**Ouvrage La modélisation des villes et des territoires**

Chapitre 5 : Méthodes d’exploration des modèles de simulation

Juste Raimbault et Denise Pumain

On rappelle d’abord dans ce chapitre à quel point les modèles de simulation sont une absolue nécessité en sciences humaines et sociales, qui ne peuvent que très exceptionnellement recourir aux méthodes des sciences expérimentales pour construire leurs savoirs. Les modèles offrent la possibilité de simuler les processus sociaux en remplaçant le jeu complexe des actions et réactions individuelles et collectives par des mécanismes mathématiques ou informatiques plus simples, permettant de faciliter la compréhension des relations entre causes et conséquences de ces interactions et d’effectuer des prévisions. La formalisation par des modèles mathématiques susceptibles d’offrir des solutions analytiques n’étant souvent guère envisageable pour donner des représentations satisfaisantes de la complexité sociale (Jensen 2018), ce sont de plus en plus des modèles informatiques à base d’agents qui sont utilisés. Pendant longtemps les capacités limitées de calcul des ordinateurs ont empêché de programmer des modèles qui prennent en compte les interactions entre de grands nombres d’entités localisées géographiquement (personnes ou territoires). En principe, ces modèles doivent nous renseigner sur les possibilités et les conditions de l’émergence de certaines situations définies à un échelon macro-géographique à partir des interactions intervenant à un niveau micro-géographique, dans des systèmes aux comportements trop complexes pour être compris par un cerveau humain. Encore faut-il pouvoir étudier le comportement dynamique de ces modèles comportant des effets de feedback non linéaires et vérifier qu’ils produisent des résultats plausibles à toutes les étapes de leur simulation. Ce travail indispensable de l’exploration de la dynamique des systèmes modélisés est resté balbutiant jusque vers la fin des années 2010. Ensuite, des algorithmes combinant des méthodes plus sophistiquées incluant des algorithmes génétiques et le recours au calcul intensif distribué ont permis un saut qualitatif important dans l’exploration et la validation des modèles, duquel résulte un tournant épistémologique pour les sciences humaines et sociales, comme l’indiquent les dernières applications réalisées avec l’aide de la plateforme OpenMOLE présentée ici.

## 1 Les sciences sociales et l’expérimentation

L’expérimentation a beaucoup aidé à construire les sciences de la nature, en ce qu’elle consiste à reproduire des processus matériels, physiques, chimiques ou biologiques, selon des dispositifs imaginés par les chercheurs pour sélectionner, souvent en les isolant, des enchaînements de faits plus simples que ceux opérant dans une réalité complexe. La confrontation des résultats de ces manipulations à des données d’observation, en totalité ou en partie étrangères à celles qui ont servi à construire le dispositif expérimental, est considérée comme apportant une preuve de la véracité ou de la justesse du raisonnement explicatif qui est à la base de la construction du modèle, plus ou moins probante en fonction de la qualité de l’ajustement entre les prédictions du modèle et les observations. On sait cependant que la justesse des prédictions d’un modèle ne suffit pas à valider totalement l’adéquation entre le mécanisme explicatif imaginé par les manipulateurs du dispositif expérimental et les processus à l’œuvre dans le système étudié, mais c’est une étape importante dans la construction de modèles et de théories enrichis par les observations.

En sciences humaines et sociales, la mise au point de dispositifs expérimentaux est très délicate car elle se heurte à de nombreux obstacles pratiques et éthiques. La critique éthique et politique impose des limites à la manipulation des personnes et au contrôle de leur liberté. Les scrupules propres à l’ontologie et la déontologie scientifiques (faisant partie de ce qu’on appelle aujourd’hui l’intégrité), n’ont certes pas empêché en pratique les manipulations, bienveillantes ou non, opérées au cours des temps historiques par des acteurs disposant d’un pouvoir politique, culturel ou économique de prendre des décisions, plus ou moins bien informées « scientifiquement » (voir à toutes époques les écrits des « conseillers du prince », tels Bodin, Machiavel, Botero… pour citer quelques-uns parmi ceux qui ont traité de l’aménagement des territoires) et de procéder à des « expérimentations » de formes de gouvernance ou d’innovations technologiques ou culturelles dont les résultats ont pu être évalués, tantôt comme bénéfiques et tantôt comme désastreux. Evaluer l’efficacité de décisions est d’ailleurs encore plus compliqué puisque souvent on ne dispose pas de contre-épreuve, de scénario alternatif, et que les acteurs eux-mêmes ne se privent pas de « biaiser » le jeu avec leurs « prophéties auto-réalisatrices », assurant par avance du résultat qu’ils escomptent (Rist, 1970). La difficulté souvent déplorée de l’évaluation des politiques publiques est ainsi accrue par l’incertitude des limites entre l’action et son contexte, dans l’espace comme dans le temps.

