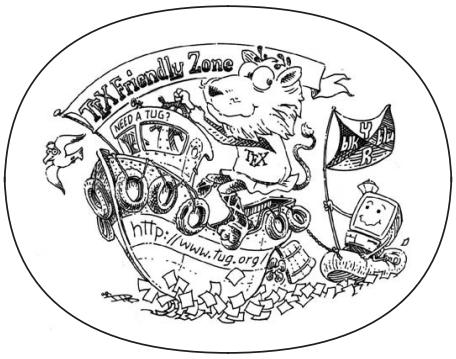
TOWARDS MODELS COUPLING URBAN GROWTH AND TRANSPORTATION NETWORK GROWTH

JUSTE RAIMBAULT



An Homage to The Elements of Typographic Style

February 2016 – version 0.1

ABSTRACT

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRONYMS

INTRODUCTION

SCIENTIFIC CONTEXT

INTERDISCIPLINARITY

COMPLEXITY HAS COME OF AGE Beyond "fashionable" positions that can be the consequence of a blind following [30], or more ambivalent, of a marketing strategy as the fight for funds is becoming a huge obstacle for research [**bollen2014funding**], Science of Complexity is taking a hole new place in the academic landscape. As an informal mix of epistemological positions, methods, and fields of applications, it relies on *unconventionnal* paradigms such as the centrality of emergence and self-organization in most of phenomena of the real world, which make it lie on a frontier of knowledge closer of us than we can think, as Laughlin develops in [48]. *Detail concepts ?*. Such concepts are indeed not new, as they were already enlighted by Anderson [**anderson1972more**]. Even cybernetics can be related to complexity by seing it as a bridge between technics and cognitive science [**wiener1948cybernetics**]. Later, synergetics [**haken1980synergetics**] paved the way for a theoretical approach of collective phenomena in physics. 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 [39], finer reflexions around the notion of model [**varenne2013modeliser**] [note : beware of a chicken-egg type problem on the relation between scientific and epistemological progress]. The theoretical and empirical potentialities of such approach play surely a role in their success, as confirmed in various domains of application (see [59] for a general survey), as for example Network Science [7] ; Neuroscience [47] ; Social Sciences ; Geography [**manson2001simplifying**][65] ; Finance with the rising importance of econophysics [83].

CONFLICTING COMPLEXITIES AND CULTURAL DIFFERENCES Yet this scientific evolution that some see as a revolution [**colander2003complexity**], or even as *a new kind of science* [91], could 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 outcoming conflicts on views and approaches. We shall now develop some concrete example that lead to such considerations when encoutered. They are of many different natures and concern different disciplines, such that it would not be honnest to assume that the issue is not general. Each come from personnal research experience

Physics reinvents geography.
Economic Geography or Geographical Economics ?
Agent-based Modeling in Economy
Finance

The drama of scientific misunderstandings is that they can indeed anhiliate 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 [2], which conclusions are presented as opposed from the one obtained by Piketty [63], follows such a scheme.

KEEP IT COMPLEX, STUPID

MEANS ARE HERE, LET USE THEM

GEOGRAPHICAL OBJECTS

Part I

THEMATIC, METHODOLOGICAL AND THEORETICAL FOUNDATIONS

1

INTERACTIONS BETWEEN NETWORK AND TERRITORY

2 QUANTITATIVE EPISTEMOLOGY

citation

INTRODUCTION

2.1 ALGORITHMIC SYSTEMATIC REVIEW

2.1.1 *In search of models of co-evolution*

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 .

Transportation networks and urban land-use are known to be strongly coupled components of urban systems at different scales [17]. One common approach is to consider them as co-evolving, avoiding misleading interpretations such as the myth of structural effect of transportation infrastructures [60]. 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.

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

2.1.2.1 *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 [19], [44] and [87] 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 [29] which includes housing market for Paris area. Diverse aspects of the same system can be translated into many models (as e. g.[88]), and traffic, residential and employment dynamics, resulting land-use evolution, influenced also by a static transportation network, are generally taken into account.

2.1.2.2 *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 [96]. In [93], economic empirical studies are positioned within other network growth approaches, such as work by physicists proposing model of geometrical network growth [8]. Analogy with biological networks was also done, reproducing typical robustness properties of transportation networks [85].

2.1.2.3 *Hybrid Approaches*

Fewer approaches coupling urban growth and network growth can be found in the literature. [9] couples density evolution with network growth in a toy model. In [73], a simple Cellular Automaton coupled with an evolutive network reproduces stylized facts of human settlements described by Le Corbusier. At a smaller scale, [1] proposes a model of co-evolution between roads and buildings, following geometrical rules. These approaches stay however limited and rare.

2.1.3 *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 [78]. Ignoring some references can even be considered as a scientific mistake in the context of emerging information systems [52]. 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 [24] 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 [58] or co-autorship networks [80], we propose to use a less explored paradigm based on text-mining introduced in [20], that obtain a dynamic mapping of scientific disciplines based on their semantic content. For our question, it has a particular interest, as we want to understand content structure of researches on the subject. We propose to apply an algorithmic method described in the following. The algorithm proceeds by iterations to obtain a stabilized corpus from initial keywords, reconstructing scientific semantic landscape around a particular subject.

2.1.3.1 *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.

2.1.3.2 *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 [57] 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 [20], calling a Python script that uses [14].

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.

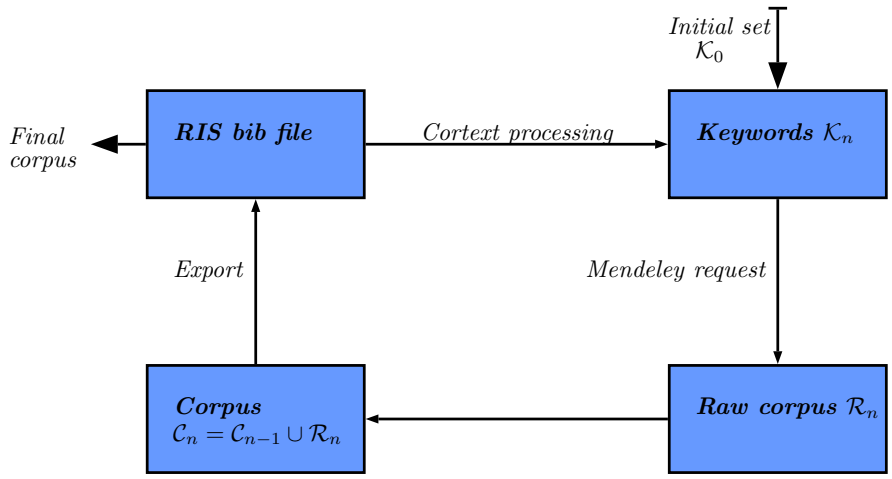


Figure 1: Global workflow of the algorithm, including implementation details : catalog request is done through Mendeley API ; final state of corpuses are RIS files.

