

Génération de Données Synthétiques Corrélées

J. Raimbault^{1,2}

`juste.raimbault@parisgeo.cnrs.fr`

¹UMR CNRS 8504 Géographie-cités

²UMR-T IFSTTAR 9403 LVMT

Journées de Rochebrune 2016

*Nature des données, théorie de la donnée I : perspective
méthodologique*

Mercredi 20 janvier 2016

Un problème fondamental...

Expérience de pensée :

Les relations skieurs/snowboarders sont devenues ces derniers jours un réel problème de société. La station dispose d'une marge de manoeuvre très faible, par exemple :

- politique de zonage (spatial ou temporel) pour limiter les interactions
- campagnes de sensibilisations pour jouer sur paramètres individuels

Comment évaluer ces mesures de manière prospective ?

→ **Modéliser !**

(exemple incongru : utilisateurs de VLib dans [Raimbault, 2015] !)

... à la solution simple...

(Suite de l'expérience)

On suppose avoir construit (avec ou sans UML) et implémenté un modèle de simulation stochastique (agent ou non), qui étant donné l'organisation spatiale d'une station et une population pratiquant les sports d'hiver (paramètres laissés à votre imagination), simule le déroulement d'une journée de ski et produit des indicateurs de performance/satisfaction.

→ *L'intégration des actions possibles de la station dans le modèle permettrait d'évaluer leur impact sur l'amélioration des conflits ski/snow ?*

mais finalement complexe ?

- L'exploration intensive des modèles de simulation (comportement statistique, analyse de sensibilité, exploration de l'espace des paramètres, calibration, etc.) est essentielle pour en tirer de la connaissance [Rey-Coyrehourcq, 2015, Banos, 2013]
- Cela inclut la sensibilité *aux conditions initiales*, ex. : topographie et configuration de la station, composition de la population ; au premier ordre (distribution des données en elle-mêmes) mais aussi *au second ordre* (covariation des données ; ex. : corrélation entre les capacités des snowboarders à prendre un téléski et la haine des skieurs).

→ d'où la nécessité de générer des *jeux de données* artificiels, ou *données synthétiques*, contrôlées au différents ordres, pour une exploration complète du modèle.

Contexte

Def. : Des *données synthétiques* sont dans notre cas des sorties de modèles génératifs (et possiblement des entrées des modèles qui les utilisent).

Méthodologie utilisée dans de nombreux domaines, par ex. évaluation thérapeutique [Abadie et al., 2010], étude des systèmes territoriaux [Moeckel et al., 2003, Pritchard and Miller, 2009], apprentissage statistique [Bolón-Canedo et al., 2013] ou la bio-informatique [Van den Bulcke et al., 2006].

Peu répandu au second ordre : exemples spécifiques comme [Ye, 2011] pour choix discrets ; méthodes pouvant être interprétées comme telles : generation de réseaux complexes [Newman, 2003].

Méthode générique proposée

\vec{X}_i processus stochastique multidimensionnel, $\mathbf{X} = (X_{i,j})$ jeu de réalisations.

But : Générer une population statistique $\tilde{\mathbf{X}} = \tilde{X}_{i,j}$ telle que :

- 1 critère de proximité aux données : étant donné une précision ε et un indicateur f , $\|f(\mathbf{X}) - f(\tilde{\mathbf{X}})\| < \varepsilon$
- 2 contrôle de la structure de corrélation estimée : $\hat{\text{Var}} \left[(\tilde{X}_i) \right] = \Sigma R$ avec R fixé.

