Raimbault *Complex Adapt Syst Model* (2019) 7:4
https://doi.org/10.1186/s40294-019-0065-y

Complex Adaptive Systems Modeling

RESEARCH Open Access

Second-order control of complex systems with correlated synthetic data

Juste Raimbault[1,2,3*]

*Correspondence:
juste.raimbault@polytechnique.edu
[1] CASA, UCL, London, UK
Full list of author information is available at the end of the article

Abstract

The generation of synthetic data is an essential tool to study complex systems, allowing for example to test models of these in precisely controlled settings, or to parametrize simulation models when data is missing. This paper focuses on the generation of synthetic data with an emphasis on correlation structure. We introduce a new methodology to generate such correlated synthetic data. It is implemented in the field of socio-spatial systems, more precisely by coupling an urban growth model with a transportation network generation model. We also show the genericity of the method with an application on financial time-series. The simulation results show that the generation of correlated synthetic data for such systems is indeed feasible within a broad range of correlations, and suggest applications of such synthetic datasets.

Keywords: Synthetic data, Statistical control, Correlations, Financial time-series, Land-use transportation interactions

Introduction

Developing methods to study complex systems, such as simulation models or data-mining techniques, often requires testbeds and benchmarks to ensure expected properties. The use of synthetic data, in the sense of statistical populations generated randomly under constraints of proximity of patterns to a studied system, is a widely used methodology tackling this issue. This approach is used in several disciplines related to complex systems such as therapeutic evaluation [1], territorial science [2, 3], machine learning [4] or bio-informatics [5].

Generation of synthetic datasets can consist in data disaggregation by producing a microscopic population with fixed macroscopic properties [6]. The creation of synthetic populations for microsimulation models is a typical example where empirical statistical distributions are reproduced [7]. In data extensive contexts, several methods have been developed and improved for a better reproduction of margin distributions [8].

Synthetic datasets can also be generated at the same scale than the targeted real dataset, with a broad range of realism levels and corresponding constraints on the generated data [9]. For example, [10] show that some datamining techniques such as decision trees can be inverted to produce datasets capturing complex non-linear patterns.

The constraints of proximity to reality of synthetic datasets will depend on expected applications. They range for example from a strong statistical fit on given indicators, to

weaker assumptions of similarity on aggregated patterns. In the case of systems where emergence plays a central role, a microscopic property does not directly imply given macroscopic patterns, and synthetic datasets may have to capture some of these. This approach therein becomes part of the complex systems simulation toolbox. Indeed, with the rise of new computational paradigms [11], data (simulated, measured or hybrid) shape our understanding of complex systems. Methodological tools for data-mining, modeling and simulation, including the generation of synthetic data, are therefore crucial to be developed.

Synthetic data and dependancy structures

Reproducing data patterns at the first order, in the sense of distribution moments, is broadly used and understood. A targeted average will be easily reproduced. Similarly, marginals are fitted when generating synthetic population. However, higher orders of data structure are more difficult to include in synthetic data generation methods. At the second order, this corresponds to a control of the covariance structure between generated variables.

Some specific examples where interdependency structure is controlled can be found. Ye [12] investigates the sensitivity of discrete choices models to the distributions of inputs and to their dependance structure. Birkin and Clarke [13] develop a generic framework to generate synthetic micro-data from heterogenous aggregated data sources, which in particular can include second-order effects in the models considered. Li et al. [14] propose to reconstruct multi-dimensional synthetic data using copulas, which capture the dependancy structure between marginal distributions. It is also possible to interpret complex networks generative models [15] as the production of an interdependence structure for a system, contained within link topology. Most methods yielding a high level of accuracy on synthetic covariance structure depend on sampling or data reconstruction methods, and need therefore large datasets.

Synthetic data and socio-spatial systems

Synthetic data with a spatial dimension, in the sense of spatial coordinates of generated data points, or more complicated spatial structures, require proper methods and paradigms. Such approaches have been proposed in disciplines such as geostatistics or Earth sciences. Robin et al. [16] describe a method to generate cross-correlated random spatial fields using Fourier transforms. Osborn et al. [17] introduce a multilevel sampling technique to produce correlated random fields. Concrete applications of such spatial synthetic data include atmospheric circulation models [18], rainfall-runoff simulations [16], or engineering [19].

In the case of socio-spatial systems, this kind of methods is less developed. Simulation approaches to spatialized social systems are already well studied by disciplines such as geosimulation [20], urban analytics [21] or theoretical and quantitative geography [22]. The use of synthetic data in these contexts is however systematically reduced to the generation of synthetic populations within agent-based models or microsimulation models, applied for example to mobility [23], land-use transport interaction models [3], or demography microsimulation models [13]. Some techniques in spatial statistics, such

as Geographically Weighted Regression [24], can also be understood as extrapolating a spatial field and thus constructing spatial synthetic data.

While several examples of stylized models initialized on synthetic configurations can be found in the literature, such as the first Simpop model [25] to simulate the dynamics of settlements at a macroscopic scale, or the SimpopNet model [26] for the co-evolution of cities and transportation networks, these are run on a single stylized synthetic configuration. There is to the best of our knowledge very few examples of works coupling a synthetic data generator with a model at an other scale than the microscopic scale of the population.

Recently, a systematic control of the effects of the initial spatial configuration on the behavior of simulation models was proposed by [27]. The aim is to be able to distinguish proper effects due to intrinsic model dynamics from particular effects due to the geographical structure of the case study. Arentze et al. [28] introduce a method to generate realistic social networks associated to a synthetic population in the geographical space. Such results are essential for the validation of conclusions obtained with modeling and simulation practices in quantitative geography. Being able to generate correlated synthetic configurations of territorial systems is thus an important development remaining to be investigated. In such systems, spatio-temporal correlation structures are a proxy to capture complex dynamics, and controlling them in synthetic data would allow better understanding of models of such systems.

Proposed approach

This literature review on different aspects of synthetic data generation unveils at least two gaps: (i) a lack of attention on controlling covariance structures when generating synthetic data; and (ii) an absence of such methods applied to the study of socio-spatial systems at aggregated scales. As spatio-temporal dependencies structures are essential in driving the dynamics of such systems [29, 30], the combination of these two aspects appears as an unexplored research problem.

We propose in this paper to study the generation of correlated synthetic data, and more particularly in the case of socio-spatial systems. We introduce here a generic methodology taking into account the dependance structure for the generation of synthetic datasets, more precisely by controlling the average of correlation matrices. It is suited to be applied on cases where microscopic data is not available and system similarity is expected on aggregated indicators.

We investigate thus the question of how to generate correlated synthetic data at aggregated levels, where constraints on macroscopic indicators are fulfilled and correlation structure is controlled. We focus on this problem in the particular case of socio-spatial systems, but keep in mind the genericity of the approach.

Our contribution is twofold: (i) we implement a generation of spatial synthetic data for socio-spatial systems, which to the best of our knowledge has never been done in that context; (ii) the method introduced is generic, and we illustrate it with an application to financial time-series.

The rest of the paper is organized as follows. The generic method to generate correlated synthetic data is first formally described. We then apply it to a generative model of territorial configurations, composed by the sequential coupling of a reaction-diffusion model

for population density with a road network generation model, and study the produced correlation patterns. We also illustrate in a following section the genericity of our method by applying it to financial time-series, which are an other example of highly complex signals for which correlations are crucial.

Method formalization

The domain-specific methods described above are too broad to be summarized within a same formalism. We therefore introduce here a generic and model-agnostic framework, focused on the control of correlations structures in synthetic data.

