

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 (), 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 à l’Aménagement du Territoire et à 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 heterogeneous, 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 automata. 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 ().

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 (*Méthode Informatique 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 (). The increase of computational possibilities has been relatively slow, allowing to consider around one thousand cities in 2007 with the Eurosims 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 an 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 characterized 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 programming 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.

Notons galement que lun des atouts principaux dOpenMOLE est l’acces transparent aux environnements de calcul haute performance (HPC). Laugmentation des moyens de calcul mentionne prcdemment peut se manifester

physiquement sous différents aspects pour le modélisateur: serveur local, cluster de calcul local, grille de calcul (mise en réseau de multiples clusters, comme la grille de calcul européenne EGI), services de cloud computing. Leur utilisation demande dans la majorité des cas des compétences informatiques avancées, généralement inaccessibles à un modélisateur géographe standard. OpenMOLE intègre une bibliothèque permettant d'accéder à la majorité de ces moyens de calcul, et leur mobilisation dans le DSL est entièrement transparente pour l'utilisateur. Celui-ci peut tester son script sur sa propre machine et passer à l'échelle sur les environnements HPC en modifiant un seul mot-clé dans celui-ci. La présentation de l'utilisation du DSL et de la mise en place de scripts n'étant pas l'objectif de ce chapitre, nous renvoyons le lecteur à la documentation en ligne d'OpenMOLE pour des exemples de scripts et d'exploration de modèles. Nous rappelons simplement les composants fondamentaux d'un script d'exploration: (i) la définition de prototypes, qui correspondent aux paramètres et aux sorties du modèle, qui prendront différentes valeurs lors de l'expérience; (ii) la définition de tâches, incluant l'exécution du modèle mais pouvant aussi par exemple être des pré- ou post-traitements - les tâches couvrant une très grande variété de langages (scala, java, NetLogo, R, Scilab, code natif comme python ou C++); (iii) la description des méthodes appliquer (exploration par échantillonnage, calibrage, recherche de diversité, etc.) qui agiront sur les valeurs des prototypes et lanceront la tâche d'évaluation considérée (le plus souvent le modèle); (iv) une spécification des données récupérées en sortie de l'exécution du script (les données de simulation étant souvent massives, une sélection par cette tâche est cruciale); et (v) la définition de l'environnement de calcul sur lequel la méthode sera lancée. La plateforme vise considérablement tendre les pratiques de la science sociale proposée par Epstein et Axtell (1996), qui envisageait chaque modèle multi-agents comme une société artificielle, engendrant des comportements macroscopiques à partir d'hypothèses mises sur les comportements microscopiques. Les expériences numériques envisageables changent d'échelle, et les questions posées au modèle de nature qualitative. Selon Clara Schmitt (2014), qui a utilisé la plateforme OpenMOLE pour développer avec Sébastien Rey-Coyrehourcq (2014) le modèle SimpopLocal destiné à simuler l'émergence d'un système de villes, le laboratoire virtuel que représente cette plateforme n'est pas seulement le modèle de simulation et les hypothèses qu'il simule (i.e. l'artificial society). Il contient aussi les méthodes, les outils et procédures de modélisation adaptés à la conception et l'exploration du modèle et dont la pratique procure autant de connaissances et de retours théoriques que la conception du modèle lui-même. Ce laboratoire virtuel s'apparente donc d'autant plus à un véritable laboratoire de recherche avec une paillasse (le modèle concevoir et explorer), les hypothèses d'un chercheur (les processus géographiques transcrits en mécanismes du modèle), des méthodes (la méthode de modélisation itérative et assistée par le calcul intensif), des outils (les procédures d'exploration automatisées et tout autre plan d'expérimentations incorporés dans OpenMOLE), le tout rassemblé dans une salle, la plateforme de modélisation SimProcess (Rey-Coyrehourcq, 2014). Par rapport aux protocoles généraux comme celui introduit par (Grimm et al., 2014) pour présenter l'ensemble des étapes de la modélisation, les principes appliqués dans OpenMOLE apportent surtout de la nouveauté en termes de capacités inédites d'exploration du comportement dynamique des modèles de simulation. Deux principales innovations consistent dans l'emploi systématique de méta-heuristiques d'optimisation, principalement des algorithmes génétiques, pour tester rapidement le plus grand nombre possible de combinaisons de valeurs des paramètres du modèle, et dans l'envoi simultané des simulations sur les multiples machines d'une grille de calcul, ce qui permet de réduire considérablement la durée des expériences qui sans cela deviendrait rapidement prohibitive.

Le choix des algorithmes génétiques comme heuristique d'optimisation est justifié par leur efficacité dans le cadre de problèmes d'optimisation multi-objectifs. Par ailleurs, le schéma de distribution en les (populations voulant indépendamment pendant une certaine durée) est particulièrement adapté à la distribution sur grille, chacun des nœuds faisant valoir une sous-population, qui est régulièrement récupérée, fusionnée dans la population globale, à partir de laquelle une nouvelle sous-population est générée et envoyée sur le nœud. Ce type d'algorithme s'étend par ailleurs relativement bien aux modèles stochastiques, même si cet aspect comporte encore un certain nombre de problèmes ouverts (Rakshit, Konar and Das, 2017). Suivant Rey-Coyrehourcq (2015), ces méthodes se situent dans le cadre plus global de l'Evolutionary Computation, et la bibliothèque scala MGO développe simultanément la plateforme et qui permet d'implémenter les algorithmes évolutionnaires, à côté connue pour être facilement tendue d'autres heuristiques en Evolutionary Computation, laissant les possibilités de méthodes incluses dans OpenMOLE totalement ouvertes.

