Génération de Données Synthétiques Correlées

Intention de communication, Journées de Rochebrune 2016

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

15 octobre 2015

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 l'évaluation thérapeutique [Abadie et al., 2010], la géographie [Moeckel et al., 2003], l'apprentissage statistique [Bolón-Canedo et al., 2013]. 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 le niveau de correlation estimé sur les données généré est maitrisé.

Description générique de la méthode Soit un ensemble de processus stochastiques X_I (l'index pouvant être le temps ou l'espace par exemple). 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 : 1. Un certain critère de proximité aux données est vérifié, i.e. $\|f(\mathbf{X}) - f(\tilde{\mathbf{X}})\| < \varepsilon$; et 2. Le niveau de correlation est controlé, i.e. pour tout R, $\mathrm{Var}\left[(\tilde{X}_i)\right] = R$, où la matrice de variance/covariance est estimée sur la population synthétique.

Application : séries temporelles financières 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 (voir matrices aléatoires [Bouchaud and Potters, 2009], analyse de réseaux [Tumminello et al., 2005]).

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 $\|\|$. Dans le cas d'estimateurs auto-regressifs multivariés, la performance dépendra des correlations respectives entre actifs

Application : données géographiques de densité et de réseaux En géographie, l'utilisation de données synthétiques est plutôt axée vers l'utilisation de population synthétiques au sein de modèles 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., 2015].

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 $n(\vec{x})$. L'utilisation d'un modèle type aggrégation-diffusion [Batty, 2006] permet de générer une distribution discrete de densité. Le modèle est calibré pour des objectifs morphologiques (entropie, hiérarchie, autocorrélation, densité) contre les valeurs calculées sur l'ensemble des grilles de taille 50km extraites de la grille européenne de densité []. D'autre part,

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*.

[Cottineau et al., 2015] 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.

[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.

[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.