Let $\vec{X}_I$ a multidimensional stochastic process (which can be indexed e.g. with time in the case of time-series, but also with space, or any other indexation). We assume to have a real dataset $\mathbf{X} = (X_{i,j})$, which is interpreted as a set of realizations of the stochastic process. We propose to generate a statistical population $\tilde{\mathbf{X}} = (\tilde{X}_{i,j})$ such that

1. A given criteria of proximity to data is verified, i.e. given a precision ε and some aggregated indicator $\vec{f}$, we have

$$\|\vec{f}(\mathbf{X}) - \vec{f}(\tilde{\mathbf{X}})\| < \varepsilon \tag{1}$$

2. The level of correlation is controlled, i.e. given a matrix $\mathbf{R}$ representing the correlation structure (any symmetric matrix with coefficients in $[-1, 1]$ and a unity diagonal), we have the estimated covariance matrix given by

$$\hat{\text{Cov}}\left[\tilde{\mathbf{X}}\right] = \mathbf{\Sigma}^T \cdot \mathbf{R} \cdot \mathbf{\Sigma} \tag{2}$$

where the standard deviation diagonal matrix $\mathbf{\Sigma}$ is estimated on the synthetic population.

The second requirement will generally be conditional to parameter values determining generation procedure, either generation models being simple or complex ($\mathbf{R}$ itself is a parameter). Formally, we can also understand synthetic processes as parametric families $\tilde{X}_i[\vec{\alpha}]$.

We propose to apply the methodology on very different examples, both typical of complex systems: territorial systems and financial high-frequency time-series. We illustrate the flexibility of the method, and claim to help building interdisciplinary bridges by methodology transposition and reasoning analogy. In the first case, morphological calibration of a population density distribution model allows respecting real data proximity. Correlations of urban form with transportation network measures are empirically obtained by exploration of coupling with a network morphogenesis model. The control is in this case indirect and the feasible space of correlations is empirically determined. In the second case, proximity to data is the equality of signals at a fundamental frequency, to which higher frequency synthetic components with controlled correlations are superposed.

Correlated population density and road network

We now apply the method to territorial systems of human settlements, in the particular case here of population distribution in correlation with road network. In this application, simulation appears as a crucial step to implement the method.

Territorial configuration model

We propose in our case to generate territorial systems summarized in a simplified way as a spatial population density $d(\vec{x})$ and a transportation network $n(\vec{x})$. Correlations we aim to control are correlations between urban morphological measures and network measures. The question of interactions between territories and networks is already well-studied [31] but remains highly complex and difficult to quantify [32]. A dynamical modeling of implied processes should shed light on these interactions [33], and [34] has investigated the concept of co-evolution within such models. We develop here in that context a simple coupling (i.e. without any feedback loop) between a population density distribution model and a network morphogenesis model.

Density model

We use a model D similar to aggregation-diffusion models [35] to generate a discrete spatial distribution of population density. A generalization of the basic model is proposed in [36], providing a calibration on morphological objectives (entropy, hierarchy, spatial auto-correlation, mean distance) against real values computed on the set of 50 km sized grid extracted from European density grid [37]. We recall here rapidly the processes included in the model. A square grid of width W, initially empty, is represented by population $(P_i(t))_{1\leq i\leq W^2}$. At each time step, until the total population reaches a fixed parameter P_m,

- total population is increased of a fixed number N_G (growth rate), following a preferential attachment such that

$$\mathbb{P}[P_i(t+1) = P_i(t) + 1 | P(t+1) = P(t) + 1] = \frac{(P_i(t)/P(t))^\alpha}{\sum (P_i(t)/P(t))^\alpha} \tag{3}$$

- a fraction β of population is diffused to four closest neighbors, what is operated n_d times

The two opposite processes of urban concentration and urban sprawl are captured by the model, what allows reproducing with a good precision a large number of existing morphologies regarding macroscopic urban form indicators.

Network model

On top of the population density model, we generate a planar transportation network with a model N at a similar scale. Several processes can be taken into account to simulate network growth [38]. Other model types could be used as well, such as biological self-generated networks [39], local network growth based on geometrical constraints optimization [40], or a more complex model based on multi-dimensional network percolation [41] which would allow the creation of loops for example. Raimbault [38] generates networks in the frame of a modular architecture, in which the choice of the network generation heuristic can be adapted to a specific need (as e.g. proximity to real data, constraints on output indicators, variety of generated forms, etc.).

We choose here an heuristic based on spatial interaction potential breakdown, which corresponds in practice to a network answering to the strongest demand patterns. The algorithm assumes realistic thematic assumptions: a connected initial network and the creation of links based on spatial interactions.

The heuristic network generation procedure is the following:

1. A fixed number N_c of centers that will be first nodes of the network are distributed given density distribution. Their spatial distribution follows a similar law than the aggregation process, i.e. the probability to be distributed in a given patch is $\frac{(P_i/P)^\alpha}{\sum(P_i/P)^\alpha}$. Population is then attributed according to Voronoi areas of centers, such that a center cumulates population of patches within its triangulation extent.
2. Centers are connected deterministically through a percolation between closest clusters: as long as the network is not fully connected, the two closest connected components, in the sense of the minimal euclidian distance between their vertices, are connected with the link realizing this distance. It yields a tree-shaped network at this stage.
3. The network is modified by adding links following a spatial interaction potential breaking. More precisely, a generalized gravity potential between two centers i and j is defined by

$$V_{ij}(d) = \left[(1-k_h) + k_h \cdot \left(\frac{P_i P_j}{P^2}\right)^\gamma\right] \cdot \exp\left(-\frac{d}{r_g(1+d/d_0)}\right) \tag{4}$$

   where d can be euclidian distance $d_{ij} = d(i,j)$ or network distance $d_N(i,j)$, $k_h \in [0,1]$ is a weight to determine the role of populations, γ gives the shape of the hierarchy across population values, r_g is a characteristic interaction distance and d_0 is a distance shape parameter.
4. A fixed number $K \cdot N_L$ of potential new links is taken among the couples having greatest euclidian distance potential ($K = 5$ is fixed).
5. Among potential links, N_L are effectively realized, that are the one with smallest rate $\tilde{V}_{ij} = V_{ij}(d_N)/V_{ij}(d_{ij})$. At this stage only the gap between euclidian and network distance is taken into account: $\tilde{V}_{ij}$ does indeed not depend on populations and is increasing with d_N at constant d_{ij}.
6. Planarity of the network is imposed by creating nodes at possible intersections formed by new links.

The nature and range of correlations produced by this model coupling, as a function of model parameters, are to be determined by simulation experiments.

Parameter space

The parameter space for the coupled model is constituted first by density generation parameters $\vec{\alpha}_D = (P_m/N_G, \alpha, \beta, n_d)$. We study for the sake of simplicity the rate between population and growth rate P_m/N_G instead of both varying, i.e. the number of steps needed to generate the distribution. These are completed by network generation parameters $\vec{\alpha}_N = (N_C, k_h, \gamma, r_g, d_0)$. We write $\vec{\alpha} = (\vec{\alpha}_D, \vec{\alpha}_N)$.

Indicators

Urban form and network structure are quantified by numerical indicators in order to modulate correlations between these. Morphology is defined as a vector $\vec{M} = (r, d, \varepsilon, a)$ giving spatial auto-correlation (Moran index), mean distance, entropy and hierarchy (see [42] for a precise definition of these indicators). Network measures $\vec{G} = (c, l, s, \delta)$ are with network denoted (V, E)

- Average centrality c defined as average *betweenness-centrality* (normalized in [0, 1]) on all links.
- Average path length l given by $\frac{1}{d_m} \frac{2}{|V|\cdot(|V|-1)} \sum_{i<j} d_N(i,j)$ with d_m normalization distance taken here as world diagonal $d_m = \sqrt{2}N$.
- Average network speed [43] which corresponds to network performance compared to direct travel, defined as $s = \frac{2}{|V|\cdot(|V|-1)} \sum_{i<j} \frac{d_{ij}}{d_N(i,j)}$.
- Network diameter $\delta = \max_{ij} d_N(i,j)$.

