Correlated Synthetic Data for Controlling Complex Systems

Juste Raimbault[a,b,1]

[a]UMR CNRS 8504 Géographie-cités, Paris, France; [b]UMR-T IFSTTAR 9403 LVMT, Champs-sur-Marne, France

This manuscript was compiled on February 3, 2017

Generation of hybrid synthetic data resembling real data to some criteria is an important methodological and thematic issue in most disciplines which 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 propose a methodology consisting in the generation of synthetic datasets on which correlation structure is controlled. The method is applied in a first example 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 model allows 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.

Synthetic Data | Statistical Control | Correlations | Financial Time-series | Land-use Transportation Interactions

The use of synthetic data, in the sense of statistical populations generated randomly under constraints of patterns proximity to the studied system, is a widely used methodology, and more particularly in disciplines related to complex systems such as therapeutic evaluation (1), territorial science (2, 3), machine learning (4) or bio-informatics (5). It can consist in data desegregation by creation of 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. These criteria will depend on expected applications and can for example vary from a restrictive statistical fit on given indicators, to weaker assumptions of similarity in aggregated patterns. In the case of chaotic systems, or systems where emergence plays a strong 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 the rise of new computational paradigms (6), 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.

Whereas first order (in the sense of distribution moments) is generally well used, it is not systematic nor simple to control generated data structure at second order, i.e. covariance structure between generated variables. Some specific examples can be found, such as in (7) where the sensitivity of discrete choices models to the distributions of inputs and to their dependance structure is examined. It is also possible to interpret complex networks generative models (8) as the production of an interdependence structure for a system, contained within link topology. We introduce here a generic method taking into account dependance structure for the generation of synthetic datasets, more precisely with the mean of controlled correlation matrices.

The rest of the paper is organized as follows. The generic method is formally described, to be then applied on very different examples both entering application frame. 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 a geographical system.

Method Formalization

Domain-specific methods aforementioned are too broad to be summarized into a same formalism. We propose a framework as generic as possible, centered on the control of correlations structure in synthetic data.

Let $\vec{X}_I$ a multidimensional stochastic process (that can be indexed e.g. with time in the case of time-series, but also space, or discrete set abstract indexation). We assume given a real dataset $\mathbf{X} = (X_{i,j})$, 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 an indicator f, we have $\|f(\mathbf{X}) - f(\tilde{\mathbf{X}})\| < \varepsilon$
2. level of correlation is controlled, i.e. given a matrix R fixing correlation structure (symmetric matrix with coefficients in $[-1, 1]$ and unity diagonal), we have $\hat{\text{Var}}\left[(\tilde{X}_i)\right] = \Sigma R \Sigma$, where the standard deviation diagonal matrix Σ is estimated on the synthetic population.

The second requirement will generally be conditional to parameter values determining generation procedure, either

Significance Statement

Authors must submit a 120-word maximum statement about the significance of their research paper written at a level understandable to an undergraduate educated scientist outside their field of speciality. The primary goal of the Significance Statement is to explain the relevance of the work in broad context to a broad readership. The Significance Statement appears in the paper itself and is required for all research papers.

Please provide details of author contributions here.

Please declare any conflict of interest here.

[1]A.O.(Author One) and A.T. (Author Two) contributed equally to this work (remove if not applicable).

[2]To whom correspondence should be addressed. E-mail: author.twoemail.com

generation models being simple or complex (R itself is a parameter). Formally, synthetic processes are parametric families $\tilde{X}_i[\vec{\alpha}]$. We propose to apply the methodology on very different examples, both typical of complex systems : 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 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 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 as feasible space is empirically determined.

Applications

Application : financial time-series.

Context. Our first field of application is that of financial complex systems, of which captured signals, financial time-series, are heterogeneous, multi-scalar and highly non-stationary (9). Correlations have already been the object of a broad bunch of related literature. For example, Random Matrix Theory allows to undress signal of 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) (10). Similarly, Complex Network Analysis on networks constructed from correlations, by methods such as Minimal Spanning Tree (11) or more refined extensions developed for this purpose (12), yielded promising results such as the reconstruction of economic sectors structure. At high frequency, the precise estimation of of interdependence parameters in the framed of fixed assumptions on asset dynamics, has been extensively studied from a theoretical point of view aimed at refinement of models and estimators (13). 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 equal otherwise*. For example, (14) obtains a bias correction for the *Hayashi-Yoshida* estimator (used to estimate integrated covariation between two brownian at high frequency in the case of asynchronous observation times) by deriving a central limit theorem for a general model that endogeneize observation times. Empirical confirmation of estimator improvement is obtained on a synthetic dataset at a fixed correlation level.

