RESEARCH

# Second-order Control of Complex Systems with Correlated Synthetic Data

Juste Raimbault

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

**Abstract**

~~Generation of hybrid synthetic data resembling real data to some criteria is an important methodological and thematic issue in most disciplines which~~ The generation of synthetic data is an essential tool to study complex systems~~. Interdependencies between constituting elements, materialized within respective relations, lead to the emergence of macroscopic patterns. Being able to control the dependance structure and level within a synthetic dataset is thus a source of knowledge on system mechanisms. We describe in this paper a methodology consisting in~~ , 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 ~~datasets on which correlation structureis controlled. The method is applied in a first example on financial~~ 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~~and allows to understand the role of interferences between components at different scales on performances of a predictive model. A second application on a geographical system is then proposed, in which the weak coupling between a population density model and a network morphogenesis modelallows to simulate territorial configurations. The calibration on morphological objective on European data and intensive model exploration unveils a large spectrum of feasible correlations between morphological and network measures. We demonstrate therein the flexibility of our method and the variety of possible applications~~ . 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 ~~, and more particularly in~~ 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]. ~~It~~

Generation of synthetic datasets can consist in data disaggregation by ~~creation of~~ producing a microscopic population with fixed macroscopic properties ~~, or in the creation of new populations at the same scale than a given sample, with criteria of proximity to the real sample~~[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].

~~The criteria to evaluate the quality of a~~ 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 dataset will depend on expected applications~~and can for example vary from a restrictive~~ . They range for example from a strong statistical fit on given indicators, to weaker assumptions of similarity ~~in~~ on aggregated patterns. In the case of systems where emergence plays a ~~strong~~ central role, a microscopic property does not directly imply given macroscopic patterns, ~~which reproduction is indeed one aim of modeling and simulation practices in complexity science. With~~ 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~~and~~ , modeling and simulation~~(~~ , including the generation of synthetic data~~)~~ , are therefore crucial to be developed.

## Synthetic data and dependancy structures

~~Whereas first order (~~Reproducing data patterns at the first order, in the sense of distribution moments~~) is generally well used , it is not systematic nor simple to control generated data structure at~~ , 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, ~~i.e.~~ this corresponds to a control of the covariance structure between generated variables.

Some specific examples where interdependency structure is controlled can be found~~, such as in [12] where~~ . [12] investigates the sensitivity of discrete choices models to the distributions of inputs and to their dependance structure~~is examined.~~ . [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. [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. ~~We~~

## 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. [16] describe a method to generate cross-correlated random spatial fields using Fourier transforms. [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, these 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. [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 ~~method~~ methodology taking into account the dependance structure for the generation of synthetic datasets, more precisely ~~with the mean of controlled~~ by controlling the average of correlation matrices. It ~~can be applied to~~ is suited to be applied on cases where microscopic data ~~on the studied system~~ is not available and system similarity is ~~targeted~~ expected on aggregated indicators. ~~Our contribution relies in this genericmethodology, and its application to two very different kind of complex systems~~

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 ~~is formally described, to be then applied on two very different examples, namely financial~~ 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 illustrate the genericity of our method by applying it to financial time-series~~and territorial spatial configurations. Each example can be read independently and illustrates potentialities of the method and possible technical limitations. We discuss then possible further developments and applications, in particular for the geographical system~~, which are an other example of highly complex signals for which correlations are crucial.

## Method Formalization

~~Domain-specific methods aforementioned~~ The domain-specific methods described above are too broad to be summarized ~~into~~ within a same formalism. We ~~propose therefore here a rather~~ 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 (~~that~~ which can be indexed e.g. with time in the case of time-series, but also with space, or ~~discrete set abstract~~ any other indexation). We assume ~~given~~ 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~~ A given criteria of proximity to data is verified, i.e. given a precision $\varepsilon$ and ~~an indicator $f$~~some aggregated indicator $\vec{f}$, we have ~~$\|f(\mathbf{X}) - f(\tilde{\mathbf{X}})\| < \varepsilon$~~

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

2 The level of correlation is controlled, i.e. given a matrix ~~$R$ fixing~~ $\mathbf{R}$ representing the correlation structure (any symmetric matrix with coefficients in $[-1,1]$ and a unity diagonal), we have ~~$\hat{\text{Var}}\left[(\tilde{X}_i)\right] = \Sigma R \Sigma,$~~ the estimated covariance matrix given by

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

where the standard deviation diagonal matrix ~~$\Sigma$~~ $\mathbf{\Sigma}$ 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 (~~$R$~~ $\mathbf{R}$ itself is a parameter). Formally, ~~synthetic processes are~~ 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~~and territorial systems~~. We illustrate the flexibility of the method, and claim to help building interdisciplinary bridges by methodology transposition and reasoning analogy. In the first case, ~~proximity~~ morphological calibration of a population density distribution model allows to respect 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. ~~It follows a logic of hybrid data for which hypothesis or model testing is done on a more realistic context than on purely synthetic data.In the second case, morphological calibration of a population density distribution modelallows to respect real dataproximity.Correlations of urban form with transportation network measures are empirically obtained by exploration of coupling with a network morphogenesis model~~

## 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. An 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 $P_w$,

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

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

- a fraction $\beta$ of population is diffused to four closest neighbors 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. [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

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

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, $\gamma$ 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 $d_{ij}$. ~~The control is in this case indirect as feasible space is empirically determined.~~

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.

## ~~Correlated financial time-series~~

*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 $\vec{\alpha} = (\vec{\alpha}_D, \vec{\alpha}_N)$.

~~Context~~

*Indicators*

~~Our first field of application are financial complex systems, of which captured signals, financial time-series~~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 *betweeness-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 $\left(\vec{M}\left[D(\vec{\alpha})\right], \vec{G}\left[N(\vec{\alpha})\right]\right)_{1\leq i\leq n}$ with the standard unbiased estimator, given by Eq. 5 below.