Conduire le changement dans la vie sociale, quelle que soit l’échelle des interventions, reste une opération coûteuse et risquée, déontologiquement peu acceptable par les scientifiques, dont bien peu finalement osent se lancer dans des projets de « recherche-action ». Une controverse a ainsi opposé pendant les années 1960 en France les tenants d’une « géographie appliquée », bonne connaisseuse du « terrain » mais parfois méthodologiquement et politiquement conservatrice à ceux d’une « géographie active » plus engagée dans la transformation de la société. Parfois, par exemple pour contribuer à la définition de la politique des métropoles d’équilibre en France (opération de la DATAR[[1]](#footnote-1) en 1964), les géographes participant aux études d’aménagement (dont Michel Rochefort en l’occurrence) s’appuyaient, sans vraiment oser le dire, sur des modèles scientifiques (dans ce cas précis la théorie des lieux centraux de Walter Christaller). Les géographes actuels affichent plus volontiers un souci d’aider à la décision de la manière la plus avisée possible selon l’état de leurs connaissances. Ils font alors souvent le choix de recourir à des modèles de simulation, *in silico*, implémentés dans des ordinateurs. La simulation informatique est ainsi devenue un substitut à l’expérimentation. Ce n’est pas un hasard si, parmi les chercheurs en sciences sociales, les géographes s’y sont intéressés très tôt : la diversité des multiples catégories de données (paysages, populations, habitats…) qu’ils manipulent pour rendre compte des aménagements apportés par les sociétés à la surface et aux interfaces terrestres, l’étendue souvent large des territoires qu’ils examinent aux échelles régionale, nationale ou mondiale, expliquent leur besoin de recourir à l’informatique pour organiser ces masses d’information et comprendre les dynamiques qu’elles représentent.

## 2 Données de la géographie et capacités informatiques

Nous considérons un modèle de simulation comme un algorithme produisant des sorties à partir de données et de paramètres en entrée. Dans un cas idéal, l’ensemble des étapes suivantes sont 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 évaluer la performance ou le comportement du modèle.
2. Évaluation des variations stochastiques : grand nombre de répétitions pour un nombre raisonnable de paramètres, établissement du nombre de répétitions nécessaire pour atteindre un certain niveau de convergence statistique.
3. Exploration directe pour une première analyse de sensibilité, si possible évaluation statistique des relations entre paramètres et indicateurs de sortie.
4. Calibrage, exploration algorithmique ciblée par l’utilisation d’algorithmes spécifiques (*Calibration Profile* (Reuillon, Schmitt at al. 2015), *Pattern Space Exploration* (Chérel, Cottineau et Reuillon 2015)).
5. Retours sur le modèle, extension et nouvelles briques de multi-modélisation, retours sur les faits stylisés et la théorie.
6. Analyses de sensibilité étendues, correspondant à des méthodes expérimentales (en cours d'élaboration et d'intégration dans la plateformeOpenMOLE), 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.

Les premiers modèles de simulation en géographie ont d’abord été calculés « à la main », dans les années 1950. Ce n’est pas un hasard si ces modèles traitent tous de faits stylisés qui traduisent les régularités les plus fréquemment observées dans l’organisation de l’espace social, et qui sont des effets de la « première loi de la géographie » résumée ainsi dès 1970 par le cartographe américain Waldo Tobler : « tout interagit avec tout, mais deux choses proches ont plus de chances d’entrer en contact que deux choses éloignées ». La puissance de l’attrait pour la proximité apparaît dans tous les processus sociaux d’aménagement de l’espace social, qui sont contraints par «  l’obligation d’espacement ». Cette expression a été proposée par Henri Reymond dès 1971 dans une formalisation des problématiques de la géographie, qui posait en premier principe que les sociétés tendent à transformer l’*étendue* terrestre, hétérogène, rugueuse et discontinue, en un *espace* organisé présentant des propriétés de plus grande homogénéité et de continuité, et faisant émerger des régularités, du fait que deux objets ne peuvent occuper la même place. Dire que *les personnes et les sociétés ont la probabilité la plus grande de choisir d’occuper et d’interagir avec les localisations les plus proches, à la fois parce qu’elles sont mieux connues et parce qu’elles permettent de réaliser des économies sur les coûts (physiques, monétaires et culturels) de franchissement de la distance*, est certainement la plus forte proposition théorique de la géographie. Elle se repère dans toutes les configurations spatiales conduisant à distinguer un centre et une périphérie, qui se manifestent à tous les échelons de l’espace géographique, du local au mondial.

Les premiers modèles de simulation en géographie ont donc touché d’abord à des processus pour lesquels le choix du plus proche, parmi les lieux avec lesquels on souhaite établir une interaction, est une constante anthropologique très dominante, que ce soit pour observer les effets d’une innovation avant de l’imiter, selon la théorie spatiale de diffusion des innovations de Torsten Hägerstrand (1952) ou encore pour le choix des lieux de destination d’une migration (Hägerstrand 1957 ; Morrill 1962 et 1963). Les modèles s’appuyant sur la proposition, dès 1954, du géographe américain Edward Ullman de construire une géographie comme la science des interactions spatiales, notamment à propos des relations commerciales, ont d’abord surtout donné lieu à des expérimentations de modèles statistiques, selon diverses formes de « modèles gravitaires ». Ceux-ci ont par la suite été intégrés dans des modèles urbains, d’abord statiques (Lowry 1964) puis dynamiques (Clarke Wilson 1983, Wilson 2014, Allen Sanglier 1981 et Allen 2012).

Une génération postérieure de modèles jouant de manière plus complexe avec les effets de la proximité a beaucoup utilisé les automates cellulaires. Les mesures de l’autocorrélation spatiale, qui traduisent de manière positive ou négative les effets d’attraction ou de concurrence liés à la proximité sont ainsi employées pour tester la plausibilité des configurations simulées pour les changements d’utilisation du sol, et en particulier la croissance urbaine (White, Engelen 1993 ; White *et al*. 2015) ou encore la propagation des épidémies dans l’espace géographique (Cliff *et al*. 2004).