We study the cross-correlation matrix $\mathrm{Cov}\left[\vec{M}, \vec{G}\right]$ between morphology and network. We estimate it on a set of n realizations at fixed parameter values $(\vec{M}[D(\vec{\alpha})], \vec{G}[N(\vec{\alpha})])_{1\leq i\leq n}$ with the standard unbiased estimator, given by Eq. 5 below.

$$\hat{\rho}[X1, X2] = \frac{\hat{C}[X1, X2]}{\sqrt{\hat{\mathrm{Var}}[X1] \cdot \hat{\mathrm{Var}}[X2]}} \tag{5}$$

The covariance is estimated by Eq. 6, where variables are indexed by t over T realizations.

$$\hat{C}[X1, X2] = \frac{1}{(T-1)} \sum_t X_1(t) X_2(t) - \frac{1}{T\cdot(T-1)} \sum_t X_1(t) \sum_t X_2(t) \tag{6}$$

The variance is estimated by Eq. 7.

$$\hat{\mathrm{Var}}[X] = \frac{1}{T} \sum_t X^2(t) - \left(\frac{1}{T} \sum_t X(t)\right)^2 \tag{7}$$

Null model

In order to provide a minimal benchmark of our correlated data generation method, we also introduce a null model to control if the produced correlation are not intrinsic to the specification of indicators for example. The procedure to generate null configuration is the following: (i) generate a random population density, by randomly selecting a proportion $r_o^{(0)}$ of occupied cell and attributing them a random density between 0 and 1; (ii) add a fixed number of network nodes $N_N^{(0)}$, either randomly in space, or following the population density with a probability of each cell proportional to its density; (iii) add a fixed number of links $N_L^{(0)}$ between random pairs of nodes; (iv) planarize the resulting network by adding nodes at link intersections. In this model, population density and network are either totally independent, or linked through network node density only. We thus expect the corresponding correlation to behave as a baseline of how correlations

between indicators behave when no particular care is given to including interaction processes.

Generating correlated synthetic data

The coupling of generative models is done both in a formal and operational way. We indeed loosely couple independent implementations. The OpenMOLE software [44] for model exploration offers a proper framework for this. Its modular workflow language allows to compose model tasks and integrate these into diverse numerical experiments. For the population density generation, we use the `scala` implementation provided by [36]. The network generation model is implemented in NetLogo [45], which offers a good compromise between performance and interactive model validation and exploration. The two models are coupled with a specific OpenMOLE script. Source code is available at https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Synthetic.

Results

The study of the density model alone is developed in [36]. It is in particular calibrated on European density grid data, on 50km width square areas with 500 m resolution for which real indicator values have been computed on the whole Europe. Furthermore, a grid exploration of model behavior yields feasible output space in reasonable parameters bounds (roughly $\alpha \in [0.5, 2], N_G \in [500, 3000], P_m \in [10^4, 10^5], \beta \in [0, 0.2], n_d \in \{1, \dots, 4\}$). The reduction of indicators space to a two dimensional plan through a Principal Component Analysis (variance explained with two components $\simeq 80\%$) allows to isolate a set of output points that covers reasonably precisely real point cloud. It confirms the ability of the model to reproduce morphologically the set of real configurations.

With a fixed population density, the conditional exploration of network generation model parameter space suggest a good flexibility on global indicators $\vec{G}$, together with good convergence properties. In order to apply the synthetic data generation method in relation with the thematical question of interactions between networks and territories, the exploration has been oriented towards the study of cross-correlations.

Given the large relative dimension of the parameter space, an exhaustive grid exploration is not possible. We use a Latin Hypercube sampling procedure with bounds given above for $\vec{\alpha}_D$ and for $\vec{\alpha}_N$, we take $N_C \in [50, 120], r_g \in [1, 100], d_0 \in [0.1, 10], k_h \in [0, 1], \gamma \in [0.1, 4], N_L \in [4, 20]$. For the number of model replications for each parameter point, less than 50 are enough to obtain confidence intervals at 95% on indicators of width less than standard deviations. For correlations a hundred give confidence intervals (obtained with Fisher method) of size around 0.4, we take thus $n = 80$ for experiments. The null model is simulated also with $n = 80$, for random and density-based node distributions, and $r_o^{(0)} \in \{0.25, 0.5, 0.75\}$, $N_N^{(0)} \in \{10, 15, 20\}$ and $N_L^{(0)} \in \{20, 30, 40\}$. Simulation results are available on the dataverse at http://dx.doi.org/10.7910/DVN/UIHBC7.

We show in Fig. 1 examples of generated territorial configurations. This visualization and some values of associated correlations already suggest that the method application yields a broad spectrum of generated correlation patterns. We obtain for example low density configurations, in aggregated or dispersed settings (top left, resp. bottom left

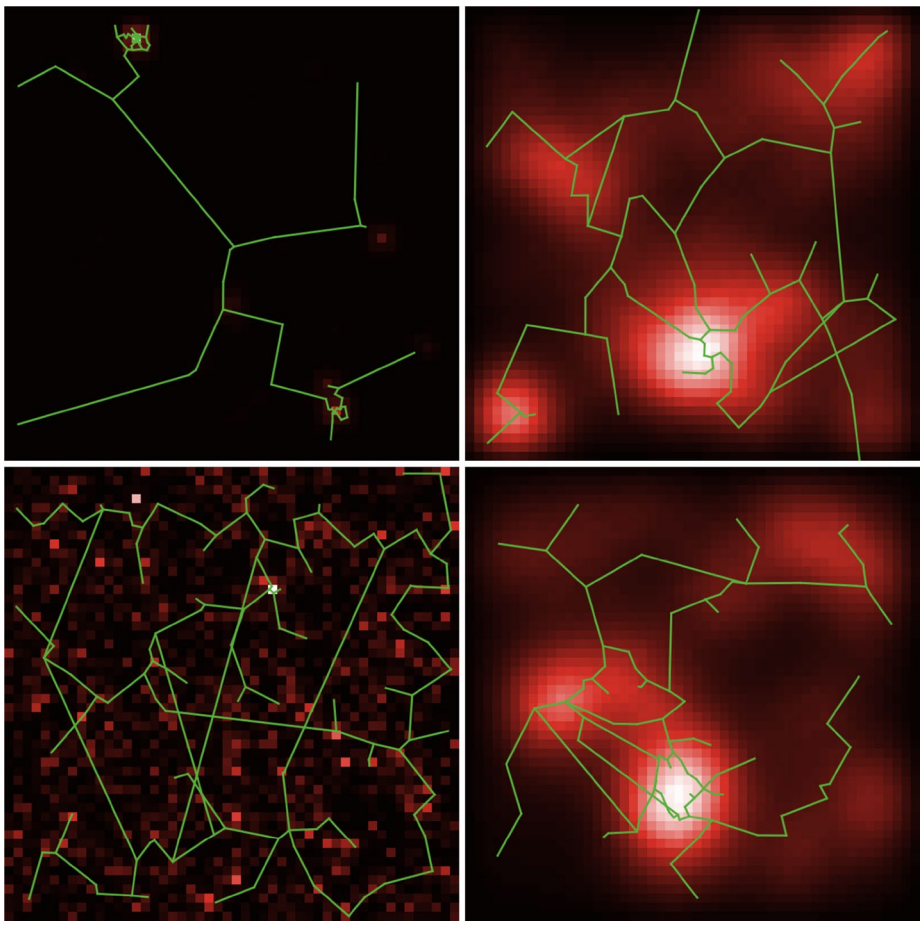

Fig. 1 Configurations obtained for parameters giving the four emphasized points in Fig. 2, in order from left to right and top to bottom. We recognize polycentric city configurations (2 and 4), diffuse rural settlements (3) and aggregated weak density area (1). See dataset for exhaustive parameter values, indicators and corresponding correlations. For example d is highly correlated with l and s ($\simeq 0.8$) in (1) but not for (3) although both correspond to rural environments; in the urban case we observe also a broad variability : $\rho[d, c] \simeq 0.34$ for (4) but $\simeq -0.41$for (2), what is explained by a stronger role of gravitation hierarchy in (2) $\gamma = 3.9, k_h = 0.7$ (for (4), $\gamma = 1.07, k_h = 0.25$), whereas density parameters are similar

panel), inducing very different types of networks. Similarly, areas with population centers which are closer to urban areas (top right and bottom right panel), can also produce different network shapes. In the latest case, increasing the role of hierarchy through γ and k_h leads from a negative correlation between average distance d and centrality c to a positive correlation. This corresponds to a transition from processes where population dispersal decreases centrality (redundant networks) to inverse processes (centralized networks).