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

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.

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

*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 500m resolution for which real indicator values have been computed on 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 $N_C \in [50, 120]$, $r_g \in [1, 100]$, $d_0 \in [0.1, 10]$, $k_h \in [0, 1]$, $\gamma \in [0.1, 4]$, $N_L$ For 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 $r_g^{(0)} \in \{0.25, 0.5, 0.75\}$, $N_N^{(0)} \in \{10, 15, 20\}$ 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 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 $\gamma$ 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 decrease centrality (redundant networks) to inverse processes (centralized networks).

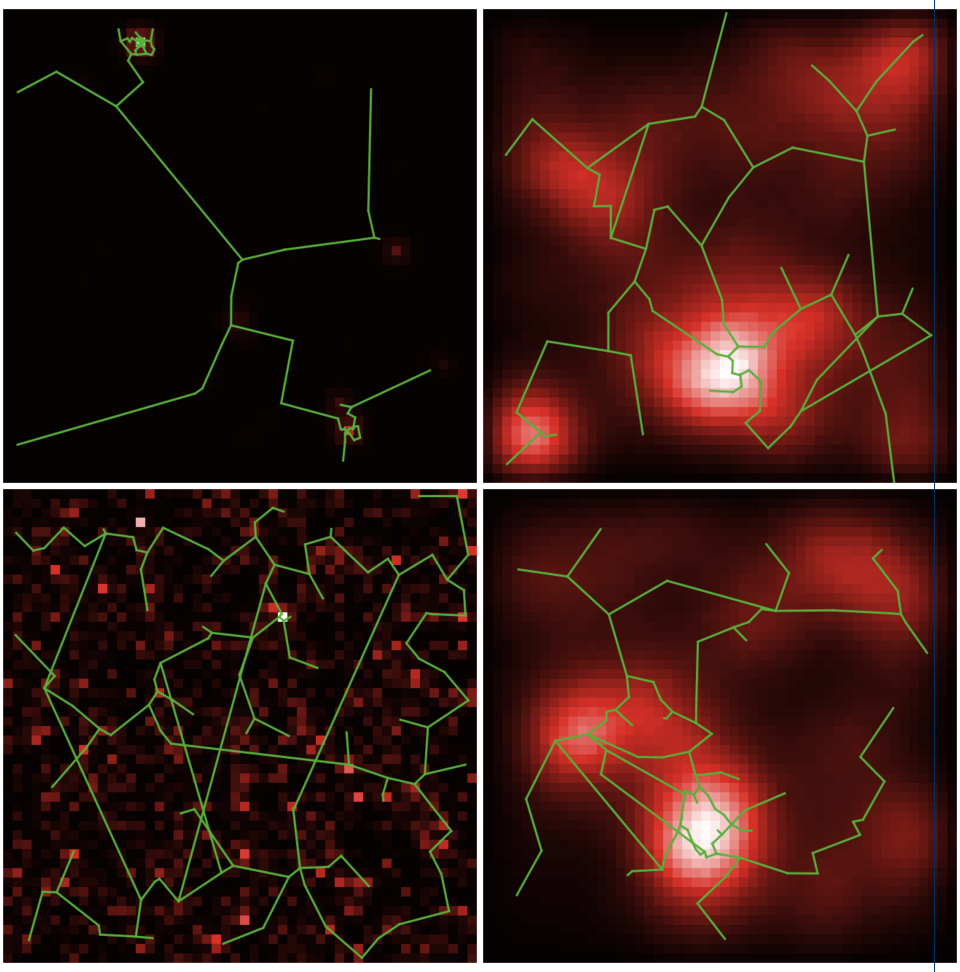

**Figure 1 Configurations obtained for parameters giving the four emphasized points in Fig. 2, in order from left to right and top to bottom.** We recognize polycentric city configurations (2 and 4), diffuse rural settlements (3) and aggregated weak density area (1). See appendice for exhaustive parameter 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.

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 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.