Mais le développement de ces modèles s’est heurté très tôt à l’obstacle des capacités de calcul des ordinateurs de l’époque, car la représentation explicite des interactions spatiales augmente comme le carré du nombre des unités géographiques considérées. Ainsi, le statisticien Christophe Terrier a dû segmenter son programme Mirabelle (Méthode Informatisée de Recherche et d'Analyse des Bassins par l'Etude des Liaisons Logement-Emploi) traitant les données du recensement de l’INSEE de 1975 avant de pouvoir simuler les découpages en bassins d’emploi des populations résidentes en fonction des navettes domicile-travail entre toutes les 36 000 communes françaises (Terrier, 1980). Le premier modèle de simulation des interactions entre des villes, destiné à reproduire leurs trajectoires démographiques et économiques influencées par les fonctions urbaines sur une période de 2000 ans sur un ordinateur de laboratoire, ne pouvait accepter qu’un maximum de 400 villes (Bura *et al*., 1996 ; Sanders *et al*., 1997). La montée en puissance des calculs informatiques a été relativement lente, autorisant seulement la considération d’environ un millier de villes en 2007 avec le modèle Eurosim (Sanders et al. 2007) ou les modèles Simpop2 appliqués par Anne Bretagnolle à l’Europe et aux Etats-Unis (Bretagnolle *et al* 2010). Surtout, l’expérimentation avec ces modèles est longtemps restée à un stade artisanal, requérant une grande application dans la modification « à la main » des valeurs des paramètres, qui étaient rarement directement observables, et qui devaient donc être estimées d’après la plausibilité des dynamiques du modèle. Or, les équations des modèles de dynamique urbaine intègrent des relations non linéaires qui produisent de nombreuses bifurcations, obligeant à de très nombreux allers-retours dans la procédure d’estimation des valeurs des paramètres (Sanders et al. 2007). Ce long travail a limité le nombre des simulations à partir duquel l’estimation obtenue pouvait être jugée comme satisfaisante, et surtout, une fois le modèle ainsi calibré, il demeurait une assez grande incertitude quant à la qualité des résultats obtenus.

## 3 Les simulations de nouvelle génération

Les années 2000 et 2010 devait modifier totalement l’environnement de travail des chercheurs, car la diffusion d’Internet puis des téléphones portables et enfin des données massives produites par toutes sortes de capteurs numériques a produit des effets en retour rapides et intenses sur la montée en puissance des capacités de calcul informatique qui elles-mêmes permettaient ces innovations technologiques disruptives. Les modèles de simulation ont pu alors intégrer des quantités considérables d’interactions entre des entités localisées caractérisées par une grande diversité d’attributs. Il y a une quinzaine d’années encore, (Gleyze 2005) était forcé de conclure que les analyses de réseau, pour les transports publics parisiens, étaient “limitées par le calcul”. Pour ne donner qu’un exemple du bond quantitatif représenté par l’accroissement des capacités de calcul et de ses conséquences sur la plus grande confiance accordée aux modèles qui en découle, on peut citer le travail pionnier en épidémiologie numérique réalisé par Eubank et al. (2004) pour simuler au moyen des modèles EpiSims et TRANSIMS les trajectoires quotidiennes, sur un réseau de transport, des déplacements d’un million et demi de personnes entre quelque 180 000 lieux de la ville virtuelle de Portland, afin de prédire les chemins de propagation d’une épidémie à partir des probabilités de rencontres interpersonnelles, dans des réseaux sociaux organisés en « petits mondes ». L’épidémie peut se propager rapidement dans une ville alors que les nombres de contacts par personne restent petits (quinze au maximum, Eliot et Daudé, 2006).

Des plateformes de simulation ont été mises au point pour que le plus grand nombre de chercheurs, même non spécialisés en informatique, puissent mettre en œuvre des modèles multi-agents. NetLogo (Tisue et Wilensky 2004) est sans doute la plus connue ; elle est généraliste et permet un accès aux simulations multi-agents sans besoin de connaissance informatiques approfondies, grâce à son langage de programmation simple et au constructeur d’interface graphique intégré. D’autres plateformes plus spécialisées, comme GAMA (Grignard *et al*. 2013), sont d’emblée construites pour proposer un couplage avec des systèmes d’information géographique. Cependant, la confiance dans les résultats issus des modèles de simulation va de pair avec une augmentation de la taille et du nombre de simulations requises, c’est-à-dire de l’ampleur des expériences numériques. Bien que des plateformes telles que NetLogo et GAMA intègrent des outils de base permettant un premier pas vers un tel « passage à l’échelle », un besoin de « méta-plateforme » dédiée à l’exploration des résultats de simulation a naturellement émergé.

**3.1 Un laboratoire virtuel : la plateforme OpenMOLE**

Depuis 2008 la plateforme OpenMOLE a été conçue pour explorer la dynamique des modèles multi-agents (Reuillon, Chuffart *et al*. 2010 ; Reuillon, Leclaire et Rey-Coyrehourcq 2013), et s’est rapidement étendue aux modèles de simulation plus généralement. Elle est issue du développement d’un précédent logiciel, SimExplorer (Amblard, 2003 ; Deffuant et al. 2003), qui offrait déjà à ses utilisateurs une interface ergonomique pour la conception de plans d’expérimentation et donnait accès au calcul distribué. OpenMOLE généralisait dans ses débuts SimExplorer en particulier en rendant possible la parallélisation massive de tâches. OpenMOLE[[2]](#footnote-2) est un outil de modélisation collaboratif en perpétuelle évolution : « Un effort permanent de généricité a permis de réaliser en quelques années une plateforme générique, pragmatique et éprouvée pour l’exploration de modèles de systèmes complexes sous forme d’un langage dédié, textuel et graphique, exposant des blocs cohérents et au bon niveau d’abstraction pour la conception d’expérimentations numériques distribuées sur des modèles de simulation » (Schmitt, 2014). 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.

