Numérique *versus* symbolique

Dialogue ontologique entre deux approches

Hélène Mathian[1], **Lena Sanders**[2]

1. UMR 5600 Environnement Ville Société ; 2. UMR 8504 Géographie-cités

RÉSUMÉ. L'objectif de cet article est de comparer une approche statistique, l'analyse des données (AD) et une approche de simulation, les systèmes multi-agents (SMA). Ces deux familles de méthodes sont a priori *considérées comme représentatives d'une approche numérique, respectivement symbolique, de la modélisation spatiale. Le cas d'application qui est mobilisé tout au long de l'article est celui de la ségrégation de l'espace scolaire en Île-de-France. En premier lieu sont explicitées et discutées les différentes étapes menant d'une question thématique à l'opérationnalisation d'une méthodologie d'analyse statistique ou de simulation destinée à analyser cette question. Pour effectuer cette comparaison, on développe un cadre conceptuel à l'interface entre les deux, qui permet de vérifier la compatibilité entre les arrières plans théoriques associés aux domaines thématiques et de modélisation en jeu. Ce cadre conceptuel prend appui sur une démarche ontologique qui est ensuite présentée. Celle-ci permet d'identifier les complémentarités entre AD et SMA et de montrer comment ces deux méthodes peuvent dialoguer dans le cadre d'une même recherche. Nous montrons combien les aspects numériques et symboliques sont finalement étroitement imbriqués au sein même de chacune de ces méthodes. Cette imbrication permet de construire une « spirale d'interactions » entre les deux familles de méthodes dont l'intérêt est illustré par les va et vient entre les phases d'analyse de structure et de simulation dynamique dans le cas de la ségrégation scolaire.*

ABSTRACT. The aim of this article is to compare a statistical approach, "geometric data analysis" (GDA), and a simulation approach, the multi-agent systems (MAS), considered as representative, respectively, of a numerical and a symbolic approach of modelling. The case study concerns segregation of scholar space in the Parisian area. First the different steps leading from a thematic question to the development of an operational model to analyze this question are presented. The central and essential role of a conceptual framework at the interface of both is shown. Indeed, before operationalisation, it is necessary to verify the compatibility between the theoretical backgrounds associated to the thematic hypotheses and the model considered. An ontological approach is then presented and used to compare GDA and MAS in order to identify their complementarities and show how these approaches can dialogue in the same research. The close interweaving between numerical and symbolic aspects in each of these approaches is shown. This leads to the construction of a "spiral of interactions" between GDA and MAS which interest is illustrated by the back and forth between modeling structure and dynamics in the case of scholar segregation.

MOTS-CLÉS : ontologie, numérique, symbolique, analyse des données, système multi-agents, ségrégation scolaire.

KEYWORDS: Ontology, Numerical, Symbolic, Data analysis, Multi-agent system, School segregation.

DOI:10.3166/RIG.31.21-45 © 2022 Lavoisier

1. Introduction

L'objectif commun des articles de ce numéro spécial est de mettre en évidence les formes de complémentarité entre les approches symboliques et numériques. Dans cet article, cet objectif prend la forme d'une comparaison entre les hypothèses, prérequis et production de connaissances apportés par une approche statistique (l'analyse des données) et une approche modélisatrice à base d'agents (les systèmes multi-agents), en prenant appui sur une approche ontologique. Dans les deux cas il s'agit de *représenter/modéliser* un phénomène d'intérêt pour décrire comment il est structuré et comprendre les processus qui sous-tendent son organisation dans l'espace. La comparaison de l'analyse des données (AD) et des systèmes multi-agents (SMA) est particulièrement stimulante dans la mesure où chacune de ces familles de méthodes a créé, lors de son introduction et diffusion au sein de la recherche en sciences sociales, un fort engouement et quasiment amené un changement de paradigme dans la modélisation dans ces domaines. Dans les années 1970, l'AD a été adoptée avec enthousiasme car elle permettait, grâce à l'ordinateur, de traiter des données abondantes issues d'enquêtes et d'observations de terrain, de mettre en évidence des régularités et des structures, sans que des hypothèses *a priori* sur les distributions statistiques de ces données ne soient imposées. Dans les années 1990, ce sont les SMA qui ont rapidement diffusé, leur succès étant dû à la possibilité d'explorer par la simulation les processus sous-jacents aux phénomènes étudiés. Ils sont fondés sur un formalisme à base de règles et permettent d'approcher les moteurs d'un changement socio-spatial. Ces deux familles de méthodes ont donc des objectifs bien différenciés, et de ce fait, leur complémentarité peut paraître d'emblée évidente. En revanche, pour évaluer dans quelle mesure ces deux approches peuvent dialoguer et se compléter sur un même cas d'étude, une comparaison de leurs fondements épistémologiques est nécessaire. Suivant la compatibilité de ces fondements, il sera en effet plus ou moins aisé de cheminer d'une approche à l'autre. La démarche la plus simple consisterait à simplement juxtaposer les résultats de chacune d'entre elles lors d'une étape finale. Une démarche plus ambitieuse repose sur un va et vient entre ces deux approches tout au long des étapes du travail de recherche. Une compatibilité épistémologique forte est alors nécessaire pour que ce va et vient prenne sens à chaque niveau où il est opéré, qu'il s'agisse, en amont, de celui des entités élémentaires en entrée de la modélisation, ou de celui, en aval, auquel les résultats de la modélisation peuvent être interprétés. A chacun de ces niveaux, ainsi qu'aux niveaux intermédiaires, nous montrons que les deux qualités « symboliques » et « numériques » interfèrent.

Pour comparer ces deux approches nous commençons par discuter des caractères numérique et symbolique de chacune d'entre elles en prenant appui sur un cas d'application, celui de la ségrégation de l'espace scolaire en Île-de-France, cas qui est mobilisé tout au long de cet article (section 1). Dans un deuxième temps, il s'agit de montrer comment sont articulées les différentes étapes constituant la chaîne de traitements à effectuer face à un questionnement thématique. A une extrémité de cette chaîne se trouvent les données et les connaissances empiriques, à l'autre,

l'opérationnalisation d'une méthodologie d'analyse statistique ou de simulation. Nous montrons notamment l'intérêt d'expliciter le cadre conceptuel de modélisation qui permet de joindre ces deux extrémités, et évoquons pour cela la place d'une démarche ontologique. L'ontologie utilisée, proposée par Pierre Livet dans la continuité des travaux de Barry Smith (2003), est ensuite présentée. Elle a été développée pour expliciter quels sont les éléments nécessaires pour étudier un phénomène spatiotemporel et modéliser son évolution, en s'assurant de la bonne compatibilité entre les concepts, les données empiriques et le cadre méthodologique choisi. Cette ontologie est utilisée ici pour comparer les deux familles de méthodes mobilisées, l'analyse de données en statistiques et les systèmes multi-agents dans le domaine de la simulation informatique (section 2). La dernière étape consiste à examiner les formes de complémentarité et de dialogue possibles entre AD et SMA pour construire de nouvelles connaissances. Le cas de la ségrégation de l'espace scolaire donnera lieu au développement de différents cadrages ontologiques et servira à illustrer comment une « spirale d'interactions » peut être construite entre les différents niveaux associés aux deux approches mobilisées (section 3).

1. Symbolique *versus* numérique, de quoi parle-t-on ?

Une première étape consiste à justifier en quoi nous considérons *a priori* ces deux familles de méthodes comme des représentantes respectives des approches numérique et symbolique, afin de pouvoir évaluer, ensuite, les implications de cette différence. D'après les définitions du CNRTL[1], le numérique « concerne des nombres, se présente sous la forme de nombres ou de chiffres, ou concerne des opérations sur des nombres » alors que le symbolique « utilise des symboles, procède par symboles ». Dans son article « Introduction à l'approche symbolique en analyse des données », Edwin Diday (1989) précise ces notions à propos des « objets » :

> « La distinction entre objets "symboliques" et "numériques" est claire dès lors que l'on considère qu'un objet est "numérique" s'il peut être représenté et utilisé comme un point de l'espace Rp considéré comme un espace vectoriel muni des opérations habituelles et qu'il est "symbolique" si ce n'est pas le cas (autrement dit, s'il est nécessaire de définir une sémantique propre au domaine d'application car la sémantique des nombres ne convient pas). »

Son propos concerne l'analyse des données, méthode classiquement utilisée pour traiter des données numériques et qu'il propose d'étendre au traitement de données plus complexes. Il évoque ainsi des objets décrits par des données composées non de simples modalités associées à des variables quantitatives ou qualitatives, mais de ce qu'il nomme une « conjonction de propriétés ». Tel est le cas, par exemple, lorsqu'une variable se présente sous la forme d'un intervalle de valeurs ou d'un

1. CNRTL (Centre national de ressources textuelles et lexicales) : www.cnrtl.fr

ensemble de modalités jointes par la conjonction « ou ». Dans l'application qui est développée par la suite, à savoir la ségrégation scolaire, la variable « établissement fréquenté par l'élève » est de type simple, qualitative, et peut donner lieu à des comptages classiques. En revanche, la variable « établissement préféré par l'élève » qui est primordiale dans la dynamique de choix des établissements, peut s'exprimer sous la forme {collège Picasso ou collège Ravel}, introduisant une information de forme plus complexe.

