For a Cautious Use of Big Data and Computation

J. Raimbault[1,2]
juste.raimbault@parisgeo.cnrs.fr

[1]UMR CNRS 8504 Géographie-cités
[2]UMR-T IFSTTAR 9403 LVMT

RGS-AC 2016

Session Geocomputation : The Next 20 years

31st September 2016

Computational power : exponential capabilities

Moore's law in Geocomputation ?

- [Gleyze, 2005] : urban network analyses, concludes that "limited by computation"
  $\rightarrow$ 10 years later : [Lagesse, 2015] !
- First Simpop models [Sanders et al., 1997] "calibrated" by hand
  $\rightarrow$ today Simpoplocal [Schmitt et al., 2014] and Marius [Cottineau et al., 2015] calibrated on grid, billions of simulations !
- Space syntax : from the theoretical origins [Hillier and Hanson, 1989] to large-scale applications [Hillier, 2016]

New and Big Data

Larger dataset can be processed, new type of data available :

- Mobility studied through various type of data : new data from transportation systems [O'brien et al., 2014], from Social Networks , other types
- Opening of "classic" dataset should allow ever more meta-analyses
- New ways to do research, more interactive, crowd-sourced science and data ? [Cottineau, 2016] ; [Chasset et al., 2016]

But to what purpose ?

[Barthelemy et al., 2013] : new data and methods, but reinvent the wheel !


SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
PHYSICS
STATISTICAL PHYSICS, THERMODYNAMICS AND

Self-organization versus top-down planning in the evolution of a city

Marc Barthelemy[1,2], Patricia Bordin[3,4], Henri Berestycki[2] & Maurizio Gribaudi[5]

But to what purpose ?

Exaggerating agent-based modeling ? Up to simulating the world at scale 1 :1 !

120 Million Agents Self-Organize into 6 Million Firms: A Model of the U.S. Private Sector

Robert L. Axtell
George Mason University
4400 University Drive
Fairfax, VA 22030 USA
+1 (703) 556-0333
rax222@gmu.edu

ABSTRACT

An agent model is described at full-scale with the U.S. private sector, consisting of some 120 million agents. Using data on the population of U.S. firms the model is calibrated to closely reproduce firm sizes, ages, growth rates, job tenure and labor flows, along with several other empirically-important facts. It consists of a coalition formation model in which the Nash equilibria are dynamically unstable for sufficiently large coalitions. When agents are free to join coalitions where they are

the economy is in general equilibrium then there is no way to realize micro-dynamics except by the imposition of external shocks. Can microeconomic models *endogenously* produce the kinds of dynamics observed empirically when the incentives agents have to change jobs are fully represented?

Here I describe a microeconomic model capable of producing, *without* exogenous shocks, firm and labor dynamics of the size and type the U.S. economy experiences. While conventional explanations for these large labor flows exist [e.g.,

But to what purpose ?

Other worrying examples :

- [Cura, 2014] : waste computational ressources to simulate mean and variance of Gibrat model (= recheck the Central Limit Theorem !), which is fully solved otherwise [Gabaix, 1999]
- Recently seen on Geotamtam : rush on new data (Pokemon Go) before thinking !
- [Louail et al., 2016] draw social equity policy recommandations by acting on mobility, from credit card transaction data but totally disconnected from Urban Form.

Theories and Computation

Claim : The computational shift [Arthur, 2015] and simulation practices will be central in geography [Banos, 2013], but are also dangerous :

- Data deluge may impose research subjects and elude theory
- Computation may elude model construction and solving

→ *Make a stronger link between computational practices, mathematics/statisti and theoretical geography* (Note : precise purpose of TQG, but seems forgotten sometimes)
→ *Theoretical and Quantitative Geography in the center of this dynamic*

Case study : Context and Rationale

Study of interactions between network and territories :

