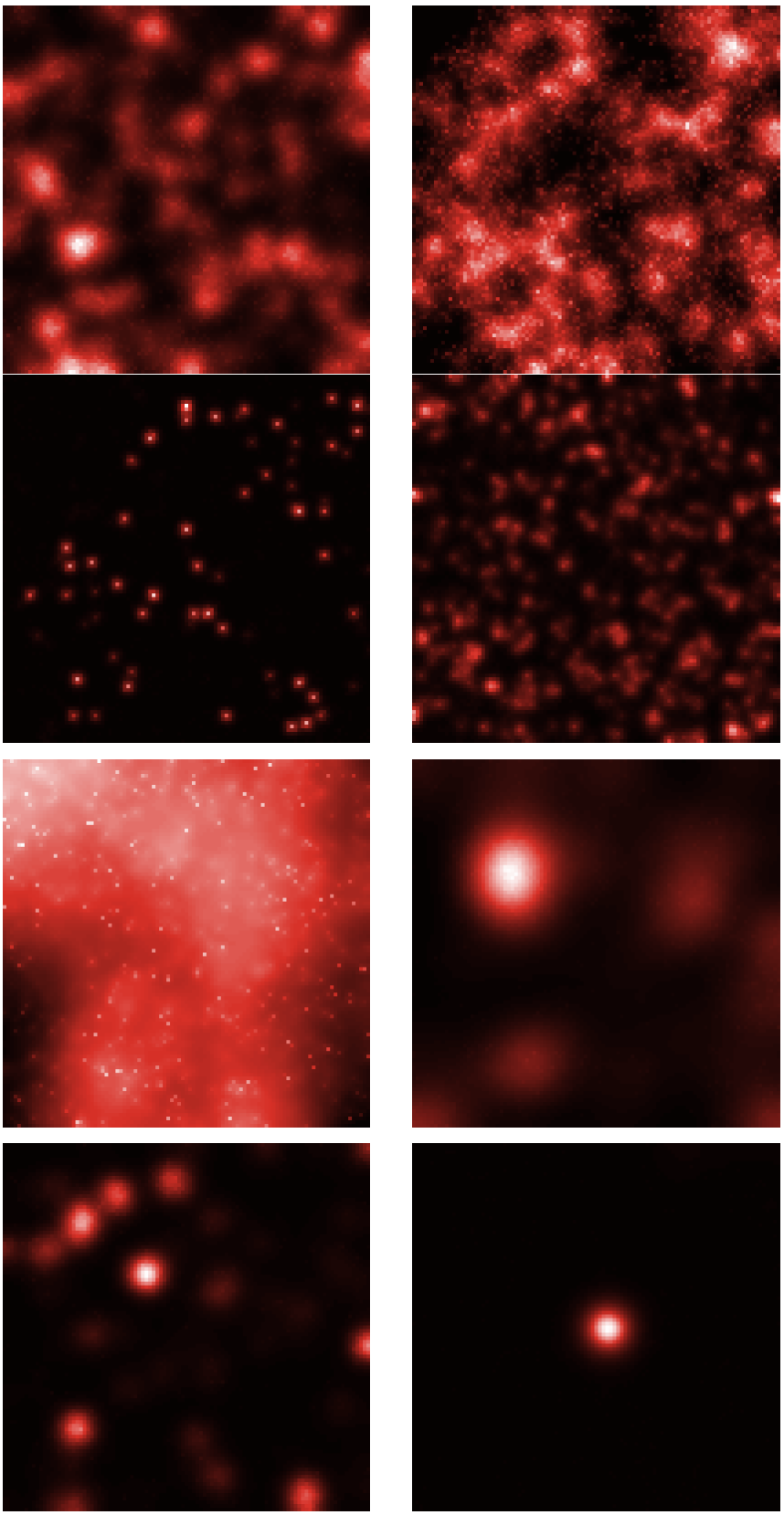


FIGURE 12 : Example of the variety of generated urban shapes

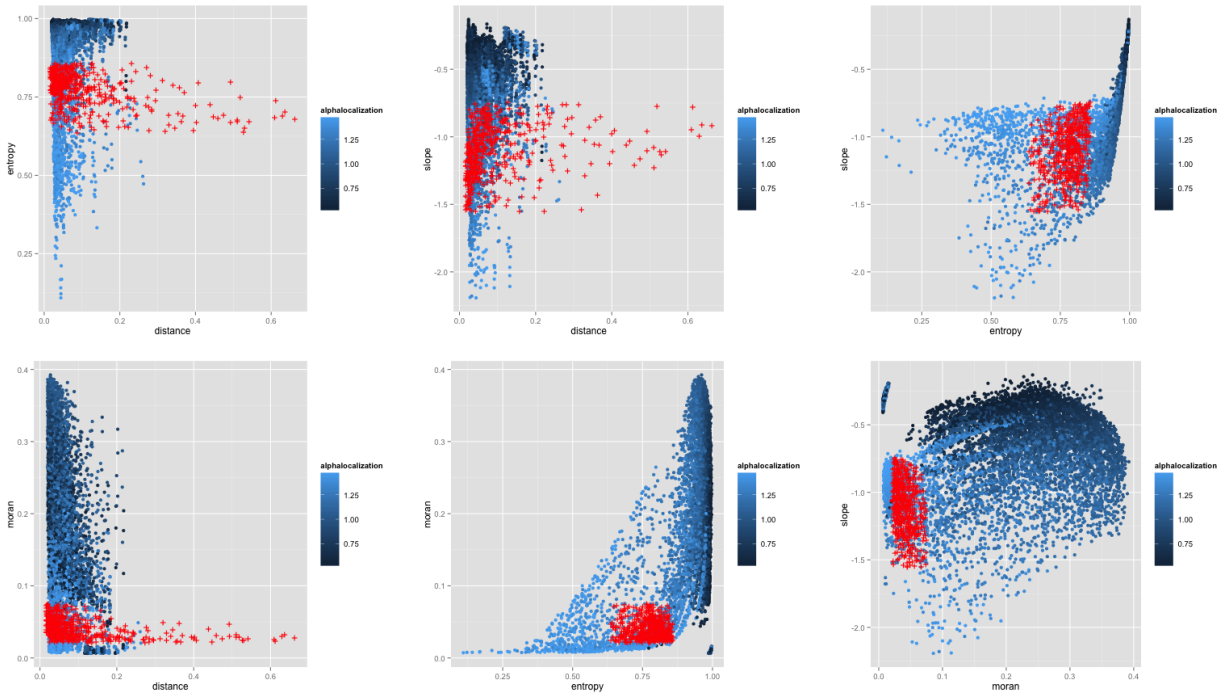


FIGURE 13 : Scatterplots of indicators distribution in an hypercube of the parameter space. We show here the influence of one parameter (localization exponent α). Red points correspond to real data.

implemented in OpenMole, and obtain in Fig. 14 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 distance between generated points and reals points. We select then the point cloud that overlaps real points in the (PC1,PC2) plan, given a distance threshold. Fig. 15 shows the points we obtain for four different values of the threshold ranging from 10^{-6} to 10^{-3}.

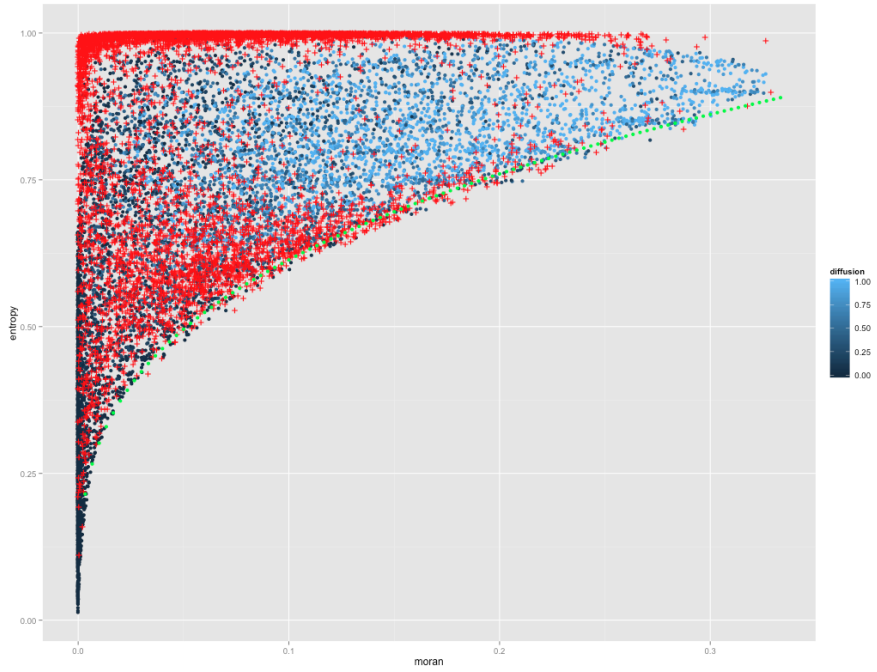


FIGURE 14 : Scatterplots of Moran against Entropy, with blue points obtained with LHS and red with PSE exploration. Lower bound is in green.

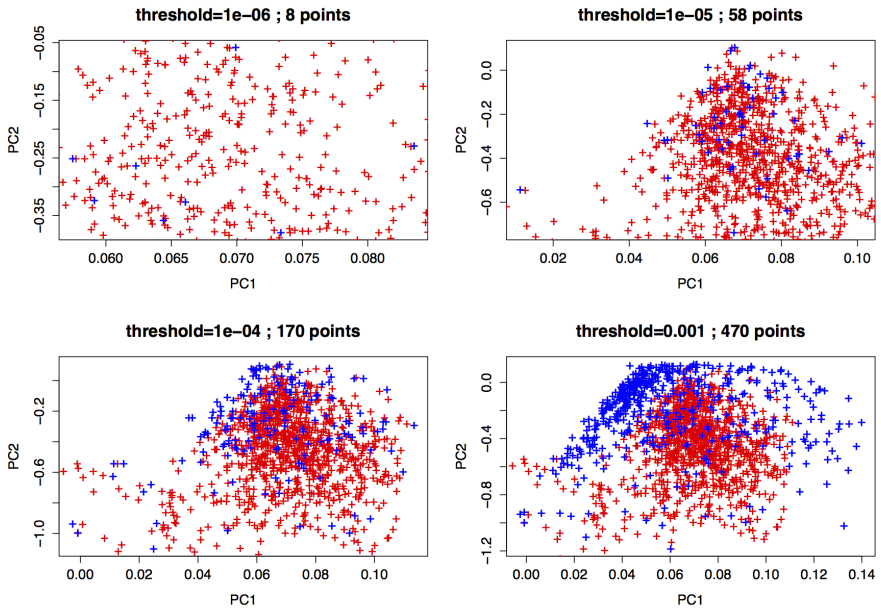


FIGURE 15 : Precise calibration of the model. The principal component analysis is conducted to maximize the spread of the differences between real data and model output, i.e. on the set $\{|R_i - M_j|\}$ where R_i is the set of real points, M_j the set of model outputs. We select then the overlapping cloud at threshold θ, by taking models output closer to real point cloud than θ in the (PC1,PC2) plan.

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.

6.1.4 *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.

Integration into a multi-scale growth model

It could be possible to couple this model with a Gibrat (or Favaro-pumain) 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.

6.2 CORRELATED GENERATION OF TERRITORIAL CONFIGURATIONS

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.