L'étude de la ségrégation scolaire est propice pour mobiliser à la fois des analyses de nature numérique et symbolique et les données dont nous disposons relèvent également de ces deux types :

– des données issues de la DEPP (Direction de l'évaluation, de la prospective et de la performance du ministère de l'Éducation nationale) portant sur les élèves des collèges de l'Île-de-France, notamment l'établissement fréquenté par l'enfant et la catégorie sociale de ses parents. Ces données sont fondamentalement numériques, même si les catégories initiales concernées sont qualitatives. Leur croisement permet par exemple de construire un tableau de contingence[2], cas le plus classique d'application de l'analyse des correspondances (AFC).

– des informations et des connaissances sur le comportement des élèves et des chefs d'établissements, en matière de choix d'établissement pour les premiers, de stratégies de recrutement des élèves pour les seconds. Ces informations, construites à partir d'entretiens (Poupeau et François, 2008), permettent par exemple de comprendre la forme des préférences des uns et des autres, informations qui peuvent être formalisées sous forme symbolique par des règles.

La nature des données étant précisée, il s'agit de s'interroger sur la forme numérique ou symbolique des traitements qui leur sont appliqués ainsi que sur celle dans laquelle sont exprimés les résultats obtenus. Diday propose ainsi de distinguer quatre cas, reposant sur le croisement des propriétés des données et de celles des traitements. Le cas le plus classique consiste à appliquer une analyse numérique sur des données numériques, cas typique des analyses factorielles (Diday, 1989 ; Balbo *et al.*, 2009). En AD, les opérations classiquement effectuées sur un tableau de contingence reposent ainsi sur des calculs de fréquences (en lignes ou en colonnes) ou de fonctions de fréquences (pour calculer, par exemple, la similarité entre les profils de lignes ou de colonnes d'un tableau de contingence). Les sorties sont sous la forme de valeurs numériques, comme par exemple les coordonnées sur les axes factoriels. Dans ce cas, les entrées, les opérations et les sorties sont toutes sous forme numérique.

Dans notre contribution, l'approche symbolique considérée est la simulation avec systèmes multi-agents (SMA), approche pour laquelle tant les données que les analyses qui en sont faites, peuvent prendre une forme relativement complexe, non résumable sous une forme numérique. Dans cette approche, en effet, les objets

2. A l'intersection de chaque ligne *i* et colonne *j* du tableau figure le nombre d'élèves de la catégorie sociale *j* fréquentant l'établissement *i.*

considérés sont localisés dans l'espace et des actions conduisant à des changements sont simulées à chaque pas de temps au niveau des objets, en fonction des caractéristiques de son environnement et des échanges qui s'opèrent avec des objets situés en d'autres lieux. L'expression de tels mécanismes nécessite un formalisme symbolique. Les principales opérations se présentent sous forme de règles qui déterminent les actions d'agents, en fonction de leur situation initiale, de leurs propriétés et des interactions qu'ont ces agents entre eux et avec leur environnement. Ces règles s'expriment souvent sous la forme symbolique « Si… Alors… Sinon... » qui permet par exemple de traduire qu'un élève de telle catégorie aura une préférence pour un établissement ayant un certain profil social. Quant aux sorties, elles peuvent être exprimées sous forme symbolique (tel élève résidant dans tel contexte et préférant tel établissement est affecté à tel autre) ou numériques (nombre d'élèves de telle catégorie sociale par établissement).

Le caractère *a priori* numérique *versus* symbolique de ces deux approches étant établi, l'objectif est d'en comparer les fondements épistémologiques en prenant appui sur une approche ontologique.

2. Un cadre ontologique pour comparer AD et SMA

Avant de mettre en œuvre la comparaison entre ces deux approches, nous explicitons l'ensemble des étapes qui rythment tout travail de modélisation, qu'il soit statistique ou informatique, en mobilisant la démarche proposée par Denis Phan (2014) qui consiste à expliciter le cadre conceptuel permettant de mettre en lien le domaine thématique et celui de la modélisation. Ce cadre conceptuel qui varie suivant les questionnements et les méthodes choisis est en effet souvent implicite. L'enjeu consiste à l'expliciter pour l'AD et les SMA. L'ontologie occupe une place centrale dans ce cadre conceptuel où elle joue un rôle de « médiation ». Nous en proposons donc une définition que nous appliquons ensuite pour comparer les approches par AD et par SMA.

2.1. Un cadre conceptuel pour lier un phénomène thématique à sa modélisation

Si on se situe du point de vue du thématicien, le point de départ d'une recherche est une question thématique. Il s'agit dans notre cas de comprendre comment émerge et s'organise une ségrégation socio-spatiale dans le domaine scolaire. Le choix d'une méthode est ensuite opéré en fonction des savoir-faire et habitudes des chercheurs, de la nature des données disponibles et des particularités de la question posée. La mise en lien de la question, des données et de la méthode peut être réalisée de différentes façons et l'absence d'explicitation des choix qui sont faits peut être source de malentendus dans un contexte où des chercheurs de disciplines différentes collaborent. La figure 1, adaptée de Phan (2014), représente les positionnements respectifs des différents domaines impliqués dans une démarche modélisatrice : à un

bout est représenté le « domaine de référence empirique » au sein duquel est formulée la question de recherche, à l'autre celui relatif aux méthodes de traitements des données associées au domaine d'intérêt, qu'elles relèvent des statistiques ou de l'informatique. Le rectangle central est le lieu du dialogue interdisciplinaire et de la mise en cohérence entre le « domaine de référence empirique » ancré dans une thématique précise et le « cadre opérationnel de développement » qui fait référence aux aspects d'opérationnalisation et d'implémentation informatique des traitements.

Le « domaine de référence empirique » (à l'extrémité gauche de la figure 1) est celui du questionnement du chercheur sur un aspect du monde empirique. Il peut s'agir d'un phénomène spatiotemporel, social ou socio-écologique. Le cas qui servira de référence dans cet article concerne l'organisation de l'espace scolaire francilien : cet espace est-il différencié ? Suivant quelle géographie ? Comment émergent ces différenciations ? Quels en sont les mécanismes producteurs ? Tels sont des exemples de questions que se posent les géographes. La nature des questions posées par les chercheurs est influencée par leur formation disciplinaire et celles posées par des sociologues, des économistes ou des psychologues seront différentes, bien que portant sur un même objet, l'espace scolaire. Ce « domaine de référence empirique » est ainsi fortement ancré dans un point de vue particulier (Müller et Diallo, 2012) qui influe sur la manière de construire le « cadre conceptuel thématique » figurant en partie centrale du schéma. Il ne s'agit pas de saisir le phénomène étudié dans toute sa « réalité » mais de mettre en œuvre une approche permettant de cibler les questions posées suivant le point de vue adopté et la nature des données disponibles.

Le « cadre opérationnel de développement » (à l'extrémité droite de la figure 1) correspond à la mise en œuvre de la démarche de modélisation mise au point dans le cadre central avec les outils appropriés, qu'ils soient informatiques, mathématiques ou statistiques. Il s'agit ici d'une part, de l'AD, dans l'esprit où elle a été développée par J.-P. Benzécri, et d'autre part, d'une approche de simulation par SMA. Ces deux approches émanent de cultures différentes, et si elles sont combinées dans certains travaux, c'est en en général la première au service de la seconde (que ce soit pour traiter des données en amont d'une modélisation SMA ou les données issues d'un tel modèle), et elles sont rarement mises en perspective face à un même objectif. Il est de ce fait intéressant de tester le rôle de médiation que peut jouer une approche ontologique pour les faire dialoguer. Chacune de ces approches méthodologiques peut en effet être utilisée pour répondre aux questions posées dans le « domaine de référence empirique » et prend elle-même appui sur un champ théorique bien défini qui réfère au « cadre conceptuel de modélisation » (à droite dans le rectangle central). Il s'agit alors de s'assurer de la compatibilité entre les grandes lignes (contraintes, présupposés) associées à ce champ théorique de modélisation et celles associées au champ disciplinaire correspondant au domaine de référence empirique. Tel est le rôle du « cadre conceptuel » développé en partie centrale de la figure 1.