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

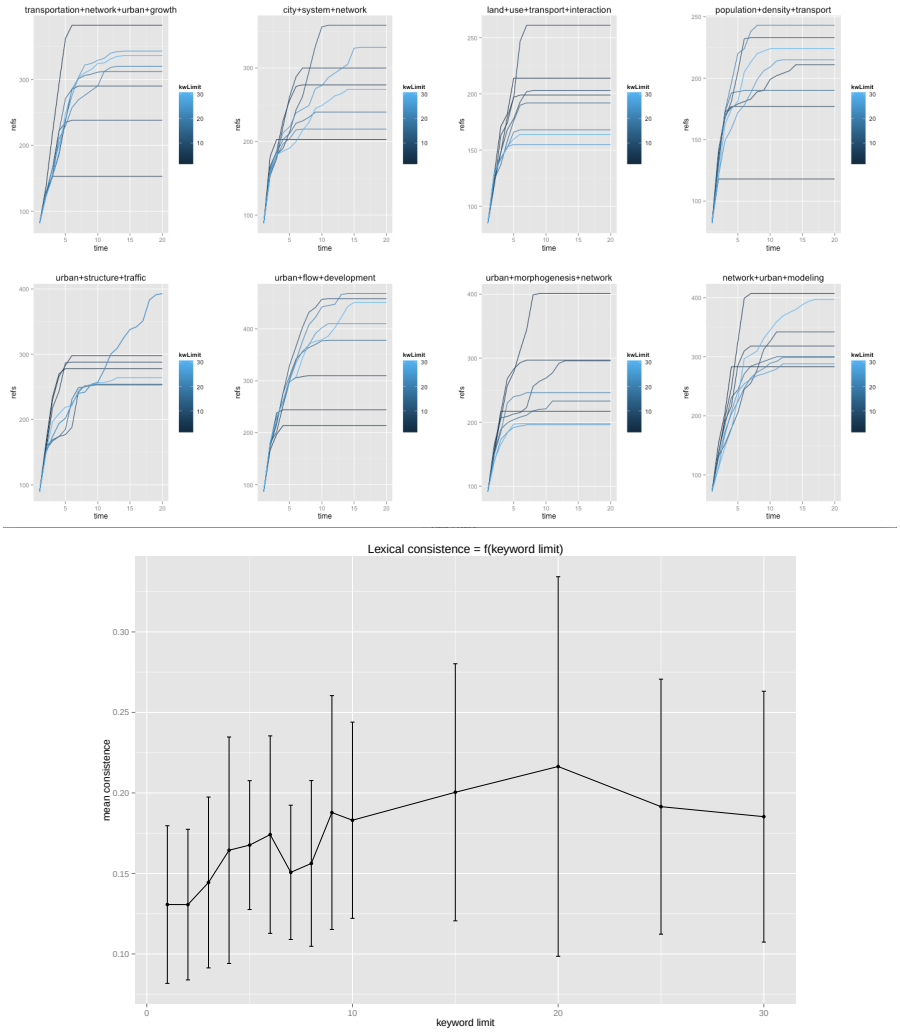


Figure 2: Convergence and sensitivity analysis. Left : Plots of number of references as a function of iteration, for various queries linked to our theme (see further), for various values of N_k (from 2 to 30). We obtain a rapid convergence for most cases, around 10 iterations needed. Final number of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

communication between scientific disciplines studying different aspects of that problems. We have proposed an algorithmic method to give elements of answers through text-mining-based corpus extraction. First numerical results seem to confirm the assumption. However, such a quantitative analysis should not be considered alone, but rather come as a back-up for qualitative studies that will be the object of further work, such as the one lead in [24], in which questionnaires with historical actors of modeling provide highly relevant information.

2.2 REFINING BIBLIOMETRICAL ANALYSIS THROUGH HYPERNETWORK ANALYSIS

2.3 TOWARDS MODELING PURPOSE AND CONTEXT AUTOMATIC EXTRACTION

3 METHODOLOGICAL DEVELOPMENTS

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 [84] 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` [74], or to integrate document creation and data analysis such as `knitr` [94]), 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 [32].

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 [90]. The model is also available online [28] and is pre-

sented 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 [72]. In the frame of that study, source code was modified and commented, which last version is available on the repository of the project[1].

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 [6], 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

1 at `https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Reproduction/UrbanSuite`

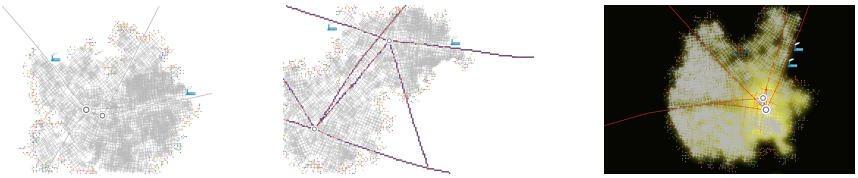


Figure 3: Example of simple improvement in visualization that can help understanding mechanisms implied in the model. *Left :* example of original output ; *Middle :* visualization of main roads (in red) and underlying patches attribution, suggesting possible implementation bias in the use of discretized 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 [39], 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. 3.

3.1.2 *On the Need of Exactitude in Model Implementation*

3.2 AN UNIFIED FRAMEWORK FOR STOCHASTIC MODELS 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 litterature, from economics to geography, including models proposed by physicists. We propose 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.

3.2.1.1 *Context*

General biblio.

Precise type of models : mathematical models ; stay to a certain level of tractability as essence of our approach is link between models. No clear definition, includes all models that can be linked in the sense of *Generalization/Particularization/Limit case/?*.

3.2.1.2 *Notations*

3.2.2 *Framework*

3.2.2.1 *Formulation*

PRESENTATION What we propose as a framework can be understood as a meta-model in the sense of [26], 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 shoud 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 [26]). It constructs a sort of model classification.

3.2.2.2 *Models Included*

The following models are included in our framework. The list is arbitrary but aims to offer a broad view of disciplines concerned

3.2.2.3 *Thematic Classification*

3.2.2.4 *Framework Formulation*

Diagram linking various models ; first embedded into time/population plane, cases Discrete/Continous. Other aspects more sparse (ex. spatialization) ; how represent it ?

3.2.3 *Models formulation*

3.2.4 *Derivations*

3.2.4.1 *Generalization of Preferential Attachment*

See [95].

3.2.4.2 *Link between Gibrat and Preferential Attachment Models*

Let consider a stricly 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 $\bar{P}_i(t) = \mathbb{E}[P_i(t)]$. Independance 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

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

3.2.4.3 *Link between Simon and Preferential Attachment*

3.2.4.4 *Link between Favaro-Pumain and Gibrat*

[35]

3.2.4.5 *Link between Bettencourt-West and Simon*

[13]

3.2.4.6 *Other Models*

[37] : Economic model giving a Simon equivalent formulation. Finds that in upper tail, proportional growth process occurs. We find the same result as a consequence of 3.2.4.3.

