Exploration methods for simulation models

Juste Raimbault et Denise Pumain

Abstract

We recall first in this chapter to what extent simulation models are an absolute necessity in social sciences and humanities, which can only very exceptionally require to experimental sciences methods to construct their knowledge. Models open the perspective to simulate social processes by replacing the complex interplay of individual and collective actions and reactions to the situations they make emerge by simpler mathematical or computational mechanisms, fostering an easier understanding of the relations between causes and consequences of these interactions and to make predictions. The formalisation through mathematical models able to offer analytical solutions being most often not possible in order to provide satisfying representations of social complexity (Jensen, 2018), computational models based on agents are more and more used. For long the limited computational capabilities of computer have forbidden to program models taking into account interactions between large numbers of entities geographically localized (individuals or territories). In principle these models should inform on the possibilities and conditions of the emergence of given configurations defined at a macro-geographical level from interactions occurring at a micro-geographical level, within systems with a too much complex behavior to be understood by a human brain. This however requires to study the dynamical behavior of these models including non-linear feedback effects and verify they produce plausible results at all stages of their simulation. This necessary stage of the exploration of the dynamics of algorithms remained rather rudimentary until the end of the last decade, when algorithms including more sophisticated methods such as evolutionary computation and the use of distributed high performance computing have allowed a significant qualitative leap forward in the validation of models, and even an epistemological turn for social sciences and humanities, as suggest the latest applications realized with the OpenMOLE platform described here.

1 Social sciences and experimentation

Experimentation played a significant role in the construction of natural sciences, since it consists in simulating material, physical, chemical or biological processes, through the use of apparatus imagined by researchers to select, often by isolating them, chains of facts that are simpler than the ones occurring in a complex reality. The confrontation of results of these experiments to observational data, partly or totally foreign to the data used to construct the experimental apparatus, is considered as bringing a proof of truth or of accuracy of the explicative reasoning at the basis of the model construction, more or less robust depending on the quality of the fit between model predictions and observations. We however know that the accuracy of a model predictions is not sufficient to fully validate the correspondance between the explicative mechanism imagined by the builders of the experimental apparatus and processes at work in the studied system, but this remains a crucial stage in the construction of models and theories enriched by observations.

In social sciences and humanities, the elaboration of experimental apparatus is highly problematic since it is confronted to numerous practical and ethical obstacles. Ethical and political critic questions the manipulation of individuals and the usurpation of their freedom. These concerns which are typical of the scientific ontology and deontology (being part of what is nowadays called integrity) have surely not avoided in practice manipulations, in a positive way or not, operated during historical times by actors with a political, cultural or economical power to make decisions which were more or less well informed “scientifically” (see at all historical periods writings by “counsellors of the prince” such as Bodin, Machiavel, Botero, etc. to give a few among the ones having dealt with the planning of territories) and to proceed to “experiments” of governance structures or of technological or cultural innovations which results could be evaluated in some case as beneficial and in others as catastrophic. The evaluation of the efficacy of decisions complicates because of the justifications brought by the actors themselves with their “self-fulfilling prophecies” (). The often recalled difficulty of the evaluation of public policies is also increased by the uncertainty in the limits between the action and its context, both in space and in time.

Driving change in social systems, whatever the scale of interventions, remains a costly and risky operation, therefore difficultly acceptable by science for deontological concerns. Very few scientists therefore engage in “research-action” projects. A controversy has thus opposed in the sixties in France the advocates of an “applied geography” with a good knowledge of the "field" but sometimes with conservative trends, to the defenders of an "active geography" which would be more implied in the transformation of society. Sometimes, for example to contribute to the definition of policies for balancing metropolitan areas in France (operation by the *Délégation l'Aménagement du Territoire et l'Action Régionale* in 1964), geographers participating in the studies such as Michel Rochefort more particularly, did rely on scientific works, without having the courage to make it explicitly open (in this specific case central place theory by Walter Christaller). Contemporary geographers are less reluctant to exhibit a concern to help decision making in the most informed way possible given the state of their knowledge. They often then make the choice to use simulation models operated in silico by computers. Computer simulation thus became a substitute to experimentation. It is not a coincidence if among researchers in social sciences, geographers have very early found an interest in it: the diversity of multiple data sources (landscapes, populations, built environment, etc.) which they use to account of modifications of terrestrial interfaces by societies, the often large spatial extent of territories they study at the regional, national, or global scales, explain their need to make use of computing to organize this large quantities of information and to understand the dynamics they represent.

2 Geographical data and computational capabilities

The first simulation models in geography were firstly computed "by hand" in the fifties. It is not a coincidence if these models all deal with stylized facts which translate the regularities most frequently observed in the organisation of social space, and which are consequences of the "first law of geography" summarized as such already in 1970 by the American geographer Waldo Tobler: "everything interacts with everything, but two closer things have more chances to make contact than two more distant things". The power of attraction by proximity occurs in all social processes transforming the social space, which are constrained by an "obligation of space". This term was forged by Henri Reymond already in 1971 in a formalisation of issues in geography, who stated as first principle that societies have the tendency to transform the surface of the Earth which is heterogenous, rough and discontinuous, in an organized space exhibiting higher homogeneity and continuity, and making regularities emerge, due to the fact that two objects can not occupy the same place. Stating that individuals and societies have the highest probability to choose occupying the closest locations, both because these are better known and also because they yield economies on costs (physical, financial, and cultural) to travel the distance, may certainly be the strongest theoretical proposal of geography. It can be identified in any spatial configuration implying to distinguish a center and a periphery, which are observed at any level of the geographical space, from the local to the global.

The first simulation models in geography have thus dealt with processes for which the choice of the closest, among the place with which an interaction is expected, is a highly salient anthropological constant, either to observe the effects of an innovation before imitating it, according to the spatial theory of the diffusion of innovations by Torsten , or for the choice of destination places for a migration (). Models already in 1954 rely on the proposal by the American geographer Edward Ullman to construct a geography as the science of spatial interactions. This concerns more particularly trade relations, which have first lead to the empirical test of statistical models, as the so-called "gravity model", before being integrated into urban models which were first static () and then dynamical ().

A later generation of models playing in a more complex way with effects of proximity has intensively used cellular automatons. Measures of spatial auto-correlation, which translate in a positive or negative way attraction or concurrency effects linked to proximity are in that context used to test the plausibility of simulated configurations for land-use changes, and in particular urban growth (), or moreover the spread of epidemics in the geographical space (Cliff et al., 2004).

But the development of these models has been very early impeded by the computational capabilities at this time, since the explicit representation of spatial interactions increases as the square of the number of geographical units considered. Therefore, the statistician Christophe Terrier had to segment his Mirabelle program (*Mthode Informatise de Recherche et d'Analyse des Bassins par l'Etude des Liaisons Logement-Emploi*) processing household survey data provided by INSEE in 1975 before being able to simulate the clustering into employment centers of resident populations as a function of work-residence commuting between all 36,000 French communes (). Our first simulation model of interactions between cities aimed at reproducing their demographic and economic trajectories influenced by urban functions on a period of 2000 years could only accept a maximum of 400 cities on a personal computer (Bura et al., 1996) (). The increase of computational possibilities has been relatively slow, allowing to consider around one thousand cities in 2007 with the Eurosim model () or the Simpop2 models applied by Anne Bretagnolle on Europe and United States (Bretagnolle and Pumain, 2010). Furthermore, the experimentation method with these models stayed at an experimental stage for long, requiring am increased attention in the modification "by hand" of parameter values, which are only very rarely directly observable, and which thus must be estimated through the plausibility of model dynamics. However, equations for urban dynamics models integrate non-linear relations which produce numerous bifurcations, forcing to laborious trial-and-error loops in the estimation procedure (). This consequent work limits the number of simulations from which the estimation obtained can be judged as satisfying, and more importantly once the model is therein calibrated, there remains a relatively high uncertainty regarding the quality of results obtained.