Regarding the generation of correlated synthetic data in itself, several results presented in Fig. 2 are worth noting. First of all, the statistical distributions of correlation coefficients (histograms, top left panel of Fig. 2) between morphology and network indicators are not systematically simple and some are bimodal. For example, the correlation $\rho[a, l]$ between hierarchy of the population distribution a and mean path length in the network l has a mode around 0 and an other around 0.6. This means that

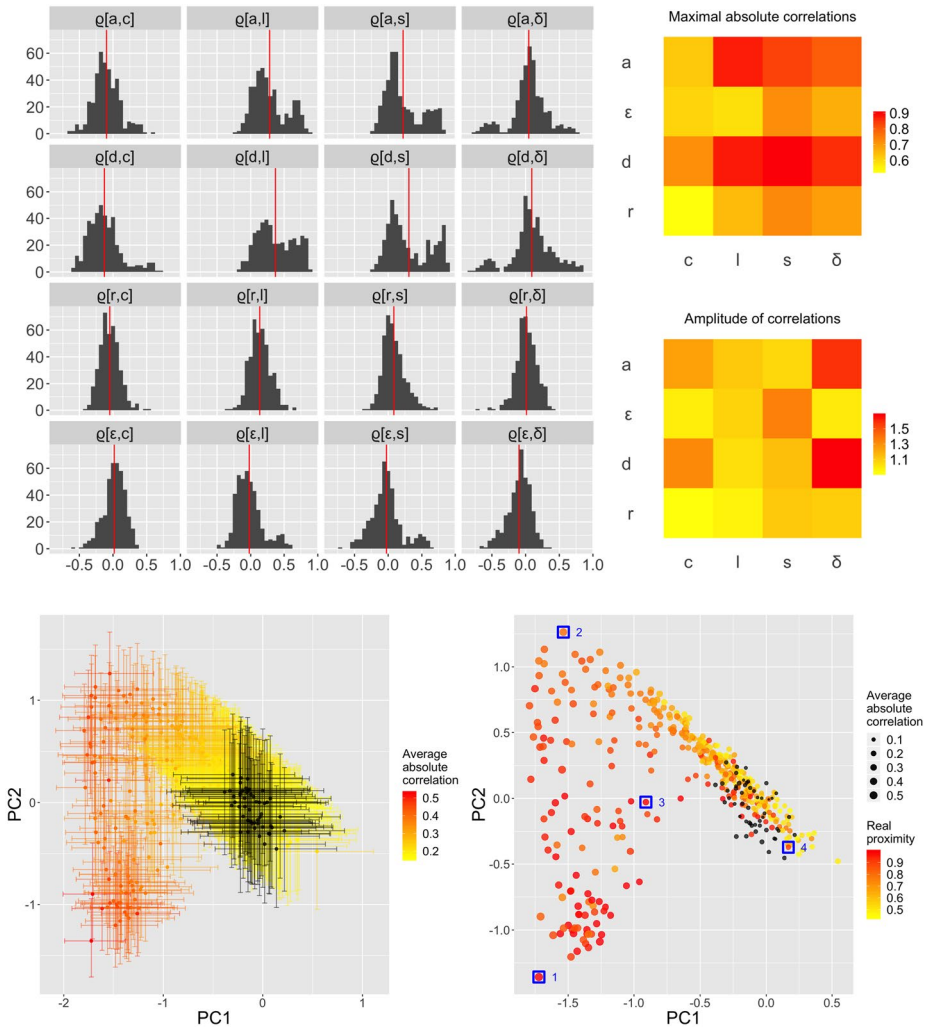

Fig. 2 Exploration of feasible space for correlations between urban morphology and network structure. (Top left) Statistical distribution of crossed-correlations between vectors $\vec{M}$ of morphological indicators (in numbering order Moran index, mean distance, entropy, hierarchy) and $\vec{N}$ of network measures (centrality, mean path length, speed, diameter). (Top right) Heatmaps for amplitude of correlations, defined as $a_{ij} = \max_k \rho_{ij}^{(k)} - \min_k \rho_{ij}^{(k)}$ and maximal absolute correlation, defined as $c_{ij} = \max_k \left|\rho_{ij}^k\right|$. (Bottom left) Projection of correlation matrices in a principal plan obtained by Principal Component Analysis on matrix population (cumulated variances: PC1 = 38%, PC2 = 68%). Error bars are initially computed as 95% confidence intervals on each matrix element (by standard Fisher asymptotic method), and boundaries of confidence intervals are transformed into the component space. Scale color gives mean absolute correlation on full matrices. Black dots and error bars correspond to the realizations of the null model. (Bottom right) Representation in the principal plan, scale color giving proximity to real data defined as $1 - \min_r \|\vec{M} - \vec{M}_r\|$ where $\vec{M}_r$ is the set of real morphological measures, point size giving mean absolute correlation. The points highlighted in blue correspond to the configurations shown in Fig. 1. Black dots correspond to the realizations of the null model

in a certain regime, these tend to decorrelate in average, while in an other regime they are strongly correlated. The latest correspond to configurations with a high Moran index and a high hierarchy, which means that more centralized urban configurations constrain the network path length through this correlation.

Second, still based on distributions in Fig. 2, but also on heatmaps for amplitude and maximal correlation (top right panel), we observe that it is possible to modulate up to a relatively high level of correlation for all indicators, since the maximal absolute correlation varies between 0.6 and 0.9. The amplitude of correlations ranges between 0.9 and 1.6, allowing thus a broad spectrum of values.

As the cross-correlation matrix is of dimension 16, we proceed to a principal component analysis on all generated correlation matrices (one matrix per row) to visualize the covered space in two dimensions. This PCA yields 38% of variance for the first component and 68% of cumulated variance for the second. We visualize the corresponding point cloud in the principal plan, with transformed confidence intervals (bottom left panel of Fig. 2) and with particular points (bottom right panel). The point cloud in the principal plan has a large extent but is not uniform: it is not possible to modulate in any direction any coefficient as they stay themselves correlated because of underlying generation processes. A more refined study at higher orders (correlation of correlations) would be necessary to precisely understand degrees of freedom in the generation of correlations. However, the covered area remains broad and confirms a rather flexible output space for generated correlations. When comparing to the null model runs (black dots and error bars), we find as expected that null model correlations are around 0 (all confidence intervals covering the origin), and that a part of the generated point cloud falls in the same area. An other important part of points fall outside the range of the null model in a statistically significant way. These are the interesting points for a possible application of the synthetic dataset, and we show thus that the method produces non-trivial and significant correlation patterns.

When evaluating the proximity of indicator values to real points (Eq. 1 in the abstract description of the method), which is given by the color level in the bottom right panel of Fig. 2, we note that the points with the highest level of correlation are also the ones which are closest to real data (red points). The points which are the farthest from real configurations are the uncorrelated ones, which also coincide with the null model. This suggests that in the frame of model hypotheses, real configurations exhibit high correlations between network properties and urban form. Raimbault [46] confirms this fact by studying real effective correlations.

Finally, some examples of configurations taken on particular points in the principal plan, highlighted in blue in Fig. 2 and described above (Fig. 1), show that similar population density profiles can yield very different correlation profiles. This confirms the flexibility of the method and the possibility to control on correlation structure.