3.2.5 *Application*

3.2.6 *Discussion*

Conclusion

3.3 ANALYTICAL SENSITIVITY OF URBAN SCALING LAWS TO SPATIAL EXTENT

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 infralinear to supralinear 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 independant 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 [37], have been shown to be a recurrent feature of urban systems, at many scales and for many types of indicators. They reside in the empirical constatation that indicators computed on elements of an urban system, that can be cities for system of cities, but also smaller entities at a smaller scale, do fit relatively well a power-law distribution as a function of entity size, i.e. that for entity i with population P_i, we have for an integrated quantity A_i, the relation $A_i \simeq A_0 \cdot \left(\frac{P_i}{P_0}\right)^{\alpha}$. Scaling exponent α can be smaller or greater than 1, leading to infra- or supralinear effects. Various thematic interpretation of this phenomena have been proposed, typically under the form of processes analysis. The economic literature has produced abundant work on the subject (see [38] for a review), but that are generally poorly spatialized, 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 [13] but are difficult to fit empirically. A interesting proposition by Pumain is that they are intrinsically due to the evolutionnary character of city systems, where complex emergent interaction between cities generate such global distributions [69]. 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 [68]. Other explanations have been proposed at other scales, such as the urban growth model at the mesoscopic scale (city scale) given in [54] 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 agaents 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 [35], 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 [55], confirmed by empirical works such as [4] 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 [25],

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 infralinear to supralinear 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. The rest of the paper is organized as follows : we formalize the simple framework used and derive an analytical relation between estimated exponent and density threshold parameter. We then present a numerical implementation of the model that confirms numerically theoretical results, explore other form of kernels that would be less tractable, and study the sensitivity along two parameters. We finally discuss the implications of our results and further work needed.

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

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

where r_i are spread parameters of kernels, d_i^0 densities at origins, $\vec{x}_i$ positions of centers. We furthemore 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 [25], 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 [46] and that it implies necessarily a power-law distribution [21]. 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

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

We obtain in a similar way the expression of $P_i(\theta)$

$$P_i(\theta) = 2\pi d_i^0 r_0^2\left[1 + \ln\left[\frac{\theta}{d_i^0}\right]\frac{\theta}{d_i^0} - \frac{\theta}{d_i^0}\right]$$

The Ordinary-Least-Square estimation, solving the problem $\inf_{\alpha,C} \|(\ln A_i(\theta) - C - \alpha \ln P_i(\theta))_i\|^2$, gives the value $\alpha(\theta) = \frac{\mathrm{Cov}[(\ln A_i(\theta))_i,(\ln P_i(\theta))_i]}{\mathrm{Var}[(\ln P_i(\theta))_i]}$. 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 $\ln P_i(\theta) = K_P - \alpha \ln i + \ln \left[\frac{\theta}{d_i^0}\right] \frac{\theta}{d_i^0}$. Developping the covariance and variance gives finally an expression of the scaling exponent as a function of θ, where $k_j, k_j{}'$ are constants obtained in the development :

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

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 [70] and using the package `kernlab` for efficient kernel mixture methods [45] are given as comments in source code[2]. We show in figure 4 example of synthetic density distributions on which the numerical study is conducted. Theoretical result obtained in Eq. 2 are studied and confronted to emprically computed values for various parameter as shown in Fig. 5.

RANDOM PERTURBATIONS The simple model used is quite reducing for maximal densities and radius distribution. We 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

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

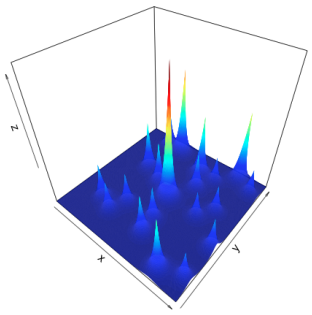


Figure 4: Example of a synthetic density distribution obtained with the exponential mixture, with a grid of size 400×400 and parameters $N = 20$, $r_0 = 10$, $P_{max} = 200$, $\alpha = 0.5$, $\theta_C = 0.01$.

Figure 5: Validation of theoretical result through numerical simulation.

- Radiuses are computed to respect rank-size law and then perturbated by a white noise.

Results shown in Fig. 6 are quantitatively different from previous one, as expected, but the same qualitative behavior is reproduced.

KERNEL TYPE We test the influence of the type of spatial kernel used on results. We test gaussian kernels and quadratic kernels with parameters within reasonable ranges analog to the exponential kernel. As shown in Fig. 7, we obtain the same qualitative results that is the significant variation of $\alpha(\theta)$ as a function of θ.

TWO-PARAMETERS PHASE DIAGRAM We introduce now a second spatial variable that has also an influence on the definition of urban entities, that is the proportion of actives working in city center, as done on empirical data in [25]. To simplify, it is used only to define urban parameter but assumed as having no influence on the local probability distribution of the amenity which stays the same function of the density. We write

3.3.5 *Discussion*

3.4 STATISTICAL CONTROL ON INITIAL CONDITIONS BY SYNTHETIC DATA GENERATION

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"

Figure 6: Variation of exponents with variable origin density and radius.

Figure 7: Scaling exponents for other kernels.

Figure 8: Two parameters phase diagram.

(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.

Let illustrate the issue with a simple example.

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.

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 [6]. 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 [81]. 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 (*Biblio Marion*). The thematic aim of the project developed in [27] is to clarify this point through a systematic model exploration. A methodological contribution is the construction of a framework allowing the analysis of the sensitivity of models to *meta-parameters*, i.e. to parameters considered as fixed initial conditions (e.g. the spatial structure for the Schelling model), or to parameters of another model generating an initial configuration, as detailed in Fig. **??** *[insert scheme describing the approach]*, where we have 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 in next section, 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 [86] , coupling is simple and no complexity is thus added. – develop

CONTEXT Let M_m a stochastic model of simulation, which inputs are to simplify initial conditions D_0 and parameters $\vec{\alpha}$, and output $M_m[\vec{\alpha}, D_0](t)$ at a given time t. We assume that it is partially data-driven in the sense that D_0 is supposed to represent a real situation at a given time, and model performance is measured by the distance of its output at final time to the real situation at the corresponding time, i.e. error function is of the form $\| \mathbb{E}[\vec{g}(M_m[\vec{\alpha}, D_0](t_f))] - \vec{g}(D_f) \|$ where $\vec{g}$ is a deterministic field corresponding to given indicators.

POSITION OF THE PROBLEM Evaluating the model on real data is rapidly limited in control possibilities, being restricted to the search of datasets allowing natural control groups. Furthermore, statistical behaviors are generally poorly characterized because of the small number of realizations. Working with synthetic data first allows to solve this issue of robustness of statistics, and then gives possibilities of control on some "meta-parameters" in the sense described before.

3.4.1 *Formal Analysis*

3.4.1.1 *Deterministic Formulation*

One has

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

→ 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.

3.4.1.2 *Stochasticity*

Dealing with stochasticity in simply coupled models → no convergence pb as $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$

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 [31]

The capture of neighborhood effects in statistical models

4

THEORETICAL FRAMEWORK

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

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 [43]. 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*[1] 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[2], 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 [66] 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 [33]. Complex System Science[3] is presented by some as a "new kind of Science" [91], and would at least be a symptom of a shift in scientific practices, from analytical and "exact" approaches to computational and evidence-based

1 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 [48].

2 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. [59], and for which study various complementary approaches may be applied, such as Dynamical Systems, Agent-based Modeling, Random Matrix Theory

3 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.

approaches [5], 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 [15].

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[4], 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 [55], but also by scientists coming from social sciences such as Batty [11]), 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 [53], 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.

4. the inclusion of the notion of emergence in order to capture multi-scale aspects of systems.

4 geographical systems in the sense of [31] 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.