(Reuillon, Leclaire et Rey-Coyrehourcq, 2013) décrivent les principes fondamentaux de la plateforme, tandis que (Pumain et Reuillon, 2017) donnent une contextualisation des différentes utilisations dans le cadre de modèles de simulation pour les systèmes de villes. Selon R. Reuillon cité par Raimbault (2017a), la philosophie d'OpenMole s'articule autour de trois axes: le modèle comme boîte noire explorer (i.e. méthodes indépendantes du modèle), l'utilisation de méthodes avancées d'exploration, et l'accès transparent aux environnements de calcul intensif. Ces différentes composantes sont en interdépendance forte, et permettent un changement de paradigme dans l'utilisation des modèles de simulation: utilisation de multi-modélisation, c'est-à-dire structure variable du modèle, comme il a été présenté au chapitre 4 (Cottineau et al., 2015), changement de la nature des questions posées au modèle (par exemple détermination complète de l'espace faisable (Chrel, Cottineau et Reuillon, 2015)), tout cela permis par l'utilisation du calcul

intensif (Schmitt et al., 2015). Les différentes méthodes disponibles dans ce cadre seront illustrées ci-dessous dans des exemples concrets. La documentation en ligne donne un aperçu global des méthodes disponibles dans la version la plus récente du logiciel et de leur articulation dans un cadre standard.

Nous considérons un modèle de simulation comme un algorithme produisant des sorties à partir de données et de paramètres en entrée. Dans ce cadre, nous rappelons que dans un cas idéal, l'ensemble des étapes suivantes devraient être nécessaires pour une utilisation robuste des modèles de simulation.

1. Identification des mécanismes principaux et des paramètres cruciaux associés, ainsi que de leur domaine de variation suggéré par leur signification thématique le cas échéant ; identification des indicateurs pour valuer la performance ou le comportement du modèle.
2. valuation des variations stochastiques : grand nombre de répétitions pour un nombre raisonnable de paramètres, tablisement du nombre de répétitions nécessaire pour atteindre un certain niveau de convergence statistique.
3. Exploration directe pour une première analyse de sensibilité, si possible valuation statistique des relations entre paramètres et indicateurs de sortie.
4. Calibrage, exploration algorithmique ciblée par utilisation d'algorithmes spécifiques (Calibration Profile (Reuillon, Schmitt et al. 2015), Pattern Space Exploration (Chrel, Cottineau et Reuillon 2015)).
5. Retours sur le modèle, extension et nouvelles briques de multi-modélisation, retours sur les faits stylistiques et la théorie.
6. Analyses de sensibilité tendues, correspondant des méthodes expérimentales en cours d'élaboration et d'intégration dans la plateforme, comme par exemple la sensibilité aux méta-paramètres et aux conditions spatiales initiales proposée par (Raimbault et al. 2018).

Le cas échéant, certaines étapes n'ont pas lieu d'être, par exemple l'évaluation de la stochasticité dans le cas d'un modèle déterministe. De même, les étapes prendront plus ou moins d'importance selon la nature de la question posée : le calibrage ne sera pas pertinent dans le cas de modèles complètement synthétiques, tandis qu'une exploration systématique d'un grand nombre de paramètres ne sera pas forcément nécessaire dans le cas d'un modèle qui a pour but d'être calibré sur des données.

Afin de mieux illustrer cette présentation générale de la plateforme et des méthodes associées, nous proposons dans la suite de cette section de développer précisément l'exemple du modèle SimpopLocal, dont la genèse a traitement à lier celle de la plateforme, et qui a été candidat pour le développement et l'application de diverses méthodes.

3.2 L'expérience Simpoplocal: simulation d'une émergence en géographie

Le modèle SimpopLocal a été conçu pour représenter l'émergence des systèmes de villes, telle qu'on a pu l'observer dans cinq ou six régions du monde, quelque 3000 ans après l'émergence de pratiques agricoles dans des sociétés sédentaires (Bairoch 1985; Marcus et al. 2008). Il s'agit bien d'expliquer l'émergence, non pas seulement de la ville, mais bien de systèmes de villes, car on sait que les villes de cette époque n'étaient jamais isolées mais déjà organisées en réseaux dans le territoire de chacune de ces civilisations antiques. Les publications les plus récentes des archéologues insistent sur une certaine continuité des processus ayant conduit la sédentarisation de populations de chasseurs-cueilleurs, regroupées en hameaux et villages, puis l'apparition de villes dans certaines de ces régions. Le développement de l'agriculture a été concomitant d'un accroissement considérable des densités de population et de la taille des groupes humains dans ces contrées (on passe de 0,1 personne par km² à 10, soit un facteur 100 entre les deux ordres de grandeur), ainsi que d'une complexification de l'organisation politique et de la division sociale du travail. Ce processus très lent d'accumulation des ressources et de concentration des populations s'effectue selon des enchaînements comportant de nombreux feedbacks, avec beaucoup de fluctuations dans la croissance, dues aux fréquents événements contraires que sont les catastrophes naturelles ou les prédations de groupes voisins. En raison de la lenteur des transformations et de leurs fréquentes interruptions, les archéologues contestent parfois désormais l'appellation de révolution néolithique qui avait été proposée en 1942 par Gordon Childe (Demoule 2018 p.159). Cependant, les géographes continuent d'identifier l'apparition des villes comme une émergence, une bifurcation pour deux raisons principales: d'une part elle ne s'est pas produite systématiquement dans toutes les régions où l'agriculture a été pratiquée, donc deux régimes d'évolution des systèmes de peuplement sont possibles et viables historiquement (des régions seulement agricoles et villageoises ont pu fonctionner pendant plusieurs siècles et subsistent aujourd'hui de façon résiduelle dans certaines forêts ou sur des îles du Pacifique par exemple), donc le régime territorial fonctionnant avec des villes constitue bien un attracteur spécifique dans la dynamique des systèmes de peuplement anciens; d'autre part, la trajectoire évolutive qui voit naître les villes traduit un changement qualitatif important (une émergence) avec un accroissement significatif de la diversité des fonctions sociales associées aux habitats et aussi un élargissement considérable dans l'échelle de la vie de relations: les échanges commerciaux qui s'y effectuent sur plus longue distance permettent ainsi aux villes d'être moins dépendantes d'un site de ressources locales comme le sont les villages agricoles et de développer les atouts d'une situation géographique exploitant les richesses d'un réseau de sites de plus en plus lointains (Reymond, 1971). Le modèle SimpopLocal vise à reproduire cet aspect remarquable de la dynamique des systèmes de peuplement, qui produit invariablement une amplification de la différenciation hiérarchique entre les habitats, définie dans la littérature comme un fait stylistique majeur: déjà dans tout système, en tout lieu et tout moment de l'histoire ou de la préhistoire, la répartition des tailles des lieux habités (mesure par la population ou l'étendue spatiale, voire la diversité des artefacts fonctionnels) est statistiquement très dissymétrique, comportant de

