Génération de Données Synthétiques Corrélées

Actes des Journées de Rochebrune 2016

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

date

Abstract

1 Introduction

L'utilisation de données synthétiques, au sens de populations statistiques d'individus générées aléatoirement sous la contrainte de reproduire certaines caractéristiques du système étudié, est une pratique méthodologique largement répandue dans de nombreuses disciplines, et particulièrement pour des problématiques liées aux systèmes complexes, telles que par exemple l'évaluation thérapeutique [Abadie et al., 2010], l'étude des systèmes territoriaux [Moeckel et al., 2003, Pritchard and Miller, 2009], l'apprentissage statistique [Bolón-Canedo et al., 2013] ou la bio-informatique [Van den Bulcke et al., 2006]. Il peut s'agir d'une désagrégation par création d'une population au niveau microscopique présentant des caractéristiques macroscopiques données, ou bien de la création de nouvelles populations au même niveau d'agrégation qu'un échantillon donné avec un critère de ressemblance aux données réelles. Le niveau de ce critère peut

Les intérêts de ces méthodes sont directement liés

Si le premier ordre est bien maitrisé, il n'a à notre connaissance pas été proposé de méthode systématique permettant un contrôle au second ordre,c'est à dire où la structure de correlation estimée sur les données générées est maitrisée. Nous proposons une telle méthode ainsi que son application à deux exemples de systèmes complexes dans des domaines relativement éloignés.

La suite de l'article est organisée de la façon suivante :

2 Formalisation de la méthode

L'ensemble des méthodologies mentionnées en introduction sont trop variées pour être résumées par un même formalisme. Nous proposons ici une formulation générique ne dépendant pas du domaine d'application, ciblée sur le contrôle de la structure de correlation des données synthétiques.

Soit un processus stochastique multidimensionnel $\vec{X}_I$ (l'ensemble d'indexation pouvant être par exemple le temps dans le cas de séries temporelles, l'espace, ou l'indexation). On se propose, à partir d'un jeu de réalisations $\mathbf{X} = (X_{i,j})$, de générer une population statistique $\tilde{\mathbf{X}} = \tilde{X}_{i,j}$ telle que

- d'une part un certain critère de proximité aux données est vérifié, i.e. étant donné une précision ε et un indicateur f, $\|f(\mathbf{X}) - f(\tilde{\mathbf{X}})\| < \varepsilon$
- d'autre part le niveau de correlation est controlé, i.e. étant donné une matrice fixant une structure de covariance R, $\mathrm{Var}\left[(\tilde{X}_i)\right] = R$, où la matrice de variance/covariance est estimée sur la population synthétique.

3 Applications

3.1 Application : séries temporelles financières

3.1.1 Contexte

Un premier domaine d'application proposé pour notre méthode est celui des séries temporelles financières, signaux typiques de systèmes complexes hétérogènes et multiscalaires [Mantegna et al., 2000] et pour lesquels les corrélations ont fait l'objet d'abondants travaux. Ainsi, l'application de la théorie des matrices aléatoires peut permettre de débruiter, ou du moins d'estimer la part de signal noyée dans le bruit, une matrice de correlations pour un grand nombre d'actifs échantillonnés à faible fréquence (retours journaliers par exemple) [Bouchaud and Potters, 2009]. De même, l'analyse de réseaux complexes construits à partir des corrélations, selon des méthodes type arbre couvrant minimal [Bonanno et al., 2001] ou des extensions raffinées pour cette application précise [Tumminello et al., 2005], ont permis d'obtenir des résultats prometteurs, tels la reconstruction de la structure économique des secteurs d'activités. A haute fréquence, l'estimation précise de paramètres d'interdépendance dans le cadre d'hypothèses fixées sur la dynamique, fait l'objet d'importants travaux théoriques dans un but de raffinement théorique des estimateurs [Barndorff-Nielsen et al., 2011]. Les résultats théoriques doivent alors être testés sur des jeux de données synthétiques, qui permettent de contrôler un certain nombre de paramètres et de s'assurer qu'un effet prédit par la théorie est bien observable *toutes choses égales par ailleurs*. Par exemple, [Potiron and Mykland, 2015] dérive une correction du biais de l'estimateur de *Hayashi-Yoshida* qui est un estimateur de la covariance de deux browniens corrélés à haute fréquence dans le cas de temps d'observation asynchrones, par démonstration d'un théorème de la limite centrale pour un modèle généralisé endogénéisant les temps d'observations. La confirmation empirique de l'amélioration de l'estimateur est alors obtenue sur un jeu de données synthétiques à un niveau de corrélation fixé.