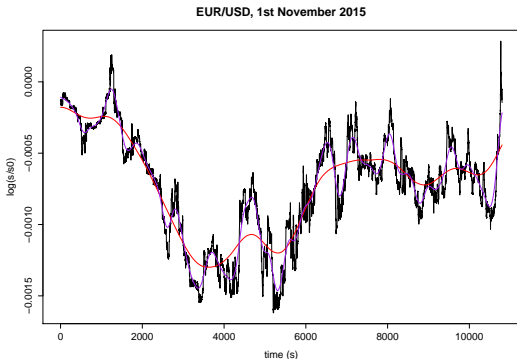
Application : Finance Quantitative

- Séries temporelles financières, signaux typiques de systèmes complexes hétérogènes et multiscalaires [Mantegna et al., 2000].
- Etude des corrélations entre actifs centrale à de nombreuses approches : Matrices aléatoires [Bouchaud and Potters, 2009], réseaux complexes [Bonanno et al., 2001] [Tumminello et al., 2005], dérivations théorique d'estimateurs spécifiques [Barndorff-Nielsen et al., 2011]
- Données synthétiques corrélées utilisées dans cas simples [Potiron and My validation de résultats théoriques, performance d'un modèle prédictif.

Formalisation

Réseau d'actif $(X_i(t))_{1 \leq i \leq N}$ interprétés comme signaux multiscalaires :
 $X_i = \sum_{\omega} X_i^{\omega}$. On notera $T_i^{\omega} = \sum_{\omega' \leq \omega} X_i^{\omega'}$.

Dynamique de Black-Scholes $dX = \sigma \cdot dW$ avec W processus de Wiener



Génération des données

Objectif : Générer \tilde{X}_i tel que avec $\omega_0 < \omega_1$,

- 1 $\text{Var}[\tilde{T}_i^{\omega_1}] = \Sigma R$ (covariance contrôlée à plus haute fréquence)
- 2 $T_i^{\omega \leq \omega_0} = \tilde{X}_i^{\omega \leq \omega_0}$ (critère de proximité au données)

→ si $dW_1 \perp\!\!\!\perp dW_2$, alors $W_2 = \rho_{12} W_1 + \sqrt{1 - \frac{\sigma_1^2}{\sigma_2^2}} \cdot \rho_{12} W_1 \perp\!\!\!\perp$ est tel que $\rho(dW_1, dW_2) = \rho_{12}$.

→ Construction des données par $\tilde{X}_i = T_i^{\omega_0} + W_i - \mathcal{F}_{\omega_0}[W_i]$ avec \mathcal{F}_{ω_0} filtre passe-bas avec $\omega_c = \omega_0$

Application

→ **Signaux** : *log-prix* et *log-retours*, définis par $X(t) := \log\left(\frac{S(t)}{S_0}\right)$ et $\Delta X(t) = X(t) - X(t-1)$, pour deux actifs du marché des devises (EUR/USD et EUR/GBP), sur 6 mois de juin 2015 à novembre 2015, filtrés à $\omega_m = 10\text{min}$

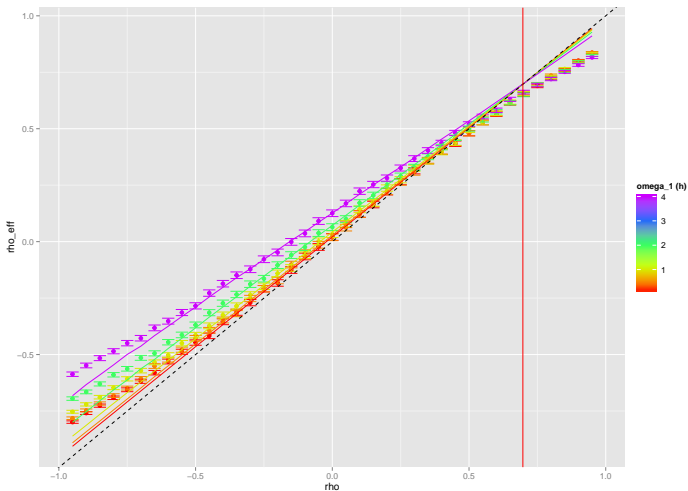
→ $\omega_0 = 24\text{h}$ et $\omega_1 = 30\text{min}, 1\text{h}, 2\text{h}$