3 A new generation of simulations

The end of the nineties was to modify completely the working environment of researchers, the diffusion of internet and then mobile phones and finally of massive data produced by diverse numerical sensors having in return rapid and intense effects on the increase of computational power which had allowed these disruptive technological innovations. Simulation models can then integrate considerable quantities of interactions between localized entities chracterized by a large diversity of attributes. Still fifteen years ago, was forced to conclude that network analysis in the case of the Parisian transportation network were "limited by computation". To give a single example of the quantitative leap forward in the increase of computational capabilities and their consequences on the higher confidence given to the models in consequence, we can mention the pioneering work in numerical epidemiology realized by Eubank et al. (2004) to simulate through the EpiSims and TRANSIMS models the daily trajectories on a transportation network of commuting of a million and a half individuals between around 180,000 places in the virtual city of Portland, in order to predict transmission pathways of an epidemics starting from interpersonal meeting probabilities in social networks organized as "small worlds". The epidemics can rapidly propagate to the whole city despite the number of contacts by individual remain low (fifteen in maximum (Eliot and Daudé, 2006)).

Simulation platforms are elaborated such that the largest number of researchers even not specialized in computer science can elaborate agent-based models. NetLogo () is amongst the most famous. It is generic and allows to access multi-agent simulations without a deep knowledge in algorithmics, thanks to its simple programming language and the integrated builder of graphic user interface. Other platforms which are more specialized such as GAMA (Grignard et al., 2013) are immediately elaborated to propose a coupling with geographic information systems. However, the confidence in results obtained from simulation models goes along with an increase in the size and the number of experiments required, i.e. of the amplitude of numerical experiments. Despite the fact that these platforms integrate basic tools for a first step towards such a change in scale, a need for a dedicated "meta-platform" has naturally emerged.

3.1 A virtual laboratory: the OpenMOLE platform

Since 2008, the OpenMOLE software has been conceived to explore the dynamics of multi-agent models (). It inherits from the development of a previous software SimExplorer () which already provided to users an ergonomic interface for the conception of experience plans and gave access to distributed computing. OpenMOLE (`https://openmole.org/`) is a collaborative modeling tool in constant evolution: "*a permanent effort for genericity has allowed to realize in a few years a pragmatic, generic, and proofed platform for the exploration of models of complex systems under the form of a dedicated language, both graphical and textual, exposing consistent blocks at the appropriated level of abstraction for the design of numerical experiments distributed on simulation models*" ().

Procedures (or workflows) proposed in OpenMOLE are described in a manner independent from the models and are thus reproducible, reusable and exchangeable between modelers. A market place is integrated to the software, similarly to the model library included in NetLogo, and allows users to collect exploration scripts that can act as template or example, in highly diverse thematic fields and for all methods and languages implemented in OpenMOLE (for example for the thematic fields calibration of geographical models, analysis of biological networks, image processing for neurosciences).

It is useful to mention the use by OpenMOLE of a Domain Specific Language (DSL) () to write exploration workflows. This practice consists in the construction of a notation and rules specific to the domain of a given problem. It is in a way a programmation language dedicated in that case to model exploration and associated methods. This language is naturally not created from scratch, but comes as an extension of the underlying language, i.e. the Scala language in the case of OpenMOLE. A reduced number of keywords and primitives fosters an easier use even for a user with no knowledge in programming, and furthermore the DSL remains highly flexible for the advanced user who can use Scala programming. According to , the DSL of OpenMOLE is one of the key elements of its genericity and accessibility.

We can also remark that one of the main assets of OpenMOLE is the transparent access to High Performance Computing environments (HPC). The increase in computational capabilities already described can in practice be implemented physically under different forms for the modeler: local server, local computation cluster, computation grid (network of multiple clusters, such as the European computing grid EGI), cloud computing services. Their use requires in most cases advanced computer science knowledge which are generally inaccessible to the standard modeler in geography. OpenMOLE integrates a library allowing to access most of these computing facilities, and their integration in the DSL is totally transparent for the user. The user script can then be tested on the local machine and then scaled on the HPC environments by modifying a single keyword in it.

The presentation of how to use the DSL and to elaborate scripts is out of the scope of this chapter, and we refer the reader to the online documentation of OpenMOLE for examples of scripts and model explorations. We simply recall the fundamental components of an exploration script: (i) the definition of prototypes, which correspond to parameters and outputs of the model, and which will take different values during the experiment; (ii) the definition of tasks, including model execution but that can also be for exemple pre- or post-processing tasks - the tasks covering a high variety of languages (scala, java, NetLogo, R, Scilab, native code such as python or C++); (iii) the description of methods to be applied (exploration by sampling, calibration, diversity search, etc.) which will act on the values of prototypes and will launch the considered evaluation task (mostly the model); (iv) a specification of the data gathered as an output of script execution (simulation data being often massive, a selection through this stage is crucial); and (v) the definition of the computation environment on which the method will be launched.

The platform aims at considerably extending practices of generative social science proposed by , which considers each multi-agent model as an artificial society, yielding macroscopic behaviors from assumptions made on microscopic behaviors. Numerical experiments that can be considered follow a change in scale, and the questions asked to the model evolve in a qualitative way. According to Clara , who used the OpenMOLE platform to develop with Sébastien the SimpopLocal model aimed at simulating the emergence of a system of cities, the virtual laboratory represented by this platform "*is not anymore only the simulation model and the hypothesis it simulated (i.e. the artificial society). It also contains the methods, tools and modeling procedures adapted to the conception and the exploration of the model and which practice creates as much knowledge and theoretical feedbacks than the conception of the model itself. This virtual laboratory is thus furthermore resembling a real research laboratory with an experimental desk (the model to conceive and explore), the assumptions of a researcher (the geographical processes translated into model mechanisms), methods (the iterative modeling method and aided by intensive computation), tools (the procedure for automatic exploration and any other experience plan integrated in OpenMOLE), all this gathered in a single room, the modeling platform SimProcess ()*".

In comparison to general protocols as the one introduced by to describe all the stage of the modeling process, principles applied in OpenMOLE mostly innovate regarding the potentialities without precedent to explore the dynamical behavior of simulation models. Two main innovations rely in the systematic application of optimization meta-heuristics, mainly genetic algorithms, to rapidly test the largest possible number of combinations for model parameter values, and in the simultaneous distribution of simulations on multiple machines of a computation grid, what allows to considerably reduce the length of experiments without which it would become quickly prohibitive.

The choice of genetic algorithms as an optimization heuristic is justified by their efficiency in the context of multi-objective optimization problems. Moreover, the island distribution scheme (populations evolving independantly during a given duration) is particularly suited to a distribution on grid, each node making a subpopulation evolve, which is regularly fetched, merged into the global population, and from which a new subpopulation is generated and sent on the node. This type of algorithms furthermore extends relatively well to stochastic models, even if this aspect still implies a certain number of open problems (). Following , these methods are situated within the larger context of Evolutionary Computation, and the scala library MGO developed simultaneously to the platform and which allows to implement evolutionary algorithms in it, has been conceived to be easily extended to other heuristics in Evolutionary Computation, opening totally the possibilities for the inclusion of new methods in OpenMOLE.