3.1.2 Formalisation

Cadre Considérons un réseau d'actifs $(X_i(t))_{1\leq i\leq N}$ échantillonnés à haute fréquence (typiquement 1s). On se place dans un cadre multi-scalaire (utilisé par exemple dans les approches par ondelettes [Ramsey, 2002] ou analyses multifractales du signal [Bouchaud et al., 2000]) pour interpréter les signaux observés comme la superposition de composantes à des multiples échelles temporelles : $X_i = \sum_\omega X_i^\omega$. Prédire l'évolution d'une composante à une échelle donnée est alors un problème caractéristique de l'étude des systèmes complexes, pour lequel l'enjeu est l'identification de régularités et leur distinction des composantes considérées comme stochastiques en comparaison. Dans un souci de simplicité, on représente un tel processus par un modèle de prédiction de tendance à une échelle temporelle ω_1 donnée, formellement un estimateur $M_{\omega_1} : (X_i(t'))_{t'<t} \mapsto \hat{X}_i(t)$ dont l'objectif est la minimisation de l'erreur sur la tendance réelle $\|X_i^\omega\|$. Dans le cas d'estimateurs auto-regressifs multivariés, la performance dépendra entre autre des correlations respectives entre actifs et il est alors intéressant d'utiliser la méthode pour évaluer celle-ci en fonction de niveaux de correlation à plusieurs échelles. On assume une dynamique de Black-Scholes [Jarrow, 1999] pour les actifs, i.e. $dX = \sigma \cdot dW$ avec W processus de Wiener, ce qui permettra d'obtenir facilement des niveaux de correlation voulus.

Génération des données Il est alors aisé de générer $\tilde{X}_i$ tel que $\mathrm{Var}\left[\tilde{X}_i^{\omega_0}\right] = \Sigma R$ (Σ variance estimée et R matrice de corrélation fixée), par la simulation de processus de Wiener au niveau de corrélation fixé et tel que $X_i^{\omega<\omega_0} = \tilde{X}_i^{\omega<\omega_0}$ (critère de proximité au données : les composantes à plus basse fréquence sont identiques). En effet, si $dW_1 \perp\!\!\!\perp dW_1^{\perp\!\!\!\perp}$, alors $W_2 = \rho_{12}W_1 + \sqrt{1-\frac{\sigma_1^2}{\sigma_2^2}\cdot\rho_{12}^2}W_1^{\perp\!\!\!\perp}$ est tel que $\rho(dW_1, dW_2) = \rho_{12}$. Les signaux suivants sont construits de la même manière par orthonormalisation de Gram. On isole alors la composante de la première fréquence $\omega_1 < \omega_0$ par filtrage, et on reconstruit les signaux synthétiques par $\tilde{X}_i = [\sum_{\omega<\omega_1} X_i^\omega] + \tilde{X}_i^{\omega_0}$.

3.1.3 Implémentation et Résultats

Méthodologie La méthode est testée sur un exemple de deux actifs du marché des devises (EUR/USD et EUR/GBP), sur une période de 6 mois de juin 2015 à novembre 2015. Le nettoyage des données[1], originellement échantillonnées à l'ordre de la seconde, consiste dans un premier temps à la détermination du support temporel commun maximal (les séquences manquantes étant alors ignorées, par translation verticale des séries, i.e. $S(t) := S(t)\cdot\frac{S(t_n)}{S(t_{n-1})}$ lorsque t_{n-1}, t_n sont les extrémités du "trou" et $S(t)$ la valeur de l'actif, ce qui revient à garder la contrainte d'avoir des retours à pas de temps similaires entre actifs). On étudie alors les *log-prix* et *log-retours*,

[1] obtenues depuis `http://www.histdata.com/`, sans licence spécifiée, les données nettoyées et filtrées à ω_m uniquement sont mises en accessibilité pour respect du copyright.