5. a central place of ontology in the definition of systems, i.e. of the sense in the real world given to its objects[5].

6. taking into account heterogeneous aspects of the same system, that could be heterogeneous components but also complementary intersecting views.

The rest of the paper 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.

Construction of the theory

Perspectives and Ontologies

The starting point of the theory construction is a perspectivist epistemological approach on systems introduced by Giere [39]. 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 [86] 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 [42] that gives a convenient way to represent it and to introduce timescales, to which is associated an ontology O in the sense of [53], i.e. a set of elements each corresponds to a *thing* 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 1 *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* $O = (O_j)_j$.

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

5 *TODO : verify positioning regarding Structural Realism ; may be close to Ontic Structural Realism [36]*

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 2 *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. 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[6]. We follow on that point the approach of Bedau on different type of emergences, in particular his definition of weak emergence given in [12]. 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[7] :

- *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 1 *A system can be partially structured by extending it with an ontology that contains (not necessarily only) relations between elements of ontologies of its perspectives. We name it the* coupling ontology *and assume its existence in the following. We assume furthermore its atomicity, i.e. if* O *is in relation with* O'*, then any subsets of* O, O' *can not be in relation, what is not restrictive as a decomposition into several independent subsets ensures it if it is not the case.*

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

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

7 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.

Proposition 1 *The following binary relationships are pre-orders on* $\mathcal{P}(\mathcal{O})$ *:*

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

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 3 *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*[8] *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 4 *A consistent part of the ontological graph is a weakly connected component of the graph. We assume for now to work on a consistent part.*

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

Assumption 2 *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* $M' = \Phi^{<-1>}(O)$*.*

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

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 [42] should allow to identify basic components corresponding to each element of the ontology, and then construct the new perspective by induction. We find however these assumptions too restrictive, as for example various ontological elements may be modeled by an irreducible machine, as a differential equations with aggregated variables. We prefer to be less restrictive and postulate the existence of the reverse mapping on some sub-ontologies, that should be in practice the ones where couplings can be effectively modeled.

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

Proposition 2 *The tree decomposition of the ontological graph in which nodes contains strongly connected components is unique. The 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 [56], 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[9] $\|\cdot\|_S$ and that the state variable has a certain level of differentiability, the *thematic scale* if defined as $\|(d^k\vec{X}(t))_k\|_S$

Application

The particular case of geographical systems

TBW

[31] $\rightarrow$ FDD proposes in that paper 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 way to define relations, may be quick to extract ontological structure. *Note : find typical emergence clues in standard relational formalizations ? would guide the application of the theory.*

Example of urban systems ; Pumain's evolutionary theory [67]

Modularity and co-evolving subsystems

TBW

Paper on modularity in dynamical system : decomposition into uncorrelated subsystems : correlation between subsystems should be positively correlated with topological distance in the tree.

Define elements of a node before merging as *strongly coupled elements*. If they are dynamic, could be a definition of *co-evolution* and co-evolving subsystems.

Discussion

TBW

LINK WITH EXISTING FRAMEWORKS Link with Cottineau-Chapron framework for multi-modeling ? $\rightarrow$ not so far if they add biblio layer.

[76] proposes the notion of "interdisciplinary coupling" $\rightarrow$ close to our notion of coupling perspectives

Link with System of Systems ? not so far are our perspective are constructed as dataflow machines ; but significantly different as notion of emergence is central. *Find paper at CSDM 2015 on endogeneous system theory*

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

CONTRIBUTIONS TO THE STUDY OF COMPLEX SYSTEMS

- Not a theory of systems (beware of cybernetics, systemics etc), but more a framework to guide research questions ? [e.g. in our cases where it will be applied : quantitative epistemo 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.] - or maybe a meta-theory which application gives a theory ?
- Emphasis on the notion of socio-technical system : crosses a social complex system approach (ontologies) with a description of technical artifacts (dataflow machines) ; "best of both worlds" ?

Part II

MODELING AND EMPIRICAL ANALYSIS

5

EMPIRICAL ANALYSIS : INSIGHTS FROM STYLIZED FACTS

5.1 STATIC CORRELATIONS OF URBAN FORM AND NETWORK SHAPE FOR EUROPEAN TERRITORIAL SYSTEMS

5.1.1 *Morphological Measures of European Population Density*

5.1.2 *Network Measures*

5.1.3 *Effective static correlations*

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

5.2.1 *Context Formalization*

5.2.1.1 *Variables*

DESCRIPTION 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

DEFINITIONS

5.2.1.2 *Accessibility*

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

EXISTENCE OF ACCESSIBILITY An elegant axiomatic definition is derived in [89]. Starting from expected properties of an accessibility function A that associate a value to *attraction* a and distance d, defined on the set of discrete spatial configurations $\mathcal{C} = \cup_{n\in\mathbb{N}}(d_i, a_i)_{1\leqslant i\leqslant n}$.

These properties include (among technical others with no thematic meaning) :

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

A canonical decomposition of any accessibility function

CONTINUOUS APPROACH AND ACCESSIBILITY POTENTIAL

5.2.2 *Statistical Tests*

5.2.2.1 *Bivariate linear models*

5.2.2.2 *Autocorrelated univariate models*

5.2.2.3 *Autocorrelated multivariate models*

5.2.2.4 *Granger causality tests*

[92] use Granger causality to link transit with land-use changes.

5.2.2.5 *Autoregressive multivariate models*

5.2.2.6 *Autoregressive autocorrelated multivariate models*

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

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

6

MODELING

INTRODUCTION

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*

6.1.2 *Model Description*

6.1.3 *The urban growth model*

Preferential attchment and diffusion, extension of [Batty, 2006]. small litt. prefAtt and then Urban sprawl, opposed forces that must shape morphology of urban systems. choice of a scale at which spatial form has a sense but city system is ''wide enough" : 50-100km, meso-scale ?

Precise formulation, description and formalization of the model. ; parameters and their possible interpretation ; def of parameter space.

6.1.4 *Indicators*

6.1.5 *Results*

Precise description of the implementation (pub openmole exploration etc, importance of intensive computation)

6.1.6 *Generation of urban patterns*

variety of genered forms, examples of extreme shapes.

Figure : Example of the variety of generated urban shapes

3.2 - Model Behavior

3.2.1 - Convergence - internal model validation

convergence properties of indicators ; number of repetitions needed for consistence of results. [histograms and stats of σ

3.2.2 - Exploration of parameter space

Grid, then LHS explorations.

Figure : Scatterplots of indicators distribution in an hypercube of the parameter space. We show here the influence of one parameter (diffusion rate). Red points correspond to real data. 3D plots as supp material ?

Figure : on 2 first PC of morpho indicators, localization of typical shapes (monocentric city, polycentric city, diffuse rural settlements, aggregated rural settlements) / comparing generated shape with a typical real one

3.2.3 - Statistical analysis

Regression indics = f(params) . TO BE DONE. interpretation ?

6.1.7 *Model Calibration*

6.1.7.1 *Real Data*

Figure : Distributions of indicators values for real dataset of european densities, computed on 50kmx50km grids, with 10km offset to avoid bord effects. [TODO : add log-normal/normal fits ?]