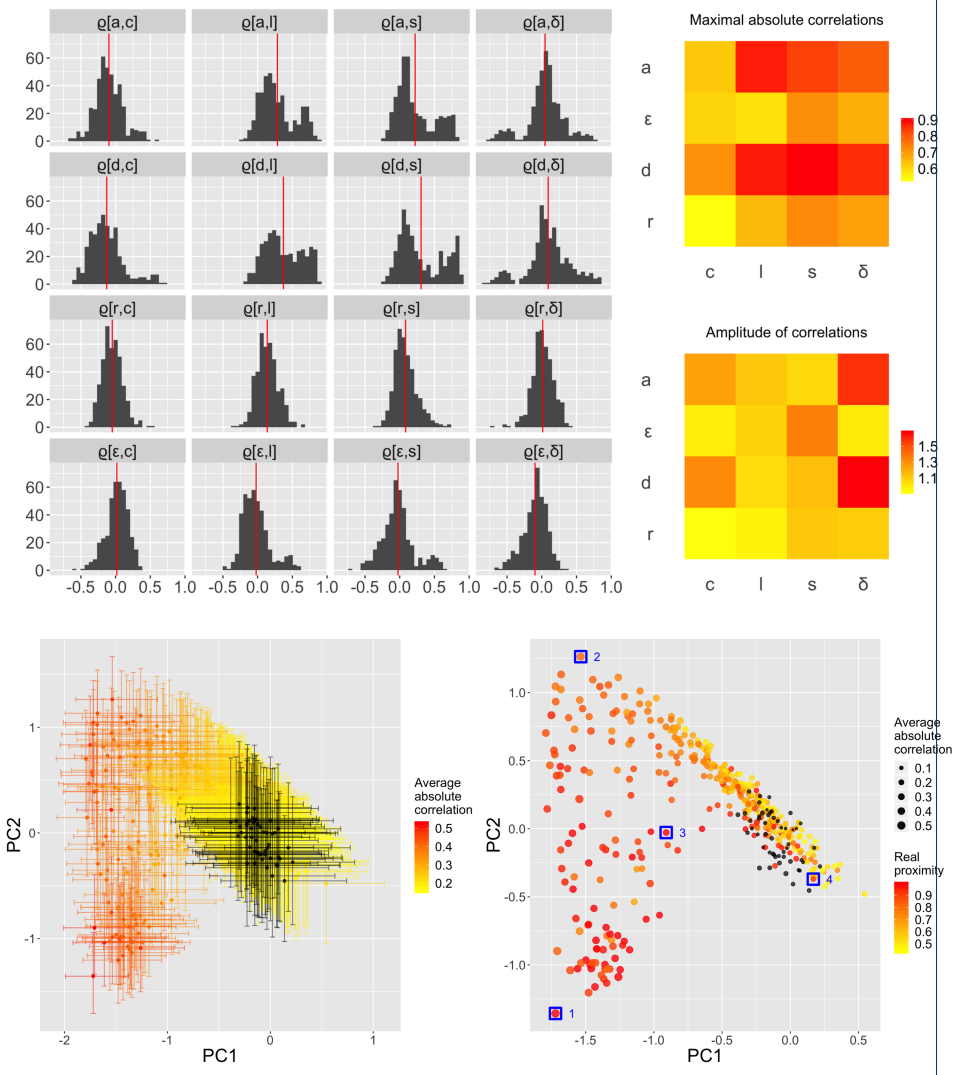


**Figure 2 Exploration of feasible space for correlations between urban morphology and network structure.** **(Top left)** Statistical distribution of crossed-correlations between vectors $\vec{M}$ of morphological indicators (in numbering order Moran index, mean distance, entropy, hierarchy) and $\vec{N}$ 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. **(Botton 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.

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 (Equation 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. [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 ~~highly~~ non-stationary [47]. Correlations ~~have already been the object of a broad part of related literature~~are broadly explored in that field. For example, Random Matrix Theory allows to distinguish signal from noise ~~, or at least to estimate the proportion of information undistinguishable from noise,~~ for a correlation matrix computed for a large number of asset with low-frequency signals~~(daily returns mostly)~~ , typically with a time scale of a day [48]. Similarly, Complex Network Analysis on networks constructed from correlations ~~, by~~ introduced methods such as Minimal Spanning Tree [49] or more refined ~~extensions developed for this purpose [50], yielded~~

~~promising results such as the reconstruction~~ topologically constrained network generation methods [50]. These provide reconstructions of economic sectors structure. At high ~~frequency~~frequencies, the precise estimation of interdependence parameters ~~in the framed of fixed assumptions on asset dynamics ,~~ 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 ~~estimator improvement~~ 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 1s). 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 $\omega$. A ~~recurrent~~ 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 $\omega_1$, formally an estimator $M_{\omega_1} : (T_i^{\omega_1}(t'))_{t'<t} \mapsto \hat{T}_i^{\omega_1}(t)$ which aims to minimize 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 ~~$\mathrm{Var}\left[\tilde{X}_i^{\omega_1}\right] = \Sigma R \Sigma$ (with $\Sigma$~~ $\mathrm{Var}\left[\tilde{X}_i^{\omega_1}\right] =^t \mathbf{\Sigma}\cdot\mathbf{R}\cdot\mathbf{\Sigma}$ (with $\mathbf{\Sigma}$ are estimated standard deviations and ~~$R$~~ $\mathbf{R}$ 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

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

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

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

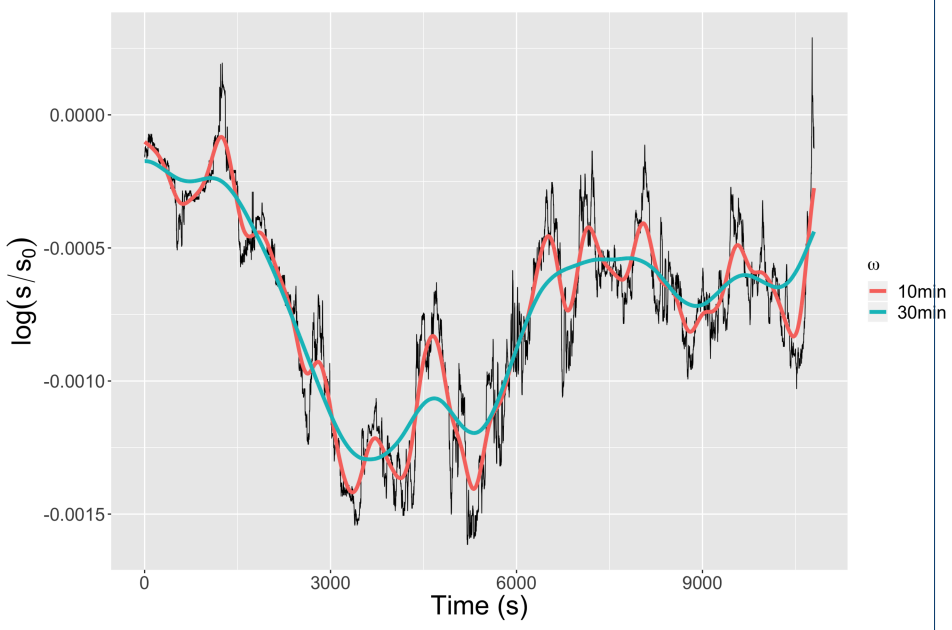


**Figure 3 Example of the multi-scalar structure of the signal, basis of the construction of synthetic signals.** *Log-prices* are represented on a time window of around 3h for November 1st 2015 for asset EUR/USD, with 10min (purple) and 30min trends. The low-frequency components are the basis to build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

frequency $\omega_1$ by filtering the signal, i.e. using signals constructed with Eq. 8 such that $\tilde{X}_i^{\omega_1} = W_i - \mathcal{F}_{\omega_0}[W_i]$ ~~(with~~, where $\mathcal{F}_{\omega_0}$ is a low-pass filter with cut-off frequency $\omega_0$~~)~~. We reconstruct then the hybrid synthetic signals by taking

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

~~Methodology~~

The method is tested on an example with two assets from foreign exchange market (EUR/USD and EUR/GBP), on a six 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 1s, 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* [55], 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 86400 for 1s sampling is reduced to a much smaller size of 432. We use a non-causal gaussian filter of total width $\omega$. We fix the fundamental frequency $\omega_0 = 24\text{h}$ and we propose to construct synthetic data at frequencies

$\omega_1 = 30\text{min}, 1\text{h}, 2\text{h}$. ~~See~~ We show in Fig. 3 ~~for~~ an example of signal structure at ~~these different scales~~ the scales $\omega_m$ and $\omega_1 = 30\text{min}$, compared with the non-filtered raw signal.