Reuillon et al. (2013) describe the fundamental principles of the platform, whereas give a contextualisation of the different uses in the frame of simulation models for systems of cities. According to R. Reuillon cited by , the philosophy of OpenMOLE is articulated around three axis: the model as a "black box" to be explored (i.e. methods which are independent from the model), the use of advanced exploration methods, and the transparent access to intensive computation environments. These different components are in a strong interdependence, and allow a paradigm shift in the use of simulation models: use of multi-modeling, i.e. variable structure of the model such as it was presented in chapter 4 (), change in the nature of questions asked to the model (for example full determination of the feasible space ()), all this allowed by the use of intensive computation (). The different methods available in that context will be illustrated below in concrete examples. The online documentation gives a broad overview of the available methods in the most recent version of the software and of their articulation within a standard context.

We consider a simulation model as an algorithm producing outputs from data and parameters as inputs. In that frame, we recall that in an ideal case, all the following stages should be necessary for a robust use of simulation models.

1. Identification of principal mechanisms and of crucial associated parameters, also with their variation range suggested by their thematic signification if they have some; identification of indicators to evaluate the performance or the behavior of the model.

2. Evaluation of stochastic variations: large number of repetitions for a reasonable number of parameters, assessment of the number of necessary repetitions to reach a certain level of statistical convergence.

3. Direct exploration for a first sensitivity analysis, if possible statistical evaluation of relations between parameters and output indicators.

4. Calibration, targeted algorithmic exploration through the use of specific algorithms (Calibration Profile (Reuillon et al., 2015), Pattern Space Exploration ()).

5. Feedbacks on the model, extension and new multi-modeling bricks, feedbacks on stylized facts and theory.

6. Extended sensitivity analysis, corresponding to experimental methods currently in construction and integration into the platform, such as for example the sensitivity to meta-parameters and to initial spatial conditions proposed by .

In some cases, some stages to not necessary take place, for example the evaluation of stochasticity in the case of a deterministic model. Similarly, each step take more or less importance depending on the nature of the question asked: calibration will not be relevant in the case of fully synthetic models, whereas a systematic exploration of a large number of parameters will not be necessary in the case of a model aimed at being calibrated on data.

In order to better illustrate this general presentation of the platform and associated methods, we propose in the following of this section to precisely develop the example of the SimpopLocal model, which genesis has been tightly linked to the one of the platform, and which has been candidate for the development and the application of diverse methods.

3.2 The SimpopLocal experiment: simulation of an emergence in geography

The SimpopLocal model has been conceived to represent emergence of systems of cities such as it has been observed in five or six regions of the world around 3000 years after the emergence of agriculture practices in sedentary societies (Marcus and Sabloff, 2008). The purpose is to explain the emergence not only of "the" city but indeed of "systems of cities" since we know that cities already at this time were never isolated but already organized as networks in the territory of each of these antic "civilisations". The most recent publications by archeologists insist in a certain continuity of processes which led to the settlement of hunter-gatherer populations, gathered in hamlets and villages and then to the apparition of cities in some of these regions. The development of agriculture has been concomitant to a considerable increase of population densities and of the size of human groups in these countries (the density switch from 0.1 person per square kilometre to to, i.e. a factor of 100 between the two orders of magnitude), ans also a complexification of the political organization and of the social division of labor. This very slow process of accumulation of ressources and of concentration of populations is done through chains including numerous feedbacks with many fluctuations in growth, due to the frequent adversary events that are natural catastrophes and the predations of neighbor groups. Because of the slow rate of transformations and their frequent interruptions, archeologists now sometimes contest the name of "neolithic revolution" which was proposed in 1942 by Gordon Childe (Demoule et al., 2018, p. 159).

However, geographers still identify the apparition of cities as an emergence, or a "bifurcation" for two main reasons: first it did not systematically occur in all regions where agriculture was practiced, therefore two evolution regimes for settlements systems are possible and historically viable (agricultural only regions and with villages may have functioned for centuries and still exist nowadays in a residual way in some forests and on pacific islands for example), therefore the territorial regime including cities indeed constitutes a specific "attractor" in the dynamic of ancient settlements systems; secondly the evolutive trajectory which led to the birth of cities translates an important qualitative change (an emergence) with a significant increase of the diversity of social functions associated to habitats and also a considerable broadening in the scale of the life of relations: commercial exchanges which are done there at a more or less high distance allow thus the cities to be less dependent of a "site" of local resources as are agricultural villages, and to develop the assets of a geographical "situation" exploiting the wealth of a network of sites more and more distant (Reymond, 1971).

The SimpopLocal model aims at reproducing this remarkable aspect of the dynamic of settlement systems, which invariably produces an amplification of the hierarchical differentiation between habitats defined in the literature as a major stylized fact: already in any system at any place and at any time of history or prehistory, the size distribution of inhabited places (measured by population or spatial extent, or even the diversity of functional artefacts) is statistically highly dissymmetrical, including numerous very small agglomerations and only few very large agglomerations following a relatively regular distribution of the type of a Zipf law or log-normal law (Fletcher, 1986; Liu, 1996). This hierarchical schema is a structural property (magnitude of the size of entities) at the macroscopic level which is relatively persistent in time whatever the local fluctuations intervening at the level of entities. The SimpopLocal model is conceived to test the hypothesis introduced in the evolutive theory of urban systems (), which explains this structural characteristic by an urban growth process in average proportional to the size attained, and its amplification through the creation of multiple technological and social innovations inducing the growth and the diversification of wealth which diffuse among the places put in relation by any sort of exchanges.

Le modle SimpopLocal sinspire dabord du modle statistique qui constitue une excellente premire approximation de lvolution des populations dans un systme de villes, en simulant la croissance urbaine comme un simple processus stochastique faisant varier la taille de chaque ville de faon proportionnelle sa taille et conduisant une distribution lognormale des populations urbaines (Gibrat 1931). La grande qualit de ce modle statistique lmentaire tient ce quil utilise comme explication de la croissance la taille dj acquise, laquelle exprime la fois la richesse accumule et la capacit dattraction et de rsilience du lieu habit (en quelque sorte il sagit dun modle selon le concept de croissance endogne des conomistes). Mais SimpopLocal est conu, comme les modles prcdents de la famille des modles multi-agents Simpop (Bura et al 1996, Sanders et al. 2007), pour pallier linsuffisante capacit du modle de Gibrat prvoir la tendance partout observe la croissance plus forte quattendu des plus grandes villes situes en tte des rseaux (Moriconi-Ebrard, 1993) et lexagration de lingalit entre les tailles des villes (Pumain 1997, Bretagnolle, Pumain 2010). Ces dviations au modle de Gibrat sont lies aux corrlations de longue porte (Rozenfeld et al. 2008), suscites par les interactions spatiales. Leffet de celles-ci amplifie la diffrenciation hirarchique entre les tailles des villes participant aux changes dans un systme urbain (Favaro et al. 2011). Les modles Simpop traduisent cet effet en introduisant, de manire exogne au modle et diffrents moments du temps de la simulation, de nouvelles fonctions urbaines qui slectionnent certaines villes ou sont captes par elles dans un processus continu dadaptation ces innovations. En comparaison des autres modles Simpop, SimpopLocal introduit deux nouveauts: il utilise une reprsentation abstraite des vagues dinnovation successives et les rassemble toutes dans un seul objet innovation. Une seconde originalit consiste rendre le processus de cration dinnovation endogne en le liant la taille du lieu habit, cense amplifier de manire non linaire lmergence de nouvelles formes techniques, sociales ou culturelles (avec une probabilit de cration variant comme le carr des populations en prsence ou en relations). Cette version plus parcimonieuse de la construction du modle permet de rduire considrablement le nombre de paramtres et autorise donc une exploration et une valuation plus systmatiques.

