Thesis Progress Meeting

J. Raimbault[1,2]

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

January 26th 2016

Achieved Work (by projects)

- Biblio/Meetings/Organisation [0.7w]
- Conference [0.7w]
- Reading Records (*Synergetics* [Sanders, 1992]) [0.2w]
- Monitorat [1,3w]
- Cybergeo Project [1w]
- Correlated Synthetic data [3w]
- Theory construction (communication JIG) [0.2w]
- BP Case Study / Spatial Econometrics [0,3w]

Context

[Introduction at Rochebrune] : imagine a model of simulation describing skiers/ snowboarders relations, measures to improve situation ? NO conclusion without model exploration, including sensitivity to ressort station spatial configuration, or to population structure, even at second order → Necessity in that case (among others) to generate synthetic data controlled at second order.

Def. : *Synthetic Data* are output of generative models (and possibly inputs of models using them).

Methodology used in various fields, e.g. therapeutic evaluation [Abadie et al., 20 territorial systems analysis [Moeckel et al., 2003, Pritchard and Miller, 2009], machine learning [Bolón-Canedo et al., 2013] or bio-informatics [Van den Bulcke

Few examples at the second order : specific examples as [Ye, 2011] for discrete choices ; methods that can be interpreted this way : generation of complex networks [Newman, 2003].

Generic Method

$\vec{X}_I$ multidimensional stochastic process, $\mathbf{X} = (X_{i,j})$ realizations.

Aim : Generate a statistical population $\tilde{\mathbf{X}} = \tilde{X}_{i,j}$ such that:

1. proximity to data : given a precision ε and an indicator f, $\|f(\mathbf{X}) - f(\tilde{\mathbf{X}})\| < \varepsilon$
2. control of the estimated correlation structure : $\hat{\mathrm{Var}}\left[(\tilde{X}_i)\right] = \Sigma R$ with R fixed.

Geographical data : Context

- In geography, generation of synthetic populations for agent-based models [Pritchard and Miller, 2009].
- Generation of spatial synthetic configuration not used (Geo. Weighted Regression [Brunsdon et al., 1998] can be interpreted this way) ; however crucial for abstract models [Schmitt, 2014]
- [Cottineau et al., 2015] recently proposed to estimate the sensitivity of spatial models of simulation to initial configuration (application to Schelling model).
- Case study : city-transportation interactions, complex to understood quantitatively [Offner, 1993, Bretagnolle, 2009] $\rightarrow$ simple model of population density and transportation network morphogenesis.

Model

Simple coupling between

- Iterative generation of a density grid by preferential attachment/diffusion [Raimbault, 2016] calibrated on morphological objectives on european density grid.

- Heuristic network generation conditional to density :
  - Distribution of a fixed number of centers preferentially following density
  - Deterministic percolation between closest neighbors
  - Breaking of interaction potentials

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

    for a fixed number of couples N_L such that $V_{ij}(d_N)/V_{ij}(d_{ij})$ is minimal among $K \cdot N_L$ strongest euclidian potentials ($K=5$ fixed)
  - Planarization

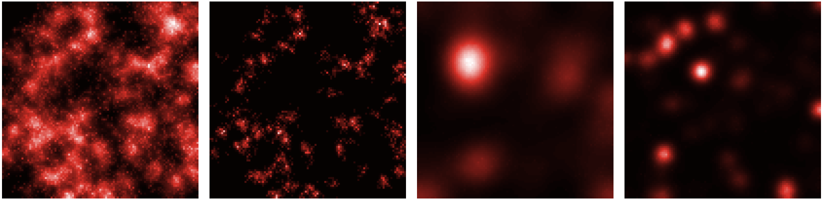
Indicators : morphology [Le Néchet, 2015] (Moran, mean distance, entropy, hierarchy) and network (centrality, mean width, speed, diameter).

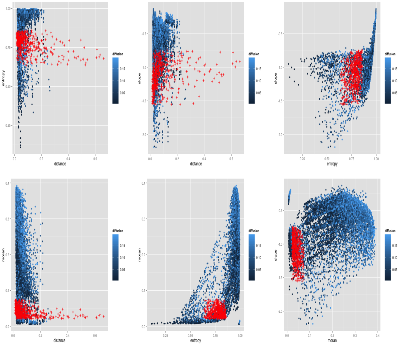
Implementation and Exploration

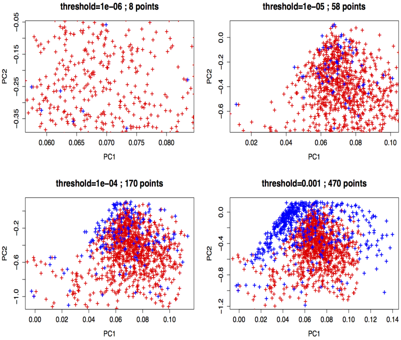
$\rightarrow$ Formal and Operational coupling : modular implementation (`scala/NetLogo`
encapsulated by `OpenMole` [Reuillon et al., 2013]