Correlated financial time-series

We also apply the method to a totally different type of system, namely financial complex systems. Financial time-series are heterogeneous, multi-scalar and non-stationary [47]. Correlations are broadly explored in that field. For example, Random Matrix Theory allows distinguishing signal from noise for a correlation matrix computed for a large number of asset with low-frequency signals, typically with a time scale of a day [48]. Similarly, Complex Network Analysis on networks constructed from correlations introduced methods such as Minimal Spanning Tree [49] or more refined topologically constrained network generation methods [50]. These provide reconstructions of economic

sectors structure. At high frequencies, the precise estimation of interdependence parameters assuming models for asset dynamics has been extensively studied from a theoretical point of view aimed at refinement of models and estimators [51]. Theoretical results must be tested on synthetic datasets as they ensure a control of most parameters in order to check that a predicted effect is indeed observable all things being otherwise equal. Empirical confirmation of the improvement of estimators is obtained on a synthetic dataset at a fixed correlation level.

We consider a network of assets $(X_i(t))_{1\leq i\leq N}$ sampled at high-frequency (typically 1 s). We use a multi-scalar framework (used e.g. in wavelet analysis approaches [52] or in multi-fractal signal processing [53]) to interpret observed signals as the superposition of components at different time scales. We thus write $X_i = \sum_\omega X_i^\omega$. We denote by $T_i^\omega = \sum_{\omega'\leq\omega} X_i^\omega$ the filtered signal at a given frequency ω. A typical problem in the study of complex systems is the prediction of a trend at a given scale. It can be viewed as the identification of regularities and their distinction from components considered as random. For the sake of simplicity, we represent such a process as a trend prediction model at a given temporal scale ω_1, formally an estimator $M_{\omega_1} : (T_i^{\omega_1}(t'))_{t'<t} \mapsto \hat{T}_i^{\omega_1}(t)$ which aims at minimizing error on the real trend $\| T_i^{\omega_1} - \hat{T}_i^{\omega_1} \|$. In the case of autoregressive multivariate estimators, the performance will depend among other parameters on respective correlations between assets. It is thus interesting to apply the method to the evaluation of performance as a function of correlation at different scales. We assume a Black–Scholes dynamic for assets [54], i.e. $dX = \sigma \cdot dW$, with W Wiener process. Such a dynamic model allows an easy modulation of correlation levels.

Data generation

We can straightforward generate $\tilde{X}_i$ such that $\mathrm{Var}\left[\tilde{X}_i^{\omega_1}\right] =^t \mathbf{\Sigma} \cdot \mathbf{R} \cdot \mathbf{\Sigma}$ (with $\mathbf{\Sigma}$ are estimated standard deviations and $\mathbf{R}$ is a fixed correlation matrix) and verifying $X_i^{\omega\leq\omega_0} = \tilde{X}_i^{\omega\leq\omega_0}$. This means in practice that the data proximity indicator is the identity of components at a lower frequency than a fundamental frequency $\omega_0 < \omega_1$. We use therefore the simulation of Wiener processes with fixed correlation. Indeed, if

$$dW_1 \perp\!\!\!\perp dW_1^{\perp\!\!\!\perp}$$

(and $\sigma_1 < \sigma_2$ indicatively, assets being interchangeable), then

$$W_2 = \rho_{12} W_1 + \sqrt{1 - \frac{\sigma_1^2}{\sigma_2^2} \cdot \rho_{12}^2} \cdot W_1^{\perp\!\!\!\perp} \tag{8}$$

is such that $\rho(dW_1, dW_2) = \rho_{12}$. Signals for other components can be constructed the same way by Gram orthonormalization. We isolate the component at the desired frequency ω_1 by filtering the signal, i.e. using signals constructed with Eq. 8 such that $\tilde{X}_i^{\omega_1} = W_i - \mathcal{F}_{\omega_0}[W_i]$, where $\mathcal{F}_{\omega_0}$ is a low-pass filter with cut-off frequency ω_0. We reconstruct then the hybrid synthetic signals by taking

$$\tilde{X}_i = T_i^{\omega_0} + \tilde{X}_i^{\omega_1} \tag{9}$$

The method is tested on an example with two assets from foreign exchange market (EUR/USD and EUR/GBP), on a 6 month period from June 2015 to November 2015. Data was obtained from http://www.histdata.com/. The data cleaning procedure, starting from original series sampled at a frequency around 1 s, consists in a first step to the determination of the minimal common temporal range (missing sequences being ignored, by vertical translation of series, i.e. $S(t) := S(t) \cdot \frac{S(t_n)}{S(t_{n-1})}$ when t_{n-1}, t_n are extremities of the "hole" and $S(t)$ value of the asset, what is equivalent to keep the constraint to have returns at similar temporal steps between assets). We study then *log-prices* and *log-returns* [47], defined by $X(t) := \log \frac{S(t)}{S_0}$ and $\Delta X(t) = X(t) - X(t-1)$. Raw data are filtered at a maximal frequency $\omega_m = 10\,\text{min}$ (which will be the maximal frequency for following treatments) for concerns of computational efficiency. As time-series are then sampled at $3 \cdot \omega_m$ to avoid aliasing, a day of size 86,400 for 1 s sampling is reduced to a much smaller size of 432. We use a non-causal gaussian filter of total width ω. We fix the fundamental frequency $\omega_0 = 24\,\text{h}$ and we propose to construct synthetic data at frequencies $\omega_1 = 30\,\text{min}, 1\,\text{h}, 2\,\text{h}$. We show in Fig. 3 an example of signal structure at the scales ω_m and $\omega_1 = 30\,\text{min}$, compared with the non-filtered raw signal.

It is crucial to consider the interference between ω_0 and ω_1 frequencies in the reconstructed signal: the correlation which is indeed estimated is

$$\rho_e = \rho\left[\Delta\tilde{X}_1, \Delta\tilde{X}_2\right] = \rho\left[\Delta T_1^{\omega_0} + \Delta\tilde{X}_1^{\omega}, \Delta T_2^{\omega_0} + \Delta\tilde{X}_2^{\omega}\right] \tag{10}$$

Assuming to be in the reasonable limit $\sigma_1 \gg \sigma_0$ (fundamental frequency small enough), that $\text{Cov}\left[\Delta\tilde{X}_i^{\omega_1}, \Delta X_j^{\omega}\right] = 0$ for all $i, j, \omega_1 > \omega$ and that returns are centered at any scale, we can develop the previous expression to compute the correction on effective

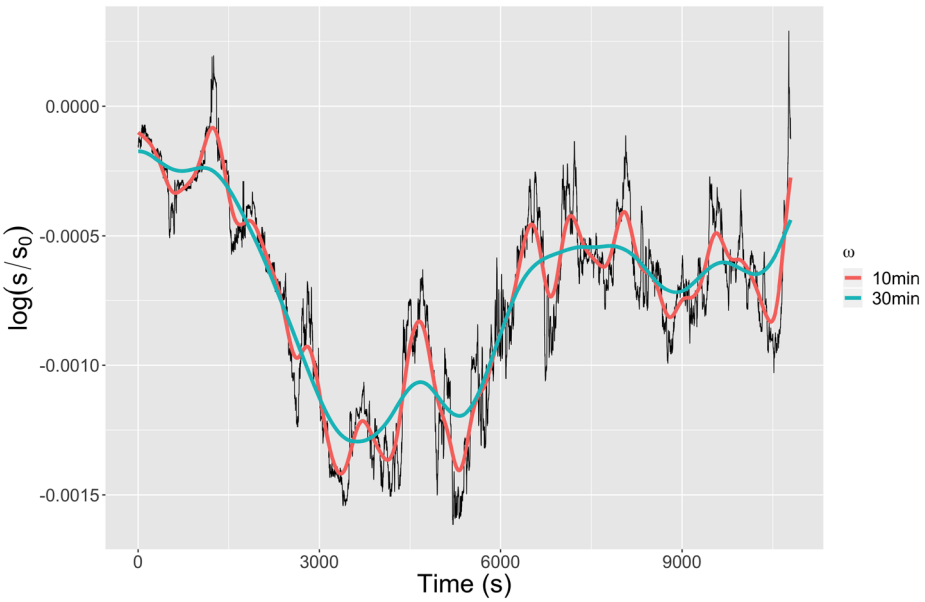

Fig. 3 Example of the multi-scalar structure of the signal, basis of the construction of synthetic signals. *Log-prices* are represented on a time window of around 3 h for November 1st 2015 for asset EUR/USD, with 10 min (purple) and 30 min trends. The low-frequency components are the basis to build synthetic data, on which noises with a controlled correlation are superposed

correlation due to interferences. We obtain at the first order the expression of effective correlation given by

$$\rho_e = [\varepsilon_1\varepsilon_2\rho_0 + \rho] \cdot \left[1 - \frac{1}{2}\left(\varepsilon_1^2 + \varepsilon_2^2\right)\right] \tag{11}$$

what corresponds to the correlation that we can effectively simulate in synthetic data.