3.3 Implmentation de SimpopLocal, de Netlogo OpenMole Simpoplocal a t initialement dvelopp avec le langage Netlogo, puis re-dvelopp avec le langage de programmation Scala. La simulation avec Netlogo a bnfici des facilits de linterface qui permet de suivre numriquement et graphiquement les modifications engendres sur les variables macroscopiques qui rsument ltat du systme, mais a montr trs vite ses limites en termes dexprimentation. La mthode manuelle de rglage des valeurs des paramtres permettait difficilement dviter les emballements de la croissance urbaine conduisant des accroissements de taille des villes bien trop normes pour lpoque historique quil tait question de simuler. La reprogrammation en Scala puis le passage sur la plateforme OpenMole devaient permettre une exploration plus prcise et complte des comportements du modle. Le modle reprsente l'volution des units de peuplement disperses dans une zone suffisamment grande pour accueillir quelques milliers d'habitants mais suffisamment limite en surface pour assurer la connexion possible entre les lieux habits en fonction des moyens de transport disponibles l'poque (par exemple il pourrait s'agir de lancienne Msopotamie ou de la mso-Amrique antique). L'espace de simulation est compos dune centaine de lieux habits. Chaque lieu est considr comme un agent fixe et est dcrit par trois attributs: l'emplacement de son habitat permanent (x, y), la taille de sa population P et les ressources disponibles dans son environnement local. La quantit de ressources disponibles R est quantifie en units dhabitants et peut tre comprise comme la capacit de charge de lenvironnement local pour soutenir une population, laquelle varie en fonction des comptences en exploitation des ressources que la population locale a acquises, grce aux innovations quelle a cres ou reues des autres lieux habits. Toutefois, lexploitation des ressources est effectue localement et le partage ou l'change de biens ou de personnes ne sont pas explicitement reprsents dans le modle. Chaque nouvelle innovation cre ou acquise par un lieu habit dveloppe ses comptences en exploitation. Lentit innovation sentend ici comme une grande invention abstraite socialement accepte, qui pourrait reprsenter une invention technique, une dcouverte, une organisation sociale, de nouvelles habitudes ou pratiques ... Chaque acquisition dinnovation par un lieu habit apporte la possibilit de surpasser ses seuils de capacit, et par consquent autorise une croissance dmographique. Le modle a t conu pour tre le plus parcimonieux possible, en minimisant le nombre des attributs des agents (qui sont des lieux habits) et les paramtres qui contrlent leur volution. On a utilis directement les ordres de grandeur moyens indiqus par les travaux des archologues pour fixer environ 4000 ans la dure de la priode de transition entre un systme de peuplement agraire et un systme de peuplement urbain, pour estimer un taux de variation moyen annuel de la population denviron 0,02On dfinit une valeur initiale pour la population et les ressources des lieux habits, puis le rseau dinteraction entre eux est cr. Ensuite, chaque tape de la simulation, les mcanismes de croissance de la population et de diffusion de l'innovation sont appliqus. Limpact des innovations sur lefficacit de lextraction des ressources est calcul. Cette boucle est itre jusqu' ce que le critre d'arrt soit atteint: dans ce cas, aprs 4000 tapes ou lorsquun nombre maximal arbitraire d'innovations a t atteint. On observe lvolution de ltat du systme de peuplement dfini au niveau macro-gographique par la distribution de la taille des lieux habits, rsume par la pente de la distribution rang-taille. Le modle utilisant certains paramtres qui sont des probabilits est stochastique, un mme jeu de valeurs de paramtres peut donner lieu des rsultats sensiblement diffrents. Une mthode automatise pour faire varier les valeurs des paramtres et interprter les rsultats obtenus a t mise au point progressivement par une collaboration entre informaticiens et gographes.

3.4 Calibrage et validation Lautomatisation de lexploration des dynamiques engendres par les modles de simulation avec la plateforme OpenMole utilise des algorithmes gntiques qui ralisent de faon systmatique les variations des valeurs des paramtres auparavant effectues la main par le chercheur. La distribution des calculs sur une infrastructure de grille (un rseau dordinateurs) permet en outre de conduire ce trs grand nombre doprations combinatoires en rduisant considrablement le temps de calcul, grce au traitement en parallle de linformation. Mais la mise en oeuvre de cette nouvelle forme de lexprimentation des modles suppose aussi une intervention du chercheur thmaticien, qui doit slectionner les objectifs prcis que son modle doit satisfaire, tandis quun raffinement supplmentaire de la mthode dexploration peut conduire un renforcement de la confiance quil accorde aux hypothses scientifiques de son modle. Le calibrage comme optimisation au moyen des algorithmes gntiques Le calibrage est une procdure qui cherche minimiser lcart (appel fitness) entre le comportement simul par le modle et le comportement observ empiriquement, en faisant varier de faon incrmentale les valeurs inconnues des paramtres du modle. Stonedahl (2011) a rappel les difficults de cette exploration qui devient vite fastidieuse lorsquelle est conduite manuellement, cause des multiples bifurcations intervenant dans des modles o la plupart des mcanismes liant les variables sont non linaires. Une exploration exhaustive de lespace des paramtres nest pas envisageable car elle exigerait des temps de calcul trop importants, en croissance exponentielle avec le nombre de ces paramtres. Comme ces procdures produisent aussi de grandes quantits de rsultats, elles exigent en outre demployer des mthodes adaptes pour traiter et visualiser les informations engendres par les simulations. Tout un ensemble de logiciels doit donc tre mis au point pour permettre au chercheur de dcouvrir les principaux schmas des dynamiques associes aux variations des paramtres de son modle. Cest l o des procdures informatiques adaptes peuvent tre utilises, en rapportant la question du calibrage un problme doptimisation. Les algorithmes gntiques ont t utiliss pour calibrer des systmes multi-agents dans plusieurs domaines, en mdecine (Castiglione et al, 2007), en cologie (Duboz et al, 2010), en conomie (Espinosa, 2012; Stonedahl et Wilensky, 2010a), ou en hydrologie (Solomatine et al, 1999). En dpit de la large utilisation des systmes multi-agents en sciences sociales, cette mthode na pas t applique trs souvent (Heppenstall et al, 2007; Stonedahl et Wilensky, 2010b). Ce type dexprience numrique exige en effet que soient dfinis des objectifs quantitatifs permettant dvaluer si les rsultats de la simulation sont compatibles avec les attentes des experts, il faut galement savoir grer lnorme charge de calcul et parvenir optimiser une fonction de fitness susceptibles de trs importantes variations stochastique (Pietro et al, 2004). Dans le cas de SimpopLocal, qui comprend 5 paramtres dont les valeurs sont inconnues (mme leurs ordres de grandeur ne peuvent pas tre estims partir de donnes empiriques), nous avons d identifier trois fonctions objectif. Celles-ci caractrisent un rsultat de simulation au niveau macro-gographique et correspondent des faits styliss dont les ordres de grandeur ont pu tre tablis partir des connaissances archologiques et historiques: la distribution finale des tailles de villes doit tre lognormale (peu diffrente dune loi de Zipf), la taille maximale quatteint la plus grande ville doit tre denviron 10 000 habitants, pour une dure de simulation quivalant 4000 ans. Cette obligation de dfinir des fonctions-objectif pourrait tre considre comme une contrainte forte sur la validit pistmologique du modle, elle semble en effet contredire lhypothse dune volution ouverte pour les systmes de villes. En fait, cette tape intermdiaire de calcul reprsente un comprim de connaissances, notre exigence a minima sur la reprsentativit et la plausibilit du comportement du modle par rapport lensemble envisageable des dynamiques des villes en systme (lpoque historique de lmergence des villes). Le rsultat en termes dvaluation des simulations doit permettre davancer dans la connaissance des processus dinteraction intra-urbains susceptibles dengendrer dette dynamique gnrale lchelon macroscopique du systme, cette reconstitution thorique sapparentant alors ce que des physiciens nomment le problme inverse. Un domaine de variation numrique assez large est tabli a priori pour chacun des cinq paramtres. Chaque jeu de paramtres, combinant une valeur pour chacun dentre eux, est valu en fonction de la sortie de simulation qu'il produit. Cette valuation mesure la proximit entre les sorties de la simulation et les fonctions objectifs dfinies pour le modle et permet ainsi de mesurer la capacit dun certain ensemble de valeurs de paramtres reproduire les faits styliss que la simulation doit approcher au mieux. Les paramtrages recevant les meilleures valuations sont ensuite utiliss comme base pour engendrer de nouveaux jeux de paramtres qui sont ensuite tests.