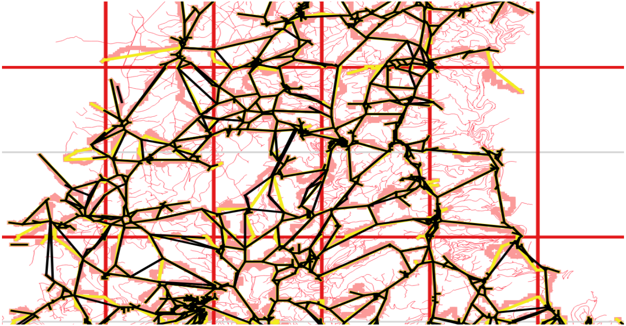
→ *searching for stylized facts, what can be learnt from static correlations between urban form and road network ?*

Theoretical Background : *A Theory of co-evolutive networked human territories* proposed in [Raimbault, 2016], that in particular postulates an important role of networks in the morphogenesis of complex adaptive urban systems that are human territories

Dataset construction

Computation of topological road network for all Europe, at 100m granularity scale (to be used consistently with population grid [EUROSTAT, 2014])

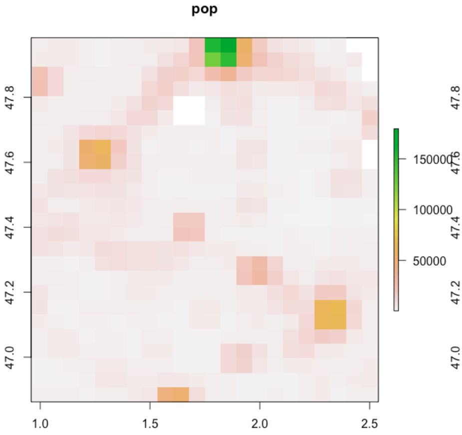
→ Import of OSM into `pgsql`, simplification at 100m granularity, topological simplification with split/merge algorithm

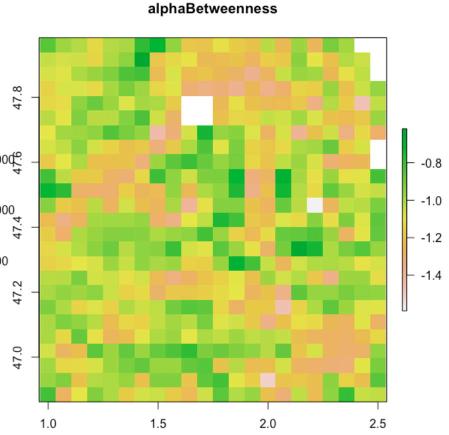


$\simeq 44 \cdot 10^6$ *links in initial OSM db,* $\simeq 61 \cdot 10^6$ *in first simplified layer,* $\simeq 21 \cdot 10^6$ *in final database*

Results : Computation of Indicators

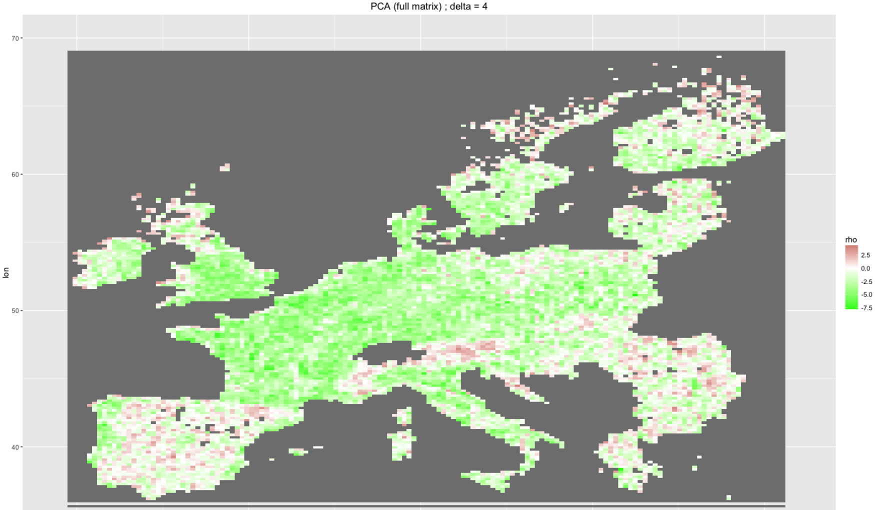
Computation of urban form indicators [Le Néchet, 2015] and network indicators on $l_0 = 10km$ side square