It is crucial to consider the interference between $\omega_0$ and $\omega_1$ frequencies in the reconstructed signal: the correlation which is indeed estimated is

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

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

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

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

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

~~Correlation is~~ Correlations are in practice estimated with a Pearson estimator, the covariances being corrected for bias, i.e.

$$\hat{\rho}[X1, X2] = \frac{\hat{C}[X1, X2]}{\sqrt{\hat{\text{Var}}\left[X1\right] \cdot \hat{\text{Var}}\left[X2\right]}}$$

~~, where $\hat{C}[X1, X2] = \frac{1}{(T-1)}\sum_t X_1(t)X_2(t) - \frac{1}{T\cdot(T-1)}\sum_t X_1(t)\sum_t X_2(t)$ and $\hat{\text{Var}}[X] = \frac{1}{T}\sum_t X^2(t) - \left(\frac{1}{T}\sum_t X(t)$~~ Eq. 5, Eq. 6 and Eq. 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 $M_{\omega_1}$ 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` [56] library for time-series models. Cleaned data and source code are available on an open `git` repository at `https://github.com/JusteRaimbault/SyntheticAsset`.

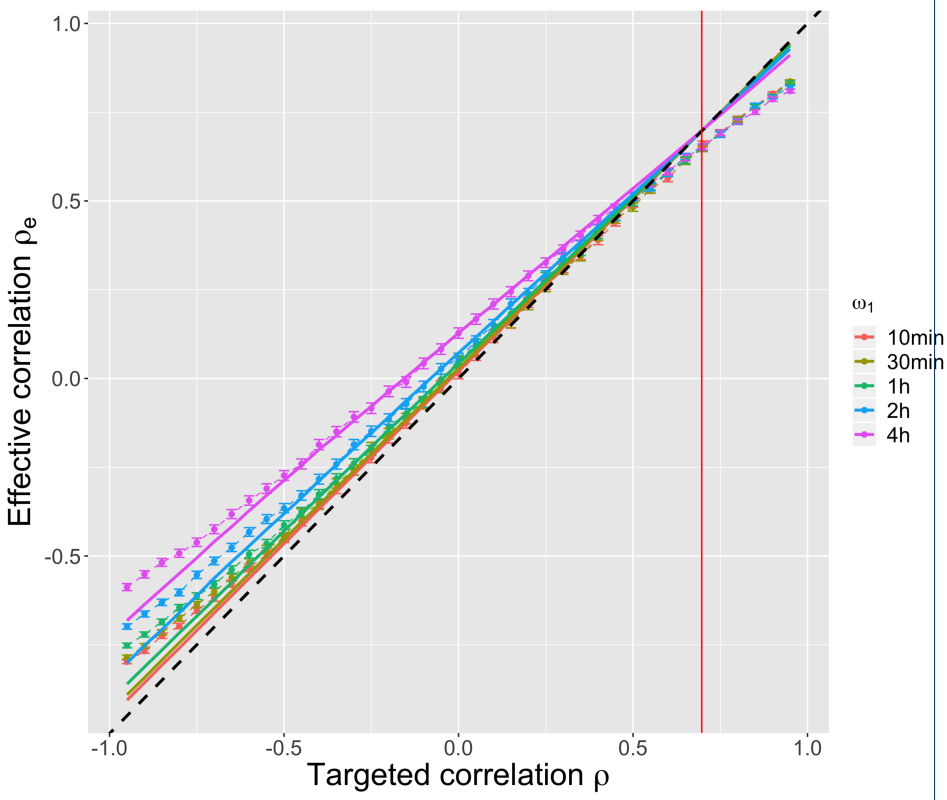


**Figure 4 Effective correlations obtained on synthetic data.** Dots represent estimated correlations on a synthetic dataset corresponding to 6 months between June and November 2015 (error-bars give 95% confidence intervals obtained with standard Fisher method); scale color gives the filtering frequency $\omega_1 = 10\text{min}, 30\text{min}, 1\text{h}, 2\text{h}, 4\text{h}$; solid lines give the theoretical values for $\rho_e$ 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 $\varepsilon_i$ 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 $\rho_0 \simeq 0.71$.

~~Results~~

~~Figure~~Fig. 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 $\rho_e$ and values predicted by Equation 11 in the interval $\rho \in [-0.5, 0.5]$. On the contrary, for larger absolute values, a deviation increasing with $|\rho|$ and as $\omega_1$ decreases : it confirms the intuition that when frequency decreases and becomes closer to $\omega_0$, interferences between the two components are not negligible anymore and invalidate independence assumptions for example.