Exploration de lespace des paramtres sous contrainte dobjectifs Le modle SimpopLocal tant stochastique, les rsultats de la simulation varient d'une simulation l'autre pour le mme paramtrage. Par consquent, l'valuation du paramtrage en fonction des trois objectifs doit prendre en compte cette variabilit. Nous avons vrifi quune centaine de simulations pour chaque jeu de paramtres suffisait saisir cette variabilit sans trop augmenter la dure du calcul. A chaque fonction-objectif correspond une mesure de lvaluation de la qualit du rsultat simul. La capacit du modle produire une distribution log-normale est mesure par lcart entre la distribution simule et une distribution log-normale thorique ayant mme moyenne et cart type selon un test de Kolmogorov-Smirnov. Lobjectif de population maximale quantifie la capacit du modle engendrer des villes plus ou moins grandes, le rsultat d'une simulation est test en calculant l'cart entre la taille de la plus grande agglomration et la valeur attendue de 10 000 habitants: [(population de la plus grande agglomration 10 000) / 10 000 —. Lobjectif de la dure de la simulation quantifie la capacit du modle gnrer une configuration attendue dans un laps de temps historiquement plausible. On calcule lcart entre le nombre ditrations de la simulation et la valeur attendue de 4000 tapes de la simulation: — (simulation dure 4000) / 4000 —. Ces trois calculs derreur sont normaliss afin de pouvoir comparer le degr de russite dune simulation vis--vis de chacun des trois objectifs. Mais lagrgation des trois calculs qui produirait une seule mesure de qualit globale ntant pas possible, un algorithme multi-objectif est ncessaire pour dterminer quelles simulations sont les plus satisfaisantes pour approcher la configuration finale souhaite. Ce type dalgorithme calcule des solutions de compromis telles quaucune ne domine toutes les autres pour tous les objectifs. Ces solutions sont appeles des compromis de Pareto et elles forment ensemble ce qui est appel un front de Pareto. Lutilisation de mthodes dexploration globales comme celle des algorithmes gntiques pour calibrer un modle multi-agent (et en particulier un modle multi-agent stochastique) implique un cot de calcul trs lev (Sharma et al, 2006). Ce type de charge est trop volumineux pour tre excut sur des ordinateurs locaux, et les supercalculateurs sont trs coteux et ne sont pas facilement disponibles dans la plupart des laboratoires. Les grilles informatiques offrent une solution pour rsoudre ces problmes de calculs intensifs. Cependant, linformatique une si grande chelle suppose dorchestrer l'excution de dizaines de milliers d'instances du modle sur des ordinateurs distribus dans le monde entier. La probabilit cumule de pannes locales et le problme de rpartir la charge de travail de faon optimale sur la grille rendent trs difficile son utilisation pour un chercheur non spcialis, comme prcis ci-dessus. Cest entre autres pour surmonter ces difficults que la plate-forme OpenMOLE a t construite (Reuillon et al, 2010; 2013). Cet exemple de la calibration du modle SimpopLocal montre bien dans quelle mesure OpenMOLE aide les modlisateurs franchir le foss technique et mthodologique qui les spare de l'informatique haute performance. Linfrastructure de la grille de calcul (EGI) nous a permis dutiliser une puissance de calcul telle quun demi milliard dexcutions du modle ont pu tre effectues pour le calibrage de SimpopLocal, lequel sans cela aurait requis quelque 20 annes de calcul avec un seul ordinateur. Le profil de calibrage, un grand saut pistmologique pour les SHS Le rsultat du processus de calibrage assure seulement que le modle peut reproduire les caractristiques stylises de lmergence dun systme de villes, avec une valuation assez prcise des valeurs des paramtres qui toutes ensemble contribuent assurer cette volution. Mais il ne dit rien de la frquence laquelle les jeux de paramtres produisent des comportements plausibles, et de quelle faon chaque paramtre contribue modifier le comportement du modle. Il serait intressant par exemple de savoir quel moment certaines valeurs de paramtre empchent le systme datteindre un comportement plausible, et de ne pas se restreindre ne connatre qu'un seul jeu de valeurs de paramtres optimales. Une nouvelle mthode a t mise au point pour reprsenter la sensibilit du modle aux variations d'un seul paramtre, indpendamment des variations de tous les autres paramtres (Reuillon et al. 2015). Au moyen dune fonction qui calcule une seule valeur numrique dcrivant la qualit du calibrage pour le modle, lalgorithme de profil calcule lerreur de calibrage la plus faible possible lorsque la valeur d'un paramtre donn est fixe et que les autres sont libres. Lalgorithme calcule cette erreur minimale pour tout le domaine de variation du paramtre tudi. Pour chaque valeur dun paramtre, lalgorithme cherche identifier les jeux de valeurs des autres paramtres qui produisent le meilleur ajustement du modle aux donnes attendues (la plus petite erreur possible). Un graphique reprsente alors les variations de cette valeur dajustement optimale en fonction des variations du paramtre tudi. Le profil de calibrage montre ainsi plusieurs formes possibles pour cette courbe. Lorsquelle prsente une nette inflexion vers les valeurs les plus basses pour lerreur de calibrage, cela pour un tout petit domaine de variation des valeurs du paramtre tudi, on peut en conclure quon a vraiment identifi lordre de grandeur du paramtre qui satisfait aux exigences en termes de comportement du modle. Si lune de ces courbes reste plate, cela indique que le paramtre na pas deffet sur le comportement du modle et peut donc en tre limin. Ainsi, dans le cas de SimpopLocal, un paramtre imagin comme le dure de vie dune innovation a t finalement exclu car des variations restaient sans effet sur la qualit dajustement du modle, toutes choses gales quant aux variations des autres paramtres (Schmitt, 2014). On a donc ici la possibilit dvaluer jusqu quel point les mcanismes imagins pour construire le modle sont non seulement suffisants, mais aussi ncessaires pour produire le comportement attendu. Certes dans les limites du cadre thorique et de la slection des faits styliss retenus, cest la premire fois que des chercheurs en SHS peuvent parvenir ce type de conclusion scientifique essentielle, grce une mthode de validation enfin efficace pour les modles de simulation multi-agents. Cest un immense progrs du point de vue pistmologique en sciences sociales certes toujours dans le cadre thorique donn par les objets, attributs et mcanismes slectionns par les chercheurs pour tre reprsentatifs du systme observ. Une forme complmentaire de validation du modle pourrait tre alors imagine si des historiens archologues tentaient de le recalibrer avec des donnes de leurs observations. En effet, le jeu de paramtres estim contient des valeurs qui engendrent bien la dynamique voulue pour un systme de peuplement mais qui ne sont pas fixes dans labsolu, elles sont relatives les unes aux autres dune part et aux donnes fictives introduites dautre part. Si lon modifie ces dernires pour les rendre compatibles avec un systme de peuplement historiquement observ, la capacit du modle simuler son dveloppement serait alors confirme, non seulement en reconstruisant les trajectoires de lvolution de la population des lieux habits considrs, mais aussi en conservant les ordres de grandeur relatifs des paramtres qui engendrent cette dynamique.