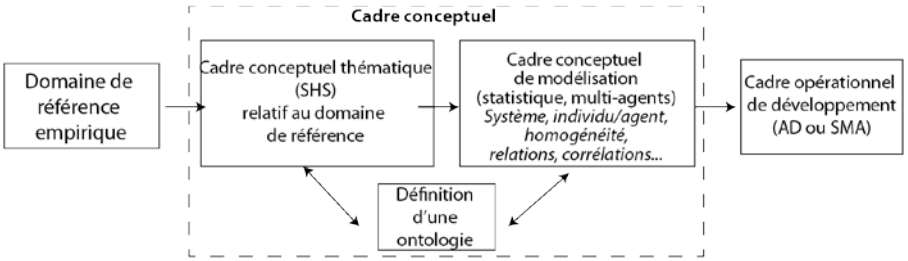

Figure 1. L'ontologie comme médiateur d'un dialogue conceptuel (d'après Phan, 2014)

Ce « cadre conceptuel » central met ainsi en relation les cadres conceptuels attachés respectivement au champ thématique associé au « domaine de référence empirique » étudié et au champ de modélisation qui sera opérationnalisé lors de l'application (respectivement l'AD et les SMA). Le cadre conceptuel thématique concerne la manière dont l'espace scolaire est appréhendé. Les questions formulées plus haut renvoient à des différenciations socio-spatiales qu'il s'agit de mettre en évidence et d'en comprendre l'émergence. Cette question est abordée différemment suivant le positionnement épistémologique des chercheurs. François *et al.* (2014) ont ainsi comparé, face à cette même question, les approches de différents chercheurs, les uns prenant appui sur l'individualisme méthodologique et prônant une « rationalité intentionnelle » des élèves comme des chefs d'établissements dans leurs choix respectifs, les autres mettant en avant l'effet des structures sociales sur les choix des élèves et chefs d'établissement. En sciences sociales, ces deux visions correspondent respectivement à celles issues des travaux de Boudon (1973) et de Bourdieu (1979). Les facteurs explicatifs, qui seront ensuite conceptualisés dans le « cadre conceptuel de modélisation », vont ainsi différer suivant le cadre conceptuel thématique adopté. Dans le premier cas, le comportement des agents, élèves comme chefs d'établissement, renvoie à un modèle décisionnel de type micro-économique alors que dans le second ces comportements dépendent des structures spatiales et sociales (François *et al*., 2014). Avant d'analyser comment ces deux positions épistémologiques différentes peuvent s'insérer dans une modélisation par AD ou SMA, il est nécessaire d'examiner à leur tour les implications théoriques de chacune de ces familles de modèles.

Le « cadre conceptuel de modélisation » associé aux SMA implique une formalisation où les concepts de système complexe, d'émergence, d'agents et d'interactions sont centraux. Les entités élémentaires modélisées sont ainsi des agents qui interagissent entre eux et avec leur environnement, ces interactions entraînant l'émergence d'une structure observable à un niveau d'organisation supérieur, que ce soit celui des établissements ou de l'espace géographique. Les

éléments constituant la ségrégation de l'espace scolaire peuvent s'exprimer et faire sens dans ce cadre conceptuel. Les agents sont alors de deux sortes, représentant respectivement les élèves et les chefs d'établissement. Ces agents interagissent, et les interactions peuvent se faire entre d'une part, les agents-élèves à travers des mécanismes d'imitation et d'autre part, les agents-élèves et les agents-chefs d'établissement à travers des mécanismes de préférences (choix d'établissement pour l'agent-élève, choix d'élèves pour l'agent chef d'établissement). Les structures qui émergent de ces interactions sont observables au niveau d'une part, des établissements, à partir de leur composition sociale, et d'autre part, de l'espace étudié (l'Île-de-France dans notre cas d'application), à partir de la répartition des établissements eux-mêmes dans l'espace en fonction de leur profil social. Le cadre de modélisation des SMA est ainsi parfaitement compatible avec le cadre conceptuel thématique à ce niveau très général. Cette compatibilité vaut aussi bien, que celui-ci s'inscrive dans l'individualisme méthodologique, ou bien dans une approche structurelle qui met en avant l'effet des structures sociales et spatiales sur les choix des individus, tel que cela a été décrit plus haut à propos du cadre conceptuel thématique. Si en revanche le formalisme de modélisation choisi avait été les équations différentielles, ceci aurait impliqué de formaliser des règles d'évolution directement au niveau d'entités agrégées[3]. Il y aurait eu une rupture entre les ancrages épistémologiques en jeu, alors que cette même approche aurait pu être parfaitement adaptée face à un autre questionnement thématique.

Le cadre conceptuel de l'AD est quant à lui attaché au champ de la statistique exploratoire, domaine que J.-P. Benzécri a développé dans les années 1970, en prenant ses distances relativement à la statistique classique reposant sur les méthodes inférentielles. Dès la première page de son livre sur l'*Analyse des correspondances* (Benzécri, 1973), il énonce son premier principe sur l'analyse des données, « Statistique n'est pas probabilité », critiquant par là la statistique en tant que discipline « riche en hypothèses qui ne sont jamais satisfaites dans la pratique. » Accordant une forte importance aux faits empiriques, il souligne : « La conception probabiliste est-elle conforme aux phénomènes étudiés ? La connaissance que nous avons de ces phénomènes permet-elle de faire des calculs ? ». Sa critique concernait tout l'arrière-plan théorique associé à la statistique inférentielle, présupposant que les variables d'intérêt suivent des lois bien définies, en général gaussiennes, alors que la réalité peut être toute autre. Il apparaît ainsi que le point de départ sur lequel il base sa conception de l'analyse des données est celui d'une incompatibilité entre le cadre conceptuel des statistiques mathématiques (avec leurs lois et leurs tests d'hypothèses) et celui des faits auxquels on s'intéresse qui sont décrits par de larges ensembles de données. En reprenant les termes de la figure 1, cette critique correspond ainsi à une incompatibilité entre le « cadre conceptuel thématique »

3. C'est par exemple le cas d'un modèle de croissance d'une population (P) en fonction du temps lorsque la capacité de l'environnement (C) est limitée. La croissance de la population suit alors une courbe logistique qui correspond à l'équation : $dP/dt = a\,P\,(1 - P/C)$ où « a » est un paramètre.

(relatif aux « phénomènes étudiés » de Benzécri) et le « cadre conceptuel du modèle » (relatif aux lois de probabilité et « hypothèses qui ne sont jamais satisfaites dans la pratique »). Le second principe qu'il énonce, « Le modèle doit suivre les données, non l'inverse », rend compte de son choix pour y remédier. Il consiste à réduire au minimum les contraintes que pourraient représenter le « cadre conceptuel de la modélisation ». Il s'agit de faire émerger (de rendre visible) la structure qui existe au sein des données, invisible en première analyse de par leur masse importante. Les seules contraintes qu'il préconise d'imposer et qui s'inscriraient dans le carré « cadre conceptuel de modélisation » sont celles de l'exhaustivité (il s'agit de prendre en compte toutes les dimensions caractérisant le phénomène d'intérêt) et de l'homogénéité de la nature de ces données. Le tableau de contingence constitué des comptages d'élèves appartenant aux différentes catégories sociales et fréquentant les différents établissements vérifie la propriété d'exhaustivité à partir du moment où l'ensemble des établissements franciliens sont considérés et où la nomenclature des catégories sociales est complète et cohérente. Les entités élémentaires comptées, les élèves, sont par ailleurs parfaitement homogènes de par leur statut même qui implique une certaine classe d'âge et l'exercice d'une pratique scolaire ainsi que par le caractère centralisé et harmonisé des données issues de la DEPP.

Cette analyse rapide des cadres conceptuels de modélisation associés à l'AD et aux SMA illustre la diversité des présupposés associés à une méthode de modélisation donnée, l'une faisant référence aux systèmes complexes et l'autre à une méthode exploratoire multidimensionnelle. Dans les deux cas cependant, une interrogation sur la nature et le sens des entités en jeu s'impose. Une approche ontologique est de ce fait appropriée pour aller plus loin dans la comparaison et l'articulation des deux méthodes de modélisation.

2.2. Un cadre ontologique pour comparer AD et SMA

Nous avons choisi de mobiliser la définition de Pierre Livet (2010) pour qui :

> « l'ontologie consiste à analyser un domaine, en identifiant les entités pertinentes (objets, propriétés, relations, événements et processus), et les opérations qui peuvent être opérées sur ces entités. »

Cette définition se place au croisement de celle du philosophe Barry Smith[4] et de l'informaticien Grüber[5]. Elle est centrée sur l'explicitation des entités qui caractérisent le domaine d'intérêt en y ajoutant un élément particulièrement

4. L'ontologie est « la science de ce qui est, des types et des structures des objets, propriétés, événements, processus et relations en tout domaine de la réalité…/… de ce qui pourrait exister » (Smith, 2003).