Results : Spatial Correlations

Computation of spatial correlation on square areas of width $\delta \cdot l_0$ *(with typically* $\delta = 4, \dots, 16$*)*



Results : Stationarity scales

Results : Stationarity scales

Locally stationary spatial correlations

$Y_i\left[\vec{x}, t\right]$ spatio-temporal stochastic process, assumptions :

1. Local spatial autocorrelation is present and bounded by l_ρ (in other words the processes are continuous in space) : at any $\vec{x}$ and t, $\left|\rho_{\|\Delta\vec{x}\|<l_\rho}\left[Y_i(\vec{x}+\Delta\vec{x}, t), Y_i(\vec{x}, t)\right]\right| > 0$.
2. Processes are locally parametrized : $Y_i = Y_i\left[\alpha_i\right]$, where $\alpha_i(\vec{x})$ varies with l_α, with $l_\alpha \gg l_\rho$.
3. Spatial correlations between processes have a sense at an intermediate scale l such that $l_\alpha \gg l \gg l\rho$.
4. Local ergodicity is present at scale l and dynamics are locally chaotic.

Stationarity and Ergodicity

- Assuming local ergodicity, spatial local stationarity implies and temporal local stationarity
- Spatial non-stationarity **at the second order** $\Longrightarrow$ temporal scale variations $\Longrightarrow$ non-ergodicity

Case study : implications

→ we show the regional nature of network-territories interactions, in particular the non-ergodicity of urban systems on these components

→ no direct results on time dynamics, but indirect : spatio-temporal processes do not have same speed and react/diffuse differently

→ still explorations to do : variable correlations areas ; same work on cities population/train network data, which are also dynamical databases : extrapolation of ergodicity parameters ?

Discussion

1. *Is a theory-free quantitative "geography possible ?*
→ close to the trap of black-box data-mining analysis ; still poor explanatory power, can exhibit relations but not reconstruct processes

2. *Is a pure computational quantitative geography possible ?*
→ even gaining 3 orders of magnitudes in computational power does not solve the dimensionality curse

In our case study : Without theory, would not know which objects/measures to look at ; multi-scale and dynamical nature ; without analytics : no conclusion from empirical analysis

Conclusion

$\rightarrow$ Nothing is really new, and more than ever we need simple but powerful theories à-la-Occam

$\rightarrow$ Need for a wise integration of new techniques/rediscovering into existing body of knowledge : multi-modeling and model families (see Rey et al. presentation) as one way to do that ?

$\rightarrow$ Need for more interdisciplinarity than ever ?

Reserve

Reserve Slides

Network Simplification Algorithm

1. Filter OSM links (highway tag) and insert into `pgsql` with `osmosis` [team, 2016]
2. First simplification : two cells of base raster (population density) are linked if and only if they are linked by OSM link (associated type/speed)
3. Topological simplification :
   - split the space into a partition, simplify within each box
   - construct independent merging subsets of the partition, merge sequentially for each subset.