4 Examples for applications of OpenMOLE: network-territories interaction models

We propose in this section to illustrate the application of the exploration methods included in OpenMOLE and intensive computation to an other thematic question, the issue of interactions between networks and territories. This question has fed numerous scientific debates fir which most of problems remain relatively open. For example, the issue of "structuring effects of transportation infrastructures" (Bonnafous and Plassard, 1974), described by Offner (1993) as a "scientific myth" invoked to justify the cost of an infrastructure through its spillovers on regional development, and which are not always observed on middle terms, can according to A. Bretagnolle in (Offner et al., 2014) be observed for broader territories and on long times, while taking into account local fluctuations in dynamics of systems of cities. The empirical difficulty to extract general stylized facts together with the conceptual difficulty of geographical entities in relations of circular causality, are avoided through the approach of modeling co-evolution between transportation networks and territories proposed by Raimbault (2018c). The results obtained are closely linked to the use of OpenMOLE and its exploration and calibration algorithms, of which we will give a few illustrations.

The application of multi-objective calibration appears to be essential for the application of models for systems of cities to real situations. For example, Raimbault (2018a) introduce a model for the evolution of a system of cities on long times which is close to the model by (Favaro and Pumain, 2011) but focuses on the effect of the physical transportation network. Growth rates of cities are determined by the superposition of several effects: (i) endogenous growth captured with a fixed growth rate corresponding to the Gibrat model; (ii) interactions between cities through a gravity model; (iii) feedback of flows circulating in the network on traversed cities. This model is calibrated in a non-stationary way in time (i.e. on temporal moving windows in order to take into account the change in nature of urban dynamics such as observed by with for example the mutations of transportation networks) on the French system of cities between 1830 and 2000. To calibrate the model, simulated populations are compared to observed populations. At this stage the use of a multi-objective calibration algorithm (the NSGA2 algorithm implemented in OpenMOLE) is crucial. Indeed, the fit can be for example computed as a mean square error in time and for all cities. However, given the disparities in city sizes due to urban hierarchy, it rapidly occurs that a mono-objective calibration on this error will focus on adjusting the size of largest cities, at the expense of most cities in the system. The addition of a second objective taken for example as a mean square error on logarithms of populations, allowing to take these into account. An important result of (Raimbault, 2018a) is then the emergence of Pareto fronts for these two objectives for all time windows considered. This shows that this type of model must be applied by making a compromise between the adjustment of population for medium-sized cities and populations for largest cities. This result is obtained thanks to the multi-objective optimization with a genetic algorithm in OpenMOLE.

An other application example for methods included in the platform which illustrate its crucial role is given by the search for co-evolution regimes. Following Raimbault (2017b), the study of lagged correlation patterns in time allows to identify typical interaction regimes between variables describing the network and variables describing the territory. More precisely, Raimbault (2018c) defines co-evolution as the existence of circular causal relationships at the level of a population of entities in a given spatial extent. In the case of networks and territories, network properties must be locally caused by the properties of territories and reciprocally. Mono-directional causalities of networks to territories correspond then to "structuring effects" mentioned above. This definition allows to capture the "congruence" Offner (1993) between these objects, in some sense their reciprocal adaptation in a dynamical way. It also yields the construction of an operational method proposed by Raimbault (2017b) which statistically investigates causality links between corresponding variables. In practice, the weak Granger causality notion is used, providing a flexibility regarding the data required and the temporal and spatial frame of estimation. This causality is in our case quantified by lagged correlations between variations of network variables (such as centralities or accessibilities) and variations of territory variables (such as population, employments, real estate transactions, etc.), and the existence of significant extrema at non zero lags gives a sense of causality. A typology of these lagged correlation profiles provides what we call "causality regimes", among which co-evolution regimes in which two variables for territory and network are in reciprocal causality.

The question is then is a given case study to identify the regimes in presence from observed data or from data simulated by a model, and particularly the regimes corresponding to a co-evolution. The demonstration of the existence of such regimes as output of a "co-evolution model" is a priori not expected, since processes included at the microscopic scale for which the influences are indeed reciprocal do not imply a reciprocal causality at the macroscopic scale of indicators, as the models considered are complex and exhibit emergence. This method is applied to a macroscopic model of co-evolution by Raimbault (2018b), which extends the model of (Raimbault, 2018a) by adding rules for the evolution of network capacities. A direct sampling which consists in a random sampling of a fixed number of parameter points (for example through Latin Hypercube Sampling maximizing the discrepancy of points), is a first experiment allowed by OpenMOLE to have an overview of the capacity of the model to produce co-evolution. This experiment provides a certain number of regimes which can potentially be produced by the model, namely 33 regimes among 729 possible regimes for the variable considered, i.e. 4,5% (in this case we consider as territory variable the populations, and as network variables the closeness centrality and the accessibility, what corresponds to six directed couples of variables, and thus $3^6 = 729$ possible configurations since each couple can exhibit a positive, negative or inexistent lagged correlation). We find among these 19 co-evolution regimes, which existence could not be intuitively predicted. The existence and the variety of these regimes is an important result, showing that it is possible to model a co-evolution, in the precise statistical sense given above.

The application of the Pattern Space Exploration algorithm (Chérel et al., 2015) with objective the diversity of produced regimes allows then to considerably extend this conclusion, since it produces 260 regimes (35,7%). This is a typical example where the strong non-linearity of outputs considered can lead to partial or even biased conclusions and where the use of a specific method is crucial. The results are then made more robust and extended thanks to the application of a specific method integrated to the OpenMOLE platform.