Formalization.

Framework 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 (15) or in multi-fractal signal processing (16)) to interpret observed signals as the superposition of components at different time scales : $X_i = \sum_\omega X_i^\omega$. We denote by $T_i^\omega = \sum_{\omega'\leq\omega} X_i^\omega$ the filtered signal at a given frequency ω. A recurrent 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 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 (18), 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 Σ estimated standard deviations and R fixed correlation matrix) and verifying $X_i^{\omega\leq\omega_0} = \tilde{X}_i^{\omega\leq\omega_0}$ (data proximity indicator : components at a lower frequency than a fundamental frequency $\omega_0 < \omega_1$ are identical). 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

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

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

$$\tilde{X}_i = T_i^{\omega_0} + \tilde{X}_i^{\omega_1} \qquad [1]$$

Implementation and Results.

Methodology The method is tested on an example with two assets from foreign exchange market (EUR/USD and EUR/GBP), in a six month period from June 2015 to November 2015. Data† cleaning, 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*, 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‡. 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 $\omega_1 = 30\text{min}, 1\text{h}, 2\text{h}$. See Fig. 1 for an example of signal structure at these different scales.

*see (17) for an extended discussion on the construction of *schema* to study complex adaptive systems (by complex adaptive systems).

†obtained from `http://www.histdata.com/`, without specified licence. For the respect of copyright, only cleaned and filtered at ω_m data are made openly available.

‡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.

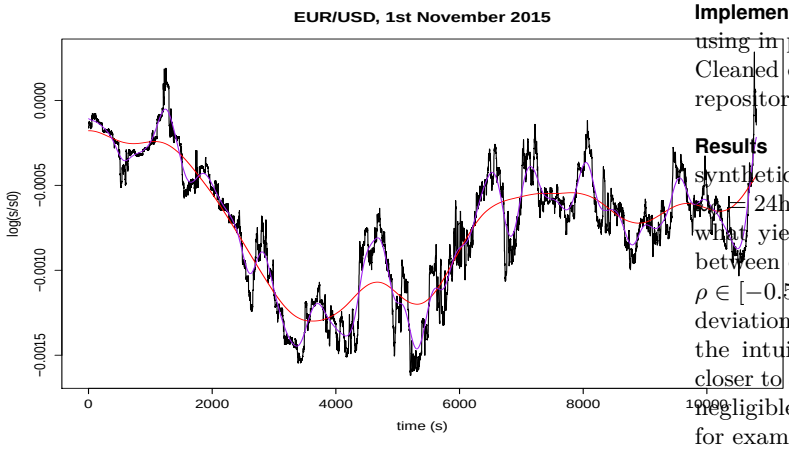


Fig. 1. Example of the multi-scalar structure of the signal, basis of the construction of synthetic signals | *Log-prices* are represented on a time window of around 3h for November 1st 2015 for asset EUR/USD, with 10min (purple) and 30min trends.

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

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

what yields in the reasonable limit $\sigma_1 \gg \sigma_0$ (fundamental frequency small enough), when $\mathrm{Cov}\left[\Delta\tilde{X}_i^{\omega_1}, \Delta X_j^{\omega}\right] = 0$ for all $i, j, \omega_1 > \omega$ and returns centered at any scale, the correction on effective correlation due to interferences : we have at first order the expression of effective correlation

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

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

Correlation is estimated by Pearson method, with estimator for covariance corrected for bias, i.e.

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

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

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[§]. More precisely, 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 $t + 1$.

[§] adaptation level staying low, as parameters T_W, p, q and model type do not vary. We are positioned within the framework of (19) which assumes a locally parametric dynamic but for which meta-parameters are fixed. We could imagine a variable T_W which would adapt for the best local fit, the same way parameters are estimated in bayesian signal processing by augmentation of the state with parameters.

Implementation Experiments are implemented in `R` language, using in particular the `MTS` (20) library for time-series models. Cleaned data and source code are openly available on the `git` repository of the project[¶].

Results Figure 2 gives effective correlations computed on synthetic data. For standard parameter values (for example $T_W = 24\mathrm{h}$, $\omega_1 = 2\mathrm{h}$ 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 2 in the interval $\rho \in [-0.5, 0.5]$. On the contrary, for larger absolute values, a deviation increasing 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.

We apply then the predictive model described above to synthetic data, in order to study its mean performance as a function of correlation between signals. Results for $\omega_1 = 1\mathrm{h}, 1\mathrm{h}30, 2\mathrm{h}$ are shown in Fig. 3. 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 he model will take them into account and will make false predictions because of the markovian character of simulated brownian[‖].