$\rightarrow$ Exploration by intensive computation on grid via `OpenMole` : calibration of density model alone ($\sim 1.5 \cdot 10^6$ runs) ; brutal exploration by LHS sampling for feasible correlations ($\sim 5 \cdot 10^4$ runs)

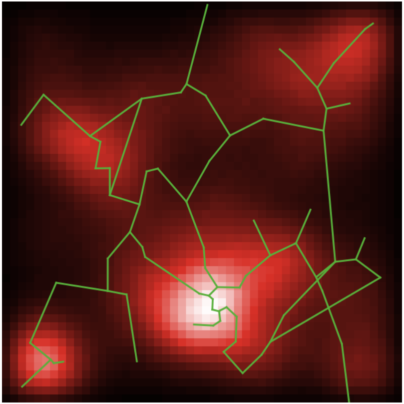
Results : Density Model alone

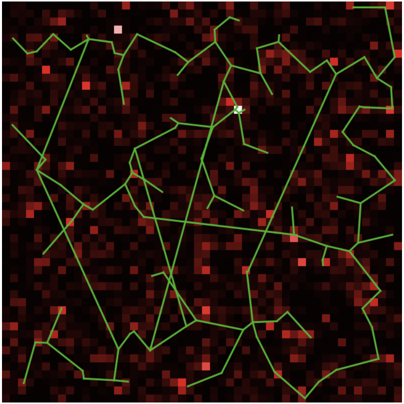




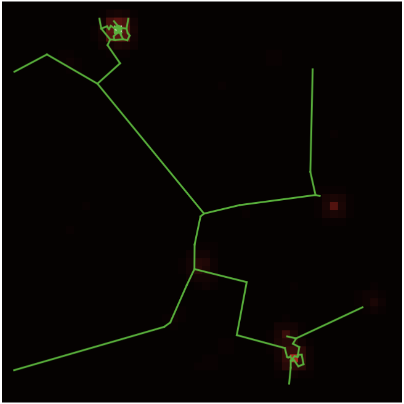


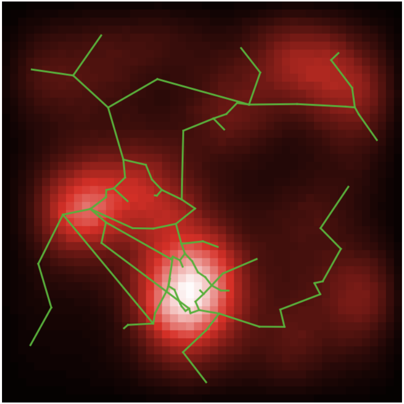
Results : examples of configurations



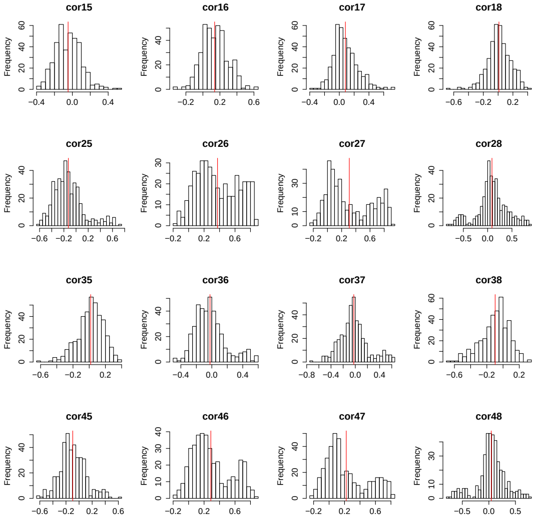


Results : examples of configurations



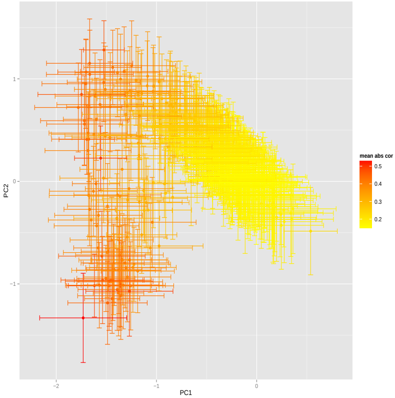


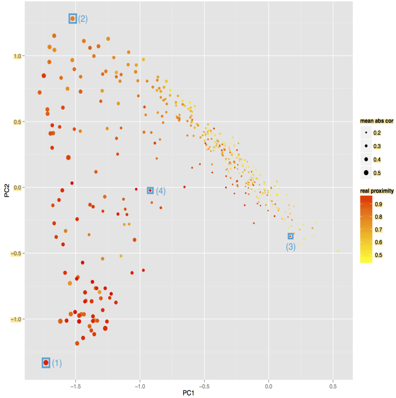
Results : cross-correlations



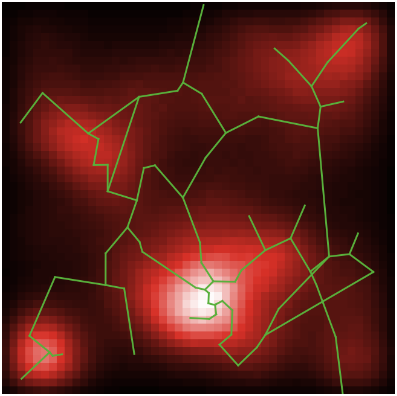
Results : feasible correlations

Mean matrices in a principal plan

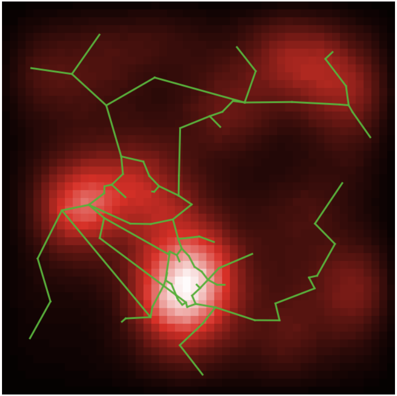




Results : exemples of correlations



$\rho[\bar{d}, \bar{c}] \simeq 0.34$



$\rho[\bar{d}, \bar{c}] \simeq -0.41$

$\rightarrow$ *gravity hierarchy more important in (1)* $\gamma = 3.9, k_h = 0.7$ *against* $\gamma = 1.07, k_h = 0.25$ *for (2)*

Applications

1. Calibration of the coupled model, street network data (edge effects!)
   $\rightarrow$ generation of correlated synthetic data corresponding to a given urban system
   $\rightarrow$ intrinsic correlations to be compared to estimated correlations between different states : non-ergodicity of urban systems [Pumain, 2012]).

2. Dynamical correlations in a strongly coupled model / spatio-temporal correlations in a strong spatial coupling.

Case study : Context

Database by Florent : main road network (route 120) in extended *Bassin Parisien* with opening dates for highways ; census data : population and employment of *communes* at dates [other data such as rail network and train timetables not used for now].