6.2.1 *Correlated geographical data of density and network*

Context

The use of synthetic data in geography is generally directed towards the generation of synthetic populations within agent-based models (mobility, *LUTI* models) [**pritchard2009advances**]. We can make a weak link with some Spatial Analysis techniques. The extrapolation of a continuous spatial field from a discrete spatial sample through a kernel density estimation for example can be understood as the creation of a synthetic dataset (even if it is not generally the initial view, as in Geographically Weighted Regression [**brunsdon1998geographically**] in which variable size kernels do not interpolate data *stricto sensu* but extrapolate abstract variables representing interaction between explicit variables). In the field of modeling in quantitative geography, *toy-models* or hybrid models require a consistent initial spatial configuration. A set of possible initial configurations becomes a synthetic dataset on which the model is tested. The first Simpop model [**sanders1997simpop**], precursor of a large family of models later parametrized with real data, could enter that frame but was studied on an unique synthetic spatialization. Similarly underlined was the difficulty to generate an initial transportation infrastructure in the case of the SimpopNet model [**schmitt2014modelisation**] although it was admitted as a cornerstone of knowledge on the behavior of the model. A systematic control of spatial configuration effects on the behavior of simulation models was only recently proposed [**cottineau2015revisiting**], approach that can be interpreted as a statistical control on spatial data. The aim is to be able to distinguish proper effects due to intrinsic model dynamics from particular effects due to the geographical structure of the case study. Such results are essential for the validation of conclusions obtained with modeling and simulation practices in quantitative geography.

Formalization

We propose in our case to generate territorial systems summarized in a simplified way as a spatial population density $d(\vec{x})$ and a transportation network $n(\vec{x})$. Correlations we aim to control are correlations between urban morphological measures and network measures. The question of interactions between territories and networks is already

well-studied [**offner1996reseaux**] but stays highly complex and difficult to quantify [**offner1993effets**]. A dynamical modeling of implied processes should shed light on these interactions ([**bretagnolle:tel-00459720**], p. 162-163). We develop in that frame a *simple* coupling (i.e. without any feedback loop) between a density distribution model and a network morphogenesis model.

DENSITY MODEL The density model is the model described and explored in the previous section. We use it for the conditional generation of network.

NETWORK MODEL On the other hand, we are able to generate a planar transportation network by a model N, at a similar scale and given a density distribution. Because of the conditional nature to the density of the generation process, we will first have conditional estimators for network indicators, and secondly natural correlations between network and urban shapes should appear as processes are not independent. The nature and modularity of these correlations as a function of model parameters are still to determine by exploration of the coupled model.

The heuristic network generation procedure is the following :

1. A fixed number N_c of centers that will be first nodes of the network si distributed given density distribution, following a similar law to the aggregation process, i.e. the probability to be distributed in a given patch is $\frac{(P_i/P)^\alpha}{\sum(P_i/P)^\alpha}$. Population is then attributed according to Voronoi areas of centers, such that a center cumulates population of patches within its extent.

2. Centers are connected deterministically by percolation between closest clusters : as soon as network is not connected, two closest connected components in the sense of minimal distance between each vertices are connected by the link realizing this distance. It yields a tree-shaped network.

3. Network is modulated by potential breaking in order to be closer from real network shapes. More precisely, a generalized gravity potential between two centers i and j is defined by

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

where d can be euclidian distance $d_{ij} = d(i,j)$ or network distance $d_N(i,j)$, $k_h \in [0,1]$ a weight to modulate role of populations, γ giving shape of the hierarchy across population values, r_g characteristic interaction distance and d_0 distance shape parameter.

4. A fixed number $K \cdot N_L$ of potential new links is taken among couples having greatest euclidian distance potential ($K = 5$ is fixed).

5. Among potential links, N_L are effectively realized, that are the one with smallest rate $\tilde{V}_{ij} = V_{ij}(d_N)/V_{ij}(d_{ij})$. At this stage only the gap between euclidian and network distance is taken into account : $\tilde{V}_{ij}$ does indeed not depend on populations and is increasing with d_N at constant d_{ij}.

6. Planarity of the network is forced by creation of nodes at possible intersections created by new links.

We insist on the fact that the network generation procedure is entirely heuristic and result of thematic assumptions (connected initial network, gravity-based link creation) combined with trial-and-error during first explorations. Other model types could be used as well, such biological self-generated networks [**tero2010rules**], local network growth based on geometrical constraints optimization [**barthelemy2008modeling**] or a more complex percolation model than the initial one that would allow the creation of loops for example. We could thus in the frame of a modular architecture, in which the choice between different implementations of a functional brick can be seen as a meta-parameter [**cottineau2015increment**] choose network generation function adapted to a specific need (as e.g. proximity to real data, constraints on output indicators, variety if generated forms, etc.).

PARAMETER SPACE Parameter space for the coupled model[1] is constituted by density generation parameters $\vec{\alpha}_D = (P_m/N_G, \alpha, \beta, n_d)$ (we study for the sake of simplicity the rate between population and growth rate instead of both varying, i.e. the number of steps needed to generate the distribution) and network generation parameters $\vec{\alpha}_N = (N_C, k_h, \gamma, r_g, d_0)$. We denote $\vec{\alpha} = (\vec{\alpha}_D, \vec{\alpha}_N)$.

INDICATORS Urban form and network structure are quantified by numerical indicators in order to modulate correlations between these. Morphology is defined as a vector $\vec{M} = (r, \bar{d}, \varepsilon, a)$ giving spatial auto-correlation (Moran index), mean distance, entropy and hierarchy (see [**le2015forme**] for a precise definition of these indicators). Network measures $\vec{G} = (\bar{c}, \bar{l}, \bar{s}, \delta)$ are with network denoted (V, E)

- Mean centrality $\bar{c}$ defined as average *betweeness-centrality* (normalized in $[0, 1]$) on all links.

1 Weak coupling allows to limit the total number of parameters as a strong coupling would involve retroaction loops and consequently associated parameters to determine their structure and intensity. In order to diminish it, an integrated model would be preferable to a strong coupling, what is slightly different in the sense where it is not possible in the integrated model to freeze one of the subsystems to obtain a model of the other subsystem that would correspond to the non-coupled model.

- Mean path length $\bar{l}$ given by $\frac{1}{d_m}\frac{2}{|V|\cdot(|V|-1)}\sum_{i<j} d_N(i,j)$ with d_m normalization distance taken here as world diagonal $d_m = \sqrt{2}N$.
- Mean network speed [**banos2012towards**] which corresponds to network performance compared to direct travel, defined as $\bar{s} = \frac{2}{|V|\cdot(|V|-1)}\sum_{i<j}\frac{d_{ij}}{d_N(i,j)}$.
- Network diameter $\delta = \max_{ij} d_N(i,j)$.

COVARIANCE AND CORRELATION We study the cross-correlation matrix $\mathrm{Cov}\left[\vec{M},\vec{G}\right]$ between morphology and network. We estimate it on a set of n realizations at fixed parameter values $(\vec{M}\left[D(\vec{\alpha})\right],\vec{G}\left[N(\vec{\alpha})\right])_{1\leqslant i\leqslant n}$ with standard unbiased estimator. We estimate correlation with associated Pearson estimator.

Implementation

Coupling of generative models is done both at formal and operational levels. We interface therefore independent implementations. The OpenMole software [**reuillon2013openmole**] for intensive model exploration offers for that the ideal frame thanks to its modular language allowing to construct *workflows* by task composition and interfacing with diverse experience plans and outputs. For operational reasons, density model is implemented in `scala` language as an OpenMole `plugin`, whereas network generation is implemented in agent-oriented language `NetLogo` [**wilensky1999netlogo**] because of its possibilities for interactive exploration and heuristic model construction. Source code is available for reproducibility on project repository[2].

Results

The study of density model alone is developed in [**raimbault2016calibration**]. It is in particular calibrated on European density grid data, on 50km width square areas with 500m resolution for which real indicator values have been computed on whole Europe. Furthermore, a grid exploration of model behavior yields feasible output space in reasonable parameters bounds (roughly $\alpha\in[0.5,2], N_G\in[500,3000], P_m\in[10^4,10^5], \beta\in[0,0.2], n_d\in\{1,\dots,4\}$). The reduction of indicators space to a two dimensional plan through a Principal Component Analysis (variance explained with two components $\simeq 80\%$) allows to isolate a set of output points that covers reasonably precisely real point cloud. It confirms the ability of the model to reproduce morphologically the set of real configurations.

At given density, the conditional exploration of network generation model parameter space suggest a good flexibility on global indicators

2 at `https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Synthetic`

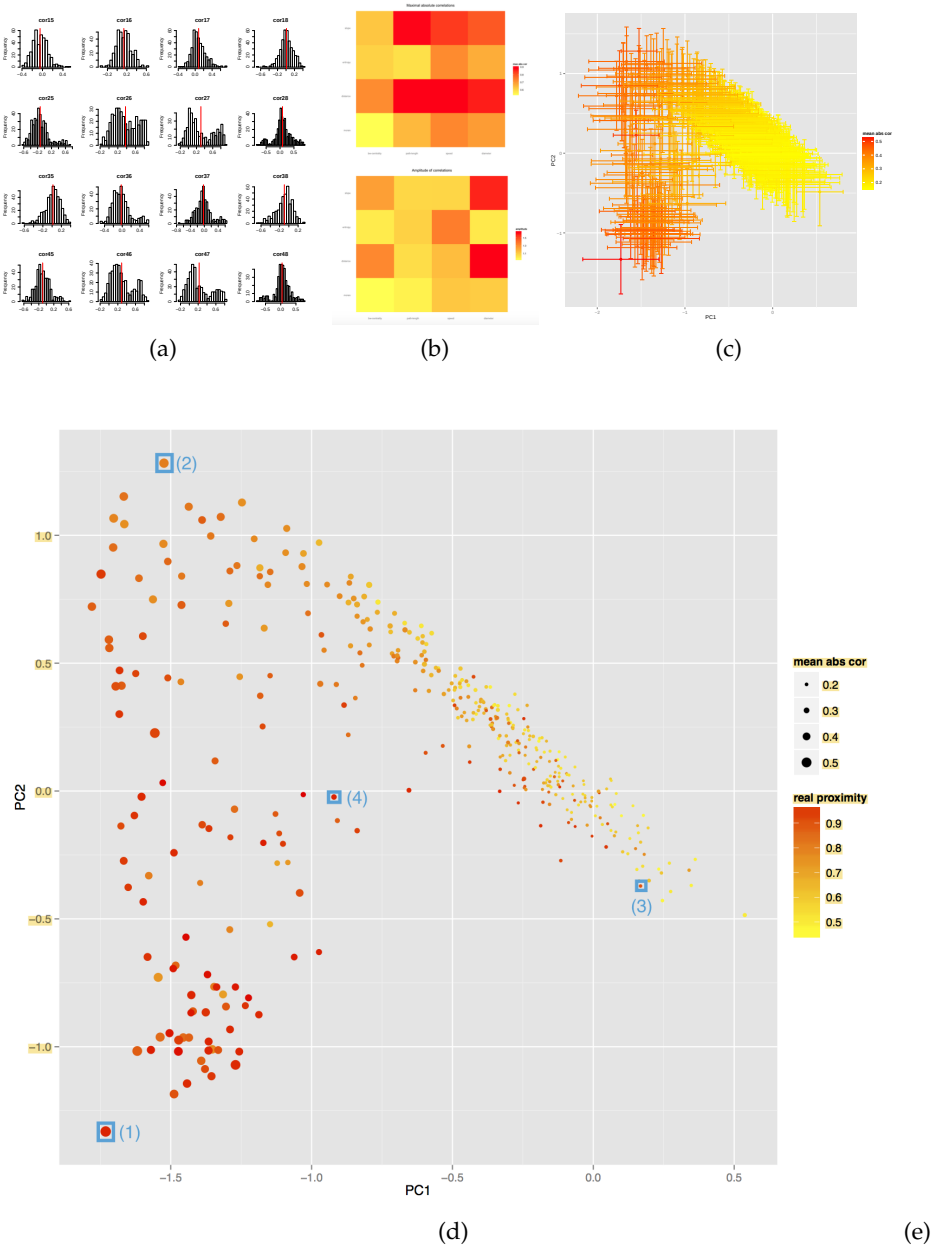


FIGURE 16 : **Exploration of feasible space for correlations between urban morphology and network structure** | **(a)** Distribution of crossed-correlations between vectors $\vec{M}$ of morphological indicators (in numbering order Moran index, mean distance, entropy, hierarchy) and $\vec{N}$ of network 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 $c_{ij} = \max_k \left|\rho_{ij}^{k}\right|$. **(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.