→ Interférence entre différentes échelles

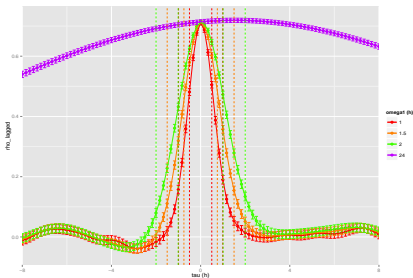
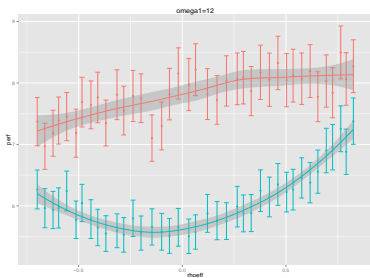
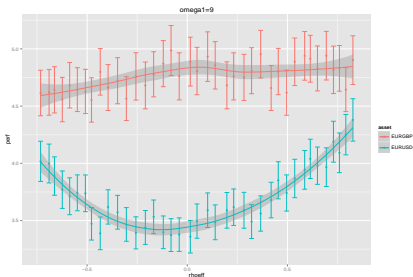
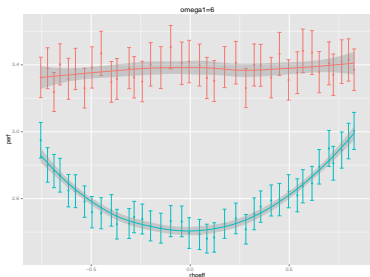
$$\rho_e \simeq [\varepsilon_1 \varepsilon_2 \rho_0 + \rho] \cdot \left[1 - \frac{1}{2} (\varepsilon_1^2 + \varepsilon_2^2) \right]$$

Résultats : Correlations effectives

Correlations simulées et correlations théoriques à différentes fréquences



Résultats : Performance d'un modèle prédictif



Données géographiques : Contexte et Objectif

- En géographie, génération de population synthétiques pour modèles agents [Pritchard and Miller, 2009].
- Génération de configurations spatiales synthétiques peu répandue (Régression Géo. Pondérée [Brunsdon et al., 1998] peut être interprétée de cette façon); pourtant essentielle même pour des modèles abstraits [Schmitt, 2014]
- [Cottineau et al., 2015] a récemment proposé d'estimer la sensibilité des modèles de simulation spatiaux à la configuration spatiale initiale (application au modèle de Schelling).
- Cas d'étude : relations ville/transports, complexe à cerner quantitativement [Offner, 1993, Bretagnolle, 2009] → modèle simple de morphogénèse de densité de population et de réseau de transport.

Modèle

Couplage simple entre

- Génération itérative d'une grille de densité par attachement préférentiel/diffusion [Raimbault, 2016], calibrée sur objectifs morphologiques sur grille européenne de densité.
- Génération heuristique de réseau conditionnelle à la densité :
 - Distribution d'un nombre fixé de centres par attachement préférentiel selon la distribution de densité
 - Percolation déterministe par plus proche voisins
 - Rupture des potentiels d'interaction

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

pour un nombre fixé de couples N_L tels que $V_{ij}(d_N)/V_{ij}(d_{ij})$ est minimal parmi $K \cdot N_L$ plus forts potentiels euclidiens ($K = 5$ fixé)