~~We apply then the~~

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 ~~mean~~ average performance as a function of correlation between signals. Results for $\omega_1 = 1\text{h}, 1\text{h}30, 2\text{h}$ are shown in Fig. 5. The a priori counter-intuitive

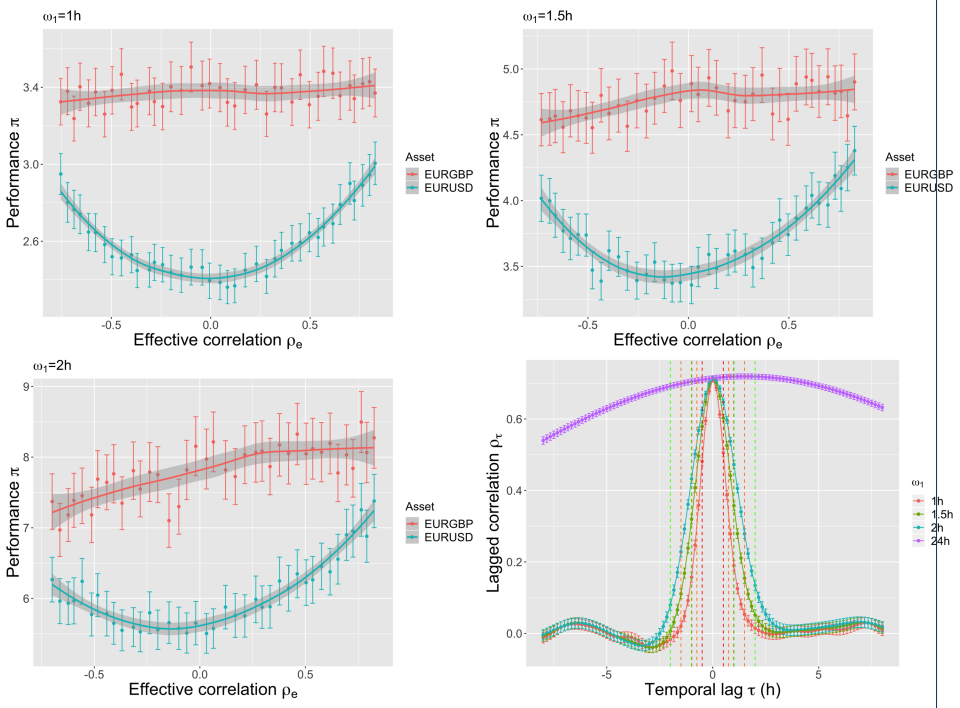


**Figure 5 Performance of a predictive model as a function of simulated correlations.** From left to right and top to bottom, ~~three first graphs~~ the plots show for each asset the normalized performance of an ARMA model ($p = 2, q = 0$), defined as $\pi = \left( \frac{1}{T} \sum_t \left( \tilde{X}_i(t) - M_{\omega_1} \left[ \tilde{X}_i \right](t) \right)^2 \right) / \sigma \left[ \tilde{X}_i \right]^2$ (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 show 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 $\tau$ units) used by the ARMA at each scale, what allows to read the corresponding lagged correlation on fundamental component.

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 he model will take them into account and will make false predictions because of the ~~markovian~~ 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 data (presence of consistent *lagged correlation* patterns, more realistic models than Black-Scholes) and apply it on more operational predictive models.

## Geographical data: correlated population density and road networkDiscussion

We now turn to a different type of system, namely geographical systems of human settlements, in the particular case here of population distribution in correlation with road network. In this exemple, no theoretical derivation can be done previously to correlated data generation, and simulation appears as a crucial step to implement the method.

### Context

The use of synthetic data in geography is generally directed towards the generation of synthetic populations within agent-based models (mobility, land-use transport interaction models) [3]. We can make a weak link with some techniques in spatial analysis. The extrapolation of a continuous spatial field from a discrete spatial sample through a kernel density estimation for example can be understood as the creation of a synthetic dataset (even if it is not generally the initial view, as in Geographically Weighted Regression [24] in which variable size kernels do not reconstruct data in a strict sense but extrapolate abstract variables representing interaction between explicit variables). In the field of modeling in quantitative geography, *toy-models* or hybrid models require a consistent initial spatial configuration. A set of possible initial configurations becomes a synthetic dataset on which the model is tested. The first Simpop model [25], precursor of a large family of models later parametrized with real data [57], could enter that frame but was studied on an unique synthetic spatialization. Similarly underlined was the difficulty to generate an initial transportation infrastructure in the case of the SimpopNet model [26] although it was admitted as a cornerstone of knowledge on the behavior of the model. A systematic control of spatial configuration effects on the behavior of simulation models was recently proposed [27], approach that can be interpreted as a statistical control on spatial data. 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. Such results are essential for the validation of conclusions obtained with modeling and simulation practices in quantitative geography.

### Territorial configuration model

#### Contributions

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 territoriesand 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]. We develop in that frame a simple coupling (i.e. without any feedback loop) between a 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 50km sized grid extracted from European density grid [37]. We recall here rapidly the processes in the model. An square grid of width $N$, initially empty, is represented by population $(P_i(t))_{1\leq i\leq N^2}$. 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 $\mathbb{P}[P_i(t+1) = P_i(t)+1 | P(t+1) = P(t)+1] = \frac{(P_i(t)/P(t))^\alpha}{\sum (P_i(t)/P(t))^\alpha}$
- a fraction $\beta$ of population is diffused to four closest neighbors is operated $n_d$ times

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

*Network model*

On the other hand, we are able to generate a planar transportation network by a model $N$, at a similar scale and given a density distribution. Because of the conditional nature to the density of the generation process, we will first have conditional estimators for network indicators, and secondly natural correlationsbetween network and urban shapes should appear as processes are not independent. The nature and modularity of these correlations as a function of model parameters are still to determine by exploration of the coupled model.

The heuristic network generation procedure is the following : A fixed number $N_c$ of centers that will be first nodes of the network si distributed given density distribution, following a similar law to 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 extent. Centers are connected deterministically by percolation between closest clusters : as soon as network is not connected, two closest connected components in the sense of minimal distance between each vertices are connected by the link realizing this distance. It yields a tree-shaped network. Network is modulated by potential breaking in order to be closer from real network shapes. More precisely, a generalized gravity potential between two centers $i$ investigated in this paper the possibility of generating synthetic data at a macrosopic 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 $j$ is defined by

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

where $d$ can be euclidian distance $d_{ij} = d(i,j)$ or network distance $d_N(i,j)$, $k_h \in [0,1]$ a weight to modulate role of populations, $\gamma$ giving shape of the hierarchy across population values, $r_g$ characteristic interaction distance and $d_0$ distance shape parameter. A fixed number $K \cdot N_L$ of potential new links is taken among couples having greatest euclidian distance potential ($K = 5$ is fixed). 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 $d_{ij}$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. Planarity of the network is imposed by creating nodes at possible intersections formed by new links. 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.