5. L'ontologie est « une spécification de la conceptualisation d'un domaine donné » (Grüber 1993).

intéressant quand il s'agit de lier le cadre conceptuel thématique à celui de la modélisation (*cf.* figure 1). Il s'agit des « opérations » qui peuvent être menées sur les entités en jeu. Cette approche de l'ontologie permet de s'assurer de la pertinence du passage du cadre conceptuel thématique au sein duquel est formulé le questionnement du chercheur en sciences sociales (ici sur la ségrégation de l'espace scolaire) au cadre conceptuel de la modélisation. La notion de « test ontologique » (Livet et Sanders, 2014), fondée sur l'opérationnalisation de la définition présentée plus haut, permet de fluidifier la circulation entre ces cadres conceptuels et de vérifier leur compatibilité lors d'une application.

Dans la suite de cette section nous prenons appui sur cette définition de l'ontologie pour comparer les deux cadres conceptuels de modélisation, l'AD d'un côté, les SMA de l'autre. Dans un premier temps nous explicitons quelles sont les *entités* en jeu, puis nous examinons quelles sont les *opérations* appliquées à ces entités. Les deux premiers termes de la définition font référence aux « objets » et aux « propriétés ». D'un point de vue thématique, les objets correspondent aux entités élémentaires qu'il s'agit d'observer et d'étudier et les propriétés à leurs caractéristiques. Examinons comment « objets » et « propriétés » sont appréhendés en AD et dans les SMA.

De manière très générale, les objets traités dans le cadre de l'AD sont de nature très variée (arbres, bâtiments, humains, établissements scolaires, villes) mais c'est toujours en nombre important que ces objets sont intéressants à analyser. C'est en effet à partir de leur diversité que l'AD peut extraire une structure qui fait sens, la faire émerger dans « l'océan des faits ». Pour Gibrat leur objectif est effectivement de « traiter simultanément de grands ensembles de faits et les confronter en vue de découvrir l'ordre global », le terme de « fait » référant au couple (individu, variable) dans le vocabulaire statistique, et correspondant ici au couple (objet, propriété). Lors de l'application d'une analyse des correspondances (AFC) au cas scolaire, les objets réfèrent d'emblée à deux échelles :

– celle élémentaire des « individus statistiques » (les élèves), qui sont les entités de base sur lesquelles se font les observations et le recueil des données ;

– celle des agrégats de ces individus statistiques au niveau des établissements scolaires.

Dans les SMA, les objets font référence aux entités de base que sont les « agents », qui ont une représentation de l'environnement dans lequel ils sont situés ainsi que des autres agents qui s'y trouvent, et qui ont la capacité d'agir. Alors que dans le cas de l'AD il s'agit toujours d'un objet qui peut être observé, qu'il s'agisse du résultat d'une expérimentation ou d'un fait existant dans le domaine empirique, dans celui des SMA, l'objet est plutôt un concept, un construit référant à un objet existant dans le domaine empirique mais en en représentant une abstraction. L'agent-élève représente ainsi un élève qui fait des choix suivant certaines règles et dont les actions dépendent à la fois de ses caractéristiques (« propriétés ») et de celles de son environnement. Il en est de même pour l'agent-établissement.

Dans le domaine statistique les propriétés font référence aux « variables » qui décrivent les « individus statistiques », c'est-à-dire les objets étudiés. En AD ces variables sont la plupart du temps nombreuses afin de cerner de la manière la plus exhaustive possible le domaine auquel on s'intéresse. Dans une analyse en composantes principales (ACP) les statuts respectifs des variables et des individus statistiques sont clairement distingués alors que dans le cas d'une analyse des correspondances (AFC) appliquée sur un tableau de contingence, les lignes et colonnes du tableau de données ont des statuts plus neutres et sont interchangeables. C'est l'interprétation qui introduit des différences sémantiques en termes d'individus statistiques (objets) et de variables (propriétés). Dans le cas, par exemple, du tableau de contingence résultant du croisement de l'établissement scolaire fréquenté par les élèves et de la catégorie sociale de leurs parents, le géographe tendra à considérer les établissements scolaires comme des individus statistiques et les catégories sociales comme des variables alors que l'inverse aurait tout autant du sens d'un point de vue statistique et ne changerait pas les résultats des calculs effectués. Le rôle des propriétés est différent dans le cadre des SMA où elles sont généralement en plus petit nombre. Il s'agit le plus souvent de quelques caractéristiques qui sont susceptibles de différencier les comportements des agents et il est intéressant pour le modélisateur de repérer d'éventuels effets seuils pouvant conduire le système à évoluer différemment suivant ces comportements.

Le terme suivant de la définition de l'ontologie mobilisée concerne les « relations ». Dans l'AD de Benzécri il s'agit de « découvrir sans parti pris, sans a priori, quels courants de lois traversent l'océan des faits » (Benzécri, 1976, cité par Cibois, 1999), ou dit autrement, de faire émerger une structure à partir des relations qui existent entre ces faits. Dans une ACP ces relations sont simplement mesurées par les corrélations entre les variables (quantitatives) caractérisant les individus statistiques. Dans une AFC, elles font référence aux similarités respectives entre les profils des lignes et des colonnes du tableau de données. Il s'agit ainsi de mesurer le degré de covariation des différentes colonnes ou des différentes lignes. Dans les deux cas (ACP et AFC) c'est une vision statistique (numérique) des relations entre des distributions qu'il faut identifier et rendre visible à partir de l'analyse des données. Dans les SMA en revanche, les relations sont d'une toute autre nature et renvoient aux interactions entre les agents et leur environnement ou entre les agents eux-mêmes. Ces interactions ne sont pas le *résultat* d'une analyse. Elles jouent au contraire un rôle moteur dans le fonctionnement de la simulation et s'expriment en général sous une forme symbolique. L'imitation, amenant un agent à agir de la même façon qu'un autre avec lequel il interagit, est un exemple simple de ce type de relation.

Les dernières entités relatives à la définition de l'ontologie concernent les événements et les processus.

Le fonctionnement des SMA repose sur la mise en œuvre de processus *bottom-up* : les mécanismes régissant les interactions entre les agents ou entre les agents et

leur environnement sont formalisés par des règles opérant au niveau même des agents. Au cours de la simulation les actions des uns vont influer sur les actions des autres et de l'ensemble de ces interactions va émerger une ou des structures observables à un niveau d'organisation supérieur. Dans le cas scolaire (François *et al.*, 2014), il s'agit par exemple de l'émergence d'une forme d'organisation spatiale plus ou moins ségréguée des établissements scolaires en fonction de leur profil social. Ces règles qui constituent le moteur de la dynamique du système modélisé sont au cœur même d'une formalisation par SMA. La dimension temporelle est ainsi intrinsèque à l'approche, chaque agent ayant la capacité d'agir à chaque pas de temps. En revanche, pour rendre compte d'un changement dans le cadre de l'AD, celui-ci doit être inscrit dans les données mêmes. C'est à travers elles seulement qu'un processus pourra être appréhendé. On peut ainsi avoir des individus statistiques associés à des dates (par exemple, lorsque ces individus sont des établissements scolaires, le collège Ravel en 2000, 2002 et 2004 peuvent constituer trois individus statistiques sur lesquels on récolte des données) et l'analyse permet d'identifier la trajectoire de cette entité dans les plans factoriels. Une autre manière de procéder consiste à construire des variables porteuses du changement (taux d'évolution par exemple) et l'analyse permet alors d'identifier les structures du changement. Alors que dans un SMA, il s'agit de formaliser le moteur même du processus grâce à des règles qui conduisent à une transformation des entités à chaque pas de temps de la simulation, dans le cas de l'AD il s'agit de rendre visible comment se structurent des changements les uns par rapport aux autres.

Une fois les différentes « entités » décrites, il reste à détailler à quoi correspondent les « opérations ». Celles-ci sont en effet indispensables pour faire « parler » des données ou mettre en œuvre les règles dans une simulation. Il s'agit de l'ensemble des manipulations effectuées sur les entités, d'une part, sur les objets et leurs propriétés dans le but de les analyser (AD) ou de les faire évoluer (SMA), d'autre part, pour mesurer ou exprimer les relations qui existent au sein du système étudié ou enfin pour caractériser la dynamique de ce système. Il peut s'agir de calculs (tableau de contingence, matrice de distance entre lignes et entre colonnes, diagonalisation de matrice, corrélations, calculs de contributions, etc.) ou de la formulation de règles (« si telle propriété et tel contexte, alors telle action de la part de l'agent »). Dans les deux cas, les opérations effectuées font émerger une structure :

– la structure de l'espace des objets et de leurs propriétés dans le cas de l'AD, rendue visible grâce à l'analyse factorielle ;

– la structure issue des interactions entre agents, observable à un niveau d'organisation supérieur dans le cas des SMA.

Dans les deux cas il s'agit de l'émergence d'une organisation, cachée dans les données pour l'AD, aboutissement du déroulé d'une suite d'actions et d'interactions pour les SMA.

3. Une spirale d'interactions entre numérique et symbolique

Après avoir comparé les cadres conceptuels de modélisation associés à l'AD et aux SMA, nous nous proposons de montrer que conceptions numériques et symboliques peuvent être articulées à plusieurs niveaux. D'une part, elles cohabitent au sein même de chacun de ces cadres, le symbolique s'invitant dans l'AD et la modélisation par SMA étant composée de plusieurs étapes numériques (sections 3.1 et 3.2). D'autre part, AD et SMA peuvent être étroitement associés et combinés au sein d'une même recherche (section 3.3).