$\rightarrow$

Indicators

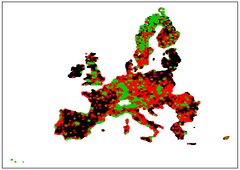
Morphological Indicators : Density, Spatial Autocorrelation, Entropy, Mean Distance, Hierarchy

Network Indicators : betweenness (mean/hierarchy), closeness (mean/hierarchy), mean link length, network performance, mean path length, diameter, components, clustering coefficient, density

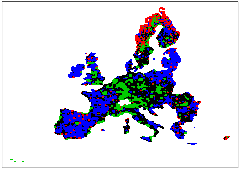
Correlation Measures : Pearson test ; correlation matrices then aggregated (mean, mean absolute, first principal component - % variance with $\delta =$)

Morphology Classification

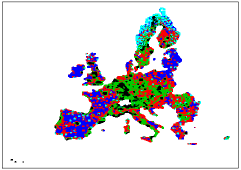
k=3 ; withinProp=0.379727175801079



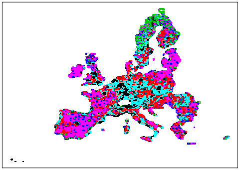
k=4 ; withinProp=0.304934256837235



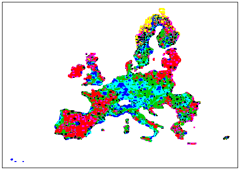
k=5 ; withinProp=0.258568287232286



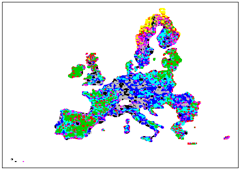
k=6 ; withinProp=0.224029913068682



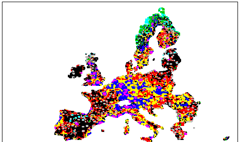
k=7 ; withinProp=0.201599568507077



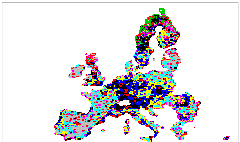
k=8 ; withinProp=0.179850720275406



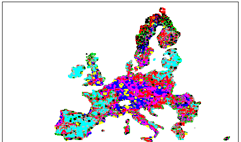
k=9 ; withinProp=0.167033521857729



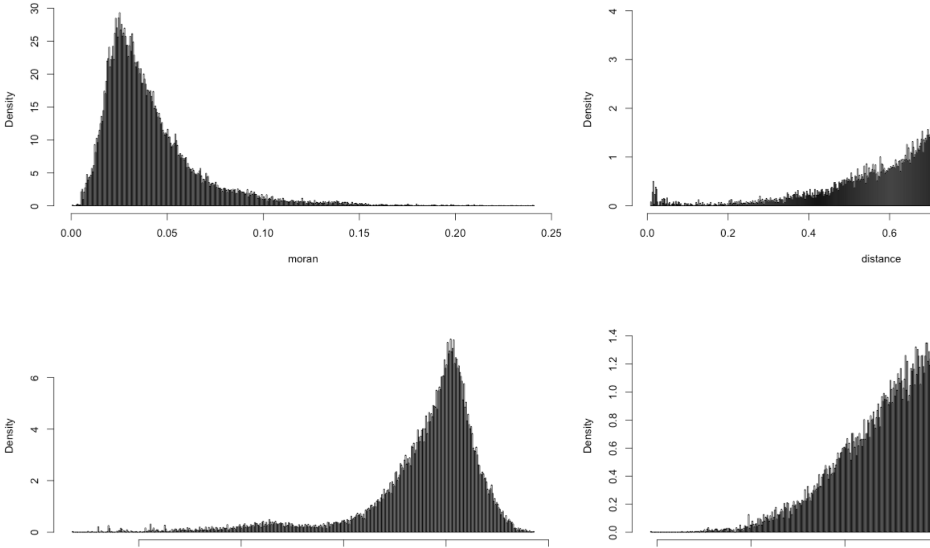
k=10 ; withinProp=0.156221620124614



k=11 ; withinProp=0.148766308022263



Morphology Distribution



Network Distribution

Theory : Pillars

1. *Networked Human Territories* $\rightarrow$ Raffestin approach to territory combined with Dupuy theory of networks.
2. *Evolutive Urban Theory* $\rightarrow$ City Systems as complex Adaptive systems, applied to human settlements in general and thus territorial systems.
3. *Urban Morphogenesis* $\rightarrow$ Morphogenesis as autonomous rules to explain growth of urban form. Used as the provider of modular decompositions.
4. *Boundaries and Co-evolution* $\rightarrow$ Co-evolution as the existence of *niche*, consequence of boundary patterns.