This case study stays a *toy-model* and has no direct practical application, but demonstrates however 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.

Application : geographical data of density and network.

Context. The use of synthetic data in geography is generally directed towards the generation of synthetic populations within agent-based models (mobility, *LUTI* models) (3). We can make a weak link with some Spatial Analysis techniques. 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 (21) in which variable size kernels do not interpolate data *stricto sensu* 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 (22), precursor of a large family of models later parametrized with real data, could enter that frame but was studied on an unique synthetic spatialization. Similarly

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

[‖] the model used has theoretically no predictive power at all on pure brownian

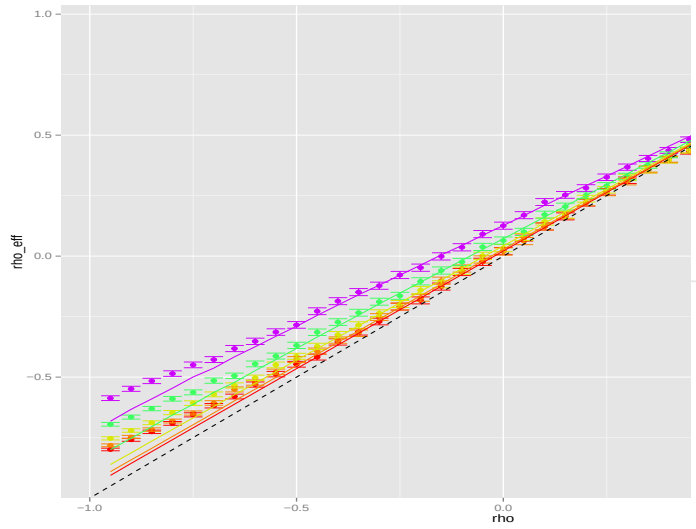


Fig. 2. Effective correlations obtained on synthetic data | Dots represent estimated correlations on a synthetic dataset corresponding to 6 months between June and November 2015 (error-bars give 95% confidence intervals obtained with standard Fisher method) ; scale color gives filtering frequency $\omega_1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

underlined was the difficulty to generate an initial transportation infrastructure in the case of the SimpopNet model (23) 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 only recently proposed (24), 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.

Formalization. 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 (25) but stays highly complex and difficult to quantify (26). A dynamical modeling of implied processes should shed light on these interactions ((27), p. 162-163). 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 (28) to generate a discrete spatial distribution of population density. A generalization of the basic model is proposed in (29), providing a calibration on morphological objectives (entropy, hierarchy, spatial auto-correlation, mean distance) against real values computed on the set of 50km