Correlations are in practice estimated with a Pearson estimator, the covariances being corrected for bias, i.e. following Eqs. 5–7.

The generated synthetic data are then used to test a toy model. We propose in particular to investigate the predictive power of a very simple linear model. The tested predictive model M_{ω_1} is a simple $ARMA$ for which parameters $p = 2, q = 0$ are fixed (as we do not create lagged correlation, we do not expect large orders of auto-regression as these kind of processes have short memory for real data; furthermore smoothing is not necessary as data are already filtered). It is however applied in an adaptive way, in the sense that given a time window T_W, we estimate for any t the model on $[t - T_W + 1, t]$ in order to predict signals at $t + 1$.

Experiments are implemented in the `R` language, using in particular the `MTS` [55] library for time-series models. Cleaned data and source code are available on an open `git` repository at https://github.com/JusteRaimbault/SyntheticAsset.

Figure 4 shows the effective correlations computed on synthetic data. For standard parameter values (for example $\omega_0 = 24\,\text{h}$, $\omega_1 = 2\,\text{h}$ and $\rho = -0.5$), we find $\rho_0 \simeq 0.71$ et $\varepsilon_i \simeq 0.3$ what yields $|\rho_e - \rho| \simeq 0.05$. We observe a good agreement between observed ρ_e and values predicted by Eq. 11 in the interval $\rho \in [-0.5, 0.5]$. On the contrary, for larger absolute values, a deviation increases with $|\rho|$ and as ω_1 decreases: it confirms the intuition that when frequency decreases and becomes closer to ω_0, interferences between the two components are not negligible anymore and invalidate independence assumptions for example.

Application to the study of a predictive model performance

The predictive model described above is then applied to synthetic data, in order to study its average performance as a function of correlation between signals. Results for $\omega_1 = 1\,\text{h}, 1\,\text{h}\ 30\,\text{min}, 2\,\text{h}$ are shown in Fig. 5. The a priori counter-intuitive result of a maximal performance at vanishing correlation for one of the assets confirms the role of synthetic data to better understand system mechanisms: the study of lagged correlations shows an asymmetry in the real data that we can understand at a daily scale as an increased influence of EUR/GBP on EUR/USD with a rough two hours lag. The existence of this *lag* allows a "good" prediction of EUR/USD thanks to fundamental component. This predictive power is perturbed by added noises in a way that increases with their correlation. The more noises correlated are, the more the model will take them into account and will make false predictions because of the Markovian character of simulated brownian (note that the model used has theoretically no predictive power at all on pure brownians).

This case study on a *toy-model* illustrates the relevance of using simulated synthetic data. Further developments can be directed towards the simulation of more realistic

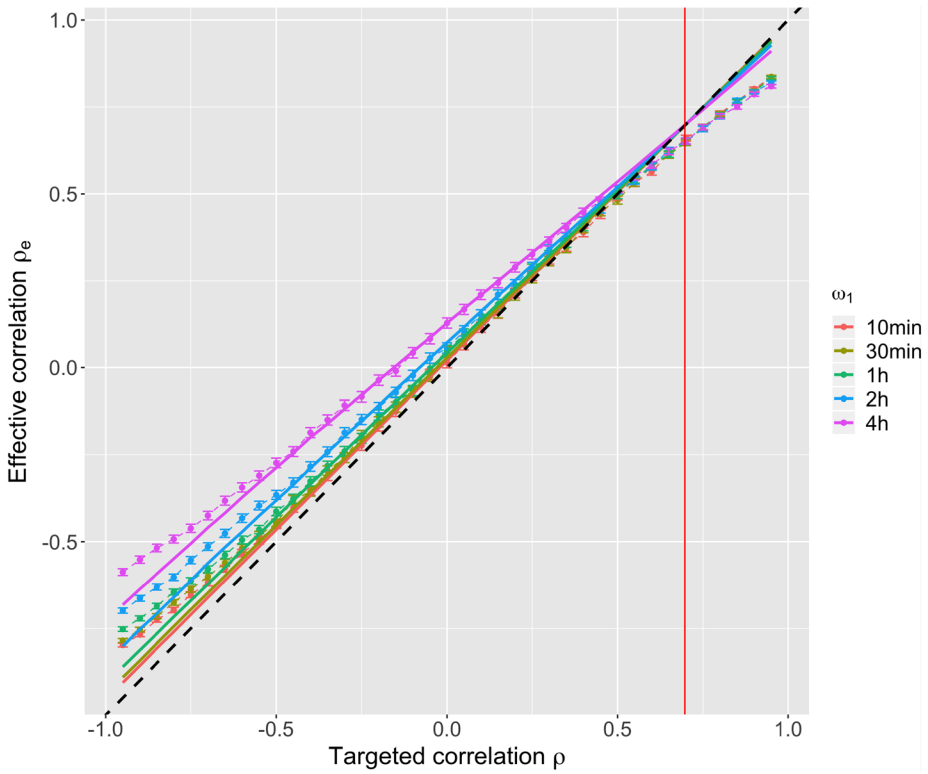

Fig. 4 Effective correlations obtained on synthetic data. Dots represent estimated correlations on a synthetic dataset corresponding to 6 months between June and November 2015 (error-bars give 95% confidence intervals obtained with standard Fisher method); scale color gives the filtering frequency $\omega_1 = 10$ min, 30 min, 1 h, 2 h, 4 h; solid lines give the theoretical values for ρ_e obtained by 11 with estimated volatilities (dotted-line diagonal for reference); vertical red line position is the theoretical value such that $\rho = \rho_e$ with mean values for ε_i on all points. We observe for high absolute correlations values a deviation from corrected values, what should be caused by non-verified independence and centered returns assumptions. Asymmetry is caused by the high value of $\rho_0 \simeq 0.71$

data (presence of consistent *lagged correlation* patterns, more realistic models than Black–Scholes) and apply it on more operational predictive models.

Discussion

Contributions

We investigated in this paper the possibility of generating synthetic data at a macroscopic level with a controlled correlation structure. The generic method we introduce can be applied to any complex system, where the proximity to real data is measured on aggregated indicators. The method was designed more particularly for socio-spatial systems. We show in the case of transportation network and territories, by exploring a weak coupled model for population density and road network generation, that varying model parameters yield a broad output space of effective correlations. Two configurations with the same first order indicator values can capture very different underlying correlations. This means that future applications to the study of upstream models to the sensitivity of spatial initial configuration, such as the one done by [27] but in which correlation structure is controlled, should be made possible by our approach.