nombreuses très petites agglomérations et seulement quelques très grandes agglomérations selon une distribution assez régulière de type loi de Zipf ou log normale (Fletcher 1986; Liu 1996). Ce schéma hiérarchique est une propriété structurelle (ordre de taille des entités) au niveau macroscopique particulièrement persistante dans le temps, quelles que soient les fluctuations locales intervenant au niveau des entités. Le modèle SimpopLocal est conçu pour tester l'hypothèse énoncée dans la théorie évolutive des systèmes urbains (Pumain 1997), qui explique cette caractéristique structurelle par un processus de croissance urbaine en moyenne proportionnelle à la taille acquise, et son amplification par la création de multiples innovations technologiques et sociales produisant l'accroissement et la diversification des richesses qui se diffusent parmi les lieux mis en relation par toutes sortes d'échanges. Le modèle SimpopLocal s'inspire d'abord du modèle statistique qui constitue une excellente première approximation de l'évolution des populations dans un système de villes, en simulant la croissance urbaine comme un simple processus stochastique faisant varier la taille de chaque ville de façon proportionnelle à sa taille et conduisant à une distribution lognormale des populations urbaines (Gibrat 1931). La grande qualité de ce modèle statistique élémentaire tient à ce qu'il utilise comme explication de la croissance la taille déjà acquise, laquelle exprime à la fois la richesse accumulée et la capacité d'attraction et de résilience du lieu habité (en quelque sorte il s'agit d'un modèle selon le concept de croissance endogène des économistes). Mais SimpopLocal est conçu, comme les modèles précédents de la famille des modèles multi-agents Simpop (Bura et al 1996, Sanders et al. 2007), pour pallier l'insuffisance de la capacité du modèle de Gibrat à prévoir la tendance partout observée à la croissance plus forte qu'attendue des plus grandes villes situées en tête des réseaux (Moriconi-Ebrard, 1993) et l'exagération de l'inégalité entre les tailles des villes (Pumain 1997, Bretagnolle, Pumain 2010). Ces déviations au modèle de Gibrat sont liées aux corrélations de longue portée (Rozenfeld et al. 2008), suscitées par les interactions spatiales. L'effet de celles-ci amplifie la différenciation hiérarchique entre les tailles des villes participant aux échanges dans un système urbain (Favaro et al. 2011). Les modèles Simpop traduisent cet effet en introduisant, de manière exogène au modèle et à différents moments du temps de la simulation, de nouvelles fonctions urbaines qui sélectionnent certaines villes ou sont captées par elles dans un processus continu d'adaptation à ces innovations. En comparaison des autres modèles Simpop, SimpopLocal introduit deux nouveautés : il utilise une représentation abstraite des vagues d'innovation successives et les rassemble toutes dans un seul objet innovation. Une seconde originalité consiste à rendre le processus de création d'innovation endogène en le liant à la taille du lieu habité, censée amplifier de manière non linéaire l'émergence de nouvelles formes techniques, sociales ou culturelles (avec une probabilité de création variant comme le carré des populations en présence ou en relations). Cette version plus parcimonieuse de la construction du modèle permet de réduire considérablement le nombre de paramètres et autorise donc une exploration et une valuation plus systématiques.

3.3 Implémentation de SimpopLocal, de Netlogo à OpenMole Simpoplocal a été initialement développé avec le langage Netlogo, puis re-développé avec le langage de programmation Scala. La simulation avec Netlogo a bénéficié des facilités de l'interface qui permet de suivre numériquement et graphiquement les modifications engendrées sur les variables macroscopiques qui résument l'état du système, mais a montré très vite ses limites en termes d'expérimentation. La méthode manuelle de réglage des valeurs des paramètres permettait difficilement d'éviter les emballements de la croissance urbaine conduisant à des accroissements de taille des villes bien trop normés pour l'époque historique qu'il s'agit de simuler. La reprogrammation en Scala puis le passage sur la plateforme OpenMole devaient permettre une exploration plus précise et complète des comportements du modèle. Le modèle représente l'évolution des unités de peuplement dispersées dans une zone suffisamment grande pour accueillir quelques milliers d'habitants mais suffisamment limitée en surface pour assurer la connexion possible entre les lieux habités en fonction des moyens de transport disponibles à l'époque (par exemple il pourrait s'agir de l'ancienne Mésopotamie ou de la Mésopotamie antique). L'espace de simulation est composé d'une centaine de lieux habités. Chaque lieu est considéré comme un agent fixe et est décrit 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 quantité de ressources disponibles R est quantifiée en unités d'habitants et peut être comprise comme la capacité de charge de l'environnement local pour soutenir une population, laquelle varie en fonction des compétences en exploitation des ressources que la population locale a acquises, grâce aux innovations qu'elle a créées ou reçues des autres lieux habités. Toutefois, l'exploitation des ressources est effectuée localement et le partage ou l'échange de biens ou de personnes ne sont pas explicitement représentés dans le modèle. Chaque nouvelle innovation créée ou acquise par un lieu habité développe ses compétences en exploitation. L'entité innovation sentend ici comme une grande invention abstraite socialement acceptée, qui pourrait représenter une invention technique, une découverte, une organisation sociale, de nouvelles habitudes ou pratiques ... Chaque acquisition d'innovation par un lieu habité apporte la possibilité de dépasser ses seuils de capacité, et par conséquent autorise une croissance démographique. Le modèle a été conçu pour être le plus parcimonieux possible, en minimisant le nombre des attributs des agents (qui sont des lieux habités) et les paramètres qui contrôlent leur évolution. On a utilisé directement les ordres de grandeur moyens indiqués par les travaux des archéologues pour fixer environ 4000 ans la durée de la période de transition entre un système de peuplement agricole et un système de peuplement urbain, pour estimer un taux de variation moyen annuel de la population d'environ 0,02. On définit une valeur initiale pour la population et les ressources des lieux habités,