Les procédures (ou *workflows*) proposées dans OpenMOLE sont décrites de manière indépendante des modèles et sont donc reproductibles, réutilisables et échangeables entre modélisateurs. Une place d’échange (*market place*) est intégrée au logiciel, à l’image de la librairie de modèles incluse dans NetLogo, et permet aux utilisateurs de bénéficier de scripts d’exploration pouvant servir de *template* ou d’exemple, dans des domaines thématiques très variés et pour l’ensemble des méthodes et langages implémentés dans OpenMOLE (par exemple pour le calibrage de modèles géographiques, l’analyse de réseaux biologiques, le traitement d’images pour les neurosciences).

OpenMOLE utilise un *Domain Specific Language* (DSL) (Van Deursen and Klint, 2002) pour l'écriture des *workflows* d’exploration. Cette pratique consiste en la création d’une notation et de règles spécifiques au domaine d’un problème donné. Il s’agit en quelque sorte d’un langage de programmation dédié, dans ce cas, à l’exploration des modèles et aux méthodes associées. Celui-ci n’est bien sûr pas créé de toutes pièces, mais vient comme une extension du langage sous-jacent, c’est-à-dire *Scala* dans le cas d’OpenMOLE. Un nombre réduit de mots-clés et de primitives rend son utilisation aisée même pour un utilisateur qui n’aurait aucune connaissance en programmation, et par ailleurs le DSL reste très flexible pour l’utilisateur avancé qui peut utiliser la programmation *Scala*. Selon Passerat-Palmbach et al. (2017), le DSL d’OpenMOLE est l’un des éléments clés de sa généricité et de son accessibilité.

Notons également que l’un des atouts principaux d’OpenMOLE est l'accès transparent aux environnements de calcul de haute performance (HPC). L’augmentation des moyens de calcul mentionnée précédemment 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 à 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 de celles-ci est cruciale) ; et (v) la définition de l’environnement de calcul sur lequel la méthode sera lancée.

La plateforme vise à considérablement étendre les pratiques de la *generative social science* proposée par Epstein et Axtell (1996), qui envisageait chaque modèle multi-agents comme une société artificielle, engendrant des comportements macroscopiques à partir d’hypothèses émises sur les comportements microscopiques. 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 plus 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 à 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 à 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 incorporé dans OpenMOLE), le tout rassemblé dans une salle, la plateforme de modélisation SimProcess [une appellation alternative du cadre dans lequel s'inscrit OpenMOLE] (Rey Coyrehourcq, 2015) ».

Par rapport aux protocoles généraux comme celui introduit par (Grimm *et al*. 2005) 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 rédhibitoire.

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 îles (populations évoluant indépendamment pendant une certaine durée) est particulièrement adapté à la distribution des calculs sur grille, chacun des nœuds de la grille faisant évoluer 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 pourrait relativement bien s’appliquer aux simulations de modèles stochastiques, même si un certain nombre de problèmes demeurent pour ce type d’application (Rakshit, Konar and Das, 2017). Selon Rey-Coyrehourcq (2015), ces méthodes se situent dans le cadre plus global de *l'Evolutionary Computation*, et la bibliothèque *scala* MGO développée simultanément à la plateforme et qui permet d’y implémenter les algorithmes évolutionnaires, a été conçue pour être facilement étendue à d’autres heuristiques en *Evolutionary Computation*, laissant totalement ouvertes la possibilité d’inclure de nouvelles méthodes dans OpenMOLE.

Selon R. Reuillon cité par Raimbault (2017a), la philosophie d’OpenMOLE s’articule autour de trois axes : le modèle comme “boîte noire” à explorer (i.e. méthodes d’exploration 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 des comportements possibles du modèle (Chérel, Cottineau et Reuillon, 2015)), tout cela permis par l’utilisation du calcul intensif (Schmitt et al., 2015). 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.

Afin d’illustrer cette présentation générale de la plateforme OpenMOLE et des méthodes associées, nous proposons dans la suite de cette section de développer précisément l’exemple du modèle SimpopLocal, dont la genèse a étroitement été liée à 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 sédentarisées (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 dès 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 à la sédentarisation de populations de chasseurs-cueilleurs, regroupées en hameaux et villages, puis à 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 à 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 uniquement 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 des interactions spatiales: les échanges commerciaux qui s’y effectuent à 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 stylisé majeur : déjà dans tout système, en tout lieu et à tout moment de l’histoire ou de la préhistoire, la répartition des tailles des lieux habités (mesurée 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 proportionnel à la taille acquise, et son amplification par la création de multiples innovations technologiques et sociétales 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 (Aghion *et al*. 1998). Mais SimpopLocal est conçu, comme les modèles précédents de la “famille” des modèles multi-agents Simpop (Bura et al 1996 ; Sanders et al. 2007 ; Pumain 2008), pour pallier l’insuffisante capacité du modèle de Gibrat à prévoir la tendance partout observée à la croissance plus forte qu’attendu des plus grandes villes situées en tête des réseaux (Moriconi-Ebrard 1993) et à 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 relation). Cette version plus parcimonieuse de la construction du modèle permet de réduire considérablement le nombre de paramètres et autorise donc une exploration et une évaluation plus systématiques.