3.3.2 - Calibration Process

Specific calib process : PCA that maximize the cumulated distance between generated points and reals points ; then select point cloud that overlaps real points in (PC1,PC2) plan, given a distance threshold.

Figure : 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.

6.1.7.2 *Calibration Results*

-> extraction of the exact parameter space covering all real situations. interpretation of its shape (correlations between parameters ?) and its volume in different directions (relative importance of parameters). [possible development : application of Calibration Profile algo

to check relative influence of parameters + ad hoc linear algebra on regression of 3.2.3 to do the same]

4 - Discussion

4.1 - Thematic interpreation of growth behavior

interpret positions of typical shapes within param space : confirms thematic interpretation of parameters. depending on results of 3.3.3, necessary and sufficient parameters to explain growth at this scale -> interpretation ?

4.2 - Integration into a multi-scale growth model

Possible coupling on a gibrat (or favaro-pumain) at europa scale (macro) (with addition of consistence on migration constraints), where meso growth rates which werer exogeneous before, are top-down determined, and bottom-up feedback is done through local aggregation level, that influence attractivity of an area.

-> Accurately calibrated spatialized urban growth model, can reproduce any (european) urban pattern. -> interpretation of parameter influence ; effective independant dimensions of the urban system at this scale.

6.2 CORRELATED GENERATION OF TERRITORIAL CONFIGURATIONS

6.2.1 *Application : geographical data of density and network*

6.2.1.1 *Context*

The use of synthetic data in geography is generally directed towards the generation of synthetic populations within agent-based models (mobility, *LUTI* models) [64]. 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 [18] 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 [79], 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 [82] 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 [27], 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.

6.2.1.2 *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 [61] but stays highly complex and difficult to quantify [60]. A dynamical modeling of implied processes should shed light on these interactions ([16], 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 We use a model D similar to aggregation-diffusion models [10] to generate a discrete spatial distribution of population density. A generalization of the basic model is proposed in [71], providing a calibration 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 [34]. More precisely, the model 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

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.

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 [**TeroAl10**], local network growth based on geometrical constraints optimization [8], 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 [26], 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

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.

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 [51] 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.
- 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 $(\vec{M}[D(\vec{\alpha})], \vec{G}[N(\vec{\alpha})])_{1\leqslant i\leqslant n}$ with standard unbiased estimator. We estimate correlation with associated Pearson estimator.

6.2.1.3 *Implementation*

Coupling of generative models is done both at formal and operational levels. We interface therefore independent implementations. The OpenMole software [75] 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` [90] because of its possibilities for interactive exploration and heuristic model construction. Source code is available for reproducibility on project repository[2].

6.2.1.4 *Results*

The study of density model alone is developed in [71]. It is in particular calibrated on European density grid data, on 50km width

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

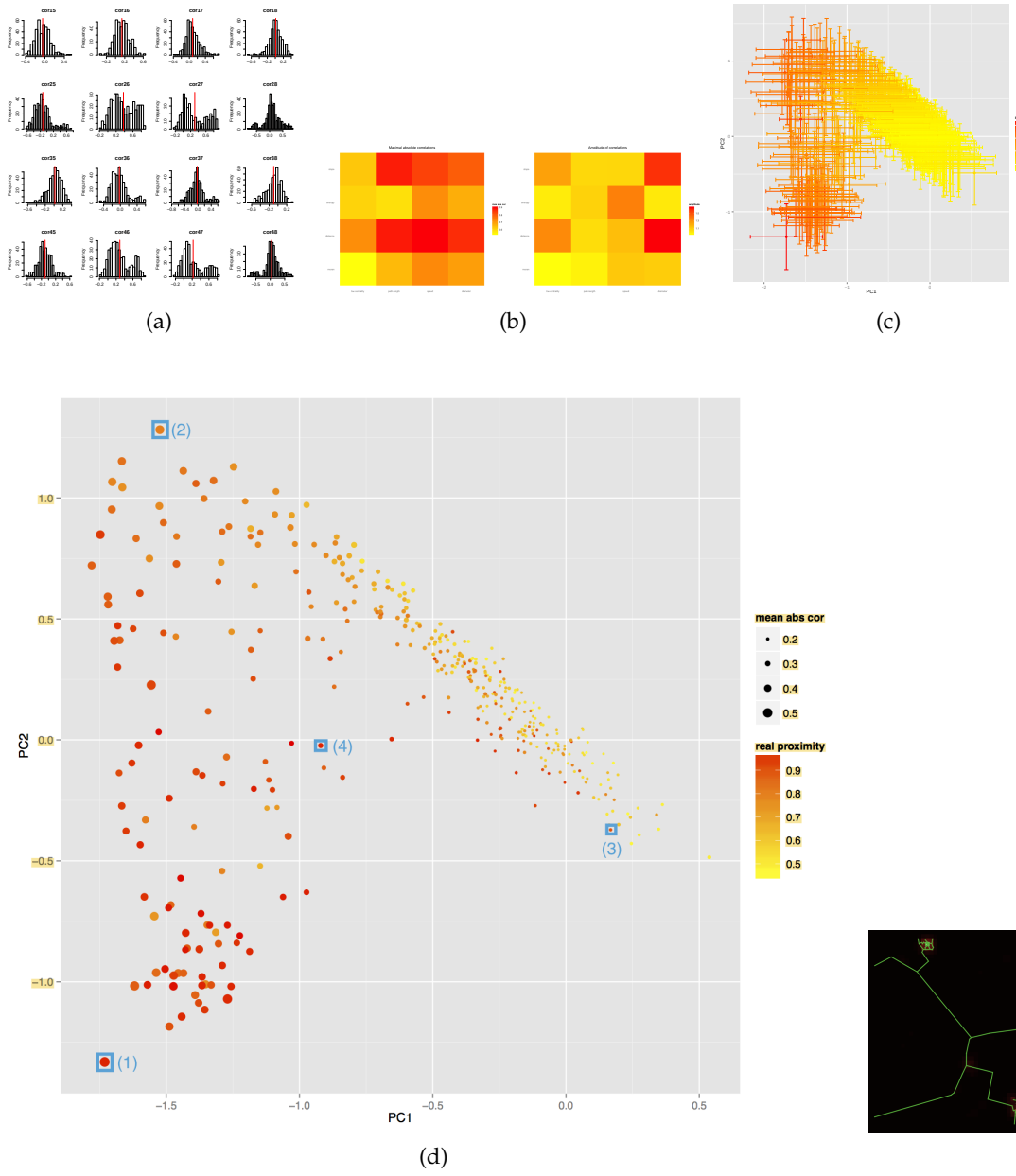


Figure 9: **Exploration of feasible space for correlations between urban morphology and network structure** | **(a)** Distribution of crossed-correlations between vectors $\vec{M}$ of morphological indicators (in numbering order Moran index, mean distance, entropy, hierarchy) and $\vec{N}$ of network 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. **(e)** 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)

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 $\vec{G}$, 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 9 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.

6.2.1.5 *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 [22].

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 [15]. 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 [41], 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 [68], there are strong chances that involved processes are different across different geographical areas (or from an other point of view that they are in an other state of meta-parameters, i.e. in an other regime) and that their interpretation as different realizations of the same stochastic process makes no sense, the impossibility of covariation estimation following. 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 [27]. 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.