We insist on the fact that the network generation procedure is entirely heuristic and result of thematic assumptions (connected initial network, gravity-based link creation) combined with trial-and-error during first explorations. 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 percolation model than the initial one that would allow the creation of loops for example. We could thus in the frame of a modular architecture, in which the choice between different implementations of a functional brick can be seen as a meta-parameter [58], choose network generation function adapted to a specific need (as e.g. proximity to real data, constraints on output indicators , variety if generated forms, etc. ).

Parameter space for the coupled model is constituted 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 instead of both varying, i.e. the number of steps needed to generate the distribution) and network generation parameters $\vec{\alpha}_N = (N_C, k_h, \gamma, r_g, d_0)$. We denote $\vec{\alpha} = (\vec{\alpha}_D, \vec{\alpha}_N)$.

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

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

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

Results

Coupling of generative models is done both at formal and operational levels. We interface therefore independent implementations. The OpenMole software [44] for intensive model exploration offers for that the ideal frame thanks to its modular language allowing to construct *workflows* by task composition and interfacing with diverse experience plans and outputs. For operational reasons, density model is implemented in `scala` language as an OpenMole plugin, whereas network generation is implemented in agent-oriented language `NetLogo` [45] because of its possibilities for interactive exploration and heuristic model construction. Source code is available for reproducibility on an open git repository at .

We recognize polycentric city configurations (2 and 4), diffuse rural settlements (3) and aggregated weak density area (1). See appendice for exhaustive parameter values, indicators and corresponding correlations. For example $\bar{d}$ is highly correlated with $\bar{l}, \bar{s}$ ($\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 : $\rho[\bar{d}, \bar{c}] \simeq 0.34$ 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.

**(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 upper bounds after transformation are taken in principal plan. Scale color gives mean absolute correlation on full matrices. **(Botton 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.

*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 500m resolution for which real indicator values have been computed on whole Europe. Furthermore, a grid exploration of model behavior yields feasible output space in reasonable parameters bounds (roughly $\alpha \in [0.5, 2], N_G \in [500, 3000], P_m \in [10^4, 10^5], \beta \in [0, 0.2], n_d \in \{1, \ldots, 4\}$). 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. For a precise study of model behavior, see appendice giving regressions analysis capturing the behavior of coupled model. In order to illustrate synthetic data generation method, 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 $N_C \in [50, 120], r_g \in [1, 100], d_0 \in [0.1, 10], k_\hbar \in [0, 1], \gamma \in [0.1, 4], N_L$ For 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. Simulation results are available on the dataverse at .~~

~~Fig. 2 gives details of experiment results, while Fig. 1 shows examples of generated configurations. Regarding the subject of correlated synthetic data generation, we can summarize the results as follows: Empirical distributions of correlation coefficients between morphology and network indicators are not simple and some are bimodal (for example $\rho_{46} = \rho[\bar{r}, \bar{l}]$ between Moran index and mean path length). it is possible to modulate up to a relatively high level of correlation for all indicators, maximal absolute correlation varying between 0.6~~ 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 ~~0.9. Amplitude of correlations varies between 0.9 and 1.6, allowing a broad spectrum of values. Point cloud in principal plan has a large extent but is not uniform : it is not possible to modulate at will 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 correlation generation. Most correlated points are also the closest to real data, what confirms the intuition and stylized fact of a strong interdependence in reality. Concrete examples taken on particular points in the principal plan show that similar density profiles can yield very different correlation profiles.~~ 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.

This case study could be refined by extending correlation control method. A precise knowledge of $N$ behavior (statistical distributions on an exhaustive grid of parameter space) conditional to $D$ would allow to determine $N^{<-1>}|D$ and have more latitude in correlation generation. We could also apply specific exploration algorithms to reach exceptional configurations realizing an expected correlation level, or at least to obtain a better knowledge of the feasible space of correlations [59].

## Discussion

### Scientific positioning

Our overall approach enters a particular epistemological frame. On the one hand the multidisciplinary aspect, and on the other hand the importance of empirical component through computational exploration methods, make this approach typical of Complex Systems science, as it is recalled by the roadmap for Complex Systems having a similar structure [60]. It combines transversal research questions (horizontal integration of disciplines) with the development of heterogeneous multi-scalar approaches which encounter similar issues as the one we proposed to tackle (vertically integrated disciplines). 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. Results presented here is an illustration of such an hybrid paradigm.

### Direct applications Future work

Starting from the second example which was limited to data generation, we propose examples of direct applications that should give an overview of the range of possibilities. The 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 [61], with however some adjustments required on the generation procedure in order to avoid edge effects due its restrictive frame. This should theoretically allow to unveil parameter sets reproducing accurately existing configurations both for urban morphology and network shape. It could be then possible to derive a "theoretical correlation" for these, as an empirical correlation is according to some theories of urban systems not computable as a unique realization of stochastic processes is observed. Because of non-ergodicity of urban systems [62], there are strong chances that involved processes are different across different geographical areas (or from an other point of view that they are in an other state of meta-parameters, i.e. in an other regime) and that their interpretation as different realizations of the same stochastic process makes no sense, the impossibility of covariation estimation following. 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. As territorial configurations emerge from spatio-temporal interdependences between components of territorial systems, this intrinsic correlation emerges the same way, and its knowledge gives information on these interdependences and thus on relations between territories and networks.