**3.3 Implémentation de SimpopLocal, de NetLogo à OpenMOLE**

SimpopLocal a été initialement développé avec le langage NetLogo, puis redéveloppé 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 énormes pour l’époque historique qu’il était question 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 méso-Amérique 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 s’entend 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 surpasser 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 à environ 4000 ans la durée de la période de transition entre un système de peuplement agraire et un système de peuplement urbain, pour estimer un taux de variation moyen annuel de la population d’environ 0,02% par an et pour considérer que la taille du plus grand lieu habité du système allait passer d’une centaine à quelque 10 000 habitants. En revanche, les valeurs de cinq autres paramètres ne pouvaient pas être estimées d’après la littérature et devaient être déduites des expériences de simulation. Il s’agit de la probabilité de création d’une innovation par interaction entre deux personnes d’un même lieu, de la probabilité de diffusion d’une innovation par interaction entre deux personnes de lieux différents, de l’intensité de l’effet dissuasif de la distance sur ces interactions entre lieux, et de l’impact d’une innovation sur la croissance de la population (qui passe par un accroissement des ressources disponibles) et de la dimension maximale possible d’un lieu habité (mesurée en termes de population ou de ressources disponibles) qui intervient dans l’équation de croissance logistique retenue comme modèle générique d’une évolution encore très fortement contrainte par les ressources locales. Les équations qui résument le modèle et les tableaux définissant précisément les paramètres et leur action sont détaillés dans (Schmitt 2014) et (Schmitt, Rey Coyrehourcq in Pumain, Reuillon, 2017, pp 21-34).

On définit une valeur initiale pour la population et les ressources des lieux habités, puis le réseau d’interaction entre eux est créé. Ensuite, à chaque étape 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 étapes ou lorsqu’un nombre maximal arbitraire d'innovations a été atteint. On observe l’évolution de l’état du système de peuplement au niveau macro-géographique par la distribution de la taille des lieux habités, résumée 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 permettent de réaliser 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. La mise en œuvre de cette nouvelle forme d’expérimentation des modèles suppose toutefois une intervention du chercheur thématicien, qui doit sélectionner les objectifs précis que son modèle doit satisfaire. En contrepartie de ce qui pourrait être considéré comme une opération trop réductrice, 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, à 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 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 en œuvre 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 l’énorme charge de calcul et parvenir à optimiser une fonction de *fitness* susceptible 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 établis 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, et la durée de simulation doit équivaloir 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 *a minima* sur la représentativité et la plausibilité du comportement du modèle par rapport à l’ensemble envisageable des dynamiques des villes en système (à l’époque historique de l’émergence des villes). Le résultat en termes d’évaluation des simulations doit permettre de progresser dans la connaissance des processus d’interaction intra et interurbains 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 établi *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 étant stochastique, les résultats de la simulation varient d'une simulation à l'autre pour le même paramétrage. Par conséquent, l'évaluation du paramétrage en fonction des trois objectifs doit prendre en compte cette variabilité. Nous avons vérifié qu’une centaine de simulations pour chaque jeu de paramètres suffisait à saisir cette variabilité sans trop augmenter la durée du calcul.

A chaque fonction-objectif correspond une mesure de l’évaluation de la qualité du résultat simulé. La capacité du modèle à produire une distribution log-normale est mesurée par l’écart entre la distribution simulée et une distribution log-normale théorique ayant même moyenne et écart type selon un test de Kolmogorov-Smirnov. L’objectif de population maximale quantifie la capacité du modèle à engendrer des villes plus ou moins grandes, le résultat d'une simulation est testé en calculant l'écart entre la taille de la plus grande agglomération et la valeur attendue de 10 000 habitants : [(population de la plus grande agglomération −10 000) / 10 000 |. L’objectif de la durée de la simulation quantifie la capacité du modèle à engendrer une configuration attendue dans un laps de temps historiquement plausible. On calcule l’écart entre le nombre d’itérations de la simulation et la valeur attendue de 4000 étapes de la simulation: | (simulation durée −4000) / 4000 |. Ces trois calculs d’erreur sont normalisés afin de pouvoir comparer le degré de réussite d’une simulation vis-à-vis de chacun des trois objectifs. Mais l’agrégation des trois calculs qui produirait une seule mesure de qualité globale n’étant pas possible, un algorithme multi-objectif est nécessaire pour déterminer quelles simulations sont les plus satisfaisantes pour approcher la configuration finale souhaitée. Ce type d’algorithme calcule des solutions de compromis telles qu’aucune ne domine toutes les autres pour tous les objectifs. Ces solutions sont appelées des compromis de Pareto et elles forment ensemble ce qui est appelé un « front de Pareto ».