La question de la ségrégation sociale permet d'illustrer les différentes formes d'articulations qui lient ces différentes approches. Sur un plan thématique il s'agit de tester si la ségrégation sociale au sein du système scolaire reproduit simplement la ségrégation sociale résidentielle ou si elle constitue un lieu de réduction ou d'aggravation de cette dernière. Ce questionnement est par essence dynamique et un enjeu consiste à tester les conditions d'émergence et d'évolution de différentes structures de ségrégations sociales au sein du système scolaire. Il est notamment intéressant d'explorer (avec un SMA) comment une certaine situation initiale (abstraite à partir d'une structure mise en évidence par une AD des données observées) peut être transformée suivant les stratégies mises en place par les différents acteurs du système scolaire (chefs d'établissement et familles).

D'un point de vue conceptuel, le système étudié est composé d'élèves et d'établissements scolaires (environ 1 120 collèges) qui sont considérés comme des « objets » en relation. Nous déclinons successivement les formalisations associées à chacune, en reprenant à dessein les termes utilisés pour la description des ontologies, de manière à les mettre en résonance d'un point de vue conceptuel.

3.1. Articulation des dimensions numérique et symbolique dans l'AD de la ségrégation scolaire

L'analyse des données nécessite de construire un « tableau » décrivant « l'ensemble des objets du système analysé ». Ici le tableau croise les élèves (en ligne) avec leur catégorie sociale (en colonne) et l'établissement fréquenté. L'objet de l'AD (ici une ACM[6]) est de décrire la structure d'association/relation entre les propriétés (caractéristiques) des objets (individus). Dans ce formalisme les n individus sont décrits chacun par une série de valeurs, identifiant par des 0 (absence) et des 1 (présence) sa catégorie d'appartenance pour chacune des deux variables, à savoir la catégorie sociale de sa famille (p1 modalités) et le collège qu'il fréquente (p2 modalités). Chaque individu est un point d'un espace de dimension $R^{p_1+p_2}$. Chaque modalité de variable est un point d'un espace de dimension R^n. Les opérations associées sont basées sur le calcul de la distance (chi2) entre les points et

6. Analyse des correspondances multiples.

celui d'un système d'axes (diagonalisation de la matrice d'inertie) permettant de synthétiser au mieux la forme du nuage de points. Une projection du nuage de points (individus et modalités) sur les axes principaux permet de rendre compte de ces proximités et d'identifier la structure globale de différenciation des élèves et par symétrie la structure d'association entre « catégories sociales » et « établissements ». Les catégories sociales les plus fréquemment co-localisées dans les mêmes établissements seront proches dans les plans d'axes et réciproquement les établissements fréquentés par les mêmes « profils sociaux » seront proches.

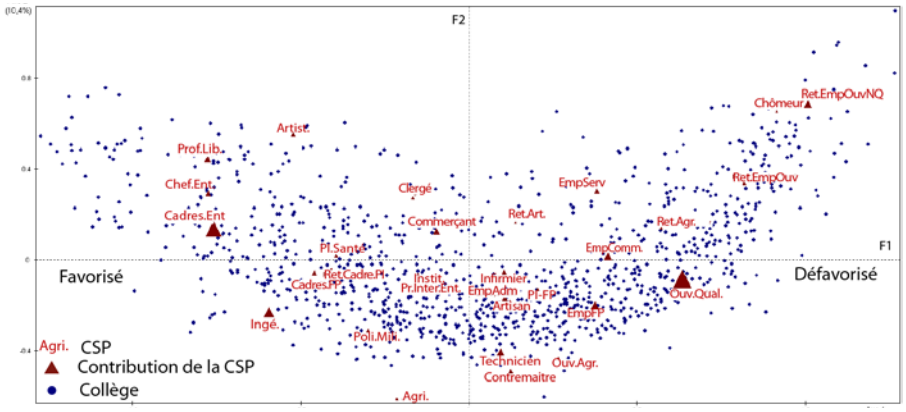

Figure 2. Résultats de l'analyse factorielle des correspondances sur les origines sociales détaillées des élèves des collèges franciliens. Lecture : les points bleus représentent les collèges (1125), les triangles rouges représentent les modalités de la variable « catégories sociales » en fonction de leur poids (effectif). (source : rapport H. Ciesielsky 2012 - Éducation nationale, 2012)

Dans cette analyse, les établissements sont formalisés comme une modalité de variable (un élève appartient ou n'appartient pas à tel établissement). Une propriété majeure de l'ACM et de la distance utilisée pour rendre compte de la structure d'association entre les modalités des catégories est la propriété d'équivalence distributionnelle. Cette propriété induit, entre autres, que le résultat obtenu à partir du tableau décrit ci-dessus serait équivalent à celui qui serait obtenu si l'on travaillait directement sur le tableau de contingence croisant les établissements (en lignes) et les catégories sociales (en colonnes) et qui décrirait donc les « établissements » par le nombre d'élèves dans chacune des catégories sociales, c'est-à-dire par son profil social. La figure 2 illustre le résultat de cette analyse. La différenciation sur le premier plan factoriel montre un grand continuum de situations sociales des établissements (points) s'organisant le long d'un arc dessiné par les différentes catégories sociales (triangles). Le long de cet arc, les catégories s'organisent selon une hiérarchie sociale, illustrant une ségrégation forte. D'un côté de ce continuum on trouve des collèges ayant une surreprésentation d'élèves de catégories telles que « chef d'entreprise », « cadre d'entreprise », « profession

libérale » alors que de l'autre côté on trouve des collèges fréquentés essentiellement par des élèves issus des catégories sociales « ouvriers ».

D'un point de vue numérique, le fait de passer d' « individus-élèves » à des « individus-établissements » relève d'une simple opération d'agrégation. Dans ce contexte, l'établissement est vu comme un contenant du point de vue du « cadre conceptuel de modélisation » (*cf.* figure 1). Le principe d'équivalence distributionnelle souligne bien la neutralité de ce statut. Dans le « cadre conceptuel thématique », en revanche, l'établissement représente bien plus qu'un simple contenant. C'est un objet complexe, lieu d'interactions entre élèves, et entretenant des relations avec les autres établissements par un jeu de concurrence ou complémentarité. Ces éléments figurent sous forme symbolique dans la représentation qu'a le thématicien du système scolaire lors de son interprétation des résultats de l'analyse statistique.

Si l'on s'intéresse maintenant à la ségrégation scolaire d'un point de vue dynamique, l'AD permet de l'explorer non pas en termes de processus, mais d'évolution des structures entre deux dates. Pour ce faire, on peut construire un tableau de données superposant les deux tableaux décrivant les profils sociaux d'un même ensemble d'établissements à chacune des deux dates. Un établissement scolaire apparaîtra ainsi deux fois, comme deux individus distincts dans le plan factoriel, et on pourra interpréter l'espacement entre leurs positions comme reflétant l'évolution entre deux états différents d'un même établissement. Cette évolution s'interprétera par rapport à la structure d'association des catégories sociales, qui définit dans ce cas une structure « moyenne » sur les deux dates. C'est ce qu'illustre la figure 3 où sont identifiés un certain nombre de collèges. Lorsque les points-dates d'un même collège sont proches, il y a peu de changement entre les deux dates. S'il y a une certaine distance entre eux, alors on interprète le déplacement comme un changement dans la spécificité sociale de cet établissement relativement à tous les autres, qu'il se dirige vers une plus grande diversité de profils (déplacement vers le centre du graphique) ou au contraire une spécialisation plus forte vers les classes défavorisées ou favorisées selon la direction du déplacement. C'est une évaluation numérique (espacement des points) qui permet d'interpréter de manière symbolique le changement entre deux dates.

Ainsi, dans la mise en œuvre d'une AD, alors que les données sont numériques et que la méthode repose sur des opérations numériques, les résultats (sorties) de l'analyse sont symboliques, et cela de deux points de vue. D'une part, c'est par l'interprétation des associations des modalités d'une variable (propriétés) mises en évidence par l'AD que le thématicien peut qualifier la structure de la division sociale (ici le gradient s'ordonne selon une certaine hiérarchie sociale). D'autre part, il va identifier des types d'établissements, au regard de leurs compositions sociales, et pouvoir les confronter aux caractéristiques des espaces résidentiels où ils sont localisés. Le résultat d'une telle analyse s'interprète donc de manière symbolique et « l'analyse des données apparaît ici comme une pratique interprétative, laissant une

place à un raisonnement analogique encadré par une interprétation statistique rigoureuse » (Lebaron, 2010). Partant d'une description numérique, l'analyse permet de mettre à jour des structures d'associations sur lesquelles vont se fonder des catégorisations abstraites, qui ne peuvent se faire qu'à l'interface des domaines conceptuels thématique et de modélisation (figure 1). Le raisonnement en termes de dynamique s'effectue aussi à cette interface comme l'illustre le propos de Bourdieu parlant des résultats de l'ACM : « elle permet de porter au jour la structure des oppositions, ou, ce qui revient au même, la structure de la distribution des pouvoirs et des intérêts spécifiques qui détermine, et explique, les stratégies des agents ». Ainsi dans notre exemple, les établissements sont appréhendés comme des individus (objets), dont la composition sociale est le résultat de processus impliquant différents acteurs et jeux d'intérêts (au niveau des familles comme au niveau des chefs d'établissements). Dans cette approche, la formalisation de la dynamique n'est pas explicite, elle est « portée au jour » par l'interprétation qui est faite du résultat des analyses en mobilisant une connaissance symbolique externe.