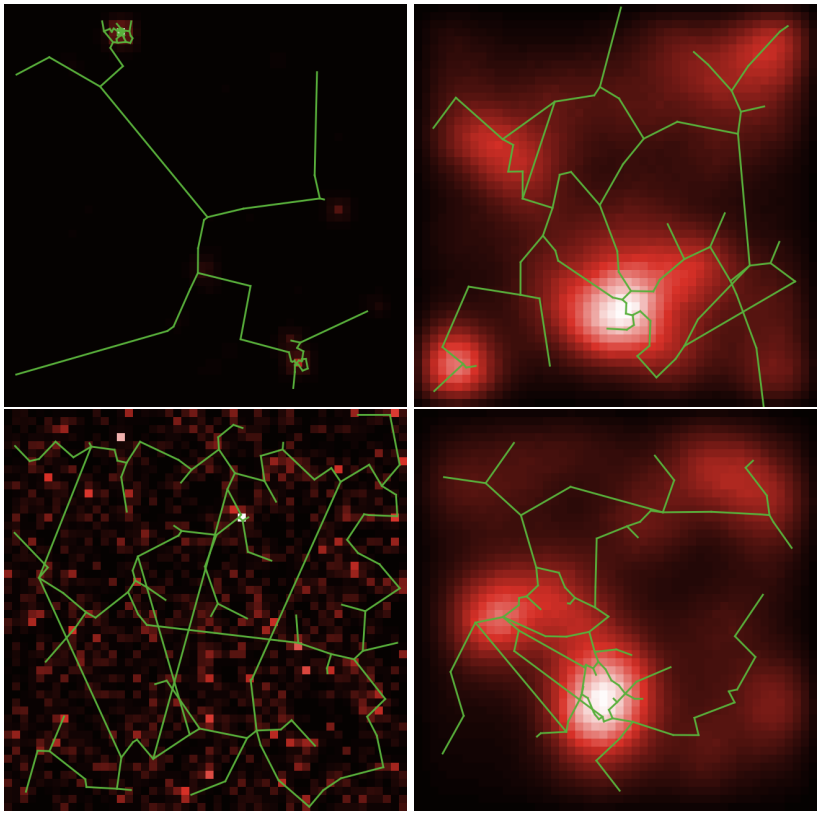


FIGURE 17 : Configurations obtained for parameters giving the four emphasized points in (d), in order from left to right and top to bottom. We recognize polycentric city configurations (2 and 4), diffuse rural settlements (3) and aggregated weak density area (1). See appendice for exhaustive parameter 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.

$\vec{G}$, together with good convergence properties. For a precise study of model behavior, see appendice giving regressions analysis capturing the behavior of coupled model. In order to illustrate synthetic data generation method, the exploration has been oriented towards the study of cross-correlations.

Given the large relative dimension of parameter space, an exhaustive grid exploration is not possible. We use a Latin Hypercube sampling procedure with bounds given above for $\vec{\alpha}_D$ and for $\vec{\alpha}_N$, we take $N_C \in [50, 120], r_g \in [1, 100], d_0 \in [0.1, 10], k_h \in [0, 1], \gamma \in [0.1, 4], N_L \in [4, 20]$. For number of model replications for each parameter point, less than 50 are enough to obtain confidence intervals at 95% on indicators of width less than standard deviations. For correlations a hundred give confidence intervals (obtained with Fisher method) of size around 0.4, we take thus $n = 80$ for experiments. Figure 16 gives details of experiment results. Regarding the subject of correlated synthetic data generation, we can sum up the main lines as following :

- Empirical distributions of correlation coefficients between morphology and network indicators are not simple and some are bimodal (for example $\rho_{46} = \rho[r, \bar{l}]$ between Moran index and mean path length).
- it is possible to modulate up to a relatively high level of correlation for all indicators, maximal absolute correlation varying between 0.6 and 0.9. Amplitude of correlations varies between 0.9 and 1.6, allowing a broad spectrum of values. Point cloud in principal plan has a large extent but is not uniform : it is not possible to modulate at will any coefficient as they stay themselves correlated because of underlying generation processes. A more refined study at higher orders (correlation of correlations) would be necessary to precisely understand degrees of freedom in correlation generation.
- Most correlated points are also the closest to real data, what confirms the intuition and stylized fact of a strong interdependence in reality.
- Concrete examples taken on particular points in the principal plan show that similar density profiles can yield very different correlation profiles.

Possible developments

This case study could be refined by extending correlation control method. A precise knowledge of N behavior (statistical distributions on an exhaustive grid of parameter space) conditional to D would allow to determine $N^{<-1>}|D$ and have more latitude in correlation generation. We could also apply specific exploration algorithms to reach exceptional configurations realizing an expected correlation level, or

at least to obtain a better knowledge of the feasible space of correlations [**10.1371/journal.pone.0138212**].

6.2.2 *Discussion*

Scientific positioning

Our overall approach enters a particular epistemological frame. On the one hand the multidisciplinary aspect, and on the other hand the importance of empirical component through computational exploration methods, make this approach typical of Complex Systems science, as it is recalled by the roadmap for Complex Systems having a similar structure [**2009arXiv0907.2221B**]. It combines transversal research questions (horizontal integration of disciplines) with the development of heterogeneous multi-scalar approaches which encounter similar issues as the one we proposed to tackle (vertically integrated disciplines). The combination of empirical knowledge obtained from data mining, with knowledge obtained by modeling and simulation is generally central to the conception and exploration of multi-scalar heterogeneous models. Results presented here is an illustration of such an hybrid paradigm.

Direct applications

Starting from the second example which was limited to data generation, we propose examples of direct applications that should give an overview of the range of possibilities.

- Calibration of network generation component at given density, on real data for transportation network (typically road network given the shape of generated networks; it should be straightforward to use OpenStreetMap open data[3] that have a reasonable quality for Europe, at least for France [**girres2010quality**], with however adjustments on generation procedure in order to avoid edge effects due its restrictive frame, for example by generating on an extended surface to keep only a central area on which calibration would be done) should theoretically allow to unveil parameter sets reproducing accurately existing configurations both for urban morphology and network shape. It could be then possible to derive a "theoretical correlation" for these, as an empirical correlation is according to some theories of urban systems not computable as a unique realization of stochastic processes is observed. Because of non-ergodicity of urban systems [**pumain2012urban**], there are strong chances that involved processes are different across different geographical areas (or from an other point of view that they are in an

3 https://www.openstreetmap.org

other state of meta-parameters, i.e. in an other regime) and that their interpretation as different realizations of the same stochastic process makes no sense, the impossibility of covariation estimation following. By attributing a synthetic dataset similar to a given real configuration, we would be able to compute a sort of *intrinsic correlation* proper to this configuration. As territorial configurations emerge from spatio-temporal interdependences between components of territorial systems, this intrinsic correlation emerges the same way, and its knowledge gives information on these interdependences and thus on relations between territories and networks.

- As already mentioned, most of models of simulation need an initial state generated artificially as soon as model parametrization is not done completely on real data. An advanced model sensitivity analysis implies a control on parameters for synthetic dataset generation, seen as model meta-parameters **[cottineau2015revisiting]**. In the case of a statistical analysis of model outputs it provides a way to operate a second order statistical control.
- We studied in the first example stochastic processes in the sense of random time-series, whereas time did not have a role in the second case. We can suggest a strong coupling between the two model components (or the construction of an integrated model) and to observe indicators and correlations at different time steps during the generation. In a dynamical spatial models we have because of feedbacks necessarily propagation effects and therefore the existence of lagged interdependences in space and time [**pigozzi1980interurban**]. It would drive our field of study towards a better understanding of dynamical correlations.

Generalization

We were limited to the control of first and second moments of generated data, but we could imagine a theoretical generalization allowing the control of moments at any order. However, as shown by the geographical example, the difficulty of generation in a concrete complex case questions the possibility of higher orders control when keeping a consistent structure model and a reasonable number of parameters. The study of non-linear dependence structures as proposed in [**chicheportiche2013nested**] is in an other perspective an interesting possible development.

6.2.3 *Conclusion*

We described a model allowing to generate synthetic datasets in which correlation structure is controlled. Its exploration shows its flexibility and the broad range of possible applications. More generally, it is

crucial to favorise such practices of systematic validation of computational models by statistical analysis, in particular for agent-based models for which the question of validation stays an open issue.

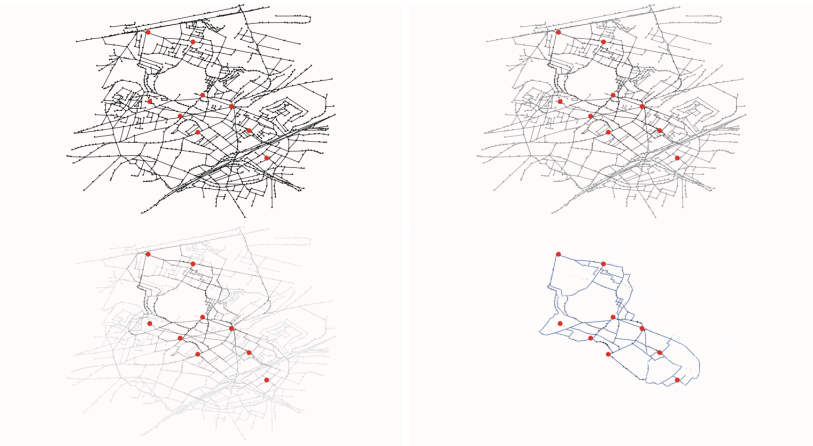


FIGURE 18 : Example of the application of the slime mould network generation model to the computation of an optimal public transportation network design.

6.3 NETWORK GROWTH MODELS : EXPLICATIVE POWER FOR VARIOUS APPROACHES

6.3.1 *Benchmarking Network growth heuristics*

Considering Network Growth in itself, many heuristics are available to generate a network under some constraints. As already developed, from economic network growth approach to local optimization heuristics, geographical mechanisms or biological network growth, each has its advantages and particularities. We plan to compare these varied methods against real network indicators values for the european road network once these will have been calculated. We present in Fig. 18 a preliminary work done in [**raimbault2015labex**] to explore implementation of the biological network growth models. Also the implementation of local optimization models was explored, typically the one described in the methodology section on reproducibility.

6.3.2 *An interdisciplinary approach to 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 discuss 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.

7

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 macroeconomic/accessibility-based model to explore transportation companies line implementation strategies are still at the state of ideas and are not described here.

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

7.1.1 *Thematic Context*

We briefly describe a simple game-theory based framework, conjointly done 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.

7.1.2 *Formalization*

The model architecture couples in a complex way a module for land-use evolution with a module for transportation network growth. Submodules, 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 $E(\vec{x}) = E(\vec{x}) + \frac{k_{ext} \cdot E_{max}}{n_{ext}}$.

- 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

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

- Trajectories then attributed by effective shortest path, and corresponding congestion c obtained (no Wardrop equilibrium).
- Speed of network is given by a BPR function $v(c) = v_0\left(1 - \frac{c}{\kappa}\right)^{\gamma_c}$. 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 $X_i(A) = A_i \cdot \sum_j E_j \exp\left(-\lambda \cdot d_{ij}\right)$ and $X_j(E) = E_j \cdot \sum_i A_i \exp\left(-\lambda \cdot d_{ij}\right)$.
- Relocations are then done deterministically following a discrete choice model :

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

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

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 ; $O(N_{inters}^3)$
- Weak component clusters and distance between clusters updated ; $O(N_{nw}^2)$
- Network distances between network patches updated, through the heuristic of only minimal connexions between clusters ; $O(N_{nw}^2)$
- Effective distances (taking paces/congestion into account) updated as minimum between euclidian time and

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

  , complexity in $O(N_{clusters}^2 \cdot N^2)$ (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 divise 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

$$\begin{aligned}\mathbb{E}[U_1] &= p_1p_2U_1(C,C) + p_1\cdot(1-p_2)U_1(C,A) + p_2\cdot(1-p_1)U_1(A,C) + (1-p_1)(1-p_2)U\\ &= p_1\cdot\left[p_2\cdot\left(\kappa\cdot\Delta X_1(Z_C^*) - \frac{\delta I}{2}\right) - \kappa\cdot\Delta X_1(Z_1^*) + I\right] + p_2\cdot\frac{\delta I}{2} + \kappa\cdot\Delta X_1(Z_1^*) - I\end{aligned}$$

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

$$\frac{\partial\,\mathbb{E}[U_1]}{\partial p_1} = 0 \iff p_2 = \frac{\delta I/2}{\Delta X_{\bar{2}}Z_C^\star - \Delta X_{\bar{2}}Z_{\bar{2}}^\star}$$

We obtain generally

$$p_i = \frac{J}{\Delta X_{\bar{i}}Z_C^\star - \Delta X_{\bar{i}}Z_{\bar{i}}^\star}$$

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 $I \leqslant \kappa\cdot\min(\Delta X_1(Z_1^*), \Delta X_2(Z_2^*))$ (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

$$U_i(C) - U_i(NC) = p_{\bar{i}}\,(\Delta X_i Z_C^\star - \Delta X_i Z_i^\star) - J$$

and therefore p_i verifies the equation that is solved numerically

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

This module is also implemented for comparison purposes.