3 https://www.openstreetmap.org

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 [23] is in an other perspective an interesting possible development.

6.2.3 *Conclusion*

We proposed an abstract method to generate synthetic datasets in which correlation structure is controlled. Its rapid implementation in two very different fields 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.

6.3 NETWORK GROWTH MODELS : EXPLICATIVE POWER FOR VARIOUS APPROACHES

7

TOWARDS MORE COMPLEX MODELS

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 which aims to be integrated as behavioral rules for governing agents in a hybrid model introduced in [49] and formalized then explored in [50]. 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 [50], 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 (gravital flows) are the first step of an iteration. The transportation infrastructure (road network) is then evolved by contructing 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 entrelaced 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 [40]. 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 [62]. Such a framework as already been used in transportation investment studies, as e.g. in [77] 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. Sub-modules, detailed in the following, include in particular a governance module that rules processes of network evolution.

7.1.2.1 *Land-use evolution*

7.1.2.2 *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 buiseness 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 infrastucture 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 proportionaly to wealth in the case of a collaboration. The payoff matrix of the game is the following, with κ a normalization constant ("price of acessibility") :

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

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 negociations take generally some time during which agents are able to find the way to optimize rationnaly their expected utility. If $\mathbb{P}[S_1 = C] = p_1$ and $\mathbb{P}[S_2 = C] = p_2$, we have

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

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

$$\frac{\partial \mathbb{E}[U_1]}{\partial p_1} = 0 \iff p_2 = \frac{\Delta X_1(Z_1^*) - \frac{I}{\kappa}}{\Delta X_1(Z_C^*) - \frac{\delta I}{2 \cdot \kappa}}$$

We obtain the same way

$$p_1 = \frac{\Delta X_2(Z_2^*) - \frac{I}{\kappa}}{\Delta X_2(Z_C^*) - \frac{\delta I}{2\cdot\kappa}}$$

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 competetive 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 negociation 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.

7.1.3 *Results*

7.1.4 *Perspectives*

Part III

TOWARDS OPERATIONAL MODELS

CONCLUSION

BIBLIOGRAPHY

[1] Merwan Achibet, Stefan Balev, Antoine Dutot, and Damien Olivier. "A Model of Road Network and Buildings Extension Co-evolution." In: *Procedia Computer Science* 32 (2014), pp. 828–833.

[2] Philippe Aghion, Ufuk Akcigit, Antonin Bergeaud, Richard Blundell, and David Hémous. *Innovation and Top Income Inequality*. 2015.

[3] Alex Anas, Richard Arnott, and Kenneth A. Small. "Urban Spatial Structure." English. In: *Journal of Economic Literature* 36.3 (1998), pp. 1426–1464. ISSN: 00220515. URL: http://www.jstor.org/stable/2564805.

[4] E. Arcaute, E. Hatna, P. Ferguson, H. Youn, A. Johansson, and M. Batty. "Constructing cities, deconstructing scaling laws." In: *ArXiv e-prints* (Jan. 2013). arXiv: 1301.1674 [physics.soc-ph].

[5] W. Brian Arthur. *Complexity and the Shift in Modern Science*. Conference on Complex Systems, Tempe, Arizona. 2015.

[6] Arnaud Banos. "Pour des pratiques de modélisation et de simulation libérées en Géographies et SHS." In: *HDR. Université Paris* 1 (2013).

[7] Albert-Laszlo Barabasi. "Linked: How everything is connected to everything else and what it means." In: *Plume Editors* (2002).

[8] Marc Barthélemy and Alessandro Flammini. "Modeling urban street patterns." In: *Physical review letters* 100.13 (2008), p. 138702.

[9] Marc Barthélemy and Alessandro Flammini. "Co-evolution of density and topology in a simple model of city formation." In: *Networks and spatial economics* 9.3 (2009), pp. 401–425.

[10] Michael Batty. "Hierarchy in cities and city systems." In: *Hierarchy in natural and social sciences*. Springer, 2006, pp. 143–168.

[11] Michael Batty. *The new science of cities*. Mit Press, 2013.

[12] Mark Bedau. "Downward causation and the autonomy of weak emergence." In: *Principia: an international journal of epistemology* 6.1 (2002), pp. 5–50.

[13] Luís MA Bettencourt, José Lobo, and Geoffrey B West. "Why are large cities faster? Universal scaling and self-similarity in urban organization and dynamics." In: *The European Physical Journal B-Condensed Matter and Complex Systems* 63.3 (2008), pp. 285–293.

[14] Steven Bird. "NLTK: the natural language toolkit." In: *Proceedings of the COLING/ACL on Interactive presentation sessions*. Association for Computational Linguistics. 2006, pp. 69–72.

[15] P. Bourgine, D. Chavalarias, and al. "French Roadmap for complex Systems 2008-2009." In: *ArXiv e-prints* (July 2009). arXiv: 0907.2221 [nlin.AO].

[16] Anne Bretagnolle. "Villes et réseaux de transport : des interactions dans la longue durée, France, Europe, États-Unis." Français. HDR. Université Panthéon-Sorbonne - Paris I, June 2009. URL: http://tel.archives-ouvertes.fr/tel-00459720.

[17] Anne Bretagnolle, Denise Pumain, and Céline Vacchiani-Marcuzzo. "The organization of urban systems." In: *Complexity perspectives in innovation and social change*. Springer, 2009, pp. 197–220.

[18] Chris Brunsdon, Stewart Fotheringham, and Martin Charlton. "Geographically weighted regression." In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 47.3 (1998), pp. 431–443.

[19] Justin S Chang. "Models of the Relationship between Transport and Land-use: A Review." In: *Transport Reviews* 26.3 (2006), pp. 325–350.

[20] David Chavalarias and Jean-Philippe Cointet. "Phylomemetic patterns in science evolution—the rise and fall of scientific fields." In: *Plos One* 8.2 (2013), e54847.

[21] Yanguang Chen. "Characterizing growth and form of fractal cities with allometric scaling exponents." In: *Discrete Dynamics in Nature and Society* 2010 (2010).

[22] Guillaume Chérel, Clémentine Cottineau, and Romain Reuillon. "Beyond Corroboration: Strengthening Model Validation by Looking for Unexpected Patterns." In: *PLoS ONE* 10.9 (Sept. 2015), e0138212. DOI: 10.1371/journal.pone.0138212. URL: http://dx.doi.org/10.1371%2Fjournal.pone.0138212.

[23] Rémy Chicheportiche and Jean-Philippe Bouchaud. "A nested factor model for non-linear dependences in stock returns." In: *arXiv preprint arXiv:1309.3102* (2013).

[24] Hadrien Commenges. "The invention of daily mobility. Performative aspects of the instruments of economics of transportation." Theses. Université Paris-Diderot - Paris VII, Dec. 2013. URL: https://tel.archives-ouvertes.fr/tel-00923682.

[25] Clémentine Cottineau. *Urban scaling: What cities are we talking about?* Presentation of ongoing work at Quanturb seminar, April 1st 2015. 2015.

[26] Clémentine Cottineau, Paul Chapron, and Romain Reuillon. "An incremental method for building and evaluating agent-based models of systems of cities." In: (2015).