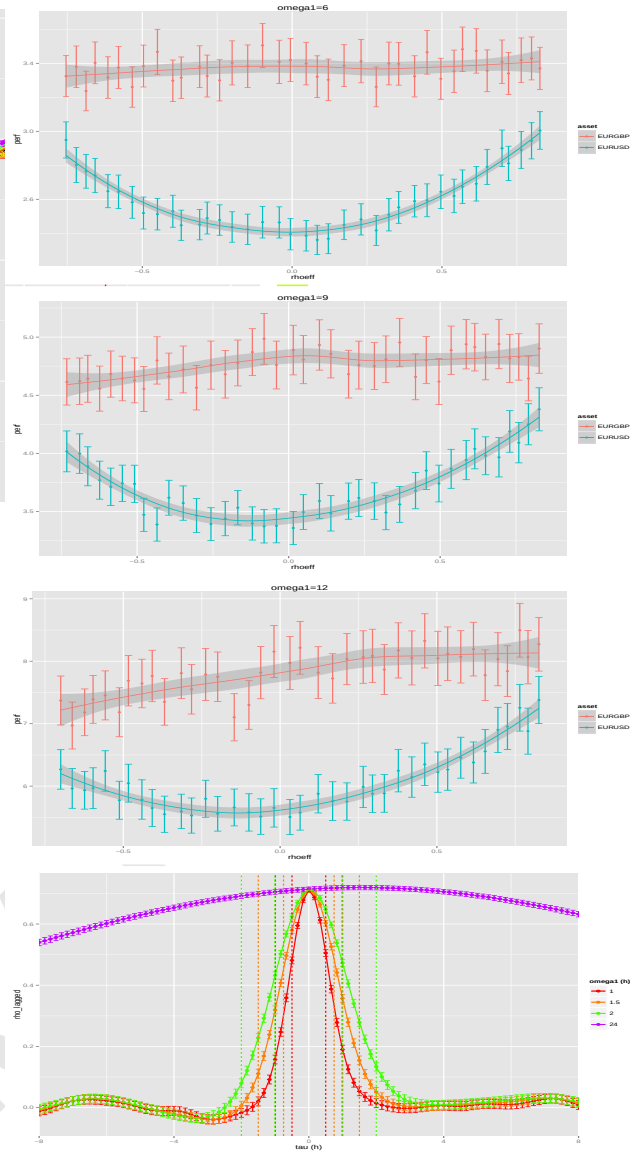


Fig. 3. Performance of a predictive model as a function of simulated correlations | From left to right and top to bottom, three first graphs show for each asset the normalized performance of an ARMA model ($p = 2, q = 0$), defined as $\pi = \left(\frac{1}{T}\sum_t \left(\tilde{X}_i(t) - M_{\omega_1}\left[\tilde{X}_i\right](t)\right)^2\right)/\sigma\left[\tilde{X}_i\right]^2$ (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 : 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 τ units) used by the ARMA at each scale, what allows to read the corresponding lagged correlation on fundamental component.

sized grid extracted from european density grid (30). More precisely, the model proceeds iteratively the following way. 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 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 β of population is diffused to four closest neighbors is operated n_d times

The two contradictory 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.

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 correlations between 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 :

1. 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.

2. 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.

3. 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 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, γ giving shape of the hierarchy across population values, r_g characteristic interaction distance and d_0 distance shape parameter.

4. A fixed number $K \cdot N_L$ of potential new links is taken among 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 forced by creation of nodes at possible intersections created by new links.

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 biological self-generated networks (**?**), local network growth based on geometrical constraints optimization (31), 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 (32), 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 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 $\vec{\alpha} = (\vec{\alpha}_D, \vec{\alpha}_N)$.

Indicators Urban form and network structure are quantified by numerical indicators in order to modulate correlations between these. Morphology is defined as a vector $\vec{M} = (r, \bar{d}, \varepsilon, a)$ giving spatial auto-correlation (Moran index), mean distance, entropy and hierarchy (see (33) 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 (34) 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 $\delta = \max_{ij} d_N(i,j)$.

Covariance and correlation 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 $(\vec{M}\left[D(\vec{\alpha})\right], \vec{G}\left[N(\vec{\alpha})\right])_{1\leq i\leq n}$ with standard unbiased estimator. We estimate correlation with associated Pearson estimator.

Implementation. Coupling of generative models is done both at formal and operational levels. We interface therefore independent implementations. The OpenMole software (35) 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

[**] Weak coupling allows to limit the total number of parameters as a strong coupling would involve retroaction loops and consequently associated parameters to determine their structure and intensity. In order to diminish it, an integrated model would be preferable to a strong coupling, what is slightly different in the sense where it is not possible in the integrated model to freeze one of the subsystems to obtain a model of the other subsystem that would correspond to the non-coupled model.

is implemented in `scala` language as an OpenMole `plugin`, whereas network generation is implemented in agent-oriented language `NetLogo` (36) because of its possibilities for interactive exploration and heuristic model construction. Source code is available for reproducibility on project repository[††].

Results. The study of density model alone is developed in (29). 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.

At given 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.