7.1.3 *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. 19 and Fig.20 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).

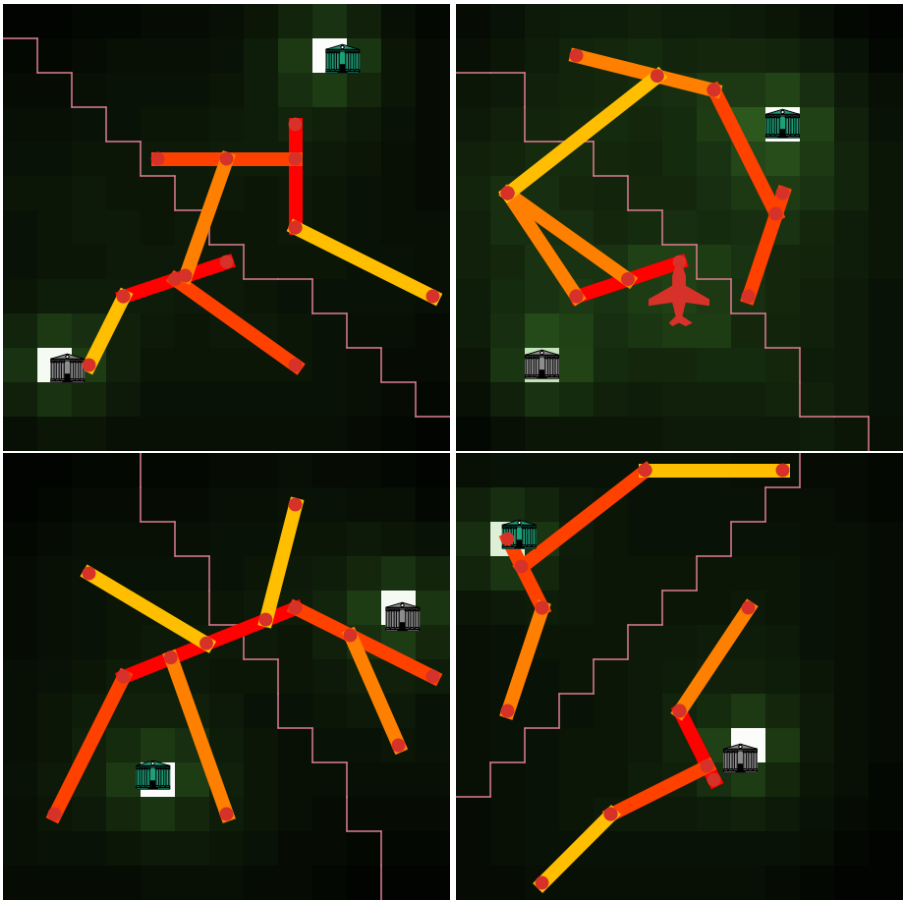


FIGURE 19 : Examples of final configurations, with or without externality, for different values of cooperation parameters.

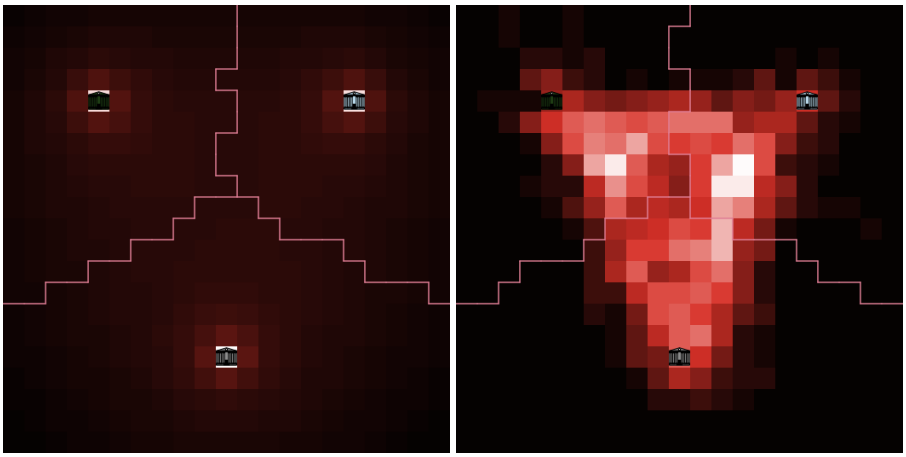


FIGURE 20 : Validation of network exploration heuristic : mean accessibility(left) and network positions on 500 realizations on the same initial configuration. The optimal distribution of network validates network generation heuristic.

Troisième partie

TOWARDS OPERATIONAL MODELS

This concluding remark, for now a brief roadmap, is one objective of our thesis as implementation of our theory ans thus is expected to become a consequent part. We conclude here this preliminary work by perspectives and roadmap.

8

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.

8.1 OBJECTIVES

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.

8.2 CASE STUDIES

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.

8.3 ROADMAP

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 6), 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.

9

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 restreignante 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.

9.1 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 coexistent aujourd'hui, parmi lesquels on trouve par exemple les modèles dynamiques de micro-simulation, généralement opposés aux techniques de basant sur l'équilibre. Tandis que la validité des modèles microscopiques a été étudiée de façon conséquente et leur application souvent questionnée, la littérature est relativement pauvre en études empiriques assurant l'hypothèse d'équilibre stationnaire du cadre de l'Equilibre Utilisateur Statique (EUS). De nombreux développements plus réalistes on été documentés dans la littérature, tels l'Equilibre Utilisateur Dynamique Stochastique (EUDS) (voir pour une description par example [**han2003dynamic**]). A un niveau intermédiaire entre les cadres statiques et stochastiques se trouve l'Equilibre Utilisateur Stochas-

tique 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 le 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 ensuite une application interactive pour l'exploration du jeu de données, dans le but de fournir une intuitions sur les motifs présents ; puis

nous donnons divers résultats d'analyses quantitatives allant dans le sens d'indices convergents pour une non-stationnarité des flots de trafic ; nous discutons finalement les implications de ces résultats et des développements possibles.

9.2 DATA COLLECTION

9.2.1 *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, executés 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].

9.2.2 *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 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

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

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.

9.3 METHODS AND RESULTS

9.3.1 *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 grace au zoom. L'application est accessible en ligne à l'adresse `http://shiny.parisgeo.cnrs.fr/transportation`. La Figure 21 présente une capture d'écran de l'interface. La conclusion majeure de l'exploration interactive des données est qu'une grande hétérogénéité spatiale et temporelle est la règle. Le motif temporel le plus récurrent, la périodicité journalière des heures de pointe, est perturbée pour une proportion non négligeable de jours. En première approximation, les heures creuses peuvent être approchées par une distribution localement stationnaire des flots, tandis que les heures de pointe sont trop courtes pour pouvoir impliquer la validation de l'hypothèse d'équilibre. Concernant l'espace, aucun motif spatial particulier n'émerge clairement. Cela signifie que dans le cas d'une validité de l'équilibre utilisateur statique, les méta-paramètres régissant son établissement doivent varier à des échelles temporelles plus courtes qu'un jour. Nous postulons au contraire que le système de traffic est loin de l'équilibre, en particulier pendant les heures de pointe pendant lesquelles des transitions de phase critiques à l'origine des embouteillages émergent.

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

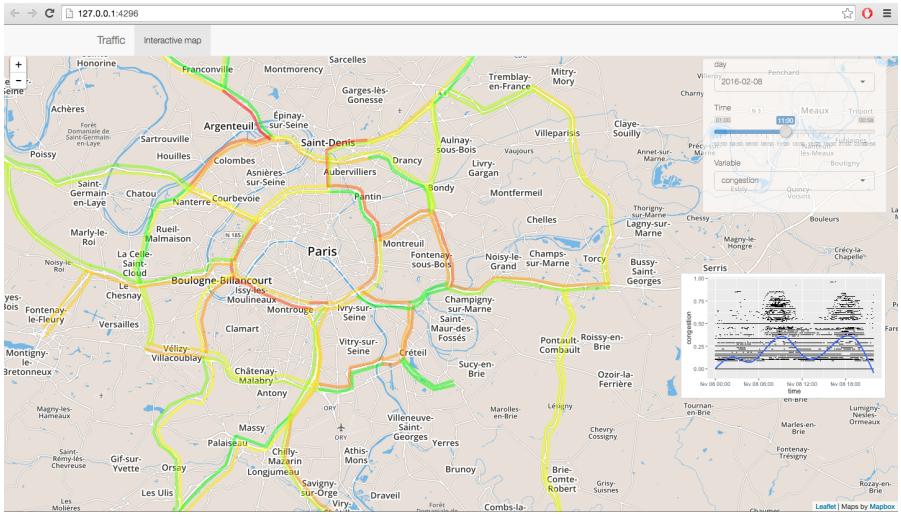


FIGURE 21

9.3.2 *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 22 montre un exemple de variation spatiale extrême d'un trajet pour une paire Origine-Destination particulière.

L'exploration systématique de la variabilité du temps de trajet sur l'ensemble du jeu de données, et des distances de trajet associées, confirme, comme présenté en figure , que la variation absolue du temps de trajet présente fréquemment une forte variation de son maximum sur l'ensemble des paires O-D, jusqu'à 25 minutes avec une moyenne temporelle locale autour de 10 minutes. La variabilité spatiale correspondante entraine des détours allant jusqu'à 35km.

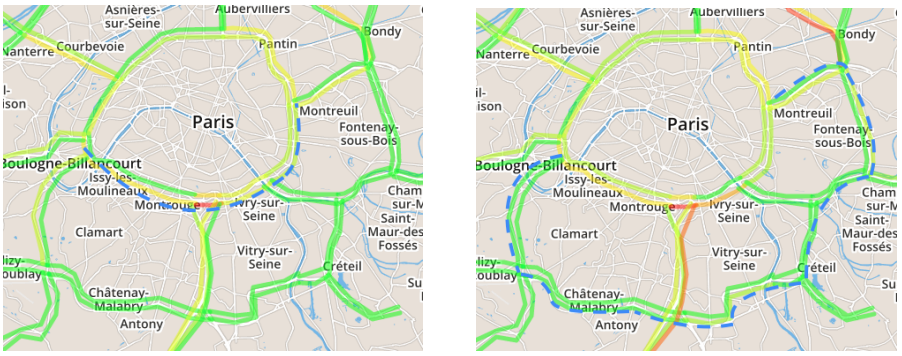


FIGURE 22

9.3.3 *Stability of Network measures*

The variability of potential trajectories observed in the previous section can be confirmed by studying the variability of network properties. In particular, network topological measures capture global patterns of a transportation network. Centrality and node connectivity measures are classical indicators in transportation network description as recalled in [**bavoux2005geographie**]. The transportation literature has developed elaborated and operational network measures, such as network robustness measures to identify critical links and measure overall network resilience to disruptions (an example among many is the Network Trip Robustness index introduced in [**sullivan2010identifying**]).