Formalisation : Dynamic transportation network $n(\vec{x}, t)$ within a dynamic territorial landscape $\vec{T}(\vec{x}, t)$, which components are population $p(\vec{x}, t)$ and employments $e(\vec{x}, t)$, discretized in space and in time, i.e. the spatial field $\vec{T}$ is summarized by $\mathbf{T} = \left(\vec{T}(\vec{x}_i, t_j)\right)_{i,j}$ with $1 \leq i \leq N$ and $1 \leq j \leq T$. To simplify, network distances sampled at same times and spatial points (support extended if not the case), given by $\mathbf{N} = \left(\vec{d}_n(\vec{x}_i, t_j)\right)_{i,j}$.

On Accessibility

Is the notion of accessibility crucial for statistical analysis ?

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

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

Then A verifies these *iff* it is of the form

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

where T is increasing with null origin, z is a *distance substitution function* (i.e. verifying axiom 2) and $\oplus$ a *standard composition* associating two attractions at zero distance to th corresponding unique one.
$\rightarrow$ *Well suited matrices of autocorrelation should capture accessibility in regressions ; or captured by non-linear regression on* $\mathbf{N}$

Accessibility as potential ?

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

Statistical Analysis

Large set of analysis to be tested (non exhaustive) :

- On data :
  - Multivariate models $\mathcal{L}\left[\mathbf{T},\mathbf{N}\right] \sim \varepsilon$
  - Autocorrelated univariate models $(\mathbf{I} - \Sigma RW)\mathbf{X} \sim \varepsilon$
  - Autocorrelated multivariate models $(\mathcal{L}' - \Sigma RW)\left[\mathbf{T}+\mathbf{N}\right] \sim \varepsilon$
  - Geographically Weighted Regression [Brunsdon et al., 1998]

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

  - Granger causality tests : [Xie and Levinson, 2009] use Granger causality to link transit with land-use changes.
- On data returns :
  - Autoregressive multivariate models