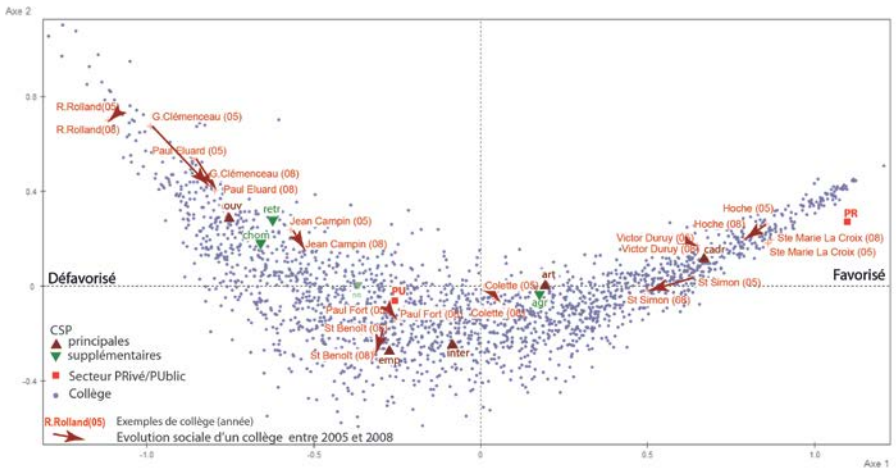

Figure 3. Évolution des profils : résultats de l'analyse factorielle des correspondances (AFC) sur les origines sociales (5 catégories) des élèves des collèges franciliens, à 2 dates, 2005 et 2008 (source : DEPP)

3.2. Articulation des dimensions numérique et symbolique dans un modèle SMA sur la ségrégation scolaire

Un modèle à base d'agents a été conçu afin de mieux comprendre comment émerge un espace scolaire ségrégué (Muller et Diallo, 2012 ; François *et al*., 2014). Comme exposé en section 2.2, l'approche à base d'agents repose essentiellement sur des connaissances symboliques. Ainsi « le modèle conceptuel du modèle est une

représentation formalisée directe du discours théorique du thématicien » (Muller et Diallo, 2012).

La base du modèle repose sur la définition des agents, objets ayant la capacité d'agir. Ici ce sont les « élèves » et « les chefs d'établissement », agents actifs, et « les établissements » qui sont des agents passifs (figure 4). Les propriétés des élèves sont leur âge, leur capital culturel, leur capital économique, leur niveau scolaire, leur nationalité et leur lieu de résidence. Les chefs d'établissement dirigent un établissement et sont dotés d'une propriété caractérisant leur comportement (« républicain », « élitiste », « pragmatique ») en termes de stratégie de recrutement des élèves. Les établissements, quant à eux, ont des caractéristiques propres : leur capacité d'accueil, leur statut « public » ou « privé », et des caractéristiques qui reflètent la stratégie du chef d'établissement, à savoir la liste des différentes filières proposées. Ils ont également des caractéristiques agrégées à partir du profil des élèves qui le fréquentent (économique, culturel, niveau scolaire et origine des élèves). La figure 4 présente le modèle conceptuel associé. On y retrouve les éléments décrits ci-dessus. Les modalités des propriétés sont des catégories symboliques. Le schéma montre bien les différents jeux de relations entre l'ensemble des entités, qu'il s'agisse d'objets, d'agents ou de propriétés. A un instant donné, un.e élève a un certain profil (économique, culturel, origine et niveau scolaire), et réside en un lieu, connaît l'existence d'un certain nombre d'établissements, en préfère certains et finalement fréquente un établissement donné. Les caractéristiques agrégées des établissements vont varier au cours du temps en fonction des choix et de l'affectation des élèves.

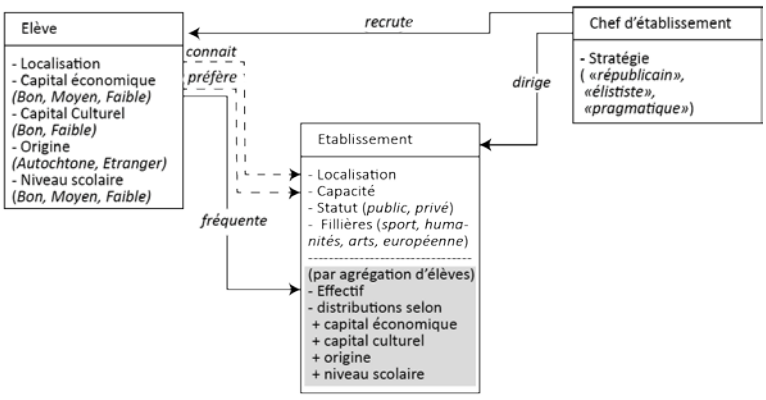

Figure 4. Objets, propriétés et relations de l'application à la ségrégation scolaire

Les relations entre les élèves et les établissements sont définies au niveau du cadre conceptuel thématique du modèle et déterminent la forme que va prendre leurs évolutions respectives. Ainsi la relation « élève fréquente un établissement »

représentée sur le schéma est le résultat d'un processus de décision qui met en relation élèves et chefs d'établissement, et qui s'appuie sur un calcul de préférences du côté des élèves, et un système de sélection du côté des chefs d'établissements.

Le modèle thématique postule que chaque élève a une connaissance partielle des établissements, et les préférences sont évaluées entre chaque élève et l'ensemble des établissements « connus » de lui. Deux règles permettent de formaliser le processus de choix d'établissement de l'élève :

– La connaissance est basée sur la proximité géographique, le rayon de connaissance variant d'un élève à l'autre en fonction de son « capital économique » et son « capital culturel ».

– La préférence se fonde sur un degré d'appétence différencié d'un type d'élève pour un type d'établissement, ce dernier étant défini par les filières proposées (sport, arts, langues, humanités) et son statut (public ou privé), l'ensemble reflétant la stratégie du chef d'établissement.

L'implémentation de la première « règle » implique essentiellement un calcul de distance entre lieux de résidence et établissements (valeur numérique). Pour la seconde règle, le choix qui a été fait est d'associer à chaque type d'élève une suite de scores reflétant ses différentes appétences pour chaque valeur possible des caractéristiques d'établissements. Le type de l'élève est déterminé par les combinaisons possibles des caractéristiques de l'élève, par exemple un capital économique moyen, un capital culturel fort, un niveau scolaire fort, et de nationalité étrangère. Le score correspond à une évaluation de l'appétence d'un type d'élève pour une filière. C'est le résultat d'un tirage aléatoire de lois normales centrées sur des valeurs d'appétences moyennes de 1, 0, ou -1. Un établissement étant défini par une liste de présences ou d'absences de filières, et un statut, la préférence d'un type d'élève pour un type d'établissement sera la somme des scores obtenus (valeur numérique) pour chacune des caractéristiques.

Concernant la formalisation du temps, il est intrinsèque au modèle, à la différence de l'AD où l'évolution est une interprétation *a posteriori* d'un jeu de données spécifique (« superposition » de données décrivant de la même manière des états à des dates différentes). Ici, à chaque itération du modèle, les profils des établissements sont recalculés, comme les préférences et les affectations à chaque nouvelle génération d'entrants au collège.

De manière générale, les règles se construisent tout d'abord à partir d'une formalisation de connaissances (discours du thématicien) sous forme de schéma conceptuel. Il est ensuite nécessaire pour l'implémenter de mobiliser un formalisme mathématique ou statistique ayant recours à des données numériques. L'exemple développé ci-dessus montre comment à partir d'une connaissance symbolique, on construit une règle à base de mesures numériques de « préférences ». Cette construction n'est pas une évidence, et la description qui vient d'en être faite peut sembler exagérément simplificatrice. Les catégories et scores utilisés sont là à titre

exploratoire et c'est à travers des tests successifs que l'on pourra avoir une idée de leur éventuel pouvoir explicatif.