[27] Clémentine Cottineau, Florent Le Néchet, Marion Le Texier, and Romain Reuillon. "Revisiting some geography classics with spatial simulation." In: *Plurimondi. An International Forum for Research and Debate on Human Settlements*. Vol. 7. 15. 2015.

[28] FD De Leon, M Felsen, and U Wilensky. "NetLogo Urban Suite-Tijuana Bordertowns model." In: *Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL* (2007). URL: http://ccl.northwestern.edu/netlogo/models/UrbanSuite-TijuanaBordertowns.

[29] Jean Delons, Nicolas Coulombel, and Fabien Leurent. "PIRANDELLO an integrated transport and land-use model for the Paris area." Aug. 2008. URL: https://halv3-preprod.archives-ouvertes.fr/hal-00319087.

[30] Lynn Dirk. "A Measure of Originality The Elements of Science." In: *Social Studies of Science* 29.5 (1999), pp. 765–776.

[31] O Dollfus and F Durand Dastès. "Some remarks on the notions of 'structure'and 'system' in geography." In: *Geoforum* 6.2 (1975), pp. 83–94.

[32] Chris Drummond. "Replicability is not reproducibility: nor is it good science." In: (2009).

[33] Gabriel Dupuy and Lucien Gilles Benguigui. "Sciences urbaines: interdisciplinarités passive, naïve, transitive, offensive." In: *Métropoles* 16 (2015).

[34] EUROSTAT. *Eurostat Geographical Data*. 2014. URL: http://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units.

[35] Jean-Marc Favaro and Denise Pumain. "Gibrat Revisited: An Urban Growth Model Incorporating Spatial Interaction and Innovation Cycles." In: *Geographical Analysis* 43.3 (2011), pp. 261–286.

[36] Roman Frigg and Ioannis Votsis. "Everything you always wanted to know about structural realism but were afraid to ask." In: *European journal for philosophy of science* 1.2 (2011), pp. 227–276.

[37] Xavier Gabaix. "Zipf's law for cities: an explanation." In: *Quarterly journal of Economics* (1999), pp. 739–767.

[38] Xavier Gabaix and Yannis M. Ioannides. "Chapter 53 The evolution of city size distributions." In: *Cities and Geography*. Ed. by J. Vernon Henderson and Jacques-François Thisse. Vol. 4. Handbook of Regional and Urban Economics. Elsevier, 2004, pp. 2341 –2378. DOI: http://dx.doi.org/10.1016/S1574-0080(04)80010-5. URL: http://www.sciencedirect.com/science/article/pii/S1574008004800105.

[39] Ronald N Giere. *Scientific perspectivism*. University of Chicago Press, 2010.

[40] Frédéric Gilli and Jean-Marc Offner. *Paris, métropole hors les murs: aménager et gouverner un Grand Paris*. Sciences Po, les presses, 2009.

[41] Jean-François Girres and Guillaume Touya. "Quality assessment of the French OpenStreetMap dataset." In: *Transactions in GIS* 14.4 (2010), pp. 435–459.

[42] Boris Golden, Marc Aiguier, and Daniel Krob. "Modeling of complex systems ii: A minimalist and unified semantics for heterogeneous integrated systems." In: *Applied Mathematics and Computation* 218.16 (2012), pp. 8039–8055.

[43] C. A. Hidalgo. "Disconnected! The parallel streams of network literature in the natural and social sciences." In: *ArXiv e-prints* (Nov. 2015). arXiv: 1511.03981 [physics.soc-ph].

[44] Michael Iacono, David Levinson, and Ahmed El-Geneidy. "Models of transportation and land use change: a guide to the territory." In: *Journal of Planning Literature* 22.4 (2008), pp. 323–340.

[45] Alexandros Karatzoglou, Alex Smola, Kurt Hornik, and Achim Zeileis. "kernlab – An S4 Package for Kernel Methods in R." In: *Journal of Statistical Software* 11.9 (2004), pp. 1–20. URL: http://www.jstatsoft.org/v11/i09/.

[46] Marie-Laurence Keersmaecker, Pierre Frankhauser, and Isabelle Thomas. "Using fractal dimensions for characterizing intra-urban diversity: The example of Brussels." In: *Geographical analysis* 35.4 (2003), pp. 310–328.

[47] Christof Koch and Gilles Laurent. "Complexity and the nervous system." In: *Science* 284.5411 (1999), pp. 96–98.

[48] Robert B Laughlin. *A different universe: Reinventing physics from the bottom down*. Basic Books, 2006.

[49] Florent Le Nechet. "Approche multiscalaire des liens entre mobilité quotidienne, morphologie et soutenabilité des métropoles européennes: cas de Paris et de la région Rhin-Ruhr." PhD thesis. Université Paris-Est, 2010.

[50] Florent Le Nechet. "Aménagement urbain et jeux d'échelles : construction à Aménagement urbain et jeux d'échelles : construction à Aménagement urbain et jeux d'échelles : construction à plusieurs niveaux d'un réseau de transport métropolitain." 2012.

[51] Florent Le Néchet. "De la forme urbaine à la structure métropolitaine: une typologie de la configuration interne des densités pour les principales métropoles européennes de l'Audit Urbain." In: *Cybergeo: European Journal of Geography* (2015).

[52] Michael Lissack. "Subliminal influence or plagiarism by negligence ? The Slodderwetenschap of ignoring the internet." In: *Journal of Academic Ethics* (2013).

[53] Pierre Livet, Jean-Pierre Muller, Denis Phan, and Lena Sanders. "Ontology, a Mediator for Agent-Based Modeling in Social Science." In: *Journal of Artificial Societies and Social Simulation* 13.1 (2010), p. 3. ISSN: 1460-7425. URL: http://jasss.soc.surrey.ac.uk/13/1/3.html.

[54] R. Louf and M. Barthelemy. "How congestion shapes cities: from mobility patterns to scaling." In: *ArXiv e-prints* (Jan. 2014). arXiv: 1401.8200 [physics.soc-ph].

[55] Rémi Louf and Marc Barthelemy. "Scaling: lost in the smog." In: *arXiv preprint arXiv:1410.4964* (2014).

[56] Steven M Manson. "Does scale exist? An epistemological scale continuum for complex human–environment systems." In: *Geoforum* 39.2 (2008), pp. 776–788.

[57] Mendeley. *Mendeley Reference Manager*. http://www.mendeley.com/. 2015.

[58] M. E. J. Newman. "Prediction of highly cited papers." In: *ArXiv e-prints* (Oct. 2013). arXiv: 1310.8220 [physics.soc-ph].

[59] MEJ Newman. "Complex systems: A survey." In: *arXiv preprint arXiv:1112.1440* (2011).

[60] Jean-Marc Offner. "Les "effets structurants" du transport: mythe politique, mystification scientifique." In: *Espace géographique* 22.3 (1993), pp. 233–242.

[61] Jean-Marc Offner and Denise Pumain. "Réseaux et territoires-significations croisées." In: (1996).

[62] Peter C Ordeshook. *Game theory and political theory: An introduction*. Cambridge University Press, 1986.

[63] Thomas Piketty. *Le capital au XXIe siècle*. Seuil, 2013.