L’utilisation de méthodes d’exploration globales comme celle des algorithmes génétiques pour calibrer un modèle multi-agent (et en particulier un modèle multi-agent stochastique) implique un coût de calcul très élevé (Sharma *et al*, 2006). Ce type de charge est trop volumineux pour être exécuté sur des ordinateurs locaux, et les supercalculateurs sont très coûteux et ne sont pas facilement disponibles dans la plupart des laboratoires. Les grilles informatiques offrent une solution pour résoudre ces problèmes de calculs intensifs. Cependant, l’informatique à une si grande échelle suppose d’orchestrer l'exécution de dizaines de milliers d'instances du modèle sur des ordinateurs distribués dans le monde entier. La probabilité cumulée de pannes locales et le problème de répartir la charge de travail de façon optimale sur la grille rendent très difficile son utilisation pour un chercheur non spécialisé, comme précisé ci-dessus. C’est entre autres pour surmonter ces difficultés que la plate-forme OpenMOLE a été construite (Reuillon et al. 2010; 2013). Cet exemple de la calibration du modèle SimpopLocal montre bien dans quelle mesure OpenMOLE aide les modélisateurs à franchir le fossé technique et méthodologique qui les sépare 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 modèle ont pu être effectuées pour le calibrage de SimpopLocal, lequel sans cela aurait requis quelque 20 années de calcul avec un seul ordinateur.

*Le profil de calibrage, un grand saut épistémologique pour les SHS*

Le résultat du processus de calibrage assure seulement que le modèle *peut* reproduire les caractéristiques stylisées de l’émergence d’un système de villes, avec une évaluation assez précise des valeurs des paramètres qui toutes ensemble contribuent à assurer cette évolution. Mais il ne dit rien de la fréquence à laquelle les jeux de paramètres produisent des comportements plausibles, et de quelle façon chaque paramètre contribue à modifier le comportement du modèle. Il serait intéressant par exemple de savoir à quel moment certaines valeurs de paramètre empêchent le système d’atteindre un comportement plausible, et de ne pas se restreindre à ne connaître qu'un seul jeu de valeurs de paramètres «optimales».

Une nouvelle méthode a été mise au point pour représenter la sensibilité du modèle aux variations d'un seul paramètre, indépendamment des variations de tous les autres paramètres (Reuillon *et al*. 2015). Au moyen d’une fonction qui calcule une seule valeur numérique décrivant la qualité du calibrage pour le modèle, l’algorithme de profil calcule l’erreur de calibrage la plus faible possible lorsque la valeur d'un paramètre donné est fixée et que les autres sont libres. L’algorithme calcule cette erreur minimale pour tout le domaine de variation du paramètre étudié. Pour chaque valeur d’un paramètre, l’algorithme cherche à identifier les jeux de valeurs des autres paramètres qui produisent le meilleur ajustement du modèle aux données attendues (la plus petite erreur possible). Un graphique représente alors les variations de cette valeur d’ajustement optimale en fonction des variations du paramètre étudié. Le profil de calibrage montre ainsi plusieurs formes possibles pour cette courbe. Lorsqu’elle présente 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 paramètre étudié, on peut en conclure qu’on a vraiment identifié l’ordre de grandeur du paramètre qui satisfait aux exigences en termes de comportement du modèle. Si l’une de ces courbes reste plate, cela indique que le paramètre n’a pas d’effet sur le comportement du modèle et peut donc en être éliminé. Ainsi, dans le cas de SimpopLocal, un paramètre imaginé comme la durée de vie d’une innovation a été finalement exclu car des variations restaient sans effet sur la qualité d’ajustement du modèle, toutes choses égales quant aux variations des autres paramètres (Schmitt, 2014). On a donc ici la possibilité d’évaluer jusqu’à quel point les *mécanismes imaginés pour construire le modèle sont non seulement suffisants, mais aussi nécessaires* pour produire le comportement attendu. Dans les limites du cadre théorique et de la sélection des faits stylisés retenus, c’est la première fois que des chercheurs en SHS peuvent parvenir à ce type de conclusion scientifique essentielle, grâce à une *méthode de validation enfin efficace pour les modèles de simulation multi-agents*. C’est un immense progrès du point de vue épistémologique en sciences sociales – certes toujours relatif au cadre théorique donné par les objets, attributs et mécanismes sélectionnés par les chercheurs pour être représentatifs du système observé.

Une forme complémentaire de validation du modèle pourrait être alors imaginée si des historiens archéologues tentaient de le recalibrer avec des données de leurs observations. En effet, le jeu de paramètres estimé contient des valeurs qui engendrent bien la dynamique voulue pour un système de peuplement mais qui ne sont pas fixées dans l’absolu. Elles sont relatives les unes aux autres d’une part et aux données fictives introduites d’autre part. Si l’on modifie ces dernières pour les rendre compatibles avec un système de peuplement historiquement observé, la capacité du modèle à simuler son développement serait alors confirmée, non seulement en reconstruisant les trajectoires de l’évolution de la population des lieux habités considérés, mais aussi en conservant les ordres de grandeur relatifs des paramètres qui engendrent cette dynamique.

## 4 Autres exemples d’applications d’OpenMOLE : modèles d’interaction réseaux-territoires

Nous proposons dans cette section d’illustrer l’application des méthodes d’exploration d’OpenMOLE et du calcul intensif à une autre question thématique, celle des interactions entre réseaux et territoires. Cette question a alimenté de nombreux débats scientifiques, pour lesquels la plupart des problèmes restent relativement ouverts. Par exemple, le problème des “effets structurants des infrastructures de transport” (Bonnafous et Plassard 1974), présenté par (Offner 1993) comme un “mythe scientifique” invoqué pour justifier le coût d’une nouvelle infrastructure par ses retombées sur le développement régional, pas toujours observées à moyen terme, peut selon A. Bretagnolle (dans Offner *et al*. 2014) être observé pour des territoires plus vastes et sur le temps long, tout en tenant compte des fluctuations locales dans les dynamiques des systèmes de villes. Le débat est ainsi alimenté par des résultats contradictoires selon les lieux et les périodes étudiés, et aussi selon les décalages temporels admis pour mesurer une relation entre une « cause » et son « effet ». La difficulté empirique d’extraire des faits stylisés généraux ainsi que la difficulté conceptuelle d'entités géographiques qui sont en fait en relations de causalités circulaires, sont contournées par l’approche de modélisation de la coévolution des réseaux de transport et des territoires proposée par Raimbault (2018b). Les résultats obtenus sont étroitement liés à l’utilisation d’OpenMOLE et de ses algorithmes d’exploration et de calibrage, dont nous allons donner quelques illustrations.