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

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

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

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

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

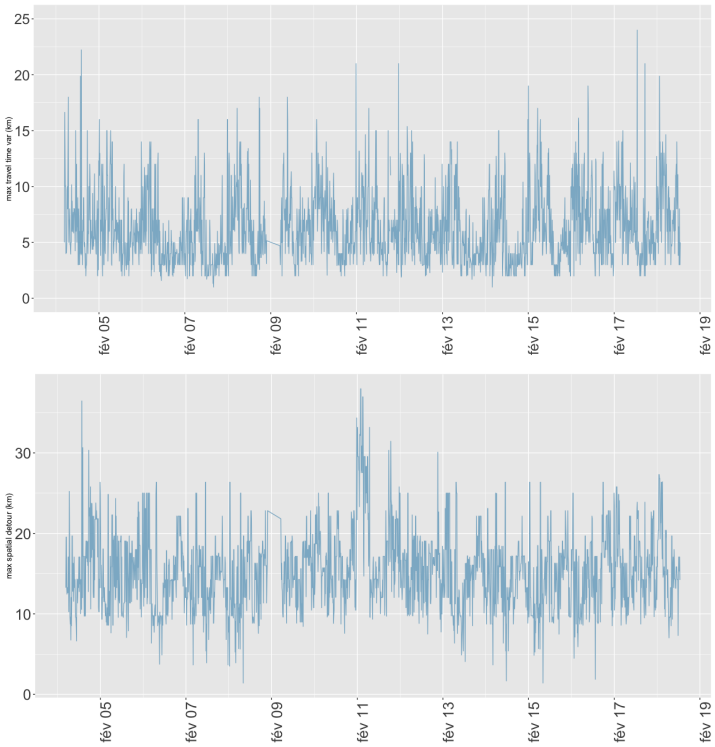


FIGURE 23 : Travel time (top) in min and corresponding travel distance (bottom) maximal variability on a two weeks sample. We plot the maximal on all OD pairs of the absolute variability between two consecutive time steps. Peak hours imply a high time travel variability up to 25 minutes and a path length variability up to 35km.

where Δt is the time step of the dataset (the smallest time window on which we can capture variability). This absolute relative variation has a direct meaning : a variation of 20% (which is attained a significant number of times as shown in Fig. 24) means that in case of a negative variation, at least this proportion of potential travels have changed route and the local potential congestion has decrease of the same proportion. In the case of a positive variation, a single node has captured at least 20% of travels. Under the assumption (that we do not try to verify in this work and assume to be also not verified as shown by [**zhu2010people**], but that we use as a tool to give an idea of the concrete meaning of betweenness variability) that users rationally take the shortest path and assuming that a majority of travels are realized such a variation in centrality imply a similar variation in effective flows, leading to the conclusion that they can not be stationary in time (at least at a scale larger than Δt) nor in space.

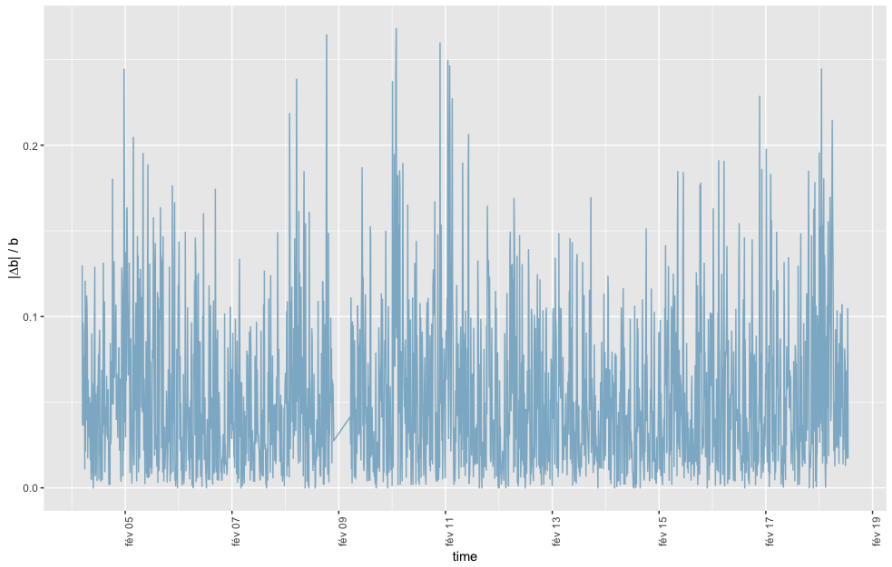


FIGURE 24 : Temporal stability of maximal betweenness centrality. We plot in time the normalized derivative of maximal betweenness centrality, that expresses its relative variations at each time step. The maximal value up to 25% correspond to very strong network disruption on the concerned link, as it means that at least this proportion of travelers assumed to take this link in previous conditions should take a totally different path.

9.3.4 *Spatial heterogeneity of equilibrium*

To obtain a different insight into spatial variability of congestion patterns, we propose to use an index of spatial autocorrelation, the Moran index (defined e.g. in [**tsai2005quantifying**]). More generally used in spatial analysis with diverse applications from the study of urban form to the quantification of segregation, it can be applied to any spatial variable. It allows to establish neighborhood relations and unveils spatial local consistence of an equilibrium if applied on localized traffic variable. At a given point in space, local autocorrelation for variable c is computed by

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

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

Figure 25 presents temporal evolution of spatial autocorrelation for congestion. As expected, we have a strong decrease of autocorrelation with distance decay parameter, for both amplitude and temporal average. The high temporal variability implies short time scales for potential stationarity windows. When comparing with congestion (fitted to plot scale for readability) for 1km decay, we observe that high correlations coincide with off-peak hours, whereas peaks involve vanishing correlations. Our interpretation, combined with the observed variability of spatial patterns, is that peak hours correspond to chaotic behaviour of the system, as jams can emerge in any link : correlation thus vanishes as feasible phase space for a chaotic dynamical system is filled by trajectories in an uniform way what is equivalent to apparently independent random relative speeds.

9.4 DISCUSSION

9.4.1 *Theoretical and practical implications of empirical conclusions*

We argue that the theoretical implications of our empirical findings do not imply in a total discarding of the Static User Equilibrium framework, but unveil more a need of stronger connections between theoretical literature and empirical studies. If each newly introduced theoretical framework is generally tested on one on more case study, there are no systematic comparisons of each on large and different datasets and on various objectives (prediction of traffic, reproduction of stylized facts, etc.) as systematic reviews are the rule in therapeutic evaluation for example. This imply however broader data and model sharing practices than the current ones. The precise knowledge of application potentialities for a given framework may induce unexpected developments such as its integration into larger models. The example of Land-use and Transportation Interaction studies (LUTI models) is a good illustration of how the SUE can still be used for larger purpose than transportation modeling. [**kryvobokov2013comparison**] describe two LUTI models, one of which includes two equilibria for four-step transportation model and for land-use evolution (households and firms relocation), the other being more dynamical. The conclusion is that each model has its own advantages regarding the pursued objective, and that the static model can be used for long time policy purposes, whereas the dynamic model provide more precise information at smaller time scale. In the first case, a more complicated transportation module would have been complicated to include, what is an advantage of the static user equilibrium.

Concerning practical applications, it seems natural that static models should not be used for traffic forecast and management at small time scales (week or day) and efforts should be made to implement

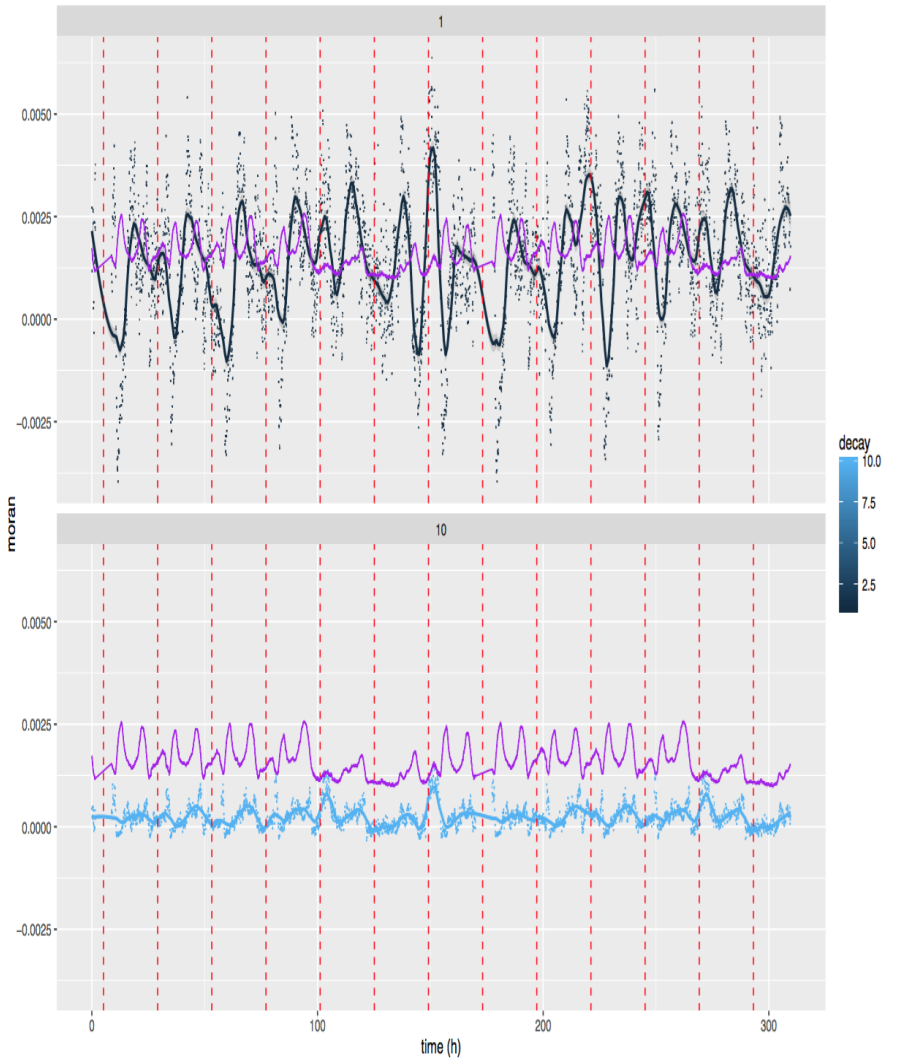


FIGURE 25 : Spatial auto-correlations for relative travel speed on two weeks. We plot for varying value of decay parameter (1,10km) values of auto-correlation index in time. Intermediate values of decay parameter yield a rather continuous deformation between the two curves. Points are smoothed with a 2h span to ease reading. Vertical dotted lines correspond to midnight each day. Purple curve is relative speed fitted at scale to have a correspondence between auto-correlation variations and peak hours.

more realistic models. However the use of models by the planning and engineering community is not necessarily directly related to academic concerns and state-of-the-art. For the particular case of France and mobility models, [**commenges2013invention**] showed that engineers had gone to the point of constructing inexistent problems and implementing corresponding models that they had imported from a totally different geographical context (planning in the United States). The use of one framework or type of model has historical reasons that may be difficult to overcome.

9.4.2 *Towards explanative interpretations of non-stationarity*

An assumption we formulate regarding the origin of non-stationarity of network flows, in view of data exploration and quantitative analysis of the database, is that the network is at least half of the time highly congested and in a critical state. The off-peak hours are the larger potential time windows of spatial and temporal stationarity, but consist in less than half of the time. As already interpreted through the behavior of autocorrelation indicator, a chaotic behavior may be at the origin of such variability in the congested hours. The same way a supercritical fluid may condense under the smallest external perturbation, the state of the link may qualitatively change with a small incident, producing a network disruption that may propagate and even amplify. The direct effect of traffic events (notified incidents or accidents) can not be studied without external data, and it could be interesting to enrich the database in that direction. It would allow establishing the proportion of disruptions that do appear to have a direct effect and quantify a level of criticality of network congestion in time, or to investigate more precise effects such as the consequences of an incident on traffic of the opposite lane.