This method furthermore allows to compare between them models with a certain confidence in the exhaustivity of solutions obtained. Raimbault (2018d) applies the same approach to the SimpopNet model introduced by Schmitt (2014) which is also a co-evolution model at the macroscopic scale exhibiting a large number of common points with the previous model in particular in the variables considered and thus in the output indicators that can be computed. A smaller number of regimes of interaction and of co-evolution is then obtained, confirming on the one hand that it is not straightforward for a model conceived for co-evolution to effectively produce co-evolution regimes, and on the other hand suggesting that stronger constraints in the evolution rules for the network induce a bigger difficulty to produce a diversity of regimes.

5 Perspectives

The elaboration of the OpenMOLE platform has created a research axis, even a research domain, with a specific positioning which one of the remarkable aspects is a high level of interdisciplinarity between social sciences and more technical disciplines such as computer science. According to Banos (2017) this leads to the production of a broader and deeper knowledge (in a way similar to the virtuous spiral between disciplinarity and interdisciplinarity described by Banos (2013)). But also through the philosophy of unique platform (described above, through the strong interaction between the three axis of model embedding, access to innovative model exploration methods, and transparent access to intensive computation environments), the perspectives opened are numerous, as much on the technical side than on the theoretical, methodological or thematic side. We give below a few examples, accounting of a current state of possible futures for OpenMOLE.

5.1 Methods

The extension of available methods is a privileged axis of research linked to the development of OpenMOLE. For example, the exhaustive resolution of inverse problems (Aster et al., 2018) is currently not included. Solving an inverse problem consists in determining all the antecedents of a given objective in the output space of the model.

Calibration algorithms solve similar problems but do not ensure the exhaustivity of the solutions produced, what can become a considerable issue in the case of equifinality (), i.e. of parameter configurations or initial conditions leading through different trajectories to an identical result. An heuristic for inverse problems inspired by the PSE mechanisms is currently being elaborated for an integration into OpenMOLE.

The use of Bayesian inference methods is also a direction developed. Indeed, in the case of strongly stochastic models, and in which the joint distributions have a non standard form, an estimation of the probability distribution of parameters can be provided by this type of method. In the case of simulation models, the method of Approximate Bayesian Computation (Csilléry et al., 2010) allows for a given observed dataset to get the probability distribution of parameters having the most likely produced it. This is therefore an extended calibration with a probabilistic knowledge produced allowing to take into account uncertainty. A specification of this method proposed by Lenormand et al. (2013) with the purpose to reduce the number of simulations in the case of models with a significant computation time, is also being adapted to parallel computation and integrated into the platform.

We can finally mention diverse methodological directions which are also being investigated: (i) the question of high dimensionality is rapidly an issue in the use of the PSE algorithm, since the number of output configurations is potentially victim of the dimensionality curse, i.e. that the time or the size of execution are an exponential function of the number of dimensions (a grid exploration is the simplest example to get a grasp on this phenomenon) - new methods combining dimensionality reduction and diversity search would allow to solve this problem and take into account a much higher richness of outputs; (ii) the question of the sensitivity to initial spatial conditions which was already mentioned (Raimbault et al., 2018) is particularly relevant for geographical models, and a scala library including synthetic generators for population configurations at different scales is currently being implemented, including for example the generators for districts studied by Raimbault and Perret (2019); (iii) the implementation of information criteria for the performance of models, already described in chapter 4, is also being studied, such as the POMIC criteria proposed by Piou et al. (2009).

5.2 Tools

All along its development, OpenMOLE has always been innovative in terms of tools used and developed. The choice of the Scala language to replace Java already in the first versions is an innovative technological choice which is particularly relevant through the functional programming but also object programming possibilities while still keeping the underlying Java infrastructure allowing a high portability without complications depending on the operating system or on the hardware, what is crucial for the distribution of computations of computations on heterogenous nodes of the computation grid. For example, properties such as trait mixing make scala particularly suited to multi-modeling (Odersky and Zenger, 2005). The possibilities offered by object programming are conserved in Scala and can be combined to the abstraction of functional programming, making it a language more powerful in this sense of flexibility than other functional languages such as Haskell (Oliveira and Gibbons, 2010). Furthermore, properties such as implicit conversions or case classes make Scala highly ergonomic for the design of DSL (Sloane, 2008), which as we already described is an essential feature of OpenMOLE.

The issue of program embedding, and by extension of model embedding, remain an active research field in particular in relation with reproducibility. The docker software which uses containers allows to wrap an execution environment in an identical manner whatever the operating system and the hardware. Hung et al. (2016) propose to couple docker with a graphical user interface for scientific reproducibility. Similar softwares such as Singularity are specifically dedicated to the reproducibility of HPC experiments (Kurtzer et al., 2017). The core of the embedding strategy taken by OpenMOLE does not rely on such a software, for example because of performance reasons, but some tasks relying on the execution of binaries or of programs with a complicated environment are embedded in OpenMOLE through a task using docker (for example for the R language task which requires the installation of a full R environment). An improvement of the integration of docker into OpenMOLE is an active research direction which is crucial for the future extension of the genericity of embeddable programs. OpenMOLE is therefore at the edge in technical research regarding scientific reproducibility. In a similar way, the question of scalability of experiments is at the core of the philosophy of the platform, and research are done for example to automatize the deployment of multiple OpenMOLE instances on a cluster and facilitate the use within communities of thematic researchers.

Conclusion

The exploration of simulation models has been progressively established in geography through the intermediary of initiatives such as the development of the OpenMOLE platform. It has been achieved in a highly interdisciplinary and reciprocal framework (win-win relations between computer scientists and geographers), but also through a novel integration of knowledge domains (Raimbault, 2017a), i.e. of empirical, theoretical and modeling knowledges, but also tools and methods which are within each of these domains in strong interaction. The OpenMOLE enterprise and its branch linked to geography in the context of the Geodivercity ERC project witnesses of a novel way to produce geographical knowledge, in a robust evidence-based manner, and suggesting the possibility of scientific proofs in social sciences. Such an emancipation remains to be propagated and the approach to be valorized to realize its potential of future direction of Quantitative and Theoretical Geography, in complementarity with new emerging disciplines of City Science and Urban Analytics described by Batty (2019), but the proof-of-concept is largely validated and provides significant evidence to social sciences to resist the colonizer hegemony of hard sciences such as physics pretending to a monopoly on evidence-based approaches to social systems (Dupuy and Benguigui, 2015).

References

Aster, R. C., Borchers, B., and Thurber, C. H. (2018). *Parameter estimation and inverse problems*. Elsevier.

Banos, A. (2013). *Pour des pratiques de modélisation et de simulation libérées en géographie et SHS*. PhD thesis, Université Paris 1 Panthéon Sorbonne.

Banos, A. (2017). Knowledge accelerator' in geography and social sciences: Further and faster, but also deeper and wider. In *Urban Dynamics and Simulation Models*, pages 119–123. Springer.

Batty, M. (2019). Urban analytics defined.

Bonnafous, A. and Plassard, F. (1974). Les méthodologies usuelles de l'étude des effets structurants de l'offre de transport. *Revue économique*, 25(25):208–232.

Bretagnolle, A. and Pumain, D. (2010). Simulating urban networks through multiscalar space-time dynamics: Europe and the united states, 17th-20th centuries. *Urban Studies*, 47(13):2819–2839.

Bura, S., Guérin-Pace, F., Mathian, H., Pumain, D., and Sanders, L. (1996). Multiagent systems and the dynamics of a settlement system. *Geographical analysis*, 28(2):161–178.