[64] David R Pritchard and Eric J Miller. "Advances in agent population synthesis and application in an integrated land use and transportation model." In: *Transportation Research Board 88th Annual Meeting*. 09-1686. 2009.

[65] Denise Pumain. "Pour une théorie évolutive des villes." In: *Espace géographique* 26.2 (1997), pp. 119–134.

[66] Denise Pumain. "Cumulativité des connaissances." In: *Revue européenne des sciences sociales. European Journal of Social Sciences* XLIII-131 (2005), pp. 5–12.

[67] Denise Pumain. "Une théorie géographique des villes." In: *Bulletin de la Société géographie de Liège* 55 (2010), pp. 5–15.

[68] Denise Pumain. "Urban systems dynamics, urban growth and scaling laws: The question of ergodicity." In: *Complexity Theories of Cities Have Come of Age*. Springer, 2012, pp. 91–103.

[69] Denise Pumain, Fabien Paulus, Céline Vacchiani-Marcuzzo, and José Lobo. "An evolutionary theory for interpreting urban scaling laws." In: *Cybergeo: European Journal of Geography* (2006).

[70] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2015. URL: http://www.R-project.org/.

[71] J. Raimbault. "Calibration of a Spatialized Urban Growth Model." In: *Working Paper, draft at https://github.com/JusteRaimbault/CityNetwork/tree/master/Do* (2016).

[72] J. Raimbault. *Vers des Modèles Couplant Développement Urbain et Croissance des Réseaux de Transports, PhD Project Description*. Tech. rep. Géographie-Cités UMR CNRS 8504/LVMT UMR-T IFSTTAR 9403, October 2014.

[73] J. Raimbault, A. Banos, and R. Doursat. "A hybrid network/grid model of urban morphogenesis and optimization." In: *Proceedings of the 4th International Conference on Complex Systems and Applications (ICCSA 2014), June 23-26, 2014, Université de Normandie, Le Havre, France; M. A. Aziz-Alaoui, C. Bertelle, X. Z. Liu, D. Olivier, eds.: pp. 51-60*. 2014.

[74] Karthik Ram. "Git can facilitate greater reproducibility and increased transparency in science." In: *Source code for biology and medicine* 8.1 (2013), p. 7.

[75] Romain Reuillon, Mathieu Leclaire, and Sebastien Rey-Coyrehourcq. "OpenMOLE, a workflow engine specifically tailored for the distributed exploration of simulation models." In: *Future Generation Computer Systems* 29.8 (2013), pp. 1981–1990.

[76] Henri Reymond and Colette Cauvin. "La logique ternaire de Stéphane Lupasco et le raisonnement géocartographique bioculturel d'Homo geographicus. L'apport de la notion de couplage transdisciplinaire dans l'approche de l'agrégation morphologique des agglomérations urbaines." In: *Cybergeo: European Journal of Geography* (2013).

[77] Athena Roumboutsos and Seraphim Kapros. "A game theory approach to urban public transport integration policy." In: *Transport Policy* 15.4 (2008), pp. 209 –215. ISSN: 0967-070X. DOI: http://dx.doi.org/10.1016/j.tranpol.2008.05.001. URL: http://www.sciencedirect.com/science/article/pii/S0967070X08000280.

[78] Gerta Rucker. "Network meta-analysis, electrical networks and graph theory." In: *Research Synthesis Methods* 3.4 (2012), pp. 312–324.

[79] Lena Sanders, Denise Pumain, Hélene Mathian, France Guérin-Pace, and Stephane Bura. "SIMPOP: a multiagent system for the study of urbanism." In: *Environment and Planning B* 24 (1997), pp. 287–306.

[80] E. Sarigöl, R. Pfitzner, I. Scholtes, A. Garas, and F. Schweitzer. "Predicting Scientific Success Based on Coauthorship Networks." In: *ArXiv e-prints* (Feb. 2014). arXiv: 1402.7268 [physics.soc-ph].

[81] Thomas C Schelling. "Dynamic models of segregation." In: *Journal of mathematical sociology* 1.2 (1971), pp. 143–186.

[82] Clara Schmitt. "Modélisation de la dynamique des systèmes de peuplement: de SimpopLocal à SimpopNet." PhD thesis. Paris 1, 2014.

[83] H Eugene Stanley, Luis A Nunes Amaral, David Canning, Parameswaran Gopikrishnan, Youngki Lee, and Yanhui Liu. "Econophysics: Can physicists contribute to the science of economics?" In: *Physica A: Statistical Mechanics and its Applications* 269.1 (1999), pp. 156–169.

[84] Victoria Stodden. "The scientific method in practice: Reproducibility in the computational sciences." In: (2010).

[85] Atsushi Tero, Seiji Takagi, Tetsu Saigusa, Kentaro Ito, Dan P. Bebber, Mark D. Fricker, Kenji Yumiki, Ryo Kobayashi, and Toshiyuki Nakagaki. "Rules for Biologically Inspired Adaptive Network Design." In: *Science* 327.5964 (2010), pp. 439–442. DOI: 10.1126/science.1177894. eprint: http://www.sciencemag.org/content/327/5964/439.full.pdf. URL: http://www.sciencemag.org/content/327/5964/439.abstract.

[86] Franck Varenne. "Framework for M&S with Agents in Regard to Agent Simulations in Social Sciences." In: *Activity-Based Modeling and Simulation* (2010), pp. 53–84.

[87] Michael Wegener and Franz Fürst. "Land-use transport interaction: state of the art." In: *Available at SSRN 1434678* (2004).

[88] Michael Wegener, Roger L Mackett, and David C Simmonds. "One city, three models: comparison of land-use/transport policy simulation models for Dortmund." In: *Transport Reviews* 11.2 (1991), pp. 107–129.

[89] Jörgen W Weibull. "An axiomatic approach to the measurement of accessibility." In: *Regional Science and Urban Economics* 6.4 (1976), pp. 357–379.

[90] Uri Wilensky. "NetLogo." In: (1999).

[91] Stephen Wolfram. *A new kind of science*. Vol. 5. Wolfram media Champaign, 2002.

[92] Feng Xie and David Levinson. "How streetcars shaped suburbanization: a Granger causality analysis of land use and transit in the Twin Cities." In: *Journal of Economic Geography* (2009), lbp031.

[93] Feng Xie and David Levinson. "Modeling the growth of transportation networks: A comprehensive review." In: *Networks and Spatial Economics* 9.3 (2009), pp. 291–307.

[94] Yihui Xie. "knitr: A general-purpose package for dynamic report generation in R." In: *R package version* 1.7 (2013).

[95] Kazuko Yamasaki, Kaushik Matia, Sergey V Buldyrev, Dongfeng Fu, Fabio Pammolli, Massimo Riccaboni, and H Eugene Stanley. "Preferential attachment and growth dynamics in complex systems." In: *Physical Review E* 74.3 (2006), p. 035103.

[96] Lei Zhang and David Levinson. "The economics of transportation network growth." In: *Essays on transport economics*. Springer, 2007, pp. 317–339.

Part IV

APPENDIX

8

ARCHITECTURE AND SOURCES FOR ALGORITHMS AND MODELS OF SIMULATION

9

TOOLS AND WORKFLOW FOR AN OPEN REPRODUCIBLE RESEARCH