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

TBW

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 Application

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 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.

3.1.2 Déclinaison de la méthode

Considérons un réseau d'actifs $(X_i(t))_{1\leq i\leq N}$ échantillonés à haute fréquence (typiquement 1s), vus comme la superposition de signaux à des multiples échelles temporelles : $X_i = \sum_\omega X_i^\omega$ sur lesquels est appliqué un modèle de prédiction de tendance à une échelle temporelle ω_0 donnée, représenté formellement comme un estimateur $M_{\omega_0} : (X_i) \mapsto \hat{X}_i$ dont l'objectif est la minimisation de l'erreur sur la tendance réelle X_i^ω. Dans le cas d'estimateurs auto-regressifs multivariés, la performance dépendra entre autre des correlations respectives entre actifs et on peut alors l'estimer en appliquant la méthode. On assume une dynamique de Black-Scholes pour les actifs : $dX = \sigma \cdot dW$ avec W processus de Wiener. 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 $W_1 \perp\!\!\!\perp W_1^{\perp\!\!\!\perp}$, alors $W_2 = \rho_{12}W_1 + \sqrt{1-\rho_{12}^2}W_1^{\perp\!\!\!\perp}$ est tel que $\rho(W_1, W_2) = \rho_{12}$. 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

La méthode est testée sur un exemple de deux actifs du marché des devises (EUR/USD et EUR/GBP), sur la période de l'année 2014, permettant d'obtenir un bruit sur les corrélations finales négligeable. Le test sur une dimension plus grande doit encore être implémenté, ainsi que l'application à l'étude de la performance de modèle prédictif.

3.1.4 Résultats

3.2 Application : données géographiques de densité et de réseaux

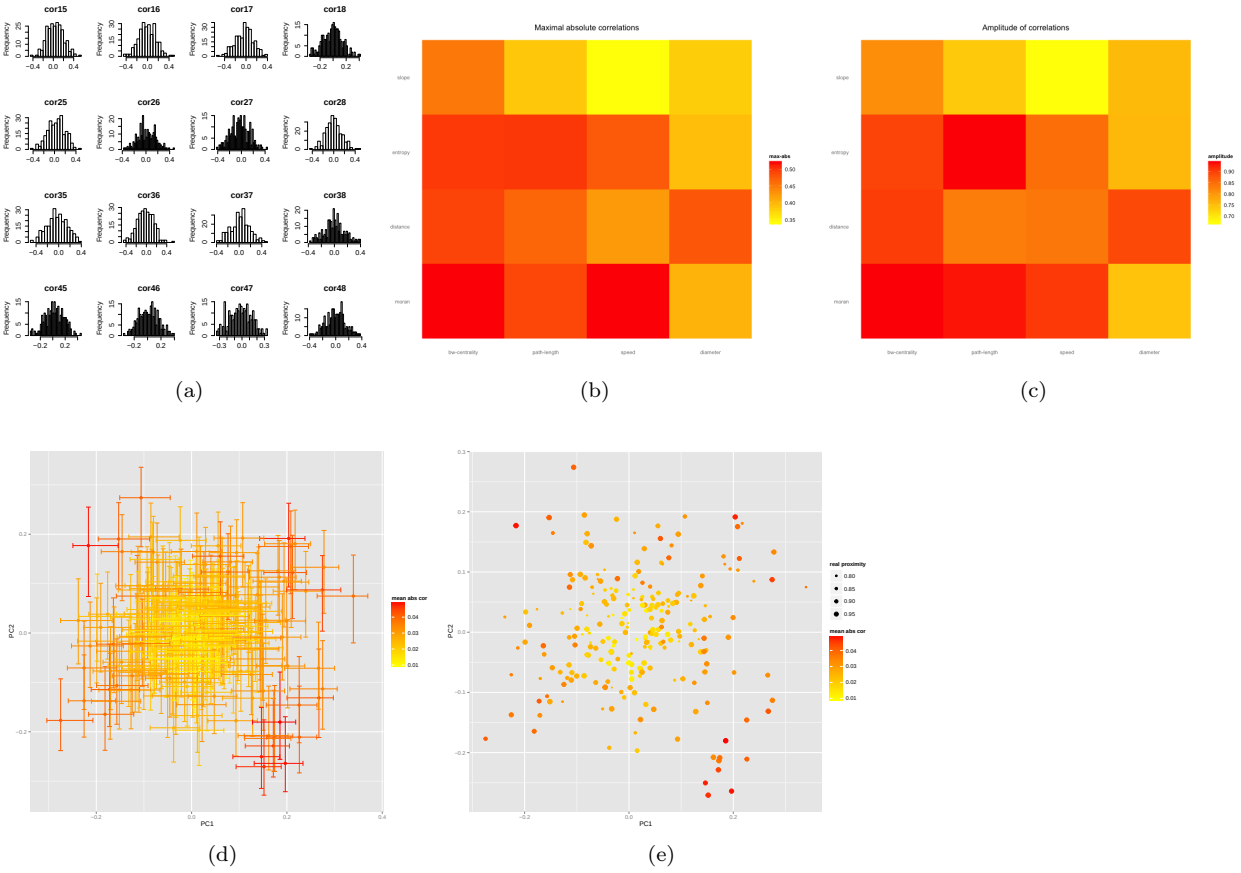
3.2.1 Contexte

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.

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.

L'utilisation d'un modèle D type aggrégation-diffusion [Batty, 2006] permet de générer une distribution discrete de densité. Le modèle est calibré pour des objectifs morphologiques M (entropie, hiérarchie, autocorrélation, densité) contre les valeurs calculées sur l'ensemble des grilles de taille 50km extraites de la grille européenne de densité [EUROSTAT, 2014]. 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. On distribue un nombre fixé de noeuds



(a) (b) (c)

(d) (e)

Figure 1

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[1].

3.2.3 Implémentation

3.2.4 Résultats

A densité fixée, les premières exploration 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.

[1] 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 fonctionelle peut être vue comme méta-paramétre [Cottineau et al., 2015a].

4 Discussion

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.

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).

[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.

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

[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.

[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)*.

[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.

[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.

[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.