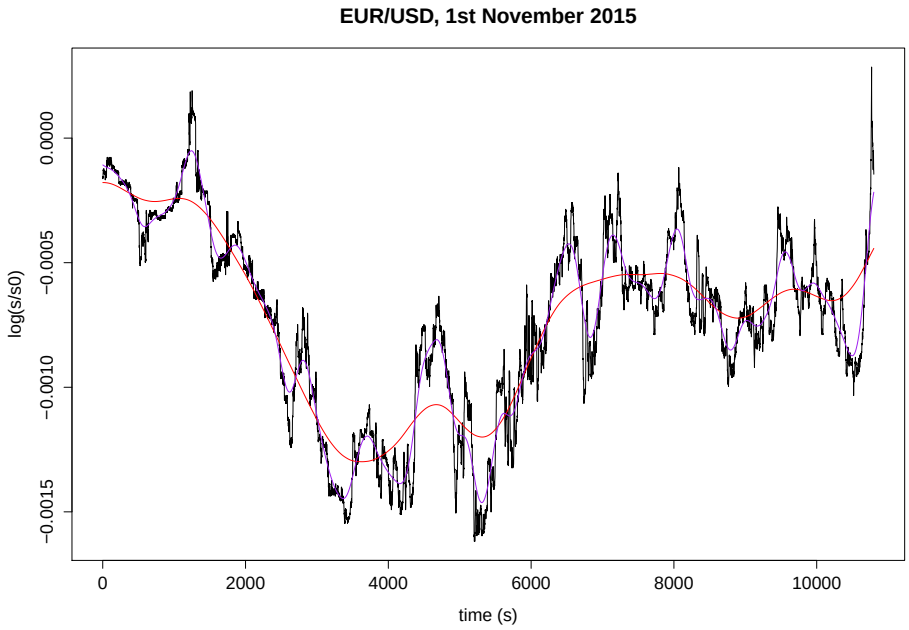


Figure 1: **Exemple de la structure multi-scalaire du signal qui sert de base à la construction des signaux synthétiques |** Les *log-prix* sont représentés sur environ 3h pour la journée du 1er novembre 2015 pour l'actif EUR/USD, ainsi que les tendances à 10min (violet) et à 30min.

définis par $X(t) := \log \frac{S(t)}{S_0}$ et $\Delta X(t) = X(t) - X(t-1)$. Les données brutes sont filtrées à une fréquence $\omega_m = 10\text{min}$ (qui sera la fréquence maximale d'étude) pour un souci de performance computationnelle. On fixe $\omega_0 = 12\text{h}$ et on se propose de construire des données synthétiques aux fréquences $\omega_1 = 30\text{min}, 1\text{h}, 2\text{h}$. Il est crucial de noter l'interférence entre les fréquences ω_0 et ω_1 dans le signal construit : la correlation effectivement estimée est

$$\rho_e = \rho\left[\Delta\tilde{X}_1, \Delta\tilde{X}_2\right] = \rho\left[\Delta X_1^{\omega_0} + \sum_{\omega_0<\omega<\omega_1} \Delta\tilde{X}_1^{\omega}, \Delta X_2^{\omega_0} + \sum_{\omega_0<\omega<\omega_1} \Delta\tilde{X}_2^{\omega}\right]$$

ce qui conduit à dériver dans la limite raisonnable $\sigma_1 \gg \sigma_0$ (fréquence fondamentale suffisamment basse), lorsque $\text{Cov}\left[\tilde{X}_i^{\omega}, X_j^{\omega_0}\right] = 0$ pour tous $i, j, \omega > \omega_0$, et les retours d'espérance nulle à toutes échelles, en notant $\rho_0 = \rho\left[\Delta X_1^{\omega_0}, \Delta X_2^{\omega_0}\right]$, $\rho = \rho\left[\sum_{\omega_0<\omega<\omega_1} \tilde{X}_1^{\omega}, \sum_{\omega_0<\omega<\omega_1} \tilde{X}_2^{\omega}\right]$, et $\varepsilon_i = \frac{\sigma(X_i^{\omega_0})}{\sigma\left(\sum_{\omega_0<\omega<\omega_1} \tilde{X}_i^{\omega}\right)}$, la correction sur la correlation effective due aux interférences : le correlation effective est alors au premier ordre :

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

Implémentation L'implémentation est faite en language R, utilisant en particulier la bibliothèque `MTS` [Tsay, 2015] pour les modèles de séries temporelles. Les données nettoyées et le code source sont disponibles de manière ouverte sur le dépôt `git` du projet[2].