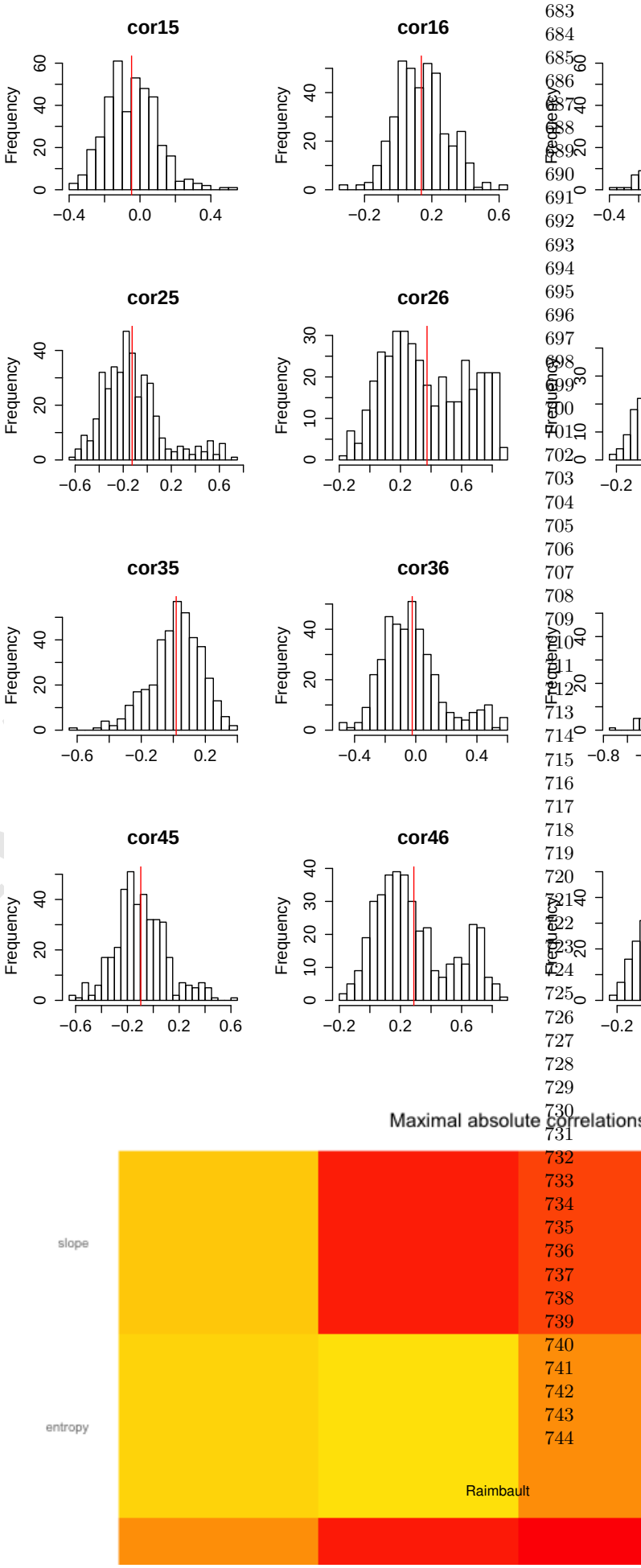
Given the large relative dimension of parameter space, an exhaustive grid exploration is not possible. We use a Latin Hypercube sampling procedure with bounds given above for $\vec{\alpha}_D$ and for $\vec{\alpha}_N$, we take $N_C \in [50, 120], r_g \in [1, 100], d_0 \in [0.1, 10], k_h \in [0, 1], \gamma \in [0.1, 4], N_L \in [4, 20]$. For 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. Figure 8 gives details of experiment results. Regarding the subject of correlated synthetic data generation, we can sum up the main lines as following :

- Empirical distributions of correlation coefficients between morphology and network indicators are not simple and some are bimodal (for example $\rho_{46} = \rho[r, \bar{l}]$ 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 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.

[††] at `https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Synthetic`



- Concrete examples taken on particular points in the principal plan show that similar density profiles can yield very different correlation profiles.

Possible developments. 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 (37).

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 (38). 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. 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.

- Calibration of network generation component at given density, on real data for transportation network (typically road network given the shape of generated networks ; it should be straightforward to use OpenStreetMap open data[‡‡] that have a reasonable quality for Europe, at least for France (39), with however adjustments on generation procedure in order to avoid edge effects due its restrictive frame, for example by generating on an extended surface to keep only a central area on which calibration would be done) 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 (40), 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.

[‡‡] https://www.openstreetmap.org

- As already mentioned, most of models of simulation 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 (24). 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 example stochastic processes in the sense of random time-series, whereas time did not have a role in the second 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 (41). It would drive our field of study towards a better understanding of dynamical 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 (42) is in an other perspective an interesting possible development.

Conclusion

We proposed an abstract method to generate synthetic datasets in which correlation structure is controlled. Its rapid 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 stays an open issue.

ACKNOWLEDGMENTS. Please include your acknowledgments here, set in a single paragraph. Please do not include any acknowledgments in the Supporting Information, or anywhere else in the manuscript.

1. Abadie A, Diamond A, Hainmueller J (2010) Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association* 105(490).
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*. No. 09-1686.
4. Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A (2013) A review of feature selection methods on synthetic data. *Knowledge and information systems* 34(3):483–519.
20. Tsay RS (2015) *MTS: All-Purpose Toolkit for Analyzing Multivariate Time Series (MTS) and Estimating Multivariate Volatility Models*. R package version 0.33.
21. Brunsdon C, Fotheringham S, Charlton M (1998) Geographically weighted regression. *Journal of the Royal Statistical Society: Series D (The Statistician)* 47(3):431–443.
22. Sanders L, Pumain D, Mathian H, Guérin-Pace F, Bura S (1997) Simpop: a multiagent system for the study of urbanism. *Environment and Planning B* 24:287–306.
23. Schmitt C (2014) Ph.D. thesis (Paris 1).
24. Cottineau C, Le Néchet F, Le Texier M, Reuillon R (2015) Revisiting some geography classics with spatial simulation in *Plurimondi. An International Forum for Research and Debate on Human Settlements*. Vol. 7.