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

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

P. Bourgine framework for Complex Adaptive Systems

Bourgine has recently developed a framework to extract patterns of Complex Adaptive Systems, using a representation theorem : any discrete stationary process is a *Hidden Markow Model* (Knight, 1975)
Given the definition of a causal state as $\mathbb{P}[future|A] = \mathbb{P}[future|B]$, the partition of system states induced by the corresponding equivalence relations allows to derive a *Recurrent Network* that is enough to determine next state of the system, as it is a *deterministic* function of previous state and hidden states [Shalizi and Crutchfield, 2001] :

$$(x_{t+1}, s_{t+1}) = F\left[(x_t, s_t)\right]$$

$\rightarrow$ *Estimation of Hidden States and of the Recurrent Function thus captures through deep learning entirely dynamical patterns of the system, i.e. full information on its dynamics and internal processes.*

Some questions for an application to Geography :

- Can the stationarity assumption be tackled through augmentation of system states ?
- Can heterogeneous and asynchronous data be used to bootstrap long time-series necessary for a correct estimation of the neural network ?

Next steps (until February 15th 2016)

- Theory exemplification, paper finalization [1w]
- Spatial Econometrics / Case study [0.5w]
- Cybergeo [0.5w]
- Wrap everything within a 1-year Memoire [1w]

References I

Abadie, A., Diamond, A., and Hainmueller, J. (2010).
Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program.
Journal of the American Statistical Association, 105(490).

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2013).
A review of feature selection methods on synthetic data.
Knowledge and information systems, 34(3):483–519.

Bretagnolle, A. (2009).
Villes et réseaux de transport : des interactions dans la longue durée, France, Europe, États-Unis.
Hdr, Université Panthéon-Sorbonne - Paris I.

References II

Brunsdon, C., Fotheringham, S., and Charlton, M. (1998).
Geographically weighted regression.
Journal of the Royal Statistical Society: Series D (The Statistician), 47(3):431–443.

Cottineau, C., Le Néchet, F., Le Texier, M., and Reuillon, R. (2015).
Revisiting some geography classics with spatial simulation.
In *Plurimondi. An International Forum for Research and Debate on Human Settlements*, volume 7.

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 III

Moeckel, R., Spiekermann, K., and Wegener, M. (2003).
Creating a synthetic population.
In *Proceedings of the 8th International Conference on Computers in Urban Planning and Urban Management (CUPUM)*.

Newman, M. E. (2003).
The structure and function of complex networks.
SIAM review, 45(2):167–256.

Offner, J.-M. (1993).
Les "effets structurants" du transport: mythe politique, mystification scientifique.
Espace géographique, 22(3):233–242.

References IV

Pritchard, D. R. and Miller, E. J. (2009).
Advances in agent population synthesis and application in an integrated land use and transportation model.
In *Transportation Research Board 88th Annual Meeting*, number 09-1686.

Pumain, D. (2012).
Urban systems dynamics, urban growth and scaling laws: The question of ergodicity.
In *Complexity Theories of Cities Have Come of Age*, pages 91–103. Springer.

Raimbault, J. (2016).
Calibration of a spatialized urban growth model.
Working Paper, draft at https://github.com/JusteRaimbault/CityNetwork/tree/master/Docs/Papers

References V

Reuillon, R., Leclaire, M., and Rey-Coyrehourcq, S. (2013).
Openmole, a workflow engine specifically tailored for the distributed exploration of simulation models.
Future Generation Computer Systems, 29(8):1981–1990.

Sanders, L. (1992).
Système de villes et synergétique.
Economica.

Schmitt, C. (2014).
Modélisation de la dynamique des systèmes de peuplement: de SimpopLocal à SimpopNet.
PhD thesis, Paris 1.

Shalizi, C. R. and Crutchfield, J. P. (2001).
Computational mechanics: Pattern and prediction, structure and simplicity.
Journal of statistical physics, 104(3-4):817–879.

References VI

Van den Bulcke, T., Van Leemput, K., Naudts, B., van Remortel, P., Ma, H., Verschoren, A., De Moor, B., and Marchal, K. (2006).
Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms.
BMC bioinformatics, 7(1):43.

Weibull, J. W. (1976).
An axiomatic approach to the measurement of accessibility.
Regional Science and Urban Economics, 6(4):357–379.

Xie, F. and Levinson, D. (2009).
How streetcars shaped suburbanization: a granger causality analysis of land use and transit in the twin cities.
Journal of Economic Geography, page lbp031.

References VII

Ye, X. (2011).
Investigation of underlying distributional assumption in nested logit model using copula-based simulation and numerical approximation.
Transportation Research Record: Journal of the Transportation Research Board, (2254):36–43.