puis le réseau d'interaction entre eux est créé. Ensuite, chaque tape de la simulation, les mécanismes de croissance de la population et de diffusion de l'innovation sont appliqués. L'impact des innovations sur l'efficacité de l'extraction des ressources est calculé. Cette boucle est itérée jusqu'à ce que le critère d'arrêt soit atteint : dans ce cas, après 4000 tapes ou lorsqu'un nombre maximal arbitraire d'innovations a été atteint. On observe l'évolution de l'état du système de peuplement défini au niveau macro-géographique par la distribution de la taille des lieux habités, résumé par la pente de la distribution rang-taille. Le modèle utilisant certains paramètres qui sont des probabilités est stochastique, un même jeu de valeurs de paramètres peut donner lieu à des résultats sensiblement différents. Une méthode automatisée pour faire varier les valeurs des paramètres et interpréter les résultats obtenus a été mise au point progressivement par une collaboration entre informaticiens et géographes.

3.4 Calibrage et validation L'automatisation de l'exploration des dynamiques engendrées par les modèles de simulation avec la plateforme OpenMole utilise des algorithmes génétiques qui réalisent de façon systématique les variations des valeurs des paramètres auparavant effectuées à la main par le chercheur. La distribution des calculs sur une infrastructure de grille (un réseau d'ordinateurs) permet en outre de conduire ce très grand nombre d'opérations combinatoires en réduisant considérablement le temps de calcul, grâce au traitement en parallèle de l'information. Mais la mise en œuvre de cette nouvelle forme de l'expérimentation des modèles suppose aussi une intervention du chercheur thématique, qui doit sélectionner les objectifs précis que son modèle doit satisfaire, tandis qu'un raffinement supplémentaire de la méthode d'exploration peut conduire à un renforcement de la confiance qu'il accorde aux hypothèses scientifiques de son modèle. Le calibrage comme optimisation au moyen des algorithmes génétiques Le calibrage est une procédure qui cherche à minimiser l'écart (appelé *fitness*) entre le comportement simulé par le modèle et le comportement observé empiriquement, en faisant varier de façon incrémentale les valeurs inconnues des paramètres du modèle. Stonedahl (2011) a rappelé les difficultés de cette exploration qui devient vite fastidieuse lorsqu'elle est conduite manuellement, cause des multiples bifurcations intervenant dans des modèles où la plupart des mécanismes liant les variables sont non linéaires. Une exploration exhaustive de l'espace des paramètres n'est pas envisageable car elle exigerait des temps de calcul trop importants, en croissance exponentielle avec le nombre de ces paramètres. Comme ces procédures produisent aussi de grandes quantités de résultats, elles exigent en outre d'employer des méthodes adaptées pour traiter et visualiser les informations engendrées par les simulations. Tout un ensemble de logiciels doit donc être mis au point pour permettre au chercheur de découvrir les principaux schémas des dynamiques associées aux variations des paramètres de son modèle. C'est là où des procédures informatiques adaptées peuvent être utilisées, en rapportant la question du calibrage à un problème d'optimisation. Les algorithmes génétiques ont été utilisés pour calibrer des systèmes multi-agents dans plusieurs domaines, en médecine (Castiglione et al, 2007), en écologie (Duboz et al, 2010), en économie (Espinosa, 2012; Stonedahl et Wilensky, 2010a), ou en hydrologie (Solomatine et al, 1999). En dépit de la large utilisation des systèmes multi-agents en sciences sociales, cette méthode n'a pas été appliquée très souvent (Heppenstall et al, 2007; Stonedahl et Wilensky, 2010b). Ce type d'expérience numérique exige en effet que soient définis des objectifs quantitatifs permettant d'évaluer si les résultats de la simulation sont compatibles avec les attentes des experts, il faut également savoir gérer une énorme charge de calcul et parvenir à optimiser une fonction de *fitness* susceptibles de très importantes variations stochastiques (Pietro et al, 2004). Dans le cas de SimpopLocal, qui comprend 5 paramètres dont les valeurs sont inconnues (même leurs ordres de grandeur ne peuvent pas être estimés à partir de données empiriques), nous avons dû identifier trois fonctions objectif. Celles-ci caractérisent un résultat de simulation au niveau macro-géographique et correspondent à des faits stylisés dont les ordres de grandeur ont pu être tablés à partir des connaissances archéologiques et historiques : la distribution finale des tailles de villes doit être lognormale (peu différente d'une loi de Zipf), la taille maximale qu'atteint la plus grande ville doit être d'environ 10 000 habitants, pour une durée de simulation équivalente à 4000 ans. Cette obligation de définir des fonctions-objectif pourrait être considérée comme une contrainte forte sur la validité épistémologique du modèle, elle semble en effet contredire l'hypothèse d'une évolution ouverte pour les systèmes de villes. En fait, cette étape intermédiaire de calcul représente un comprimé de connaissances, notre exigence est minimale sur la représentativité et la plausibilité du comportement du modèle par rapport à l'ensemble envisageable des dynamiques des villes en système (l'époque historique de l'émergence des villes). Le résultat en termes d'évaluation des simulations doit permettre d'avancer dans la connaissance des processus d'interaction intra-urbains susceptibles d'engendrer cette dynamique générale. À l'échelon macroscopique du système, cette reconstitution théorique s'apparentant alors à ce que des physiciens nomment le problème inverse. Un domaine de variation numérique assez large est tablé a priori pour chacun des cinq paramètres. Chaque jeu de paramètres, combinant une valeur pour chacun d'entre eux, est évalué en fonction de la sortie de simulation qu'il produit. Cette évaluation mesure la proximité entre les sorties de la simulation et les fonctions objectifs définies pour le modèle et permet ainsi de mesurer la capacité d'un certain ensemble de valeurs de paramètres à reproduire les faits stylisés que la simulation doit approcher au mieux. Les paramétrages recevant les meilleures évaluations sont ensuite utilisés comme base pour engendrer de nouveaux jeux de paramètres qui sont ensuite testés.