Résultats La figure 2 donne les correlations effectives calculées sur les données synthétiques.

Pour des valeurs standard des paramètres (par exemple pour $\omega_0 = 24\text{h}$, $\omega_1 = 2\text{h}$ et $\rho = -0.5$, on a $\rho_0 \simeq 0.71$ et $\varepsilon_i \simeq 0.3$ et ainsi $|\rho_e - \rho| \simeq 0.13$)

[2] at `https://github.com/JusteRaimbault/SynthAsset`

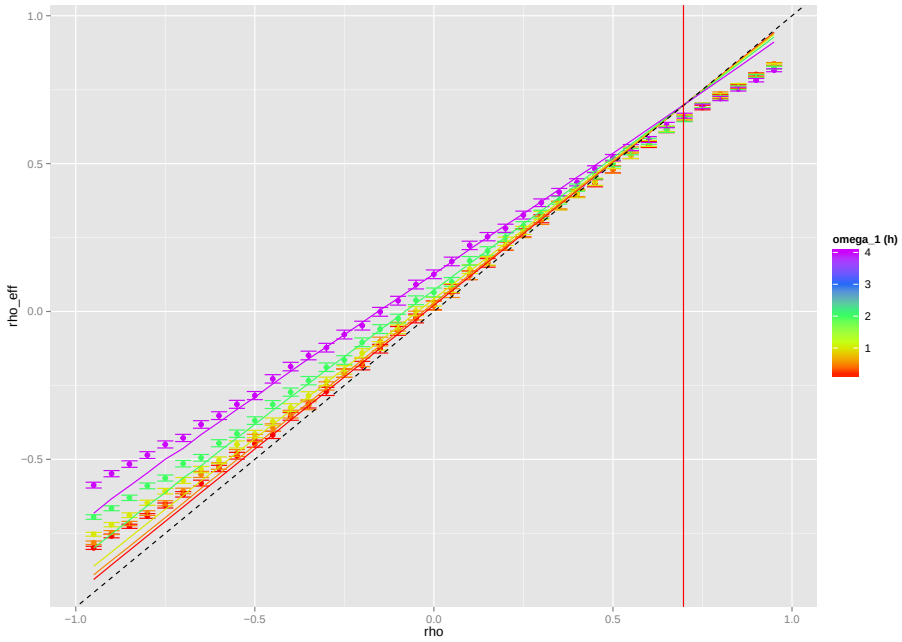


Figure 2: **Correlations effectives obtenues sur les données synthétiques |** Les points représentent les correlations estimées sur une génération d'un jeu de données synthétiques correspondant aux 6 mois de juin à novembre 2015 (barres d'erreurs obtenue par méthode de Fisher standard) ; l'échelle de couleur donne la fréquence de filtrage $\omega_1 = 10\text{min}, 30\text{min}, 1\text{h}, 2\text{h}, 4\text{h}$; les courbes sont les valeurs théoriques de ρ_e obtenues par 1 avec les volatilités estimées (diagonale en pointillés pour référence) ; l'abscisse de la ligne rouge est la valeur théorique telle que $\rho = \rho_e$ avec les valeurs moyennes de ε_i sur l'ensemble des points. On observe dans les fortes correlations une déviation des valeurs corrigées, qui peut être dû aux hypothèses d'indépendance ou d'espérance nulle non vérifiées. La dissymétrie de la courbe est causée par la forte valeur positive de $\rho_0 \simeq 0.71$.

Figure 3

3.2 Application : données géographiques de densité et de réseaux

3.2.1 Contexte

En géographie, l'utilisation de données synthétiques est plus généralement axée vers l'utilisation de population synthétiques au sein de modèles basés agents (mobilité, modèles $LUTI$) [Pritchard and Miller, 2009]. Il a récemment été proposé de contrôler systématiquement les effets de la configuration spatiale sur le comportement de modèles de simulation spatialisés [Cottineau et al., 2015b], méthodologie pouvant être interprétée comme un contrôle par données statistiques spatiales. L'enjeu est de pouvoir alors distinguer effets propres dus à la dynamique intrinsèque du modèle, d'effet particuliers dus à la structure géographique du cas d'application. Celui-ci est crucial pour la validation des conclusions issues des pratiques de modélisation et simulation en géographie quantitative.