- Planarification

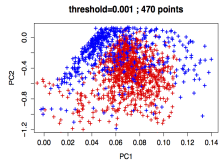
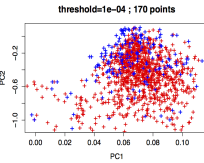
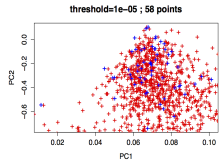
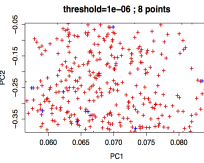
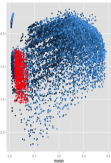
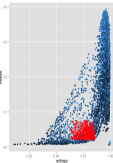
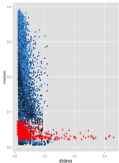
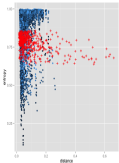
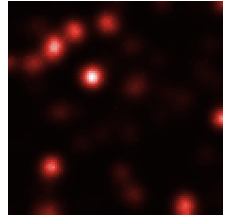
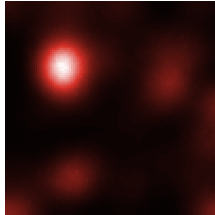
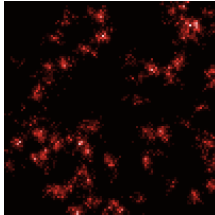
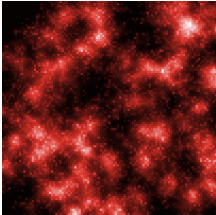
Indicateurs : morphologie [Le Néchet, 2015] (Moran, distance moyenne, entropie, hiérarchie) et réseau (centralité, largeur moyenne, vitesse, diamètre).

Implémentation et exploration

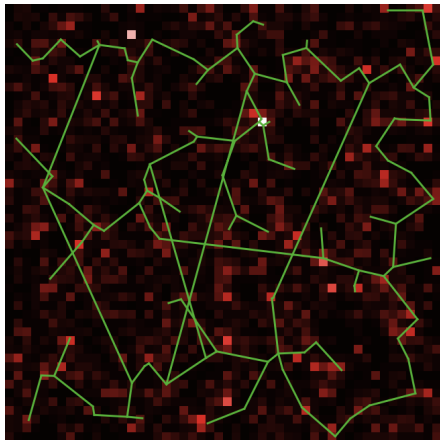
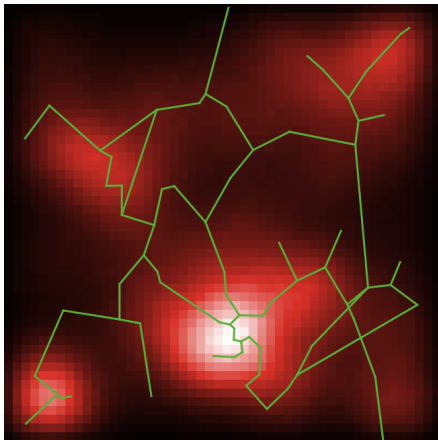
→ Couplage formel et opérationnel : implémentation modulaire (scala/NetLogo) encapsulée par OpenMo1e [Reuillon et al., 2013]

→ Exploration par calcul intensif sur grille de calcul via OpenMo1e : calibration du modèle de densité seul sur grille de densité européenne ($\sim 1.5 \cdot 10^6$ runs) ; exploration brute par criblage LHS pour corrélations faisables ($\sim 5 \cdot 10^4$ runs)

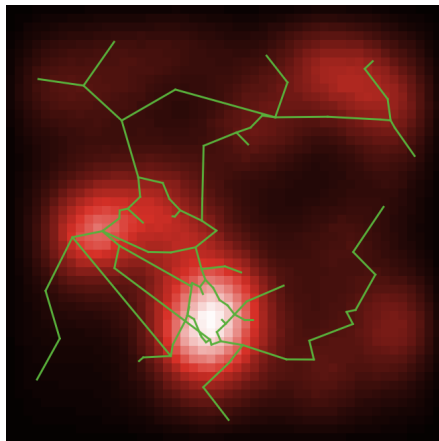
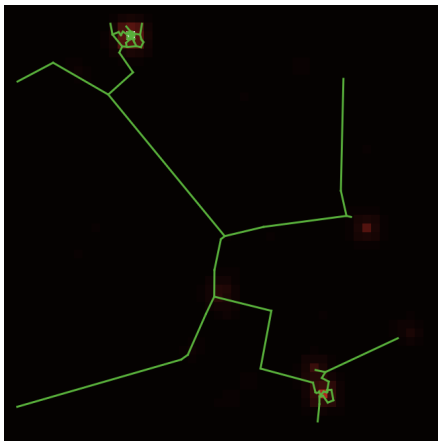
Résultats : Modèle de densité seul