Exploration de l'espace des paramètres sous contrainte d'objectifs Le modèle 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 qu’une 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 l’valuation 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. L’objectif de population maximale quantifie la capacit du modle engendrer des villes plus ou moins grandes, le rsultat d’une simulation est test en calculant lcart entre la taille de la plus grande agglomration et la valeur attendue de 10 000 habitants: $[(\text{population de la plus grande agglomration} - 10\,000) / 10\,000]$. L’objectif 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 d’iterations de la simulation et la valeur attendue de 4000 tapes de la simulation: $[(\text{simulation dure} - 4000) / 4000]$. Ces trois calculs de erreur sont normaliss afin de pouvoir comparer le degr de russite d’une simulation vis--vis de chacun des trois objectifs. Mais la rggration des trois calculs qui produirait une seule mesure de qualit globale n’tant pas possible, un algorithme multi-objectif est ncessaire pour dterminer quelles simulations sont les plus satisfaisantes pour approcher la configuration finale souhaite. Ce type d’algorithme calcule des solutions de compromis telles qu’aucune 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. L’utilisation de mthodes d’exploration 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, l’informatique une si grande chelle suppose d’orchestrer l’exécution 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. C’est 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. L’infrastructure de la grille de calcul (EGI) nous a permis d’utiliser une puissance de calcul telle qu’un demi milliard d’exécutions 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 l’mergence d’un 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 d’atteindre 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 d’une fonction qui calcule une seule valeur numrique dcrivant la qualit du calibrage pour le modle, l’algorithme de profil calcule l’erreur de calibrage la plus faible possible lorsque la valeur d’un paramtre donn est fixe et que les autres sont libres. L’algorithme calcule cette erreur minimale pour tout le domaine de variation du paramtre tudi. Pour chaque valeur d’un paramtre, l’algorithme 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 d’ajustement optimale en fonction des variations du paramtre tudi. Le profil de calibrage montre ainsi plusieurs formes possibles pour cette courbe. Lorsqu’elle prsente une nette inflexion vers les valeurs les plus basses pour l’erreur de calibrage, cela pour un tout petit domaine de variation des valeurs du paramtre tudi, on peut en conclure qu’on a vraiment identifi l’ordre de grandeur du paramtre qui satisfait aux exigences en termes de comportement du modle. Si l’une de ces courbes reste plate, cela indique que le paramtre n’a pas d’effet 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 d’une innovation a t finalement exclu car des variations restaient sans effet sur la qualit d’ajustement du modle, toutes choses gales quant aux variations des autres paramtres (Schmitt, 2014). On a donc ici la possibilit d’valuer jusqu’au quel point les mcanismes imaginés 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, c’est 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. C’est 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 representatifs 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 Exemples dapplications dOpenMOLE: modles dinteraction rseaux-territoires Nous proposons dans cette section dillustrer lapplication des mthodes dexploration dOpenMOLE et du calcul intensif une autre question thmatique, celle des interactions entre rseaux et territoires. Cette question a aliment de nombreux dbats scientifiques, pour lesquels la plupart des questions restent relativement ouvertes. Par exemple, le problme des effets structurants des infrastructures de transport (Bonnafois et Plassard, 1974), prsent par (Offner, 1993) comme un mythe scientifique invoqu pour justifier le cot dune nouvelle infrastructure par ses retombes sur le dveloppement rgional, pas toujours observes moyen terme, peut selon A. Bretagnolle dans (Offner et al., 2014) tre observ pour des territoires plus vastes et sur le temps long, tout en tenant compte des fluctuations locales dans les dynamiques des systmes de villes. La difficult empirique dextraire des faits styliss gnraux ainsi que la difficult conceptuelle d’entits gographiques en relations de causalits circulaires, sont contournes par lapproche de modlisation de la co-volution des rseaux de transport et des territoires propose par Raimbault (2018b). Les rsultats obtenus sont troitement lis lutilisation dOpenMOLE et de ses algorithmes dexploration et de calibrage, dont nous allons donner quelques illustrations. Lapplication de calibrage multi-objectif s’avre essentielle pour lapplication des modles de systmes de villes des situations relles. Par exemple, (Raimbault, 2018a) introduit un modle d’volution dun systme de villes sur le temps long, proche du modle de (Favaro et Pumain, 2011), mais se concentrant sur leffet du rseau de transport physique. Les taux de croissance des villes sont dtermins par la superposition de plusieurs effets: (i) croissance endogène capture par un taux de croissance fixe correspondant au modle de Gibrat; (ii) interactions entre villes par un modle gravitaire; (iii) rtroaction des flux circulant dans le rseau sur les villes traverses. Ce modle est calibr de manire non stationnaire dans le temps (cest--dire sur des fenêtres temporelles glissantes, afin de prendre en compte le changement de nature des dynamiques urbaines, comme observ par Bretagnolle et Franc (2018) avec par exemple les mutations des rseaux de transport), sur le systme de villes franais entre 1830 et 2000. Pour calibrer le modle, les populations simules sont compares aux populations observes. ce stade, lutilisation dun algorithme de calibrage multi-objectif (l’algorithme NSGA2 implment dans OpenMOLE) est essentielle. En effet, lajustement peut par exemple seffectuer sur une erreur carre moyenne dans le temps et pour lensemble des villes. Cependant, vu les disparits de taille des villes lies la hirarchie urbaine, il merge rapidement quune optimisation mono-objectif sur cette erreur s’attellera ajuster la taille des plus grandes villes, au dtriment de la majorit des villes du systme. Lajout dun second objectif, pris par exemple comme une erreur carre moyenne sur les logarithmes des populations, permet de prendre celles-ci en compte. Un rsultat important de (Raimbault, 2018a) est alors l’mergence de fronts de Pareto pour ces deux objectifs, pour lensemble des fenêtres temporelles considres. Cela montre que ce type de modle doit tre appliqu en faisant un compromis entre lajustement des populations pour les villes moyennes et des populations pour les plus grandes villes. Ce rsultat est permis grce loptimisation multi-objectif par algorithme gnrique dOpenMOLE. Un autre exemple dapplication des mthodes de la plateforme qui illustre son rle crucial est donn par la recherche de rgimes de co-volution. Suivant Raimbault (2017b), l’tude des motifs de corrélation retarde dans le temps permet disoler des rgimes typiques dinteraction entre variables de rseau et variables de territoires. Plus prcisment, Raimbault (2018b) dfinit la co-volution comme lexistence de relations circulaires causales, au niveau dun ensemble dentits dans une certaine emprise spatiale. Dans le cas des rseaux et des territoires, les propriéts des rseaux doivent tre localement causes par celles des territoires, et rciproquement. Des causalits unidirectionnelles des rseaux vers les territoires correspondent alors aux effets structurants mentionns ci-dessus. Cette dfinition permet de capturer la congruence (Offner, 1993) entre ces objets, en quelque sorte leur adaptation rciproque de manire dynamique. Elle permet aussi la construction dune mthode oprationnelle propose par Raimbault (2017b), qui cherche statistiquement des liens de causalit entre variables correspondantes. En pratique, la notion faible de causalit de Granger est mobilise, permettant une flexibilit au regard des donnes ncessaires et du cadre temporel et spatial destination. Cette causalit est dans notre cas quantifie par les corrélation retardes entre variations des variables de rseau (comme les centralits ou l’accessibilit) et variations des variables de territoires (comme population, emplois, transactions immobilires, etc.), et lexistence de maxima significatifs des retards non nuls donne une direction de causalit. Une typologie de ces profils de corrélation retardes fournit ce quon nomme des rgimes de causalit, parmi lesquels des rgimes de co-volution o deux variables territoire et rseau sont en causalit rciproque. La question est