Chérel, G., Cottineau, C., and Reuillon, R. (2015). Beyond corroboration: Strengthening model validation by looking for unexpected patterns. *PLoS ONE*, 10(9):e0138212.

Cliff, A. D., Haggett, P., Smallman-Raynor, M., et al. (2004). *World atlas of epidemic diseases*. Arnold London.

Csilléry, K., Blum, M. G., Gaggiotti, O. E., and François, O. (2010). Approximate bayesian computation (abc) in practice. *Trends in ecology & evolution*, 25(7):410–418.

Demoule, J.-P., Garcia, D., and Schnapp, A. (2018). *Une histoire des civilisations*. Paris, La Dcouverte, INRAP.

Dupuy, G. and Benguigui, L. G. (2015). Sciences urbaines: interdisciplinarités passive, naïve, transitive, offensive. *Métropoles*, (16).

Eliot, E. and Daudé, É. (2006). Diffusion des épidémies et complexités géographiques. perspectives conceptuelles et méthodologiques. *Espace populations sociétés. Space populations societies*, (2006/2-3):403–416.

Eubank, S., Guclu, H., Kumar, V. A., Marathe, M. V., Srinivasan, A., Toroczkai, Z., and Wang, N. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180.

Favaro, J.-M. and Pumain, D. (2011). Gibrat revisited: An urban growth model incorporating spatial interaction and innovation cycles. : . *Geographical Analysis*, 43(3):261–286.

Fletcher, R. (1986). Settlement archaeology: World-wide comparisons. *World Archaeology*, 18(1):59–83.

Grignard, A., Taillandier, P., Gaudou, B., Vo, D. A., Huynh, N. Q., and Drogoul, A. (2013). Gama 1.6: Advancing the art of complex agent-based modeling and simulation. In *International Conference on Principles and Practice of Multi-Agent Systems*, pages 117–131. Springer.

Hung, L.-H., Kristiyanto, D., Lee, S. B., and Yeung, K. Y. (2016). Guidock: using docker containers with a common graphics user interface to address the reproducibility of research. *PloS one*, 11(4):e0152686.

Jensen, P. (2018). *Pourquoi la société ne se laisse pas mettre en équations*. Le Seuil.

Kurtzer, G. M., Sochat, V., and Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. *PloS one*, 12(5):e0177459.

Lenormand, M., Jabot, F., and Deffuant, G. (2013). Adaptive approximate bayesian computation for complex models. *Computational Statistics*, 28(6):2777–2796.

Liu, L. (1996). Settlement patterns, chiefdom variability, and the development of early states in north china. *Journal of anthropological archaeology*, 15(3):237–288.

Marcus, J. and Sabloff, J. A. (2008). *The ancient city: New perspectives on urbanism in the old and new world*. School for Advanced Research on the.

Odersky, M. and Zenger, M. (2005). Scalable component abstractions. In *ACM Sigplan Notices*, volume 40, pages 41–57. ACM.

Offner, J.-M. (1993). Les "effets structurants" du transport: mythe politique, mystification scientifique. *L'espace géographique*, pages 233–242.

Offner, J.-M., Beaucire, F., Delaplace, M., Frémont, A., Ninot, O., Bretagnolle, A., and Pumain, D. (2014). Les effets structurants des infrastructures de transport. *Espace Geographique*, 43(1):p–51.

Oliveira, B. C. and Gibbons, J. (2010). Scala for generic programmers: comparing haskell and scala support for generic programming. *Journal of functional programming*, 20(3-4):303–352.

Piou, C., Berger, U., and Grimm, V. (2009). Proposing an information criterion for individual-based models developed in a pattern-oriented modelling framework. *Ecological Modelling*, 220(17):1957–1967.

Raimbault, J. (2017a). An applied knowledge framework to study complex systems. In *Complex Systems Design & Management*, pages 31–45.

Raimbault, J. (2017b). Identification de causalites dans des donnees spatio-temporelles. In *Spatial Analysis and GEOmatics 2017*.

Raimbault, J. (2018a). Indirect evidence of network effects in a system of cities. *Environment and Planning B: Urban Analytics and City Science*, page 2399808318774335.

Raimbault, J. (2018b). Modeling the co-evolution of cities and networks. *arXiv preprint arXiv:1804.09430*.

Raimbault, J. (2018c). Modélisation des interactions entre réseaux de transport et territoires: une approche par la co-évolution. In *JJC Pacte-Citeres-Les capacités transformatives des réseaux dans la fabrique des territoires*.

Raimbault, J. (2018d). Unveiling co-evolutionary patterns in systems of cities: a systematic exploration of the simpopnet model. *arXiv preprint arXiv:1809.00861*.

Raimbault, J., Cottineau, C., Texier, M. L., Néchet, F. L., and Reuillon, R. (2018). Space matters: extending sensitivity analysis to initial spatial conditions in geosimulation models. *arXiv preprint arXiv:1812.06008*.

Raimbault, J. and Perret, J. (2019). Generating urban morphologies at large scales. *arXiv preprint arXiv:1903.06807*.

Reuillon, R., Leclaire, M., and Rey-Coyrehourcq, S. (2013). Openmole, a workflow engine specifically tailored for the distributed exploration of simulation models. *Future Generation Computer Systems*, 29(8):1981–1990.

Reuillon, R., Schmitt, C., De Aldama, R., and Mouret, J.-B. (2015). A new method to evaluate simulation models: The calibration profile (cp) algorithm. *Journal of Artificial Societies and Social Simulation*, 18(1):12.

Reymond, H. (1971). Pour une problmatique thorique. In Isnard, H., Racine, J. B., and Reymond, H., editors, *Problématiques de la géographie*, volume 29. Presses Universitaires de France-PUF.

Schmitt, C. (2014). *Modélisation de la dynamique des systèmes de peuplement: de SimpopLocal à SimpopNet*. PhD thesis, Université Panthéon-Sorbonne-Paris I.

Sloane, T. (2008). Experiences with domain-specific language embedding in scala. In *Domain-Specific Program Development*, page 7.

Stonedahl F, Wilensky U, 2010b, Evolutionary robustness checking in the artificial Anasazi model, in Proceedings of the AAAI Fall Symposium on Complex Adaptive Systems: Resilience, Robustness, and Evolvability (AAAI Press, Menio Park, CA), 120129. Taillandier P., Grignard A., Gaudou B. et Drogoul A. 2014, Des donnes gographiques la simulation base dagents: application de la plate-forme GAMA, Cybergeo : European Journal of Geography, 671, URL: http://journals.openedition.org/cybergeo/26263; DOI: 10.4000/cybergeo.26263 Tisue, S., Wilensky, U. 2004. Netlogo: A simple environment for modeling complexity. In International conference on complex systems (Vol. 21, pp. 16-21). Terrier C. 1980, MIRABELLE, Courrier des Statistiques, 73. Van Deursen, A., Klint, P. (2002). Domain-specific language design requires feature descriptions. Journal of Computing and Information Technology, 10(1), 1-17. White, R., Engelen, G. 1993. Cellular automata and fractal urban form: a cellular modelling approach to the evolution of urban land-use patterns. Environment and planning A, 25(8), 1175-1199. White R. Engelen G. Uljee I. 2015, Modeling cities and regions as complex systems. From Theory to Planning Applications. Cambridge (Mass.), MIT Press, 330 p. Wilson, A. G. 2014. Complex spatial systems: the modelling foundations of urban and regional analysis. Routledge.