9.4.3 *Possible developments*

Further work may be planned towards a more refined assessment of temporal stability on a region of the network, i.e. the quantitative investigation of consideration of peak stationarity given above. To do so we propose to compute numerically Liapounov stability of the dynamical system ruling traffic flows using numerical algorithms such as described by [**goldhirsch1987stability**]. The value of Liapounov exponents provides the time scale by which the unstable system runs out of equilibrium. Its comparison with peak duration and average travel time, across different spatial regions and scales should provide more information on the possible validity of the local stationarity assumption. This technique has already been introduced at an other scale in transportation studies, as e.g. [**tordeux2016jam**] that study

the stability of speed regulation models at the microscopic scale to avoid traffic jams.

Other research directions may consist in the test of other assumptions of static user equilibrium (as the rational shortest path choice, which would be however difficult to test on such an aggregated dataset, implying the use of simulation models calibrated and cross-validated on the dataset to compare assumptions, without necessarily a direct clear validation or invalidation of the assumption), or the empirical computation of parameters in stochastic or dynamical user equilibrium frameworks.

9.5 CONCLUSION

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

10

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

Multi-objective evaluation is a necessary aspect when managing complex systems, as the intrinsic complexity of a system is generally closely linked to the potential number of optimization objectives. However, an evaluation makes no sense without its robustness being given (in the sense of its reliability). Statistical robustness computation methods are highly dependent of underlying statistical models. We propose a formulation of a model-independent framework in the case of integrated aggregated indicators (multi-attribute evaluation), that allows to define a relative measure of robustness taking into account data structure and indicator values. We implement and apply it to a synthetic case of urban systems based on Paris districts geography, and to real data for evaluation of income segregation for Greater Paris metropolitan area. First numerical results show the potentialities of this new method. Furthermore, its relative independence to system type and model may position it as an alternative to classical statistical robustness methods.

10.1 INTRODUCTION

10.1.1 *General Context*

Multi-objective problems are organically linked to the complexity of underlying systems. Indeed, either in the field of *Complex Industrial Systems*, in the sense of engineered systems, where construction of Systems of Systems (SoS) by coupling and integration often leads to contradictory objectives [**marler2004survey**], or in the field of *Natural Complex Systems*, in the sense of non engineered physical, biological or social systems that exhibit emergence and self-organization properties, where objectives can e.g. be the result of heterogeneous interacting agents (see [**newman2011complex**] for a large survey of systems concerned by this approach), multi-objective optimization can be explicitly introduced to study or design the system but is often already implicitly ruling the internal mechanisms of the system. The case of socio-technical Complex Systems is particularly interesting as, following [**haken2003face**], they can be seen as hybrid systems embedding social agents into "technical artifacts" (sometimes to an unexpected degree creating what PICON describes as *cyborgs* [**picon2013smart**]), and thus cumulate propensity to be at the origin of multi-objective

issues[1]. The new notion of *eco-districts* [**souami2012ecoquartiers**] is a typical example where sustainability implies contradictory objectives. The example of transportation systems, which conception shifted during the second half of the 20th century from cost-benefit analysis to multi-criteria decision-making, is also typical of such systems [**bavoux2005geographie**]. Geographical system are now well studied from such a point of view in particular thanks to the integration of multi-objective frameworks within Geographical Information Systems [**carver1991integrating**]. As for the micro-case of eco-districts, meso and macro urban planning and design may be made sustainable through indicators evaluation [**jegou2012evaluation**].

A crucial aspect of an evaluation is a certain notion of its reliability, that we call here *robustness*. Statistics naturally include this notion since the construction and estimation of statistical models give diverse indicators of the consistence of results [**launer2014robustness**]. The first example that comes to mind is the application of the law of large numbers to obtain the *p-value* of a model fit, that can be interpreted as a confidence measure of estimates. Besides, confidence intervals and *beta-power* are other important indicators of statistical robustness. Bayesian inference provide also measures of robustness when distribution of parameters are sequentially estimated. Concerning multi-objective optimization, in particular through heuristic algorithms (for example genetic algorithms, or operational research solvers), the notion of robustness of a solution concerns more the stability of the solution on the phase space of the corresponding dynamical system. Recent progresses have been done towards unified formulation of robustness for a multi-objective optimization problem, such as [**deb2006introducing**] where robust Pareto-front as defined as solutions that are insensitive to small perturbations. In [**1688537**], the notion of degree of robustness is introduced, formalized as a sort of continuity of other solutions in successive neighborhood of a solution.

However, there still lack generic methods to estimate robustness of an evaluation that would be model-independent, i.e. that would be extracted from data structure and indicators but that would not depend on the method used. Some advantages could be for example an *a priori* estimation of potential robustness of an evaluation and thus to decide if the evaluation is worth doing. We propose here a framework answering this issue in the particular case of Multi-attribute evaluations, i.e. when the problem is made unidimensional by objectives aggregation. It is data-driven and not model-driven in the sense that robustness estimation does not depend on how indicators are compu-

1 We design by *Multi-Objective Evaluation* all practices including the computation of multiple indicators of a system (it can be multi-objective optimization for system design, multi-objective evaluation of an existing system, multi-attribute evaluation; our particular framework corresponds to the last case).

ted, as soon as they respect some assumptions that will be detailed in the following.

10.1.2 *Proposed Approach*

OBJECTIVES AS SPATIAL INTEGRALS We assume that objectives can be expressed as spatial integrals, so it should apply to any territorial system and our application cases are urban systems. It is not that restrictive in terms of possible indicators if one uses suitable variables and integrated kernels : in a way analog to the method of geographically weighted regression [**brunsdon1998geographically**], any spatial variable can be integrated against regular kernels of variable size and the result will be a spatial aggregation which sense depends on kernel size. The example we use in the following such as conditional means or sums suit well the assumption. Even an already spatially aggregated indicator can be interpreted as a spatial indicator by using a Dirac distribution on the centroid of the corresponding area.

LINEARLY AGGREGATED OBJECTIVES A second assumption we make is that the multi-objective evaluation is done through linear aggregation of objectives, i.e. that we are tackling a multi-attribute optimization problem. If $(q_i(\vec{x}))_i$ are values of objectives functions, then weights $(w_i)_i$ are defined in order to build the aggregated decision-making function $q(\vec{x}) = \sum_i w_i q_i(\vec{x})$, which value determines then the performance of the solution. It is analog to aggregated utility techniques in economics and is used in many fields. The subtlety lies in the choice of weights, i.e. the shape of the projection function, and various approaches have been developed to find weights depending on the nature of the problem. Recent work [**dobbie2013robustness**] proposed to compare robustness of different aggregation techniques through sensitivity analysis, performed by Monte-Carlo simulations on synthetic data. Distribution of biases where obtained for various techniques and some showed to perform significantly better than others. Robustness assessment still depended on models used in that work.

The rest of the paper is organized as follows : section 2 describes intuitively and mathematically the proposed framework ; section 3 then details implementation, data collection for case studies and numerical results for an artificial intra-urban case and a metropolitan real case ; section 4 finally discuss limitations and potentialities of the method.

10.2 FRAMEWORK DESCRIPTION

10.2.1 *Intuitive Description*

We describe now the abstract framework allowing theoretically to compare robustnesses of evaluations of two different urban systems. Our framework is a generalization of an empirical method proposed in **[ecodistrictReport]** besides a more general benchmarking study on indicator sense and relevance in a sustainability context. Intuitively, it relies on empirical base resulting from the following axioms :

- Urban systems can be seen from the information available, i.e. raw data describing the system. As a data-driven approach, this raw data is the basis of our framework and robustness will be determined by its structure.
- From data are computed indicators (objective functions). We assume that a choice of indicators is an intention to translate particular aspects of the system, i.e. to capture a realization of an "urban fact" (*fait urbain*) in the sense of MANGIN **[mangin1999projet]** - a sort of stylized fact in terms of processes and mechanisms, having various realizations on spatially distinct systems, depending on each precise context.
- Given many systems and associated indicators, a common space can be built to compare them. In that space, data represents more or less well real systems, depending e.g. on initial scale, precision of data, missing data. We precisely propose to capture that through the notion of point cloud discrepancy, which is a mathematical tool coming from sampling theory expressing how a dataset is distributed in the space it is embedded in **[dick2010digital]**.

Synthesizing these requirements, we propose a notion of *Robustness* of an evaluation that captures both, by combining data reliability with relative importance,

1. *Missing Data* : an evaluation based on more refined datasets will naturally be more robust.
2. *Indicator importance* : indicators with more relative influence will weight more on the total robustness.

10.2.2 *Formal Description*

INDICATORS Let $(S_i)_{1 \leqslant i \leqslant N}$ be a finite number of geographically disjoints territorial systems, that we assume described through raw data and intermediate indicators, yielding $S_i = (\mathbf{X}_i, \mathbf{Y}_i) \in \mathcal{X}_i \times \mathcal{Y}_i$ with $\mathcal{X}_i = \prod_k \mathcal{X}_{i,k}$ such that each subspace contain real matrices :

$\mathcal{X}_{i,k} = \mathbb{R}^{n^X_{i,k} p^X_{i,k}}$ (the same holding for $\mathcal{Y}_i$). We also define an ontological index function $I_X(i,k)$ (resp. $I_Y(i,k)$) taking integer values which coincide if and only if the two variables have the same ontology in the sense of [**livet2010**], i.e. they are supposed to represent the same real object. We distinguish "raw data" $\mathbf{X}_i$ from which indicators are computed via explicit deterministic functions, from "intermediate indicators" $\mathbf{Y}_i$ that are already integrated and can be e.g. outputs of elaborated models simulating some aspects of the urban system. We define the partial characteristic space of the "urban fact" by

$$(\mathcal{X}, \mathcal{Y}) \underset{\text{def}}{=} \left(\prod \tilde{\mathcal{X}}_c\right) \times \left(\prod \tilde{\mathcal{Y}}_c\right) = \left(\prod_{\mathcal{X}_{i,k} \in \mathcal{D}_\mathcal{X}} \mathbb{R}^{p^X_{i,k}}\right) \times \left(\prod_{\mathcal{Y}_{i,k} \in \mathcal{D}_\mathcal{Y}} \mathbb{R}^{p^Y_{i,k}}\right) \tag{10}$$

with $\mathcal{D}_\mathcal{X} = \{\mathcal{X}_{i,k} | I(i,k) \text{ distincts}, n^X_{i,k} \text{ maximal}\}$ (the same holding for $\mathcal{Y}_i$). It is indeed the abstract space on which indicators are integrated. The indices c introduced as a definition here correspond to different indicators across all systems. This space is the minimal space common to all systems allowing a common definition for indicators on each.

Let $\mathbf{X}_{i,c}$ be the data canonically projected in the corresponding subspace, well defined for all i and all c. We make the key assumption that all indicators are computed by integration against a certain kernel, i.e. that for all c, there exists H_c space of real-valued functions on $(\tilde{\mathcal{X}}_c, \tilde{\mathcal{Y}}_c)$, such that for all $h \in H_c$:

1. h is "enough" regular (tempered distributions e.g.)
2. $q_c = \int_{(\tilde{\mathcal{X}}_c, \tilde{\mathcal{Y}}_c)} h$ is a function describing the "urban fact" (the indicator in itself)

Typical concrete example of kernels can be :

- A mean of rows of $\mathbf{X}_{i,c}$ is computed with $h(x) = x \cdot f_{i,c}(x)$ where $f_{i,c}$ is the density of the distribution of the assumed underlying variable.
- A rate of elements respecting a given condition C, $h(x) = f_{i,c}(x)\chi_{C(x)}$
- For already aggregated variables $\mathbf{Y}$, a Dirac distribution allows to express them also as a kernel integral.