Enfin, le modèle produit tout un ensemble de données numériques. Il est possible d'extraire des suivis au niveau de chaque élève, de chaque établissement et des propriétés observées au niveau macroscopique du système scolaire. La figure 5 illustre ainsi un exemple d'analyse de l'évolution de la mixité au cours du temps (axe des X) calculée à partir des données sur les établissements, extraites des sorties du modèle. Il s'agit de l'indice de Hoover, mesurant la concentration moyenne de chaque modalité (sociales, scolaire, origine) à l'intérieur des établissements. Un tel modèle produit des données numériques qu'il n'aurait pas été possible d'observer de manière empirique. Ces différentes visualisations numériques d'expérimentations à base symbolique permettent d'explorer le résultat des mécanismes mis en œuvre et leurs interactions.

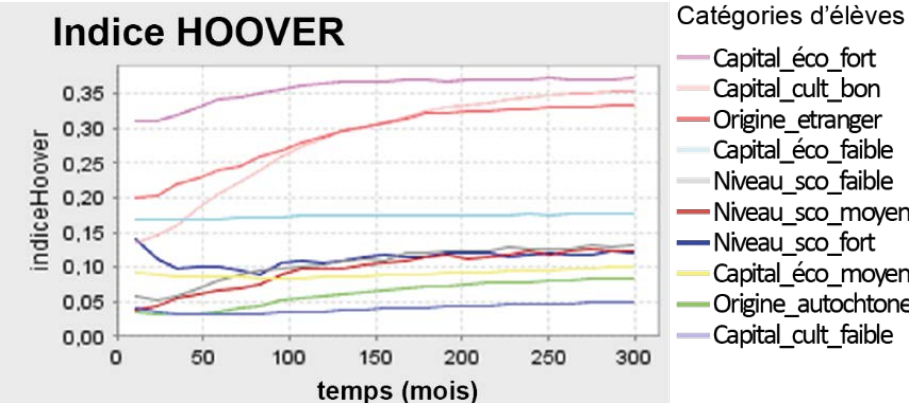

Figure 5. Exemple d'analyse de la mixité à l'intérieur des établissements à partir des sorties du modèle (l'indice de Hoover exprime pour chaque catégorie la part d'élèves de cette catégorie qu'il faudrait déplacer d'un établissement à l'autre pour aboutir à l'équirépartition)

Au terme de cette comparaison entre l'AD et les SMA, nous avons illustré comment des approches, étiquetées respectivement numérique et symbolique, font en fait cohabiter des considérations numériques et symboliques, et ce, que l'on considère les **entrées** du modèle, les **traitements** qui sont faits ou les **sorties**. Le tableau 1 synthétise cette comparaison.

Tableau 1. Forme numérique vs *symbolique des entrées, traitements et sorties correspondant à une AD et à un SMA, appliqués au cas de la ségrégation scolaire*

	Entrées	**Traitements**	**Sorties**
AD	Observables **numériques**	**Numériques :** – calcul des similarités entre établissements, – calcul des axes factoriels	**Numériques** : coordonnées sur les axes factoriels, contributions, plans factoriels ; **Symboliques** : interprétation des structures de différenciation
SMA	Connaissances **symboliques** (discours formalisé des experts)	**Symboliques** : règles régissant les comportements des agents ; **Numériques** : paramétrages pour implémenter les règles (calcul de scores…)	**Numériques** : trajectoires individuelles et agrégées **Symboliques** : interprétation des différentiels entre scénarios

3.3. Une spirale d'interactions entre analyse des données et système multi-agent

L'intérêt de mettre en relation ces deux approches est de montrer quelles formes peut prendre le dialogue entre les deux approches tant du point de vue opérationnel que thématique. Ces formes sont schématisées graphiquement dans la figure 6, en intégrant les trois étapes fondamentales évoquées plus haut, les entrées (E), les traitements (T) et les sorties (S). Leur enchaînement est matérialisé de manière linéaire par des flèches. A gauche on trouve l'approche AD et à droite l'approche SMA. Dans chaque approche sont identifiés systématiquement les statuts numériques/symboliques des « matériaux » en jeu, le symbole de « tableau de données » désignant le numérique, et celui du « personnage » faisant référence à un niveau d'abstraction et d'interprétation qui renvoie à des connaissances symboliques. Le temps est signifié par une flèche positionnée à l'étape où il est formalisé et où il intervient : dans les entrées-sorties pour l'AD, intrinsèque à l'étape de traitement dans les SMA.

Le fait que chaque approche articule du numérique et du symbolique permet d'envisager un dialogue entre les deux approches. C'est l'expérience qui a été faite sur la ségrégation scolaire illustrée figure 6. Les flèches en tiretés désignent le lien entre le résultat que produit une approche et son utilisation à une étape de l'autre approche. L'interprétation de l'évolution de la ségrégation sociale (en bas à gauche), par exemple, peut servir à formaliser des règles sur les stratégies des chefs d'établissement (en haut à droite). Réciproquement, les données issues du modèle

SMA (en bas à droite) peuvent être analysées par une AD (en haut à gauche) pour interpréter le scénario en cours.

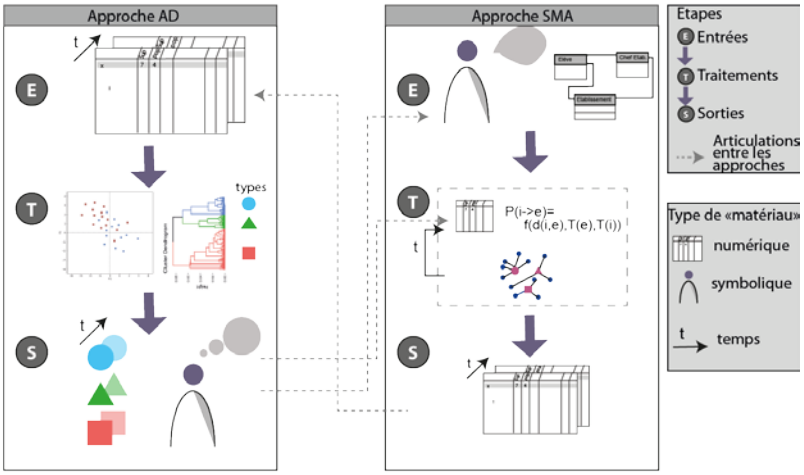

Figure 6. Cohabitation du numérique (AD) et du symbolique (SMA) à l'intérieur de chaque approche et articulation entre les deux approches

La figure 7 présente un exemple concret d'articulation entre les deux approches. A gauche on a le résultat d'une typologie sociale de l'espace résidentiel (AFC + CAH des catégories de revenus des résidents par commune en Île-de-France). Les classes de la typologie sont constituées d'agrégats de communes se ressemblant. La signification de chaque classe est définie sur une base numérique, à savoir le profil moyen des communes constituant la classe (figure 7a). L'interprétation symbolique de ces classes a permis de repérer les grandes lignes des divisions socio-spatiales d'un espace métropolitain. Celles-ci ont servi pour construire la situation initiale stylisée sur laquelle appliquer les règles formalisées au sein du SMA. Les agents-élèves ont ensuite été distribués dans l'espace résidentiel, en fonction de leurs propriétés (figure 7b). Deux propriétés de l'agent-élève, « capital culturel » et « capital économique », non directement observables, ont pu être identifiées lors de l'interprétation des résultats de l'AFC effectuée sur les catégories sociales (figure 3). Elles sont davantage porteuses de sens dans les différences de comportement de choix des élèves que la variable « catégorie sociale ». Pierre Bourdieu exprime ainsi la dimension heuristique de l'AFC : « j'utilise beaucoup l'analyse des correspondances, parce que je pense que c'est une procédure relationnelle dont la philosophie exprime pleinement ce qui selon moi constitue la réalité sociale. » (Bourdieu[7] *in* Lebaron 2015). Dans un SMA, on va

7. dans la préface à l'édition allemande du *Métier de sociologue.*

ainsi pouvoir mobiliser des propriétés (variables) jugées explicatives dans la « réalité sociale » telle qu'elle est interprétée par l'expert thématicien, même si elles ne sont pas observables dans le domaine empirique. Tel est ainsi le cas du « capital culturel » et du « capital économique » des élèves dont la combinaison est, suivant l'interprétation que l'expert thématicien a fait à partir d'AFC sur les catégories sociales, motrice dans la dynamique de la ségrégation scolaire.