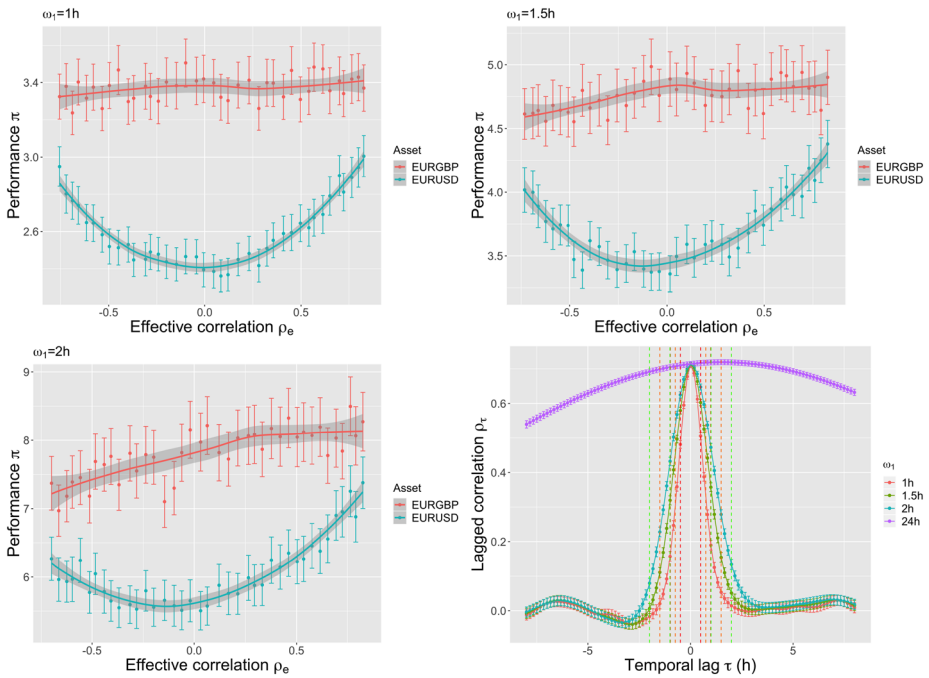

Fig. 5 Performance of a predictive model as a function of simulated correlations. From left to right and top to bottom, the plots show for each asset the normalized performance of an ARMA model ($p = 2, q = 0$), defined as $\pi = \left(\frac{1}{T} \sum_t \left(\tilde{X}_i(t) - M_{\omega_1} \left[\tilde{X}_i \right](t) \right)^2 \right) / \sigma \left[\tilde{X}_i \right]^2$ (95% confidence intervals computed by $\pi = \bar{\pi} \pm (1.96 \cdot \sigma[\pi])/\sqrt{T}$, local polynomial smoothing to ease reading). It is interesting to note the U-shape for EUR/USD, due to interference between components at different scales. Correlation between simulated noises deteriorates predictive power. The study of lagged correlations (here $\rho[\Delta X_{\text{EURUSD}}(t), \Delta X_{\text{EURGBP}}(t - \tau)]$) on real data clarifies this phenomenon: the fourth graph shows an asymmetry in curves at any scale compared to zero lag ($\tau = 0$) what leads fundamental components to increase predictive power for the dollar, amelioration then perturbed by correlations between simulated components. Dashed lines show time steps (in equivalent τ units) used by the ARMA at each scale, what allows to read the corresponding lagged correlation on fundamental component

We postulate that the method can be also applied in other fields where similar constraints can be of interest. Indeed, in the context of financial data, considering data proximity indicators based on low-frequency components of signals, we showed how correlation can be controlled and even analytically predicted to a certain extent. Our work recalls thus the interest in generating hybrid data, and is differentiated from most work where only the microscopic level is taken into account.

As already mentioned, most of simulation models need an initial state generated artificially as soon as model parametrization is not done completely on real data. An advanced model sensitivity analysis implies a control on parameters for synthetic dataset generation, seen as model meta-parameters [27]. In the case of a statistical analysis of model outputs it provides a way to operate a second order statistical control.

Future work

Regarding the application to geographical data, the calibration of the network generation component at given density, on real data for transportation network, is a potential development. It would be relevant typically on road networks given the shape of generated

networks, what should be possible using OpenStreetMap open data which have a reasonable quality for Europe [56]. This should theoretically allow unveiling parameter sets reproducing accurately existing configurations both for urban morphology and network shape. By attributing a synthetic dataset similar to a given real configuration, we would be able to compute a sort of *intrinsic correlation* proper to this configuration.

We studied in the second example stochastic processes in the sense of random time-series, whereas time did not have a role in the first case. We can suggest a strong coupling between the two model components (or the construction of an integrated model) and to observe indicators and correlations at different time steps during the generation. In dynamical spatial models the existence of lagged interdependences in space and time [29] is an important feature of complex dynamics. This would provide a better understanding of the link between static and dynamic correlations.

We were limited to the control of first and second moments of generated data, but we could imagine a theoretical generalization allowing the control of moments at any order. However, as shown by the geographical example, the difficulty of generation in a concrete complex case questions the possibility of higher orders control when keeping a consistent structure model and a reasonable number of parameters. The study of non-linear dependence structures as proposed in [57] is in an other perspective an interesting possible development.

We could also apply specific exploration algorithms to explore more exhaustively the feasible correlation space. Such an algorithm based on Novelty Search has been introduced by [58]. Coupling it with our method would allow establishing the full range of feasible correlations for a given generation model.

Conclusion

We proposed an abstract method to generate synthetic datasets in which correlation structure is controlled, but the empirical data required can be sparse or targeting macroscopic aggregated criteria. Its implementation in two very different fields shows its flexibility and the broad range of possible applications. More generally, it is crucial to favorise such practices of systematic validation of computational models by statistical analysis, in particular for agent-based models for which the question of validation remains an open issue.

Furthermore, our overall approach enters a particular epistemological frame. On the one hand it has a strong multidisciplinary aspect, and on the other hand the importance of empirical component through computational exploration methods make this approach typical of Complex Systems science [59]. The combination of empirical knowledge obtained from data mining, with knowledge obtained by modeling and simulation is generally central to the conception and exploration of multi-scalar heterogeneous models. The method and results presented here are an illustration of such an hybrid paradigm.

Acknowledgements

Results obtained in this paper were computed on the vo.complex-system.eu virtual organization of the European Grid Infrastructure (http://www.egi.eu). The author thanks the European Grid Infrastructure and its supporting National Grid Initiatives (France-Grilles in particular) for providing the technical support and infrastructure. The author thanks the organizers and participants of Journées de Rochebrune 2016 for which the work was originally conceived. The author also thanks E. Marandon (L2 Technologies) for the original idea of correlated financial signals.

Authors' contributions

JR designed the study, did the analysis and wrote the paper. The author read and approved the final manuscript.

Funding
This work is part of DynamiCity, a FUI project funded by BPI France, Auvergne-Rhône-Alpes region, Ile-de-France region and Lyon metropolis. This work was also funded by the Urban Dynamics Lab Grant EPSRC EP/M023583/1.

Availability of data and materials
All data and code used in this study, including simulation results, are openly available on git repositories at https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Synthetic and https://github.com/JusteRaimbault/SyntheticAsset. Large dataset are available on the dataverse repository at http://dx.doi.org/10.7910/DVN/UIHBC7.

Competing interests
The author declares no competing interests.

Author details
[1] CASA, UCL, London, UK. [2] UPS CNRS 3611 ISCPIF, Paris, France. [3] UMR CNRS 8504 Géographie-cités, Paris, France.

Received: 7 August 2019 Accepted: 11 November 2019
Published online: 20 November 2019