AGGREGATION Weighting objectives in multi-attribute decision-making is indeed the crucial point of the processes, and numerous methods are available (see [**wang2009review**] for a review for the particular case of sustainable energy management). Let define weights for the linear aggregation. We assume the indicators normalized, i.e. $q_c \in$

$[0,1]$, for a more simple construction of relative weights. For i, c and $h_c \in H_c$ given, the weight $w_{i,c}$ is simply constituted by the relative importance of the indicator $w^L_{i,c} = \frac{\hat{q}_{i,c}}{\sum_c \hat{q}_{i,c}}$ where $\hat{q}_{i,c}$ is an estimator of q_c for data $\mathbf{X}_{i,c}$ (i.e. the effectively calculated value). Note that this step can be extended to any sets of weight attributions, by taking for example $\tilde{w}_{i,c} = w_{i,c} \cdot w'_{i,c}$ if $\mathbf{w}'$ are the weights attributed by the decision-maker. We focus here on the relative influence of attributes and thus choose this simple form for weights.

ROBUSTNESS ESTIMATION The scene is now set up to be able to estimate the robustness of the evaluation done through the aggregated function. Therefore, we apply an integral approximation method similar to methods introduced in [**varet2010developpement**], since the integrated form of indicators indeed brings the benefits of such powerful theoretical results. Let $\mathbf{X}_{i,c} = (\vec{X}_{i,c,l})_{1 \leqslant l \leqslant n_{i,c}}$ and $D_{i,c} = \mathrm{Disc}_{\tilde{\mathfrak{X}}_c, L^2}(\mathbf{X}_{i,c})$ the discrepancy of data points cloud[2] [**niederreiter1972discrepancy**]. With $h \in H_c$, we have the upper bound on the integral approximation error

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

where K is a constant independent of data points and objective function. It directly yields

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

Assuming the error reasonably realized ("worst case" scenario for knowledge of the theoretical value of aggregated function), we take this upper bound as an approximation of its magnitude. Furthermore, taking normalized indicators implies $|||h_c||| = 1$. We propose then to compare error bounds between two evaluations. They depend only on data distribution (equivalent to *statistical robustness*) and on indicators chosen (sort of *ontological robustness*, i.e. do the indicators have a real sense in the chosen context and do their values make sense), and are a way to combine these two type of robustnesses into a single value.

2 The discrepancy is defined as the L2-norm of local discrepancy which is for normalized data points $\mathbf{X} = (x_{ij}) \in [0,1]^d$, a function of $\mathbf{t} \in [0,1]^d$ comparing the number of points falling in the corresponding hypercube with its volume, by $\mathrm{disc}(\mathbf{t}) = \frac{1}{n} \sum_i \mathbb{1}_{\prod_j x_{ij} < t_j} - \prod_j t_j$. It is a measure of how the point cloud covers the space.

We thus define a *robustness ratio* to compare the robustness of two evaluations by

$$R_{i,i'} = \frac{\sum_c w_{i,c} \cdot D_{i,c}}{\sum_c w_{i',c} \cdot D_{i',c}} \tag{11}$$

The intuitive sense of this definition is that one compares robustness of evaluations by comparing the highest error done in each based on data structure and relative importance.

By taking then an order relation on evaluations by comparing the position of the ratio to one, it is obvious that we obtain a complete order on all possible evaluations. This ratio should theoretically allow to compare any evaluation of an urban system. To keep an ontological sense to it, it should be used to compare disjoints sub-systems with a reasonable proportion of indicators in common, or the same subsystem with varying indicators. Note that it provides a way to test the influence of indicators on an evaluation by analyzing the sensitivity if the ratio to their removal. On the contrary, finding a "minimal" number of indicators each making the ratio strongly vary should be a way to isolate essential parameters ruling the sub-system.

10.3 RESULTS

IMPLEMENTATION Preprocessing of geographical data is made through QGIS [**qgis2011quantum**] for performance reasons. Core implementation of the framework is done in R [**team2000r**] for the flexibility of data management and statistical computations. Furthermore, the package `DiceDesign` [**franco20092**] written for numerical experiments and sampling purposes, allows an efficient and direct computation of discrepancies. Last but not least, all source code is openly available on the `git` repository of the project[3] for reproducibility purposes [**ram2013git**].

10.3.1 *Implementation on Synthetic Data*

We propose in a first time to illustrate the implementation with an application to synthetic data and indicators, for intra-urban quality indicators in the city of Paris.

DATA COLLECTION We base our virtual case on real geographical data, in particular for *arrondissements* of Paris. We use open data available through the OpenStreetMap project [**bennett20100penstreetmap**] that provides accurate high definition data for many urban features. We use the street network and position of buildings within the city

3 at https://github.com/JusteRaimbault/RobustnessDiscrepancy

of Paris. Limits of *arrondissements*, used to overlay and extract features when working on single districts, are also extracted from the same source. We use centroids of buildings polygons, and segments of street network. Dataset overall consists of around 200k building features and 100k road segments.

VIRTUAL CASES We work on each district of Paris (from the 1st to the 20th) as an evaluated urban system. We construct random synthetic data associated to spatial features, so each district has to be evaluated many time to obtain mean statistical behavior of toy indicators and robustness ratios. The indicators chosen need to be computed on residential and street network spatial data. We implement two mean kernels and a conditional mean to show different examples, linked to environmental sustainability and quality of life, that are required to be maximized. Note that these indicators have a real meaning but no particular reason to be aggregated, they are chosen here for the convenience of the toy model and the generation of synthetic data. With $a \in \{1 \dots 20\}$ the number of the district, $A(a)$ corresponding spatial extent, $b \in B$ building coordinates and $s \in S$ street segments, we take

- Complementary of the average daily distance to work with car per individual, approximated by, with $n_{cars}(b)$ number of cars in the building (randomly generated by associated of cars to a number of building proportional to motorization rate α_m 0.4 in Paris), d_w distance to work of individuals (generated from the building to a uniformly generated random point in spatial extent of the dataset), and d_{max} the diameter of Paris area, $\bar{d}_w = 1 - \frac{1}{|b \in A(a)|} \cdot \sum_{b \in A(a)} n_{cars}(b) \cdot \frac{d_w}{d_{max}}$
- Complementary of average car flows within the streets in the district, approximated by, with $\varphi(s)$ relative flow in street segment s, generated through the minimum of 1 and a log-normal distribution adjusted to have 95% of mass smaller than 1 what mimics the hierarchical distribution of street use (corresponding to betweenness centrality), and $l(s)$ segment length, $\bar{\varphi} = 1 - \frac{1}{|s \in A(a)|} \cdot \sum_{s \in A(a)} \varphi(s) \cdot \frac{l(s)}{\max(l(s))}$
- Relative length of pedestrian streets $\bar{p}$, computed through a randomly uniformly generated dummy variable adjusted to have a fixed global proportion of segments that are pedestrian.

As synthetic data are stochastic, we run the computation for each district $N = 50$ times, what was a reasonable compromise between statistical convergence and time required for computation. Table 1 shows results (mean and standard deviations) of indicator values and robustness ratio computation. Obtained standard deviation confirm that this number of repetitions give consistent results. Indicators obtained through a fixed ratio show small variability what may a limit

Arrdt	$< \bar{d}_w > \pm\sigma(\bar{d}_w)$	$< \bar{\varphi} > \pm\sigma(\bar{\varphi})$	$< \bar{p} > \pm\sigma(\bar{p})$	$R_{i,1}$
1 th	0.731655 ± 0.041099	0.917462 ± 0.026637	0.191615 ± 0.052142	1.000000 ± 0.000000
2 th	0.723225 ± 0.032539	0.844350 ± 0.036085	0.209467 ± 0.058675	1.002098 ± 0.039972
3 th	0.713716 ± 0.044789	0.797313 ± 0.057480	0.185541 ± 0.065089	0.999341 ± 0.048825
4 th	0.712394 ± 0.042897	0.861635 ± 0.030859	0.201236 ± 0.044395	0.973045 ± 0.036993
5 th	0.715557 ± 0.026328	0.894675 ± 0.020730	0.209965 ± 0.050093	0.963466 ± 0.040722
6 th	0.733249 ± 0.026890	0.875613 ± 0.029169	0.206690 ± 0.054850	0.990676 ± 0.031666
7 th	0.719775 ± 0.029072	0.891861 ± 0.026695	0.209265 ± 0.041337	0.966103 ± 0.037132
8 th	0.713602 ± 0.034423	0.931776 ± 0.015356	0.208923 ± 0.036814	0.973975 ± 0.033809
9 th	0.712441 ± 0.027587	0.910817 ± 0.015915	0.202283 ± 0.049044	0.971889 ± 0.035381
10 th	0.713072 ± 0.028918	0.881710 ± 0.021668	0.210118 ± 0.040435	0.991036 ± 0.038942
11 th	0.682905 ± 0.034225	0.875217 ± 0.019678	0.203195 ± 0.047049	0.949828 ± 0.035122
12 th	0.646328 ± 0.039668	0.920086 ± 0.019238	0.198986 ± 0.023012	0.960192 ± 0.034854
13 th	0.697512 ± 0.025461	0.890253 ± 0.022778	0.201406 ± 0.030348	0.960534 ± 0.033730
14 th	0.703224 ± 0.019900	0.902898 ± 0.019830	0.205575 ± 0.038635	0.932755 ± 0.033616
15 th	0.692050 ± 0.027536	0.891654 ± 0.018239	0.200860 ± 0.024085	0.929006 ± 0.031675
16 th	0.654609 ± 0.028141	0.928181 ± 0.013477	0.202355 ± 0.017180	0.963143 ± 0.033232
17 th	0.683020 ± 0.025644	0.890392 ± 0.023586	0.198464 ± 0.033714	0.941025 ± 0.034951
18 th	0.699170 ± 0.025487	0.911382 ± 0.027290	0.188802 ± 0.036537	0.950874 ± 0.028669
19 th	0.655108 ± 0.031857	0.884214 ± 0.027816	0.209234 ± 0.032466	0.962966 ± 0.034187
20 th	0.637446 ± 0.032562	0.873755 ± 0.036792	0.196807 ± 0.026001	0.952410 ± 0.038702

TABLE 2 : Numerical results of simulation for each district with N = 50 repetitions. Each toy indicator value is given by mean on repetitions and associated standard deviation. Robustness ratio is computed relative to first district (arbitrary choice). A ratio smaller than 1 means that integral bound is smaller for upper district, i.e. that evaluation is more robust for this district. Because of the small size of first district, we expected a majority of district to give ratio smaller than 1, what is confirmed by results, even when adding standard deviations.

of this toy approach. However, we obtain the interesting result that a majority of districts give more robust evaluations than 1st district, what was expected because of the size and content of this district : it is indeed a small one with large administrative buildings, what means

less spatial elements and thus a less robust evaluation following our definition of the robustness.

10.3.2 *Application to a Real Case : Metropolitan Segregation*

The first example was aimed to show potentialities of the method but was purely synthetic, hence yielding no concrete conclusion nor implications for policy. We propose now to apply it to real data for the example of metropolitan segregation.

DATA We work on income data available for France at an intra-urban level (basic statistical units IRIS) for the year 2011 under the form of summary statistics (deciles if the area is populated enough to ensure anonymity), provided by INSEE[4]. Data are associated with geographical extent of statistical units, allowing computation of spatial analysis indicators.

INDICATORS We use here three indicators of segregation integrated on a geographical area. Let assume the area divided into covering units S_i for $1 \leqslant i \leqslant N$ with centroids (x_i, y_i). Each unit has characteristics of population P_i and median income X_i. We define spatial weights used to quantify strength of geographical interactions between units i, j, with d_{ij} euclidian distance between centroids : $w_{ij} = \frac{P_i P_j}{(\sum_k P_k)^2} \cdot \frac{1}{d_{ij}}$ if $i \neq i$ and $w_{ii} = 0$. The normalized indicators are the following

- Spatial autocorrelation Moran index, defined as weighted normalized covariance of median income by $\rho = \frac{N}{\sum_{ij} w_{ij}} \cdot \frac{\sum_{ij} w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}$
- Dissimilarity index (close to Moran but integrating local dissimilarities rather than correlations), given by $d = \frac{1}{\sum_{ij} w_{ij}} \sum_{ij} w_{ij} \left| \tilde{X}_i - \tilde{X}_j \right|$ with $\tilde{X}_i = \frac{X_i - \min(X_k)}{\max(X_k) - \min(X_k)}$
- Complementary of the entropy of income distribution that is a way to capture global inequalities $\varepsilon = 1 + \frac{1}{\log(N)} \sum_i \frac{X_i}{\sum_k X_k} \cdot \log\left(\frac{X_i}{\sum_k X_k}\right)$

Numerous measures of segregation with various meanings and at different scales are available, as for example at the level of the unit by comparison of empirical wage distribution with a theoretical null model [**louf2015patterns**]. The choice here is arbitrary in order to illustrate our method with a reasonable number of dimensions.