3.2.2 Formalisation

Dans notre cas, nous proposons de générer des systèmes de villes représentés par une densité spatiale de population $d(\vec{x})$ et la donnée d'un réseau de transport $n(\vec{x})$, représenté de façon simplifiée, pour lesquels on serait capable de contrôler les correlations entre mesures morphologiques de la densité urbaine et caractéristiques du réseau. Nous proposons un couplage *simple*

Modèle de densité L'utilisation d'un modèle D type aggrégation-diffusion [Batty, 2006] permet de générer une distribution discrete de densité. Dans [], une généralisation de ce modèle est calibré pour des objectifs morphologiques M (entropie, hiérarchie, auto-corrélation spatiale, distance moyenne) contre les valeurs réelles calculées sur l'ensemble des grilles de taille 50km extraites de la grille européenne de densité [EUROSTAT, 2014]. Plus précisément, le modèle fonctionne de manière itérative de la façon suivante. Une grille initialement vide de côté N, est représentée par la données des populations $(P_i(t))_{1\leq i\leq N^2}$. A chaque pas de temps, jusqu'à ce que la population atteigne une valeur fixée P_m,

- la population totale $P(t)$ est augmentée d'un nombre fixé N_G (taux de croissance), suivant un attachement préférentiel tel que $\mathbb{P}[P_i(t+1) = P_i(t)+1 | P(t+1) = P(t)+1] = \frac{(P_i(t)/P(t))^\alpha}{\sum (P_i(t)/P(t))^\alpha}$
- une diffusion d'une fraction β de la population aux 4 plus proches voisins est effectuée n_d fois

Les deux processus antagonistes de concentration et d'étalement urbain sont capturés par le modèle, ce qui permet de reproduire assez fidèlement un grand nombre de morphologies existantes.

Modèle de réseau D'autre part, on est capable de générer par un modèle N un réseau de transport planaire à une échelle équivalente, étant donné une distribution de densité. La génération du réseau étant conditionnée à la donnée de la densité, les estimateurs des indicateurs de réseau seront nécessairement conditionnels d'une part, et d'autre part les formes urbaines et du réseau devraient nécessairement être correlées, les processus n'étant pas indépendants. La nature et la modularité de ces correlations selon la variation des paramètres des modèles restent à déterminer par l'exploration du modèle couplé.

La procédure de génération heuristique de réseau est la suivante :

- Une nombre fixé N_c

On distribue un nombre fixé de noeuds de manière aléatoire en suivant la loi de probabilité spatiale donnée par les valeurs de densité, puis un algorithme déterministe de connexification permet d'obtenir un réseau arborescent. Le réseau est ensuite étendu par la création de boucles locales dans un rayon de voisinage r_l ainsi que de raccourcis à une plus grande échelle r_g, aléatoirement selon un processus de rupture des potentiels gravitaires[3].

Espace des paramètres

3.2.3 Implémentation

Le couplage des modèles génératifs est effectué à la fois au niveau formel et au niveau opérationnel

[3] Notons que ce choix est heuristique, et que d'autres types de modèles type réseau biologique auto-généré [Tero et al., 2006] par exemple pourraient également être envisagés, dans l'idée d'une architecture modulaire où le choix entre différentes implémentations d'une brique fonctionnelle peut être vue comme méta-paramètre [Cottineau et al., 2015a].

3.2.4 Résultats

L'étude du modèle de densité est développée dans [].

A densité fixée, les premières explorations de l'espace des paramètres du modèle de réseau synthétique suggèrent une assez bonne flexibilité sur des indicateurs globaux G (diamètre, longueur cumulée, centralité moyenne, degré moyen). L'exploration systématique via le logiciel OpenMole [Reuillon et al., 2013] par calcul intensif est un travail en cours, ainsi que la calibration contre les mesures réelles calculées sur l'ensemble de l'Europe sur des zones identiques au modèle de densité, via les données de réseau routier d'OpenStreetMap. La connaissance très fine ainsi obtenue du comportement de N (distribution statistiques sur une grille fine de l'espace des paramètres à trois dimensions), devrait permettre, étant donné une population de configuration de densités $\tilde{D}$, de déterminer via $N^{<-1>}$ une population de réseau $\tilde{N}$ telle que $\mathrm{Cov}[M, G]$ a une structure fixée (via la détermination de la valeur des paramètres à utiliser pour chaque individu de $\tilde{D}$). On pourra éventuellement appliquer des algorithmes plus fins d'exploration pour atteindre des configurations exceptionelles réalisant un niveau de corrélation voulu [Chérel et al., 2015]. Les indicateurs globaux devraient ainsi être corrélés à un niveau contrôlé, tandis que les densités et réseaux restent cohérents dans l'espace de par la forme du réseau, conditionnelle à la distribution de densité. Les applications géographiques potentielles de cette implémentation de la méthode incluent le contrôle statistique de l'effet des corrélations entre ville et réseaux sur des modèles de simulation spatiaux par exemple.