L’application d’un calibrage multi-objectif s'avère essentielle pour expérimenter des modèles de systèmes de villes correspondant à des situations réelles. Par exemple, (Raimbault, 2018a) introduit un modèle d'évolution d’un système de villes sur le temps long, proche du modèle de (Favaro et Pumain, 2011), mais se concentrant sur l’effet du réseau de transport physique. Les taux de croissance des villes sont déterminés par la superposition de plusieurs effets : (i) croissance endogène capturée par un taux de croissance fixe correspondant au modèle de Gibrat ; (ii) interactions entre villes par un modèle gravitaire ; (iii) rétroaction des flux circulant dans le réseau sur les villes traversées. Ce modèle est calibré de manière non stationnaire dans le temps (c’est-à-dire sur des fenêtres temporelles glissantes, afin de prendre en compte le changement de nature des dynamiques urbaines, comme observé par Bretagnolle et Franc (2018) avec par exemple les mutations des réseaux de transport), sur le système de villes français entre 1830 et 2000. Pour calibrer le modèle, les populations simulées sont comparées aux populations observées. À ce stade, l’utilisation d’un algorithme de calibrage multi-objectif (l’algorithme NSGA2 implémenté dans OpenMOLE) est essentielle. En effet, l’ajustement peut par exemple s’effectuer sur une erreur carrée moyenne dans le temps et pour l’ensemble des villes. Cependant, vu les disparités de taille des villes liées à la hiérarchie urbaine, il émerge rapidement qu’une optimisation mono-objectif sur cette erreur reviendra à ajuster la taille des plus grandes villes, au détriment de la majorité des villes du système. L’ajout d’un second objectif, pris par exemple comme une erreur carrée moyenne sur les logarithmes des populations, permet de prendre celles-ci en compte. Un résultat important de (Raimbault, 2018a) est alors l'émergence de fronts de Pareto pour ces deux objectifs, pour l’ensemble des fenêtres temporelles considérées. Cela montre que ce type de modèle doit être appliqué en faisant un compromis entre l’ajustement des populations pour les villes moyennes et des populations pour les plus grandes villes. Ce résultat est permis grâce à l’optimisation multiobjectif par algorithme génétique d’OpenMOLE.

Un autre exemple d’application des méthodes de la plateforme qui illustre son rôle crucial est donné par la recherche de régimes de coévolution, c’est-à-dire de différentes formes de corrélation dans le temps entre le développement d’un réseau et la croissance d’une ville ou d’une région. Si la corrélation garde le même signe, cela signifie que la croissance du territoire peut précéder et donc causer celle du réseau, ou inversement, alors que si elle varie dans le temps, ses changements permettent de détecter des causalités circulaires et donc une coévolution entre réseau et territoire. Suivant Raimbault (2017b), l'étude des motifs de corrélation retardée dans le temps permet d’isoler des régimes typiques d’interaction entre variables de réseau et variables de territoires. Plus précisément, Raimbault (2018b) définit la coévolution comme l’existence de relations circulaires causales, au niveau d’un ensemble d’entités dans une certaine emprise spatiale. Dans le cas des réseaux et des territoires, les propriétés des réseaux doivent être localement causées par celles des territoires, et réciproquement. Des causalités unidirectionnelles des réseaux vers les territoires correspondent alors aux “effets structurants” mentionnés ci-dessus. Cette définition de la coévolution permet de capturer la “congruence” (Offner 1993) entre ces objets, en quelque sorte leur adaptation réciproque de manière dynamique. Elle permet aussi la construction d’une méthode opérationnelle proposée par Raimbault (2017b), qui cherche statistiquement des liens de causalité entre variables correspondantes. En pratique, la notion faible de *causalité de Granger[[3]](#footnote-3)* est mobilisée, permettant une flexibilité au regard des données nécessaires et du cadre temporel et spatial d’estimation. Cette causalité est quantifiée par les corrélations retardées entre variations des variables de réseau (comme les centralités ou l'accessibilité) et variations des variables de territoires (comme la population, l’emploi, les transactions immobilières, etc.). L’existence de maxima significatifs à des retards non nuls (corrélation plus forte entre les valeurs de l’une et l’autre croissance avec un décalage dans le temps) donne une direction de causalité (la croissance d’une variable de réseau précède et donc entraîne celle d’une variable de territoire, ou inversement, pour la région et la période considérées). Une typologie de ces profils de corrélations retardées fournit ce qu’on nomme des “régimes de causalité”, parmi lesquels des régimes de coévolution où deux variables « territoire » et « réseau » sont en causalité réciproque.