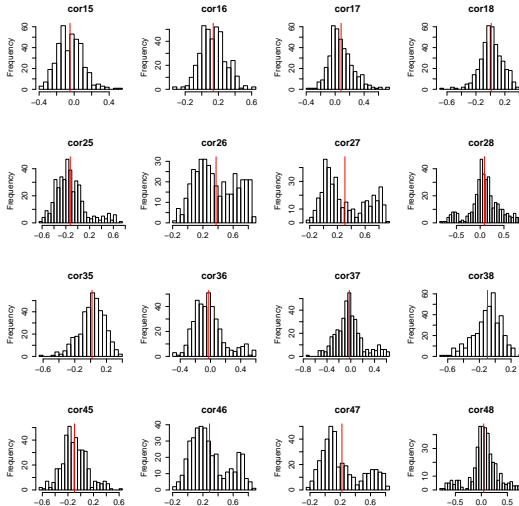
Résultats : exemples de configurations



Résultats : exemples de configurations

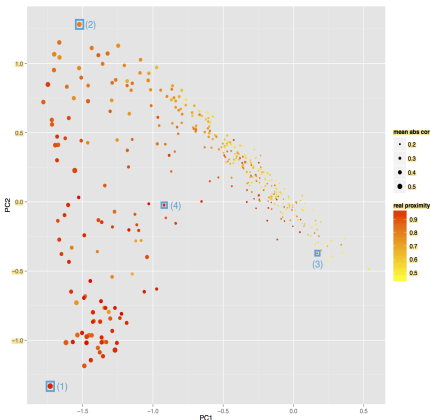
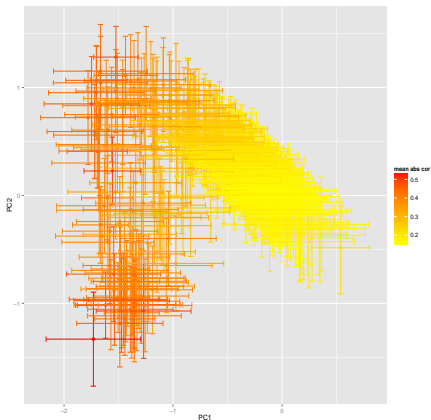


Résultats : corrélations croisées

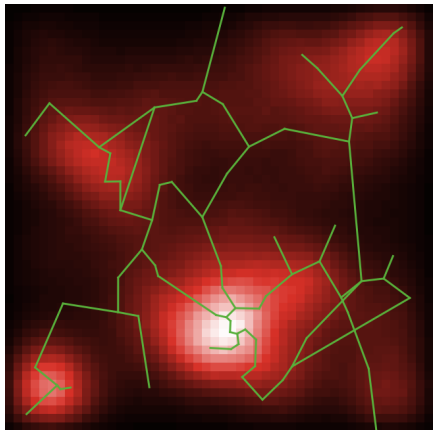


Résultats : corrélations faisables

Matrices moyennes dans un plan principal



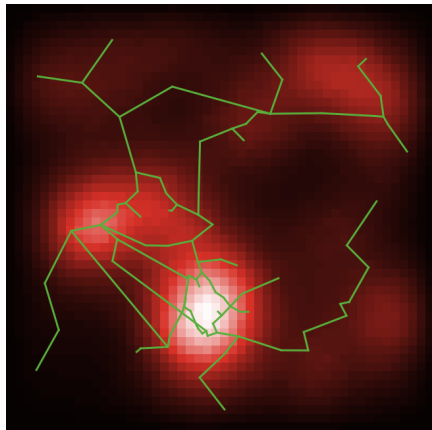
Résultats : exemples de corrélations



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

→ plus forte prégnance de la hiérarchie gravitaire dans (1)