Theory : Specification

- Previous def. of territorial systems
- Modular decomposition and stationarity : existence of scales
- Feedback loops between and inside scales yield weak emergence, thus complexity
- Morphogenesis gives modular decomposition and co-evolution
- **Main assumption.** Necessity of Networks : networks are necessary component of co-evolutive niches.

References I

Arthur, W. B. (2015).
Complexity and the shift in modern science.
Conference on Complex Systems, Tempe, Arizona.

Banos, A. (2013).
Pour des pratiques de modélisation et de simulation libérées en géographies et shs.
HDR. Université Paris, 1.

Barthelemy, M., Bordin, P., Berestycki, H., and Gribaudi, M. (2013).
Self-organization versus top-down planning in the evolution of a city.
Scientific reports, 3.

Chasset, P.-O., Commenges, H., Cottineau, C., and Raimbault, J. (2016).
cybergeo20 v1.0.

References II

Cottineau, C. (2016).
MetaZipf. (Re)producing knowledge about city size distributions.
ArXiv e-prints.

Cottineau, C., Reuillon, R., Chapron, P., Rey-Coyrehourcq, S., and Pumain, D. (2015).
A modular modelling framework for hypotheses testing in the simulation of urbanisation.
Systems, 3(4) :348–377.

Cura, R. (2014).
Gibrat population growth simulator.

EUROSTAT (2014).
Eurostat geographical data.

References III

Gabaix, X. (1999).
Zipf's law for cities : an explanation.
Quarterly journal of Economics, pages 739–767.

Gleyze, J.-F. (2005).
La vulnérabilité structurelle des réseaux de transport dans un contexte de risques.
PhD thesis, Université Paris-Diderot-Paris VII.

Hillier, B. (2016).
The fourth sustainability, creativity : Statistical associations and credible mechanisms.
In *Complexity, Cognition, Urban Planning and Design*, pages 75–92. Springer.

References IV

Hillier, B. and Hanson, J. (1989).
The social logic of space.
Cambridge university press.

Lagesse, C. (2015).
Read Cities through their Lines. Methodology to characterize spatial graphs.
ArXiv e-prints.

Le Néchet, F. (2015).
De la forme urbaine à la structure métropolitaine : une typologie de la configuration interne des densités pour les principales métropoles européennes de l'audit urbain.
Cybergeo : European Journal of Geography.

References V

Louail, T., Lenormand, M., Arias, J. M., and Ramasco, J. J. (2016).
Crowdsourcing the robin hood effect in cities.
arXiv preprint arXiv :1604.08394.

O'brien, O., Cheshire, J., and Batty, M. (2014).
Mining bicycle sharing data for generating insights into sustainable transport systems.
Journal of Transport Geography, 34 :262–273.

Raimbault, J. (2016).
Towards Models Coupling Urban Growth and Transportation Network Growth. First year preliminary memoire. DOI : http ://dx.doi.org/10.5281/zenodo.60538.
PhD thesis, Université Paris-Diderot - Paris VII.

References VI

Sanders, L., Pumain, D., Mathian, H., Guérin-Pace, F., and Bura, S. (1997).
Simpop : a multiagent system for the study of urbanism.
Environment and Planning B, 24 :287–306.

Schmitt, C., Rey-Coyrehourcq, S., Reuillon, R., and Pumain, D. (2014).
Half a billion simulations : Evolutionary algorithms and distributed computing for calibrating the simpoplocal geographical model.

team, O. (2016).
Osmosis.
http ://wiki.openstreetmap.org/wiki/Osmosis.