alors dans un cas d'étude donné identifier les régimes présents à partir de données observées ou de données simulées par un modèle, et notamment ceux qui correspondent à une co-volution. La démonstration de l'existence de tels régimes en sortie d'un modèle de co-volution n'est pas a priori attendue, puisque les processus inclus à l'échelle microscopique où les influences sont en effet réciproques n'impliquent pas une causalité réciproque à l'échelle macroscopique des indicateurs, puisque les modèles considérés sont complexes et témoignent d'une émergence. Cette méthode est appliquée à un modèle macroscopique de co-volution par Raimbault (2019a), qui tend le modèle de Raimbault (2018a) par l'ajout de règles d'évolution des capacités des liens du réseau. Un échantillonnage direct, qui consiste en un tirage aléatoire d'un nombre fixe de points de paramètres (par exemple par échantillonnage Hypercube Latin maximisant la répartition des points), est une première expérience permise par OpenMOLE pour avoir un aperçu de la capacité du modèle à produire de la co-volution. Celui-ci permet d'isoler un certain nombre de régimes pouvant être potentiellement produits par le modèle (33 régimes pour 729 régimes possibles pour les variables considérées, i.e. 4,5). L'application de l'algorithme Pattern Space Exploration (Cherel, Reuillon and Cottineau, 2015) avec comme objectif la diversité des régimes produits permet alors de considérablement tendre cette conclusion, puisque celui-ci produit 260 régimes (35,7). Cette méthode permet par ailleurs de comparer entre eux des modèles avec une certaine confiance dans l'exhaustivité des solutions obtenues. Raimbault (2019b) applique la même démarche au modèle SimpopNet introduit par Schmitt (2014), qui est également un modèle de co-volution à l'échelle macroscopique et présentant un grand nombre de points communs avec le modèle précédent notamment dans les variables considérées et donc les indicateurs de sortie calculables. Il est alors obtenu un nombre plus faible de régimes d'interaction et de régimes de co-volution, confirmant d'une part qu'il n'est pas immédiat pour un modèle conçu pour la co-volution de faire effectivement merger des régimes de co-volution, et suggérant par ailleurs que des contraintes plus fortes dans les règles d'évolution du réseau induisent une plus grande difficulté à produire une diversité de régimes.

5 Perspectives

L'élaboration de la plateforme OpenMOLE a créé un axe, voire un domaine de recherche original, avec un positionnement spécifique dont l'un des aspects remarquables est un haut niveau d'interdisciplinarité entre sciences humaines et disciplines plus techniques comme l'informatique. Selon Banos (2017) cela conduit à la production d'une connaissance plus large et plus profonde (l'image de la spirale vertueuse de Banos (2013) entre disciplinarité et interdisciplinarité). Mais aussi, avec la philosophie de plateforme unique (voquée ci-dessus, par l'interaction forte entre les trois axes de développement des modèles, d'accès aux méthodes d'exploration innovantes, et d'accès transparent aux environnements de calcul intensifs), les perspectives ouvertes sont nombreuses, tant sur le plan technique que sur celui théorique, méthodologique ou thématique. Nous en donnons ci-dessous quelques illustrations, rendant compte d'un état présent des futurs possibles pour OpenMOLE.