$\gamma = 3.9, k_h = 0.7$ contre $\gamma = 1.07, k_h = 0.25$ pour (2)



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

Applications

- **Cas général** : Sensibilité d'un modèle aux structures de corrélations ; contrôle statistique sur corrélations.
- **Exemple géographique** :
 - 1 Calibration du modèle couplé, données de réseau routier (γ effets de bord !) → génération de données synthétiques correspondant à un système urbain donné → corrélations intrinsèque à une configuration (à comparer à corrélations estimées entre différents états : non-ergodicité des systèmes urbains [Pumain, 2012]).
 - 2 Corrélations dynamiques dans un modèle fortement couplé / corrélations spatio-temporelle dans un couplage fort spatial.
- Idées pour vos domaines ?

Généralisation

Possibilités de généralisation :

- Structures de dépendances non-linéaires
[Chicheportiche and Bouchaud, 2013]
- Contrôle des moments croisés à tout ordre ? Difficile car augmentation rapide du nombre de paramètres nécessaires ; peu d'intérêt ?

Conclusion

- Positionnement scientifique : multidisciplinarité (intégration horizontale des systèmes complexes) et modèles computationnels hétérogènes ou multi-échelle (intégration verticale)
- Méthode générique appliquée de deux façons : méthode hybride dérivation analytique/simulation ; exploration intensive d'un modèle de simulation.
- Retour au ski : type d'approche essentiel pour validation et application de modèles de simulation. Ski ou Snow ?

Réserve

Slides de réserve

Modèle de génération de réseau I

- 1 Un nombre fixé N_c de centres qui seront les premiers noeuds du réseau est distribué selon la distribution de densité, suivant une loi similaire à celle d'agrégation, i.e. la probabilité d'être distribué sur une case est $\frac{(P_i/P)^\alpha}{\sum (P_i/P)^\alpha}$. La population est ensuite répartie selon les zones de Voronoi des centres, un centre cumulant la population des cases dans son emprise.
- 2 Les centres sont connectés de façon déterministe par percolation entre plus proches clusters : tant que le réseau n'est pas connexe, les deux composantes connexes les plus proches au sens de la distance minimale entre chacun de leurs sommets sont connectées par le lien réalisant cette distance. On obtient alors un réseau arborescent.

Modèle de génération de réseau II

- 3 Le réseau est alors modulé par ruptures de potentiels afin de se rapprocher de formes réelles. Plus précisément, un potentiel d'interaction gravitaire généralisé entre deux centres i et j est défini par

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

où d peut être la distance euclidienne $d_{ij} = d(i, j)$ ou la distance par le réseau $d_N(i, j)$, $k_h \in [0, 1]$ un poids permettant de changer le rôle des population dans le potentiel, γ régissant la forme de la hiérarchie selon les valeurs des populations, r_g distance caractéristique de décroissance et d_0 paramètre de forme.

- 4 Un nombre $K \cdot N_L$ de nouveaux liens potentiels est pris comme les couples ayant le plus grand potentiel pour la distance euclidienne ($K = 5$ est fixé).
- 5 Parmi les liens potentiels, N_L sont effectivement réalisés, qui sont ceux ayant le plus faible rapport $V_{ij}(d_N)/V_{ij}(d_{ij})$: à cette étape seul l'écart entre distance euclidienne et distance par le réseau compte, ce rapport ne dépendant plus des populations et étant croissant en d_N à d_{ij} fixé.

Modèle de génération de réseau III

- 6 Le réseau est planarisé par création de noeuds aux intersections éventuelles créées par les nouveaux liens.