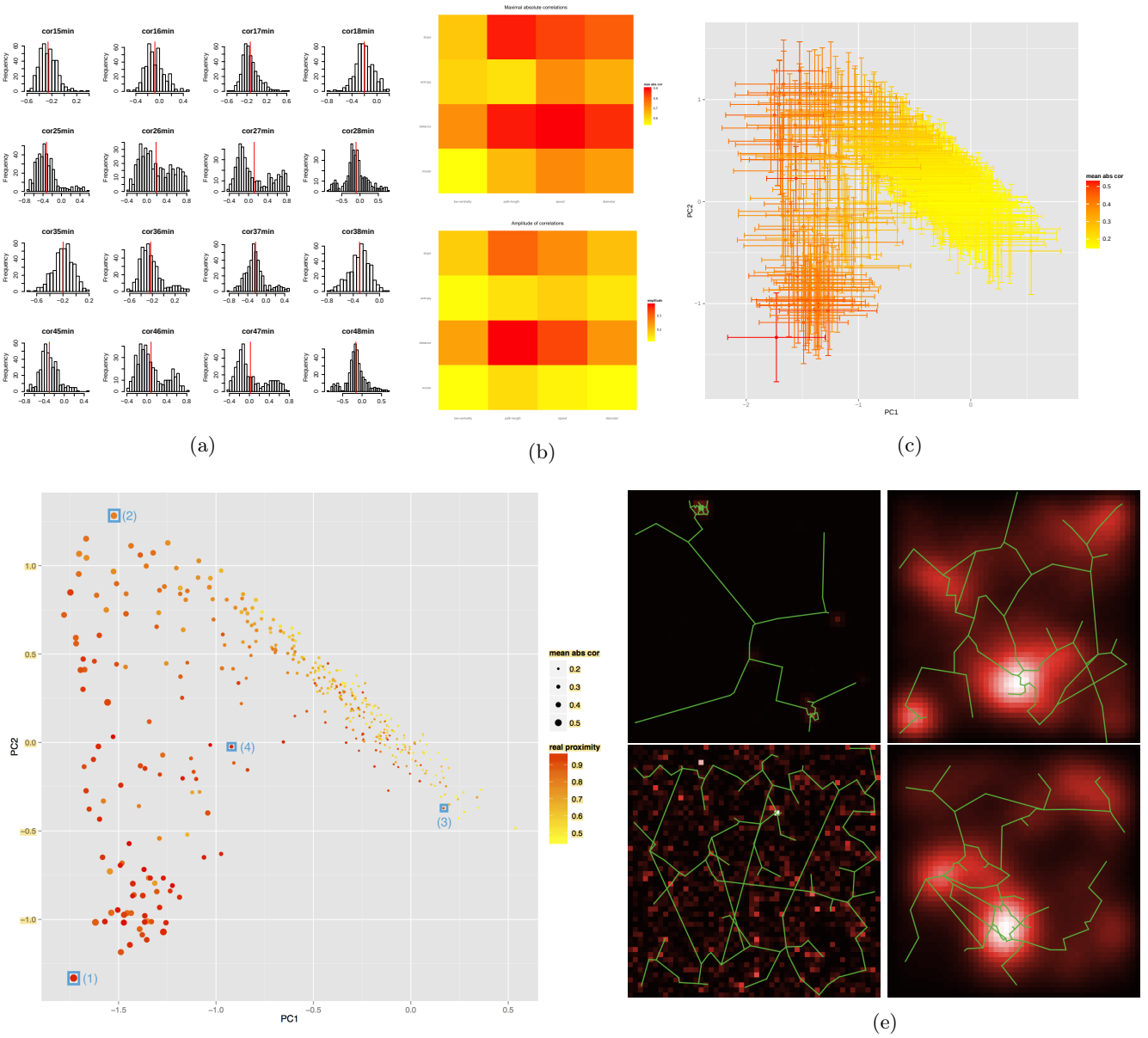
4 Discussion

4.1 Domaines potentiels d'application

4.2 Positionnement

5 Conclusion

On a ainsi proposé une méthode abstraite de génération de données synthétiques corrélées à un niveau contrôlé. Son implémentation partielle dans deux domaines très différents montre sa flexibilité et l'éventail des applications potentielles. De manière générale, il est essentiel de généraliser de telles pratiques de validation systématique de modèles par étude statistique, en particulier pour les modèles agents pour lesquels la question de la validation reste encore relativement ouverte.



(a) (b) (c)

(d) (e)

Figure 4: **Exploration de l'espace des corrélations entre forme urbaine et réseau | (a)** Distribution des corrélations croisées entre les vecteurs $\vec{M}$ des indicateurs morphologiques (dans l'ordre index de moran, distance moyenne, entropie, hiérarchie) et $\vec{N}$ des mesures de réseau (centralité, longueur moyenne, vitesse, diamètre). **(b)** Amplitude des corrélations, définie comme $a_{ij} = \max_k \rho_{ij}^{(k)} - \min_k \rho_{ij}^{(k)}$, et correlation absolue maximale, $c_{ij} = \max_k ||$. **(c)** Représentation des matrices dans un plan principal obtenu par Analyse en Composantes Principales sur la population des matrices (variances cumulées: PC1=38%, PC2=68%). Les barres d'erreur sont calculées initialement par des intervalles de confiance à 95% sur chaque élément de la matrice (par une méthode de Fisher asymptotique standard), puis les bornes supérieures sont prises dans le plan principal. L'échelle de couleur donne la correlation absolue moyenne sur l'ensemble des variables. **(d)**

References

[Abadie et al., 2010] 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).

[Barndorff-Nielsen et al., 2011] 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.

[Batty, 2006] Batty, M. (2006). Hierarchy in cities and city systems. In *Hierarchy in natural and social sciences*, pages 143–168. Springer.

[Bolón-Canedo et al., 2013] 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.

[Bonanno et al., 2001] 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.