~~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.~~

We studied in the ~~first~~ second example stochastic processes in the sense of random time-series, whereas time did not have a role in the ~~second~~ 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 ~~a~~ dynamical spatial models ~~we have because of feedbacks necessarily propagation effects and therefore~~ the existence of lagged interdependences in space and time [29] ~~. It would drive our field of study towards~~ is an important feature of complex dynamics. This would provide a better understanding of ~~dynamical~~ the link between static and dynamic correlations.

~~Generalization~~

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 [63] 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 [59]. 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 [60]. 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 is an illustration of such an hybrid paradigm.

**Availability of data and material**

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

**Competing interests**

The author declares to have no competing interests.

**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.

**Authors' contributions**

JR designed the study, did the analysis and wrote the paper.

**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 ). We thank the European Grid Infrastructure and its supporting National Grid Initiatives (France-Grilles in particular) for providing the technical support and infrastructure.

**References**

1. Abadie, A., Diamond, A., Hainmueller, J.: Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. Journal of the American Statistical Association **105**(490) (2010)
2. Moeckel, R., Spiekermann, K., Wegener, M.: Creating a synthetic population. In: Proceedings of the 8th International Conference on Computers in Urban Planning and Urban Management (CUPUM) (2003)
3. Pritchard, D.R., Miller, E.J.: Advances in agent population synthesis and application in an integrated land use and transportation model. In: Transportation Research Board 88th Annual Meeting (2009)
4. Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A.: A review of feature selection methods on synthetic data. Knowledge and information systems **34**(3), 483–519 (2013)
5. Van den Bulcke, T., Van Leemput, K., Naudts, B., van Remortel, P., Ma, H., Verschoren, A., De Moor, B., Marchal, K.: Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. BMC bioinformatics **7**(1), 43 (2006)
6. Beckman, R.J., Baggerly, K.A., McKay, M.D.: Creating synthetic baseline populations. Transportation Research Part A: Policy and Practice **30**(6), 415–429 (1996)
7. Müller, K., Axhausen, K.W.: Population synthesis for microsimulation: State of the art. Arbeitsberichte Verkehrs-und Raumplanung **638** (2010)
8. Barthelemy, J., Toint, P.L.: Synthetic population generation without a sample. Transportation Science **47**(2), 266–279 (2013)
9. Hoag, J.E.: Synthetic Data Generation: Theory, Techniques and Applications. University of Arkansas, ??? (2008)
10. Eno, J., Thompson, C.W.: Generating synthetic data to match data mining patterns. IEEE Internet Computing **12**(3), 78–82 (2008)
11. Arthur, W.B.: Complexity and the Shift in Modern Science. Conference on Complex Systems, Tempe, Arizona (2015)
12. Ye, X.: Investigation of underlying distributional assumption in nested logit model using copula-based simulation and numerical approximation. Transportation Research Record: Journal of the Transportation Research Board (2254), 36–43 (2011)
13. Birkin, M., Clarke, M.: Synthesis—a synthetic spatial information system for urban and regional analysis: methods and examples. Environment and planning A **20**(12), 1645–1671 (1988)
14. Li, H., Xiong, L., Jiang, X.: Differentially private synthesization of multi-dimensional data using copula functions. In: Advances in Database Technology: Proceedings. International Conference on Extending Database Technology, vol. 2014, p. 475 (2014). NIH Public Access
15. Newman, M.E.: The structure and function of complex networks. SIAM review **45**(2), 167–256 (2003)
16. Robin, M., Gutjahr, A., Sudicky, E., Wilson, J.: Cross-correlated random field generation with the direct fourier transform method. Water Resources Research **29**(7), 2385–2397 (1993)
17. Osborn, S., Vassilevski, P.S., Villa, U.: A multilevel, hierarchical sampling technique for spatially correlated random fields. SIAM Journal on Scientific Computing **39**(5), 543–562 (2017)
18. Gourdji, S., Hirsch, A., Mueller, K., Yadav, V., Andrews, A., Michalak, A.: Regional-scale geostatistical inverse modeling of north american co 2 fluxes: a synthetic data study. Atmospheric Chemistry and Physics **10**(13), 6151–6167 (2010)
19. Zhao, T., Wang, Y.: Simulation of cross-correlated random field samples from sparse measurements using bayesian compressive sensing. Mechanical Systems and Signal Processing **112**, 384–400 (2018)
20. Benenson, I., Torrens, P.: Geosimulation: Automata-based Modeling of Urban Phenomena. John Wiley & Sons, ??? (2004)
21. Batty, M.: The New Science of Cities. MIT Press, ??? (2013)
22. Pumain, D.: An evolutionary theory of urban systems. In: International and Transnational Perspectives on Urban Systems, pp. 3–18. Springer, ??? (2018)
23. Banos, A., Chardonnel, S., Lang, C., Marilleau, N., Thévenin, T.: 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 (2005)
24. Brunsdon, C., Fotheringham, S., Charlton, M.: Geographically weighted regression. Journal of the Royal Statistical Society: Series D (The Statistician) **47**(3), 431–443 (1998)
25. Sanders, L., Pumain, D., Mathian, H., Guérin-Pace, F., Bura, S.: Simpop: a multiagent system for the study of urbanism. Environment and Planning B **24**, 287–306 (1997)