La question est alors, pour un cas d'étude donné, d’identifier les régimes présents à partir de données observées ou de données simulées par un modèle, et notamment ceux qui correspondent à une coévolution. La démonstration de l’existence de tels régimes en sortie d’un “modèle de coévolution” 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. En effet, les modèles considérés sont complexes et témoignent d’une émergence. La méthode des corrélations retardées est appliquée à un modèle macroscopique de coévolution par Raimbault (2019a), qui étend 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 valeurs ou « 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 coévolution. Cette expérience 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% (dans ce cas on considère comme variable de territoire les populations, et comme variables de réseau la centralité de proximité et l'accessibilité, ce qui correspond à six couples dirigés de variables, et donc 3^6=729 régimes possibles, chaque couple pouvant présenter une corrélation retardée positive, négative, ou inexistante). On trouve parmi les 33 régimes isolés 19 régimes de coévolution, dont l’existence ne pouvait pas être intuitivement prédite. L’existence et la variété de ces régimes est un résultat important, montrant qu’il est possible de modéliser une coévolution, au sens statistique donné précédemment.

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 étendre cette conclusion, puisque celui-ci produit 260 régimes de coévolution (35,7%). Il s’agit d’un exemple typique où la forte non-linéarité des sorties considérées peut mener à des conclusions partielles voire biaisées et où l’utilisation d’une méthode spécifique est cruciale. Les résultats sont rendus plus robustes et étendus, grâce à l’application d’une méthode spécifique intégrée à la plateforme OpenMOLE.

Cette méthode permet par ailleurs de comparer entre eux des modèles. Ainsi, Raimbault (2019b) a appliqué cette méthode au modèle SimpopNet introduit par Schmitt (2014), qui est également un modèle de coévolution à l'échelle macroscopique et qui présente un grand nombre de points communs avec le modèle de coévolution de Raimbault (2019a), notamment dans les variables considérées et donc les indicateurs de sortie calculables. On obtient alors avec SimpopNet un nombre plus faible de régimes d’interaction et de régimes de coévolution, confirmant d’une part qu’il n’est pas immédiat pour un modèle conçu pour la coévolution (SimpopNet) de faire effectivement émerger des régimes de coévolution, et suggérant par ailleurs que ses 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 permis la création d’un axe, voire d’un domaine de recherche original, 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 (à l’image de la spirale vertueuse de Banos (2013) entre disciplinarité et interdisciplinarité). De plus, la philosophie de plateforme unique (évoquée ci-dessus, par l’interaction forte entre les trois axes d’embarquement des modèles, d'accès à des méthodes d’exploration innovantes, et d'accès transparent aux environnements de calcul intensifs), ouvre des perspectives nombreuses, tant sur le plan technique que 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

L’intégration de nouvelles méthodes d’exploration est un axe privilégié du 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'équifinalité (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 (*Pattern Space Exploration*) est actuellement en cours d'élaboration pour une intégration dans OpenMOLE.

L’utilisation de méthodes d'inférence Bayésiennes est une autre piste en cours d’exploration. En effet, dans le cas de modèles fortement stochastiques 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 ces données. Il s’agit ainsi d’un calibrage étendu, 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 encore diverses directions méthodologiques qui sont aussi 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 des fonctions exponentielles du nombre de dimensions - la réduction du nombre de configurations simulées soumises à l’algorithme PSE via l’imposition d’un critère de diversité des configurations retenues permettrait de résoudre ce problème et de 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 à l'étude, comme le critère POMIC proposé par Piou et al. (2009).

### 5.2 Outils

Au cours 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 dès les premières versions d’OpenMOLE, 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 à tout système d’exploitation et tout type de matériel informatique, ce qui est crucial pour la distribution des calculs sur des nœuds de grille hétérogènes. Par exemple, des propriétés comme le mixage de trait 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 class* 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 *container*, permet d’embarquer un environnement d'exécution à l’identique quel que soit le système d’exploitation et le matériel informatique. Hung *et al.* (2016) proposent 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 (Singularity), toutefois, certaines tâches reposant sur l'exécution de binaires ou de programme à l’environnement compliqué 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 généricité 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 l’utilisation au sein de communautés de thématiciens.

## Conclusion

L’exploration des modèles de simulation s’est poursuivie en géographie par l'intermédiaire d’initiatives telles que le développement de la plateforme OpenMOLE. Ce développement a été mené dans un cadre hautement interdisciplinaire et réciproque (relation gagnant-gagnant entre informaticiens et géographes (Pumain, 2019), 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 Geodivercity, témoigne d’une nouvelle façon de produire des connaissances géographiques, rendant envisageable la production de preuves scientifiques en sciences sociales. Elle donne ainsi des arguments de poids à la géographie et, plus largement, aux sciences humaines et sociales, face aux sciences dures telles que la physique, qui prétendent à un monopole quant à l'identification de preuves relatives au fonctionnement des systèmes sociaux (Dupuy and Benguigui, 2015). Il reste maintenant à promouvoir cette posture de recherche, ainsi que les outils et méthodes qui la rendent possible, en complémentarité avec les nouvelles disciplines émergentes de *City Science* et *Urban Analytics* décrites par Batty (2019) au sein du champ de la Géographie Théorique et Quantitative.

1. Délégation à l’Aménagement du Territoire et à l’Action Régionale [↑](#footnote-ref-1)
2. <https://openmole.org/> [↑](#footnote-ref-2)
3. La causalité faible au sens de Granger se déduit de corrélations mesurées sur des séries temporelles, on fait l’hypothèse d’une telle causalité lorsque une variable offre des résultats utiles pour prédire la valeur d’une autre variable. [↑](#footnote-ref-3)