5.1 Méthodes d'extension des méthodes mises à disposition

est un axe privilégié de la recherche lié au développement d'OpenMOLE. Par exemple, la résolution exhaustive de problèmes inverses (Aster et al. 2018) n'est actuellement pas incluse. La résolution d'un problème inverse consiste à déterminer l'ensemble des antécédents d'un objectif donné dans l'espace de sortie du modèle. Les algorithmes de calibrage résolvent des problèmes similaires mais ne garantissent pas l'exhaustivité des solutions produites, ce qui peut considérablement poser problème en cas d'incertitude (Rey-Coyrehourcq, 2015), i.e. de configurations de paramètres ou de conditions initiales conduisant par des trajectoires différentes à un résultat identique. Une heuristique de problème inverse s'inspirant des mécanismes de PSE est actuellement en cours d'élaboration pour une intégration dans OpenMOLE. L'utilisation de méthodes d'inférence Bayésiennes est également une piste développée. En effet, dans le cas de modèles fortement stochastiques, et où les distributions jointes ont des formes non standard, une estimation de la distribution de probabilité des paramètres peut être fournie par ce type de méthodes. Dans le cas des modèles de simulation, la méthode d'Approximate Bayesian Computation (Csilléry et al. 2010) permet, pour un jeu de données observées, de fournir la distribution de probabilité des paramètres ayant le plus probablement conduit à celles-ci. Il s'agit ainsi d'un calibrage tendu, avec une connaissance produite probabiliste permettant de prendre en compte l'incertitude. Une spécification de cette méthode proposée par Lenormand et al. (2013), destinée à réduire le nombre de simulations dans le cas de modèles au temps de calcul significatif, est également en cours d'adaptation au calcul parallèle et d'intégration dans la plateforme. Signalons finalement diverses directions méthodologiques également en cours d'investigation: (i) la question de la haute dimensionnalité pose rapidement problème dans l'utilisation de l'algorithme PSE, puisque le nombre de configurations de sortie est potentiellement soumis à la malédiction de la dimension (curse of dimensionality) c'est-à-dire que le temps ou la taille d'exécution sont exponentiels en le nombre de dimension (une exploration par grille est l'exemple le plus simple pour se donner une idée de ce phénomène) - de nouvelles méthodes combinant réduction de dimension et recherche de diversité permettraient de résoudre ce problème et prendre en compte une richesse de sorties bien plus grande; (ii) la question de la sensibilité aux conditions spatiales initiales déjà mentionnée (Raimbault et al., 2018), est particulièrement pertinente pour les modèles géographiques, et une librairie Scala incluant des générateurs synthétiques de configurations de peuplement à différentes échelles est actuellement en cours d'élaboration, incluant par exemple les générateurs de quartiers étudiés par (Raimbault and Perret, 2019); (iii) l'implémentation de critères d'information pour la performance

des modèles, déjà mentionnés dans le chapitre 4 et qui sont une pierre angulaire des démarches de multi-modélisation, est aussi l’étude, comme le critère POMIC proposé par Piou et al. (2009).

5.2 Outils Au long de son développement, OpenMOLE a toujours été à la pointe en termes d’outils utilisés et développés. Le choix du langage Scala pour remplacer Java dans les premières versions, est un choix technologique innovant et particulièrement pertinent par les possibilités de programmation fonctionnelle mais aussi de programmation objet qu’il apporte, tout en gardant l’infrastructure sous-jacente de Java permettant une grande portabilité sans complications selon le système d’exploitation ou le hardware, ce qui est crucial pour la distribution des calculs sur des nœuds de grille hétérogènes. Par exemple, des propriétés comme le mixage de traits rendent Scala particulièrement pertinent pour la multi-modélisation (Odersky and Zenger, 2005). Les possibilités offertes par la programmation objet sont conservées dans Scala, et peuvent être combinées à l’abstraction de la programmation fonctionnelle, en faisant un langage plus puissant en ce sens de flexibilité que d’autres langages fonctionnels comme Haskell (Oliveira and Gibbons, 2010). Par ailleurs, des propriétés comme les conversions implicites ou les *case classes* rendent Scala ergonomique pour l’élaboration de DSL (Sloane, 2008), qui comme nous l’avons déjà mentionné est un aspect essentiel d’OpenMOLE. Les questions d’embarquement de programmes, et par extension de modèles, restent un domaine de recherche actif notamment en lien avec la reproductibilité. Le programme Docker, qui utilise des conteneurs, permet d’embarquer un environnement d’exécution identique quel que soit le système d’exploitation et le hardware. Hung et al. (2016) propose de coupler Docker à une interface graphique pour la reproductibilité scientifique. Des programmes similaires comme Singularity sont spécifiquement dédiés à la reproductibilité d’expériences HPC (Kurtzer et al., 2017). Le cœur de la stratégie d’embarquement d’OpenMOLE ne repose pas sur un tel programme, par exemple pour des questions de performance, mais certaines tâches reposant sur l’exécution de binaires ou de programmes dans un environnement complexe sont embarquées dans OpenMOLE par une tâche utilisant Docker (par exemple pour la tâche pour le langage R qui demande l’installation d’un environnement R complet). Une amélioration de l’intégration de Docker dans OpenMOLE est un axe de recherche actif et crucial pour l’extension future de la genericité des programmes embarquables. OpenMOLE se place ainsi à la pointe de la recherche technique en termes de reproductibilité scientifique. De la même manière, la question de la scalabilité des expériences est au cœur de la philosophie de la plateforme, et des recherches sont menées par exemple pour automatiser le déploiement de multiples instances d’OpenMOLE sur un cluster et faciliter leur utilisation au sein de communautés de thématiques. Conclusion L’exploration des modèles de simulation s’est poursuivie en géographie par l’intermédiaire d’initiatives comme le développement de la plateforme OpenMOLE. Celle-ci s’est menée dans un cadre hautement interdisciplinaire et réciproque (relation gagnant-gagnant entre informaticiens et géographes), mais aussi au travers d’une intégration inédite des domaines de connaissance (Raimbault, 2017a), c’est-à-dire des connaissances empiriques, théoriques et de modélisation, mais aussi les outils et méthodes, qui sont dans chacun de ces domaines en interaction forte. L’aventure OpenMOLE, et sa branche liée à la géographie dans le cadre de l’ERC Geodiversity, témoigne d’une nouvelle façon de produire des connaissances géographiques, résolument evidence-based, rendant envisageable la production de preuves scientifiques en sciences sociales. Cette manœuvre reste à être propagée et la démarche à être valorisée pour réaliser son potentiel de direction future de la Géographie Théorique et Quantitative, en complémentarité avec les nouvelles disciplines émergentes de City Science et Urban Analytics décrites par Batty (2019), mais la preuve de concept est largement valide et donne des arguments de poids aux sciences humaines pour résister à l’hégémonie colonisatrice de sciences dures comme la physique prétendant un monopole sur les approches evidence-based des systèmes sociaux (Dupuy and Benguigui, 2015).