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

[Bouchaud et al., 2000] Bouchaud, J.-P., Potters, M., and Meyer, M. (2000). Apparent multifractality in financial time series. *The European Physical Journal B-Condensed Matter and Complex Systems*, 13(3):595–599.

[Chérel et al., 2015] Chérel, G., Cottineau, C., and Reuillon, R. (2015). Beyond corroboration: Strengthening model validation by looking for unexpected patterns. *PLoS ONE*, 10(9):e0138212.

[Cottineau et al., 2015a] Cottineau, C., Chapron, P., and Reuillon, R. (2015a). An incremental method for building and evaluating agent-based models of systems of cities.

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

[EUROSTAT, 2014] EUROSTAT (2014). Eurostat geographical data.

[Jarrow, 1999] Jarrow, R. A. (1999). In honor of the nobel laureates robert c. merton and myron s. scholes: A partial differential equation that changed the world. *The Journal of Economic Perspectives*, pages 229–248.

[Mantegna et al., 2000] 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 et al., 2003] 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)*.

[Potiron and Mykland, 2015] Potiron, Y. and Mykland, P. (2015). Estimation of integrated quadratic covariation between two assets with endogenous sampling times. *arXiv preprint arXiv:1507.01033*.

[Pritchard and Miller, 2009] 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.

[Ramsey, 2002] Ramsey, J. B. (2002). Wavelets in economics and finance: Past and future. *Studies in Nonlinear Dynamics & Econometrics*, 6.

[Reuillon et al., 2013] 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.

[Tero et al., 2006] Tero, A., Kobayashi, R., and Nakagaki, T. (2006). Physarum solver: a biologically inspired method of road-network navigation. *Physica A: Statistical Mechanics and its Applications*, 363(1):115–119.

[Tsay, 2015] Tsay, R. S. (2015). *MTS: All-Purpose Toolkit for Analyzing Multivariate Time Series (MTS) and Estimating Multivariate Volatility Models*. R package version 0.33.

[Tumminello et al., 2005] 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 et al., 2006] 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.