26. Schmitt, C.: Modélisation de la dynamique des systèmes de peuplement: de simpoplocal à simpopnet. PhD thesis, Paris 1 (2014)
27. Raimbault, J., Cottineau, C., Le Texier, M., Le Néchet, F.L., Reuillon, R.: Space matters: extending sensitivity analysis to initial spatial conditions in geosimulation models. Journal of Artificial Societies and Social Simulation **22**(4) (2019)
28. Arentze, T., van den Berg, P., Timmermans, H.: Modeling social networks in geographic space: approach and empirical application. Environment and Planning A **44**(5), 1101–1120 (2012)
29. Pigozzi, B.W.: Interurban linkages through polynomially constrained distributed lags. Geographical Analysis **12**(4), 340–352 (1980)
30. Chen, Y.: Urban gravity model based on cross-correlation function and fourier analyses of spatio-temporal process. Chaos, Solitons & Fractals **41**(2), 603–614 (2009)
31. Offner, J.-M., Pumain, D.: Réseaux et territoires-significations croisées (1996)
32. Offner, J.-M.: Les "effets structurants" du transport: mythe politique, mystification scientifique. Espace géographique **22**(3), 233–242 (1993)
33. Bretagnolle, A.: Villes et réseaux de transport : des interactions dans la longue durée, France, Europe, États-Unis. Hdr, Université Panthéon-Sorbonne - Paris I (June 2009)
34. Raimbault, J.: Caractérisation et modélisation de la co-évolution des réseaux de transport et des territoires. PhD thesis, Université Paris 7 Denis Diderot (2018)
35. Batty, M.: Hierarchy in cities and city systems. In: Hierarchy in Natural and Social Sciences, pp. 143–168. Springer, ??? (2006)
36. Raimbault, J.: Calibration of a density-based model of urban morphogenesis. PloS one **13**(9), 0203516 (2018)
37. EUROSTAT: Eurostat Geographical Data. http://ec.europa.eu/eurostat/web/gisco (2014)
38. Raimbault, J.: Multi-modeling the morphogenesis of transportation networks. In: Artificial Life Conference Proceedings, pp. 382–383 (2018). MIT Press
39. Tero, A., Takagi, S., Saigusa, T., Ito, K., Bebber, D.P., Fricker, M.D., Yumiki, K., Kobayashi, R., Nakagaki, T.: Rules for biologically inspired adaptive network design. Science **327**(5964), 439–442 (2010)
40. Courtat, T., Gloaguen, C., Douady, S.: Mathematics and morphogenesis of cities: A geometrical approach. Physical Review E **83**(3), 036106 (2011)
41. Raimbault, J.: Multi-dimensional urban network percolation. arXiv preprint arXiv:1903.07141 (2019)
42. Le Néchet, F.: De la forme urbaine à la structure métropolitaine: une typologie de la configuration interne des densités pour les principales métropoles européennes de l'audit urbain. Cybergeo: European Journal of Geography (2015)
43. Banos, A., Genre-Grandpierre, C.: Towards new metrics for urban road networks: Some preliminary evidence from agent-based simulations. In: Agent-based Models of Geographical Systems, pp. 627–641. Springer, ??? (2012)
44. Reuillon, R., Leclaire, M., Rey-Coyrehourcq, S.: Openmole, a workflow engine specifically tailored for the distributed exploration of simulation models. Future Generation Computer Systems **29**(8), 1981–1990 (2013)
45. Wilensky, U.: Netlogo (1999)
46. Raimbault, J.: An urban morphogenesis model capturing interactions between networks and territories. In: The Mathematics of Urban Morphology, pp. 383–409. Springer, ??? (2019)
47. Mantegna, R.N., Stanley, H.E., *et al.*: An Introduction to Econophysics: Correlations and Complexity in Finance vol. 9. Cambridge university press Cambridge, ??? (2000)
48. Bouchaud, J.P., Potters, M.: Financial Applications of Random Matrix Theory: a short review. ArXiv e-prints (2009). 0910.1205
49. Bonanno, G., Lillo, F., Mantegna, R.N.: Levels of complexity in financial markets. Physica A Statistical Mechanics and its Applications **299**, 16–27 (2001). cond-mat/0104369
50. Tumminello, M., Aste, T., Di Matteo, T., Mantegna, R.N.: A tool for filtering information in complex systems. Proceedings of the National Academy of Sciences of the United States of America **102**, 10421–10426 (2005)
51. Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A., Shephard, N.: 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 (2011)
52. Ramsey, J.B.: Wavelets in economics and finance: Past and future. Studies in Nonlinear Dynamics & Econometrics **6** (2002)
53. Bouchaud, J.-P., Potters, M., Meyer, M.: Apparent multifractality in financial time series. The European Physical Journal B-Condensed Matter and Complex Systems **13**(3), 595–599 (2000)
54. Jarrow, R.A.: 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, 229–248 (1999)
55. Mantegna, R.N., Stanley, H.E.: Introduction to Econophysics: Correlations and Complexity in Finance. Cambridge university press, ??? (1999)
56. Tsay, R.S.: MTS: All-Purpose Toolkit for Analyzing Multivariate Time Series (MTS) and Estimating Multivariate Volatility Models. (2015). R package version 0.33. http://CRAN.R-project.org/package=MTS
57. Pumain, D.: Multi-agent system modelling for urban systems: The series of simpop models. In: Agent-based Models of Geographical Systems, pp. 721–738. Springer, ??? (2012)
58. Cottineau, C., Chapron, P., Reuillon, R.: Growing models from the bottom up. an evaluation-based incremental modelling method (ebimm) applied to the simulation of systems of cities. Journal of Artificial Societies and Social Simulation **18**(4), 9 (2015)
59. Chérel, G., Cottineau, C., Reuillon, R.: Beyond corroboration: Strengthening model validation by looking for unexpected patterns. PLoS ONE **10**(9), 0138212 (2015)
60. Bourgine, P., Chavalarias, D., al.: French Roadmap for complex Systems 2008-2009. arXiv preprint arXiv:0907.2221 (2009)
61. Girres, J.-F., Touya, G.: Quality assessment of the french openstreetmap dataset. Transactions in GIS **14**(4), 435–459 (2010)

62. Pumain, D.: Urban systems dynamics, urban growth and scaling laws: The question of ergodicity. In: Complexity Theories of Cities Have Come of Age, pp. 91–103. Springer, ??? (2012)
63. Chicheportiche, R., Bouchaud, J.-P.: A nested factor model for non-linear dependences in stock returns. arXiv preprint arXiv:1309.3102 (2013)