References

1. Abadie A, Diamond A, Hainmueller J (2010) Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. J Am Stat Assoc 105(490):493–505
2. Moeckel R, Spiekermann K, Wegener M (2003) Creating a synthetic population. In: Proceedings of the 8th international conference on computers in urban planning and urban management (CUPUM)
3. Pritchard DR, Miller EJ (2009) Advances in agent population synthesis and application in an integrated land use and transportation model. In: Transportation research board 88th annual meeting
4. Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A (2013) A review of feature selection methods on synthetic data. Knowl Inf Syst 34(3):483–519
5. Van den Bulcke T, Van Leemput K, Naudts B, van Remortel P, Ma H, Verschoren A, De Moor B, Marchal K (2006) Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. BMC Bioinform 7(1):43
6. Beckman RJ, Baggerly KA, McKay MD (1996) Creating synthetic baseline populations. Transp Res A Policy Pract 30(6):415–429
7. Müller K, Axhausen KW (2010) Population synthesis for microsimulation: state of the art. Arbeitsberichte Verkehrs- und Raumplanung. https://doi.org/10.1016/j.trpro.2016.11.078
8. Barthelemy J, Toint PL (2013) Synthetic population generation without a sample. Transp Sci 47(2):266–279
9. Hoag JE (2008) Synthetic data generation: theory. Techniques and applications. University of Arkansas, Ann Arbor
10. Eno J, Thompson CW (2008) Generating synthetic data to match data mining patterns. IEEE Internet Comput 12(3):78–82
11. Arthur WB (2015) Complexity and the shift in modern science. In: Conference on complex systems, Tempe, Arizona
12. Ye X (2011) Investigation of underlying distributional assumption in nested logit model using copula-based simulation and numerical approximation. Transp Res Rec 2254:36–43
13. Birkin M, Clarke M (1988) Synthesis—a synthetic spatial information system for urban and regional analysis: methods and examples. Environ Plan A 20(12):1645–1671
14. Li H, Xiong L, Jiang X (2014) Differentially private synthesization of multi-dimensional data using copula functions. In: Advances in database technology: proceedings. International conference on extending database technology, vol. 2014. NIH Public Access, p 475
15. Newman ME (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256
16. Robin M, Gutjahr A, Sudicky E, Wilson J (1993) Cross-correlated random field generation with the direct fourier transform method. Water Resour Res 29(7):2385–2397
17. Osborn S, Vassilevski PS, Villa U (2017) A multilevel, hierarchical sampling technique for spatially correlated random fields. SIAM J Sci Comput 39(5):543–562
18. Gourdji S, Hirsch A, Mueller K, Yadav V, Andrews A, Michalak A (2010) Regional-scale geostatistical inverse modeling of North American CO_2 fluxes: a synthetic data study. Atmos Chem Phys 10(13):6151–6167
19. Zhao T, Wang Y (2018) Simulation of cross-correlated random field samples from sparse measurements using Bayesian compressive sensing. Mech Syst Signal Process 112:384–400
20. Benenson I, Torrens P (2004) Geosimulation: automata-based modeling of urban phenomena. Wiley, Chichester
21. Batty M (2013) The new science of cities. MIT Press, Cambridge
22. Pumain D (2018) An evolutionary theory of urban systems. International and transnational perspectives on urban systems. Springer, Singapore, pp 3–18
23. Banos A, Chardonnel S, Lang C, Marilleau N, Thévenin T (2005) Simulating the swarming city: a mas approach. In: Proceedings of the 9th international conference on computers in urban planning and urban management, pp 29–30
24. Brunsdon C, Fotheringham S, Charlton M (1998) Geographically weighted regression. J R Stat Soc Ser D (The Statistician) 47(3):431–443
25. Sanders L, Pumain D, Mathian H, Guérin-Pace F, Bura S (1997) Simpop: a multiagent system for the study of urbanism. Environ Plan B 24:287–306
26. Schmitt C (2014) Modélisation de la dynamique des systèmes de peuplement: de simpoplocal à simpopnet. Ph.D. thesis, Paris 1

27. Raimbault J, Cottineau C, Le Texier M, Le Néchet FL, Reuillon R (2019) Space matters: extending sensitivity analysis to initial spatial conditions in geosimulation models. J Artif Soc Soc Simul 22(4):10
28. Arentze T, van den Berg P, Timmermans H (2012) Modeling social networks in geographic space: approach and empirical application. Environ Plan A 44(5):1101–1120
29. Pigozzi BW (1980) Interurban linkages through polynomially constrained distributed lags. Geogr Anal 12(4):340–352
30. Chen Y (2009) Urban gravity model based on cross-correlation function and fourier analyses of spatio-temporal process. Chaos Solitons Fractals 41(2):603–614
31. Offner J-M, Pumain D (1996) Réseaux et territoires-significations croisées. Editions de l'Aube, La Tour d'Aigues
32. Offner J-M (1993) Les "effets structurants" du transport: mythe politique, mystification scientifique. Espace Géographique 22(3):233–242
33. Bretagnolle A (2009) Villes et réseaux de transport: des interactions dans la longue durée, France, Europe, États-Unis. Hdr, Université Panthéon-Sorbonne - Paris I
34. Raimbault J (2018) Caractérisation et modélisation de la co-évolution des réseaux de transport et des territoires. Ph.D. thesis, Université Paris 7 Denis Diderot
35. Batty M (2006) Hierarchy in cities and city systems. Hierarchy in natural and social sciences. Springer, Dordrecht, pp 143–168
36. Raimbault J (2018) Calibration of a density-based model of urban morphogenesis. PLoS ONE 13(9):0203516
37. EUROSTAT: Eurostat geographical data. http://ec.europa.eu/eurostat/web/gisco (2014)
38. Raimbault J (2018) Multi-modeling the morphogenesis of transportation networks. In: Artificial life conference proceedings. MIT Press, pp 382–383
39. Tero A, Takagi S, Saigusa T, Ito K, Bebber DP, Fricker MD, Yumiki K, Kobayashi R, Nakagaki T (2010) Rules for biologically inspired adaptive network design. Science 327(5964):439–442
40. Courtat T, Gloaguen C, Douady S (2011) Mathematics and morphogenesis of cities: a geometrical approach. Phys Rev E 83(3):036106
41. Raimbault J (2019) Multi-dimensional urban network percolation. arXiv preprint arXiv: 1903.07141
42. 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 Eur J Geogr. https://doi.org/10.4000/cybergeo.26753
43. Banos A, Genre-Grandpierre C (2012) Towards new metrics for urban road networks: some preliminary evidence from agent-based simulations. Agent-based models of geographical systems. Springer, Dordrecht, pp 627–641
44. Reuillon R, Leclaire M, Rey-Coyrehourcq S (2013) Openmole, a workflow engine specifically tailored for the distributed exploration of simulation models. Future Gener Comput Syst 29(8):1981–1990
45. Tisue S, Wilensky U (2004) Netlogo: a simple environment for modeling complexity. In: International conference on complex systems. New England Complex Systems Institute, Boston, pp 16–21
46. Raimbault J (2019) An urban morphogenesis model capturing interactions between networks and territories. The mathematics of urban morphology. Springer, Cham, pp 383–409
47. Mantegna RN, Stanley HE (2000) An Introduction to econophysics: correlations and complexity in finance. Cambridge University Press, Cambridge
48. Bouchaud JP, Potters M (2009) Financial applications of random matrix theory: a short review. ArXiv e-prints. arxiv: 0910.1205
49. Bonanno G, Lillo F, Mantegna RN (2001) Levels of complexity in financial markets. Phys A Stat Mech Appl 299:16–27. arxiv: cond-mat/0104369
50. Tumminello M, Aste T, Di Matteo T, Mantegna RN (2005) A tool for filtering information in complex systems. Proc Natl Acad Sci USA 102:10421–10426
51. Barndorff-Nielsen OE, Hansen PR, Lunde A, Shephard N (2011) Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. J Econom 162:149–169
52. Ramsey JB (2002) Wavelets in economics and finance: past and future. Stud Nonlinear Dyn Econom. https://doi.org/10.2202/1558-3708.1090
53. Bouchaud J-P, Potters M, Meyer M (2000) Apparent multifractality in financial time series. Eur Phys J B Condens Matter Complex Syst 13(3):595–599
54. Jarrow RA (1999) In honor of the nobel laureates Robert C. Merton and Myron S. Scholes: a partial differential equation that changed the world. J Econ Perspect 13:229–248
55. Tsay RS (2015) MTS: all-purpose toolkit for analyzing multivariate time series (MTS) and estimating multivariate volatility models. R package version 0.33. http://CRAN.R-project.org/package=MTS
56. Girres J-F, Touya G (2010) Quality assessment of the french openstreetmap dataset. Trans GIS 14(4):435–459
57. Chicheportiche R, Bouchaud J-P (2015) A nested factor model for non-linear dependencies in stock returns. Quant Finance 15(11):1789–1804
58. Chérel G, Cottineau C, Reuillon R (2015) Beyond corroboration: strengthening model validation by looking for unexpected patterns. PLoS ONE 10(9):0138212
59. Bourgine P, Chavalarias D et al (2009) French roadmap for complex systems 2008–2009. arXiv preprint arXiv :0907.2221

Publisher's Note