Paramètres

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

Indicateurs

On quantifie la forme urbaine et la forme du réseau, dans le but de moduler la corrélation entre ces indicateurs. La forme est définie par un vecteur $\vec{M} = (r, \bar{d}, \varepsilon, a)$ donnant auto-corrélation spatiale (indice de Moran), distance moyenne, entropie, hiérarchie (voir [Le Néchet, 2015] pour une définition précise de ces indicateurs). Les mesures de la forme du réseau $\vec{G} = (\bar{c}, \bar{l}, \bar{s}, \delta)$ sont, avec le réseau noté (V, E) ,

- Centralité moyenne \bar{c} , définie comme la moyenne de la *betweenness-centrality* (normalisée dans $[0, 1]$) sur l'ensemble des liens.
- Longueur moyenne des chemins \bar{l} définie par $\frac{1}{d_m} \frac{2}{|V| \cdot (|V| - 1)} \sum_{i < j} d_N(i, j)$ avec d_m distance de normalisation prise ici comme la diagonale du monde $d_m = \sqrt{2N}$.
- Vitesse moyenne [Banos and Genre-Grandpierre, 2012], qui correspond à la performance du réseau par rapport au trajet à vol d'oiseau, définie par $\bar{s} = \frac{2}{|V| \cdot (|V| - 1)} \sum_{i < j} \frac{d_{ij}}{d_N(i, j)}$.
- Diamètre du réseau $\delta = \max_{ij} d_N(i, j)$

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).

 Banos, A. (2013).

Pour des pratiques de modélisation et de simulation libérées en géographies et shs.




HDR. Université Paris, 1.

 Banos, A. and Genre-Grandpierre, C. (2012).

Towards new metrics for urban road networks : Some preliminary evidence from agent-based simulations.

In Agent-based models of geographical systems, pages 627–641.
Springer.

References II

-  Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N. (2011).
Multivariate realised kernels : consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading.
Journal of Econometrics, 162 :149–169.
-  Bolón-Canedo, V., Sánchez-Marño, N., and Alonso-Betanzos, A. (2013).
A review of feature selection methods on synthetic data.
Knowledge and information systems, 34(3) :483–519.
-  Bonanno, G., Lillo, F., and Mantegna, R. N. (2001).
Levels of complexity in financial markets.
Physica A Statistical Mechanics and its Applications, 299 :16–27.

References III



Bouchaud, J. P. and Potters, M. (2009).
Financial Applications of Random Matrix Theory : a short review.
ArXiv e-prints.





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



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.

References IV

-  Chicheportiche, R. and Bouchaud, J.-P. (2013).
A nested factor model for non-linear dependences in stock returns.
arXiv preprint arXiv :1309.3102.
-  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 V



Mantegna, R. N., Stanley, H. E., et al. (2000).

An introduction to econophysics : correlations and complexity in finance, volume 9.

Cambridge university press Cambridge.



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.

References VI



Offner, J.-M. (1993).

Les "effets structurants" du transport : mythe politique, mystification scientifique.

Espace géographique, 22(3) :233–242.



Potiron, Y. and Mykland, P. (2015).

Estimation of integrated quadratic covariation between two assets with endogenous sampling times.

arXiv preprint arXiv :1507.01033.



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.

References VII



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. (2015).

User-based solutions for increasing level of service in bike-sharing transportation systems.

In *Complex Systems Design & Management*, pages 31–44. Springer.






Raimbault, J. (2016).

Calibration of a spatialized urban growth model.

Working Paper, draft at <https://github.com/JusteRaimbault/CityNetwork/tree/master/Docs/Papers/Dens>

References VIII

-  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.
-  Rey-Coyrehourcq, S. (2015).
Une plateforme intégrée pour la construction et Une plateforme intégrée pour la construction et l'évaluation de modèles de simulation en géographie.
PhD thesis, Université Paris 1 Panthéon-Sorbonne.
-  Schmitt, C. (2014).
Modélisation de la dynamique des systèmes de peuplement : de SimpopLocal à SimpopNet.
PhD thesis, Paris 1.

References IX



Tumminello, M., Aste, T., Di Matteo, T., and Mantegna, R. N. (2005).

A tool for filtering information in complex systems.

Proceedings of the National Academy of Sciences of the United States of America, 102 :10421–10426.



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.

References X



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.