References

- 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.
- 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.
- 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.
- Lowry I.S. 1964 A model of metropolis. Pittsburgh Regional Planning Association, Rand Corporation.
- Moriconi-Ebrard F. 1993, L’urbanisation du monde. Paris, Anthropos.
- Morrill, R. L., 1962, The Development of Models of Migration and the Role of Electronic Processing Machines, Entretiens de Monaco, Human Displacement,

Monaco, M., Morrill, R. L. 1963, The distribution of migration distances. Papers of the Regional Science Association. 11, 1, 73-84.

Odersky, M., Zenger, M. (2005, October). Scalable component abstractions. In ACM Sigplan Notices (Vol. 40, No. 10, pp. 41-57). ACM.

Offner, J. M., 1993. Les effets structurants du transport: mythe politique, mystification scientifique. L'espace géographique, 233-242.

Offner, J. M., Beaucire, F., Delaplace, M., Frmont, A., Ninot, O., Bretagnolle, A., Pumain, D., 2014. Les effets structurants des infrastructures de transport. Espace Géographique, 43(1), p-51.

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

Passerat-Palmbach, J., Reuillon, R., Leclaire, M., Makropoulos, A., Robinson, E. C., Parisot, S., Rueckert, D. (2017). Reproducible Large-Scale Neuroimaging Studies with the OpenMOLE Workflow Management System. Frontiers in neuroinformatics, 11, 21.

Pietro A D, While L, Barone L, 2004, Applying evolutionary algorithms to problems with noisy, time-consuming fitness functions, Proceedings of the 2004 Congress on Evolutionary Computation 2 12541261

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

Pumain D. Reuillon R. 2017, Urban Dynamics and Simulation Models. Springer, International. Lecture Notes in Morphogenesis, 123 p. ISBN: 978-3-319-46495-4, DOI 10.1007/978-3-319-46497-8_3.

Raimbault, J. 2017(a). An Applied Knowledge Framework to Study Complex Systems. In Complex Systems Design Management 45).

Raimbault, J. 2017(b). Identification de causalités dans des données spatio-temporelles. In Spatial Analysis and GEomatics 2017.

Urban Analytics and City Science, 2399808318774335.

Raimbault, J. 2018(b). Modélisation des interactions entre réseaux de transports : une approche par la co-evolution. In JJCPacte – Citeres – Les capacités transformatives des réseaux dans la fabrication des territoires. Revue d'évolution of cities and networks. Forthcoming in Handbook of Cities and Networks, Rozenblat C., Niel Z., eds. Edward Elgar Publishing.

evolutionary patterns in systems of cities : a systematic exploration of the SimpopNet model. arXiv preprint arXiv : 1809.00861.

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

Raimbault, J., Perret, J. (2019). Extending sensitivity analysis to initial spatial conditions in geosimulation models. arXiv preprint arXiv : 1812.06008.

Rakshit, P., Rist R.C. 1970, Student social class and teacher expectations : the self-fulfilling prophecy in ghetto education. Harvard Educational Review, 40(4), 451.

Reuillon R., Leclaire M. Rey Coyrehourcq S. 2013. OpenMOLE, a workflow engine specifically tailored for the distributed exploration of complex systems. The calibration profile (cp) algorithm. Journal of Artificial Societies and Social Simulation, 18(1), 12.

Rey Coyrehourcq, S. 2014, Urban simulation process. Université Paris I Panthéon Sorbonne, Thèse de doctorat.

Reymond H. 1971, Pour une problématique théorique, in Isnard H. 1970, La ville et son territoire, 18707.

Sanders L. Pumain D. Mathian H. Gurin - Pace F. Bura S. 1997, SIMPOP : a multi-agents system for the study of urbanism. Journal of Artificial Societies and Social Simulation, 10(4), 305.

Sanders L., Favaro JM., Glisse B., Mathian H., Pumain D. 2007, Artificial intelligence and collective agents : the EUROSIM project. In Proceedings of the European Conference on Artificial Intelligence, 10-14 September 2007, Valencia, Spain, 315.

Sloane, T. (2008). Experiences with domain-specific language embedding in Scala. In Domain-Specific Program Development: From Languages to Tools, NorthWestern University of Illinois. Evanston, IL, http://forrest.stonedahl.com/thesis/forrest_toledo.pdf.

Resilience, Robustness, and Evolvability (AAAI Press, Menlo Park, CA), 120129.

Taillandier P., Grignard A., Gaudou B. et Drogonzoni A. 2017, Application de l'analyse de sensibilité à la modélisation de la dynamique des populations. European Journal of Geography, 671, URL : http://journals.openedition.org/ejg/10.4000/cybergeo.26263

Tisue, S., Wilensky, U. 2004. Netlogo : A simple environment for modeling complexity. In International Conference on Artificial Intelligence, 21.

Terrier C. 1980, MIRABELLE, Courrier des Statistiques, 73.

Van Deursen, A., Klint, P. (2002). Domain-specific languages for software development. In Proceedings of the 17th 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 adaptive systems. The modelling foundation of urban and regional analysis. Routledge.