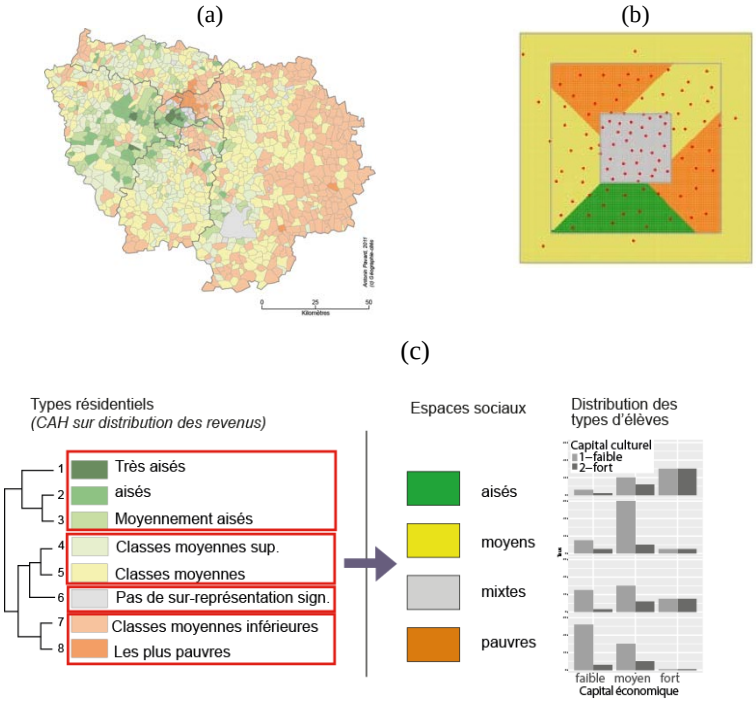

Figure 7. Articulation entre les résultats d'une typologie effectuée par AD et la construction d'une situation spatiale initiale stylisée en entrée du modèle SMA

Conclusion

Si l'on repositionne l'ensemble de la démarche relativement à la figure 1, nous avons montré comment s'articule le cadre conceptuel thématique avec le cadre conceptuel de modélisation de chacune des deux méthodes AD et SMA. Nous avons ainsi précisé au §2.1 le cadre épistémologique sur lequel sont fondées les hypothèses du modèle conceptuel du thématicien, prenant appui sur les travaux de Bourdieu, et qui mettent en avant le rôle des structures sociales sur les choix des élèves et des chefs d'établissement. Ce positionnement épistémologique contribue à la bonne articulation entre les deux approches. Si d'autres choix épistémologiques avaient été faits, l'interprétation symbolique des résultats d'une typologie n'aurait pas

forcément pu servir de référentiel à la construction abstraite des entités du modèle SMA. C'est ainsi la compatibilité conceptuelle entre les différents éléments représentés sur la figure 1 qui permet de valider l'articulation entre des méthodes différentes et non la convergence des résultats obtenus. Glaser et Strauss (1967) le soulignent de manière très claire :

> « La production de théorie ne s'appuie pas sur le fait, mais sur la catégorie conceptuelle (ou une propriété conceptuelle de la catégorie) qui en a été extraite. Un concept peut être produit à partir d'un fait qui devient alors simplement un élément dans un univers de nombreux indicateurs possibles pour le concept et de données pouvant lui être associées » (Glaser et Strauss, 1967).

Finalement, l'articulation entre les approches AD et SMA telle qu'elle a été formalisée et illustrée, permet de dessiner une spirale méthodologique, qui fait alterner approche numérique et approche symbolique. L'expérience de cette spirale est particulièrement riche dans la mesure où elle permet de tester deux formes d'exploration qui se répondent : d'une part, des structures d'associations pour en comprendre les mécanismes et d'autre part, des mécanismes qui conduisent à ces structures. Ainsi l'AD permet « d'aider à trouver et départager les candidats à l'explication sociologique » (Lebaron, 2010). Ces structures d'associations sont réinterprétées dans le cadre conceptuel thématique pour être ensuite reliées à des processus. De même, en SMA, les résultats obtenus par expérimentation de différentes règles ou paramètres permettent de cibler des « candidats à l'explication ». Le but d'un modèle SMA est en effet d'identifier un « candidat explicatif » possible pour rendre compte de la genèse, de la structure et de la dynamique du phénomène étudié. Cependant, plutôt que de prétendre qu'il s'agit de la seule explication possible, la démarche consiste à tester quelles sont les « conditions de possibilités » du phénomène étudié (Livet *et al*., 2014). Les deux méthodes, pourtant ancrées dans les arrières plans théoriques très différents de la statistique et de la simulation informatique, ont ainsi en commun cette ambition de proposer un « candidat explicatif » au phénomène étudié tout en maintenant une position expérimentale ouverte de révision et de possibilité d'évolution des modèles explicatifs proposés. Ce cheminement de l'une à l'autre méthode, suivant une approche dynamique en spirale permet de relier les deux points de vue et d'approcher progressivement les processus sous-tendant la dynamique étudiée. Chacune des approches permet en quelque sorte de valoriser la connaissance apportée par l'autre.

Remerciements

Les auteures souhaitent vivement remercier les deux évaluateurs/évaluatrices pour leurs suggestions et questions, celles-ci leur ayant permis d'expliciter un certain nombre de points.

Bibliographie

Balbo F., Saunier J., Diday E., Pinson S. (2009). De l'utilisation de l'analyse de données symboliques dans les systèmes multi-agents. *Actes de la conférence « Extraction et gestion des connaissances » (EGC'09)*, Strasbourg, 27-30 janvier.

Glaser B., Strauss A. (1967). *The discovery of grounded theory*, trad. de J.-L. Fabiani, 1995, *Revue Enquête.*

Benzécri J.-P. (1973). *L'analyse des données, tome 2 : L'analyse des correspondances*, Bordas.

Boudon R. (1973). *L'inégalité des chances: la mobilité sociale dans les sociétés industrielles*, Paris, Éditions Armand Colin.

Bourdieu P. (1979). *La Distinction : Critique sociale du jugement*, Paris, Éditions de Minuit.

Cibois Ph. (1999). Modèle linéaire contre modèle logistique en régression sur données qualitatives. *Bulletin de Méthodologie Sociologique*, n° 64, p. 5-24.

Diday E. (1989). Introduction à l'approche symbolique en analyse des données. *RAIRO, Recherche opérationnelle*, tome 23, n° 2, p. 193-236.

François J.-C., Mathian H., Sanders L., Bulle N., Waldeck R., Phan D. (2014). Modélisation par SMA de la structuration sociale de l'espace scolaire : une ontologie intégrée mais non réductrice de plusieurs points de vue méthodologiques. *Ontologies et modélisation par SMA en SHS*, Phan D. (dir.), Londres/Paris, Hermès-Lavoisier, p. 461-475 .

Gruber T.R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, vol. 5, n° 2, p. 199-220

Lebaron F. (2010). L'analyse géométrique des données dans un programme de recherche sociologique: les cas de la sociologie de Bourdieu. *Revue MODULAD*, n° 42.

Lebaron F. (2015). L'espace social. Statistique et analyse géométrique des données dans l'œuvre de Pierre Bourdieu. Chapitre 3. *La méthodologie de Pierre Bourdieu en action. Espace culturel, espace social et analyse des données*, F. Lebaron éd., Paris, Dunod, coll. « Psycho Sup », p. 43-58.

Livet P., Müller J.-P., Phan D., Sanders L. (2010). Ontology, a mediator for agent-based modeling in social science. *Journal of Artificial Societies and Social Simulation*, vol. 13, n° 1, 3, http://jasss.soc.surrey.ac.uk/13/1/3.html

Livet P., Sanders L. (2014). Le « test ontologique » : un outil de médiation pour la modélisation agent. *Ontologies et modélisation par SMA en SHS*, Phan D. (dir.), Hermès-Lavoisier, Londres-Paris, p. 95-110.

Livet P., Phan D., Sanders L. (2014). *D*iversité et complémentarité des modèles multi-agents en sciences sociales. *Revue française de sociologie*, 2014/4, vol. 55, p. 689-729. https://www.cairn-int.info/revue-francaise-de-sociologie-2014-4-page-689.htm

Mathian H., Sanders L. (2014). *Objets géographiques et processus de changement. Approches spatio-temporelles*, ISTE Éditions. ISBN 978-1-78405-031-3.

Müller J.-P., Diallo A. (2012). Vers une méthode multi-point de vue de modélisation multi-agent. *Systèmes multi-agents : ouverture, autonomie et co-évolution*. *Actes des JFSMA'12*

(Journées francophones sur les systèmes multi-agents), 17-19 octobre, Honfleur (France), Toulouse : Cépaduès, p. 33-42. ISBN 978-2-36493-037-7.

Phan D., Müller J-P., Sibertin Blanc C., Ferber J., Livet P. (2014). Introduction à la modélisation par SMA en SHS : comment fait-on une ontologie ? *Ontologies et modélisation par SMA en SHS*, sous la dir. de Phan D., Traité RTA, série « Informatique et systèmes d'information », Paris, Hermès Science Publications, p. 21-51. ISBN 978-2-7462-3207-5.

Poupeau F., François J. C. (2008). *Le sens du placement. Ségrégation résidentielle et ségrégation scolaire*, Raisons d'agir, coll. "Cours et travaux", EAN: 9782912107435.

Smith B. (2003). Ontology, *Blackwell Guide to the Philosophy of Computing and Information*, Floridi L. (Ed.), Blackwell, Oxford, p. 155-166.