4 http://www.insee.fr

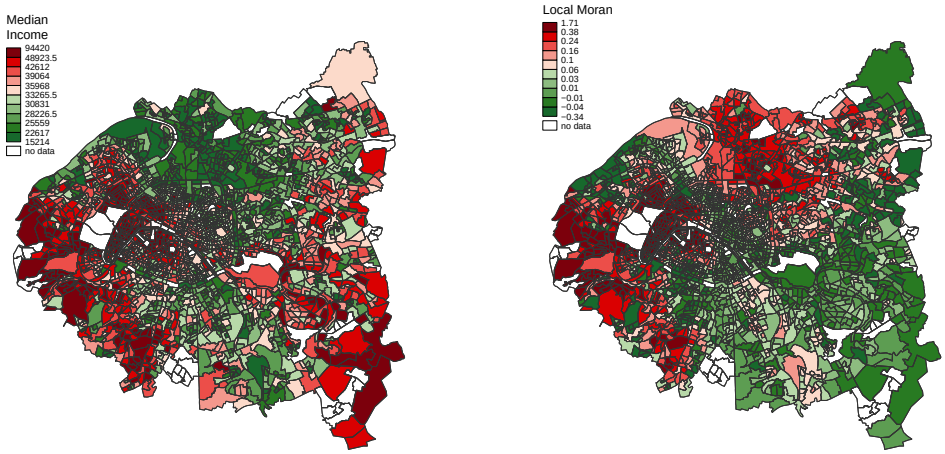


FIGURE 26 : **Maps of Metropolitan Segregation.** Maps show yearly median income on basic statistical units (IRIS) for the three departments constituting mainly the Great Paris metropolitan area, and the corresponding local Moran spatial autocorrelation index, defined for unit i as $\rho_i = N/\sum_j w_{ij} \cdot \frac{\sum_j w_{ij}(X_j - \bar{X})(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2}$. The most segregated areas coincide with the richest and the poorest, suggesting an increase of segregation in extreme situations.

RESULTS We apply our method with these indicators on the Greater Paris area, constituted of four *départements* that are intermediate administrative units. The recent creation of a new metropolitan governance system [**gilli2009paris**] underlines interrogations on its consistence, and in particular on its relation to intermediate spatial inequalities. We show in Fig. 1 maps of spatial distribution of median income and corresponding local index of autocorrelation. We observe the well-known West-East opposition and district disparities inside Paris as they were formulated in various studies, such as [**gueroisdynamique**] through the analysis of real estate transactions dynamics. We then apply our framework to answer a concrete question that has implications for urban policy : *how are the evaluation of segregation within different territories sensitive to missing data ?* To do so, we proceed to Monte Carlo simulations (75 repetitions) during which a fixed proportion of data is randomly removed, and the corresponding robustness index is evaluated with renormalized indicators. Simulations are done on each *department* separately, each time relatively to the robustness of the evaluation of full Greater Paris. Results are shown in Fig. 2. All areas present a slightly better robustness than the reference, what could be explained by local homogeneity and thus more fiable segregation values. Implications for policy that can be drawn are for example direct comparisons between areas : a loss of 30% of information on 93 area corresponds to a loss of only 25% in 92 area. The

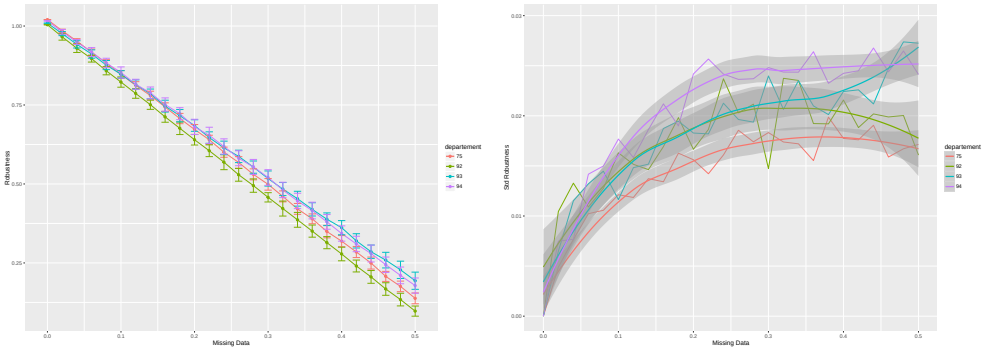


FIGURE 27 : **Sensitivity of robustness to missing data.** *Left.* For each department, Monte Carlo simulations (N=75 repetitions) are used to determine the impact of missing data on robustness of segregation evaluation. Robustness ratios are all computed relatively to full metropolitan area with all available data. Quasi-linear behavior translates an approximative linear decrease of discrepancy as a function of data size. The similar trajectory of poorest departments (93,94) suggest the correction to linear behavior being driven be segregation patterns. *Right.* Corresponding standard deviations of robustness ratios. Different regimes (in particular 93 against others) unveil phase transitions at different levels of missing data, meaning that the evaluation in 94 is from this point of view more sensitive to missing data.

first being a deprived area, the inequality is increased by this relative lower quality of statistical information. The study of standard deviations suggest further investigations as different response regimes to data removal seem to exist.

10.4 DISCUSSION

10.4.1 *Applicability to Real situations*

IMPLICATIONS FOR DECISION-MAKING The application of our method to concrete decision-making can be thought in different ways. First in the case of a comparative multi-attribute decision process, such as the determination of a transportation corridor, the identification of territories on which the evaluation may be flawed (i.e. has a poor relative robustness) could allow a more refined focus on these and a corresponding revision of datasets or an adapted revision of weights. In any case the overall decision-making process should be made more reliable. A second direction lays in the spirit of the real application we have proposed, i.e. the sensitivity of evaluation to various parameters such as missing data. If a decision appears as reliable because data have few missing points, but the evaluation is very sensitive to it, one will be more careful in the interpretation of results and taking the final decision. Further work and testing will

however be needed to understand framework behavior in different contexts and be able to pilot its application in various real situations.

INTEGRATION WITHIN EXISTING FRAMEWORKS The applicability of the method on real cases will directly depend on its potential integration within existing framework. Beyond technical difficulties that will surely appear when trying to couple or integrate implementations, more theoretical obstacles could occur, such as fuzzy formulations of functions or data types, consistency issues in databases, etc. Such multi-criteria framework are numerous. Further interesting work would be to attempt integration into an open one, such as e.g. the one described in [**tivadar2014oasis**] which calculates various indices of urban segregation, as we have already illustrated the application on metropolitan segregation indexes.

AVAILABILITY OF RAW DATA In general, sensitive data such as transportation questionnaires, or very fine granularity census data are not openly available but provided already aggregated at a certain level (for instance French Insee Data are publicly available at basic statistical unit level or larger areas depending on variables and minimal population constraints, more precise data is under restricted access). It means that applying the framework may imply complicated data research procedure, its advantage to be flexible being thus reduced through additional constraints.

10.4.2 *Validity of Theoretical Assumptions*

A possible limitation of our approach is the validity of the assumption formulating indicators as spatial integrals. Indeed, many socio-economic indicators are not necessarily depending explicitly on space, and trying to associate them with spatial coordinates may become a slippery slope (e.g. associate individual economic variables with individual residential coordinates will have a sense only if the use of the variable has a relation with space, otherwise it is a non-legitimate artifact). Even indicators which have a spatial value may derive from non-spatial variables, as [**kwan1998space**] points out concerning accessibility, when opposing integrated accessibility measures with individual-based non necessarily spatial-based (e.g. individual decisions) measures. Constraining a theoretical representation of a system to fit a framework by changing some of its ontological properties (always in the sense of real meaning of objects) can be understood as a violation of a fundamental rule of modeling and simulation in social science given in [**banos2013HDR**], that is that there can be an universal "language" for modeling and some can not express some systems, having for consequence misleading conclusion due to ontology breaking in the case of an over-constrained formulation.

10.4.3 *Framework Generality*

We argue that the fundamental advantage of the proposed framework is its generality and flexibility, since robustness of the evaluations are obtained only through data structure if ones relaxes constraints on the value of weight. Further work should go towards a more general formulation, suppressing for example the linear aggregation assumption. Non-linear aggregation functions would require however to present particular properties regarding integral inequalities. For example, similar results could search in the direction of integral inequalities for Lipschitzian functions such as the one-dimensional results of [**dragomir1999ostrowski**].

CONCLUSION

We have proposed a model-independent framework to compare the robustness of multi-attribute evaluations between different urban systems. Based on data discrepancy, it provide a general definition of relative robustness without any assumption on model for the system, but with limiting assumptions that are the need of linear aggregation and of indicators being expressed through spatial kernel integrals. We propose a toy implementation based on real data for the city of Paris, numerical results confirming general expected behavior, and an implementation on real data for income segregation on Greater Paris metropolitan areas, giving possible insights into concrete policy questions. Further work should be oriented towards sensitivity analysis of the method, application to other real cases and theoretical assumptions relaxation, i.e. the relaxation of linear aggregation and spatial integration.

ACKNOWLEDGMENTS

The author would like to thank Julien Keutchayan (Ecole Polytechnique de Montréal) for suggesting the original idea of using discrepancy, and anonymous reviewers for the useful comments and insights.

Quatrième partie

APPENDIX

11

GENERATION OF CORRELATED SYNTHETIC DATA

12 ARCHITECTURE AND SOURCES FOR ALGORITHMS AND MODELS OF SIMULATION

You must not be afraid of putting code in your thesis, code is not dirty
- Alexis Drogoul

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.

12.1 ALGORITHMIC SYSTEMATIC REVIEW

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 cortext currently being replaced by Python scripts.

ARCHITECTURE Classical object oriented, see code.

ADDITIONAL SCRIPTS R for result exploration and visualization.

12.2 INDIRECT BIBLIOMETRICS

OBJECTIVE Hypernetworks analysis of cybergeo journal.

LOCATION CLOSED (shared repository).
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 -

12.3 DENSITY URBAN GROWTH

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.

12.4 CORRELATED DATA GENERATION

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.

12.5 LUTECIA MODEL

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.

12.6 NETWORK ANALYSIS

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 manoeuvers.

ADDITIONAL SCRIPTS -

13 TOOLS AND WORKFLOW FOR AN OPEN REPRODUCIBLE RESEARCH

Open for Discovery
- PLoS

We briefly evoke here tools or workflows currently under development or testing, aimed at easing an open reproducible research and making it more transparent.

13.1 NETLOGO DOCUMENTATION GENERATOR

Documentation generation is central for reproducibility as it can automatize implementation description. NetLogo does not provide a documentation generator and we 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`.

13.2 GIT AS A REPRODUCIBILITY TOOL

The use if `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.

13.3 GIT-DATA

`git-data` is a shell based (experimental) git extension, available at `https://github.com/JusteRaimbault/gitdata`, that allows automatized 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.

13.4 TOWARDS A GIT-COMPATIBLE FIGURES METADATA HANDLER

The issue of meta-data for figures is a crucial issue, as it is often difficult to keep a trace of all parameter values that have generated it, along with the corresponding code. Tricks may furthermore happen in script environments such as R or python when variables are accidentally modified without code modification. Keeping an exhaustive trace of the exact dataset, code and history that has generated a precise figure is a necessary condition for exact reproducibility. We are elaborating a git-compatible tool that would automatically handle these metadata, for example by branching and associating the unique commit hash to the figure. 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 reproducting environment.

13.5 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.