

Évaluer et paramétrer un modèle de simulation complexe en situation d'inter-disciplinarité

Version 2019-11-24

- 31/10/2019 : Fusion anciens chapitre 3 (évaluation) et 4 (paramétrage)
- 08/11/2019 : Fin reprise + Impression Lena
- 11/11/2019 : Reprises commentaires Lena du 24/11/2017 + finalisation
- 19/11/2019 : Dernier round reprises Lena
- 24/11/2019 : Fin reprise + envoi relecture :
 - Seb : Intro + 3.1 + 3.2.1 et 3.2.2 (pas la 3.2.3) + Conclusion
 - Clem : Intro + Partie 3.3 + Conclusion
 - Paul G. : 3.2.3
 - => Retour si possible le jeudi 05/12/2019

Sommaire

Introduction	2
3.1 Comment évaluer un modèle?	3
3.1.1 Évaluation, validation, vérification...	4
3.1.2 Les étapes de l'évaluation d'un modèle	8
3.1.3 Une évaluation de la plausibilité d'un modèle : la « <i>face validation</i> »	14
3.1.4 Vers une évaluation visuelle	23
3.2 Une grille d'analyse composée d'indicateurs de sortie	27
3.2.1 Indices et indicateurs	27
3.2.2 Hiérarchiser et catégoriser les indicateurs	33
3.2.3 Les indicateurs et dimensions de SimFeodal	38
3.3 Paramétrage du modèle SimFeodal	48
3.3.1 Les paramètres	48
3.3.2 Le paramétrage	59
Conclusion	77

Introduction

Dans le chapitre précédent, nous avons présenté un modèle, SimFeodal, dont l'objectif est de répondre à un questionnement thématique (ref 2.1.2) portant les phénomènes de polarisation, de hiérarchisation et de fixation de l'habitat paysan entre les IX^e et XIII^e siècles. Pour juger de la capacité du modèle à reproduire ces processus empiriques, on on procède à une « évaluation » du modèle. Ce type d'évaluation prend appui sur un ensemble de critères : plus ceux-ci sont remplis, meilleure sera l'évaluation, et plus le modèle sera jugé satisfaisant.



L'évaluation de modèle est un domaine incontournable et fortement étudié dans le champ de la modélisation, quel que soit le type de modèles sur cette évaluation porte. Dans le cas de SimFeodal, modèle descriptif et complexe, l'évaluation ne peut se faire de manière formelle, c'est-à-dire analytique, tant les interactions entre agents et mécanismes sont nombreuses, non linéaires, et donc non prédictibles. L'évaluation ne peut donc être qu'expérimentale, en analysant le comportement du modèle sur la base de ses sorties.



SimFeodal pose de plus le problème d'être un modèle basé sur des connaissances expertes tôt que sur des données directement intégrables ou confrontables. Cela en complexifie l'évaluation, car les approches classiques sont peu adaptées à ce type de modélisation interdisciplinaire guidé par la théorie et où les données empiriques sont lacunaires. Il est en effet difficile de quantifier le comportement attendu du modèle.




Les difficultés liées à l'évaluation du modèle sont renforcées par le type de construction mis en place pour SimFeodal, qui est un modèle exploratoire dans lequel l'évaluation n'a pas vocation à valider une version définitive et statique du modèle, mais au contraire à guider son amélioration par des ajustements successifs.




Dans le chapitre précédent, on décrivait par exemple une « version » du modèle, intitulée « version 6.3 », ce qui montre qu'auparavant, il y a nécessairement eu au moins 5 versions préalables, et sans doute bien plus de sous-versions. Ce chapitre vise à présenter le processus d'amélioration du modèle que nous nommons « paramétrage », c'est-à-dire l'approche théorique et empirique qui a guidé l'évolution de SimFeodal, à partir des retours réguliers de l'évaluation du modèle.

Dans ce chapitre, nous présenterons en premier lieu la démarche globale d'évaluation, et en particulier une proposition de méthode fondée sur l'évaluation visuelle – adaptée aux modèles du type de SimFeodal. Nous spécifierons cette approche en l'appliquant au cas de SimFeodal, c'est-à-dire en présentant les critères retenus pour l'évaluation du modèle. Nous décrirons ensuite le paramétrage du modèle, son articulation avec l'évaluation, et nous mènerons une analyse rétrospective de l'évolution de SimFeodal.

3.1 Comment évaluer un modèle ?

Depuis les travaux précurseurs en simulation informatique (NAYLOR et FINGER 1967 ; HERMANN 1967 ; SARGENT 1979) jusqu'aux recherches contemporaines (AMBLARD, ROUCHIER et BOMMEL 2006 ; BANOS 2013 ; AUGUSIAK, BRINK et GRIMM 2014 ; REY-COYREHOURCQ 2015), la plupart des chercheurs ont toujours mis en avant qu'un modèle de simulation non évalué n'avait ni utilité – pour NAYLOR et FINGER 1967 notamment –, ni validité. Sans caricaturer ces écrits, on peut noter que tous cantonnent les modèles non évalués à des « jeux » ou encore, pour les plus modérés, à des outils uniquement pédagogiques. 

Comme indiqué dans le **chapitre 1**, nous considérons SimFeodal comme un modèle résolument exploratoire, en particulier par la dimension heuristique qui le caractérise : les résultats produits par le modèle n'ont pas vocation à être mobilisés directement, ce sont des supports à la synthèse et à la ré-organisation de connaissances expertes sur le cas d'étude analysé. On pourrait dès lors se passer d'en mener une évaluation quelconque. 

Il nous semble pour autant que l'exercice intellectuel que constitue la (co-)construction d'un modèle de simulation perdrait de son intérêt intrinsèque s'il ne donnait lieu à des procédures, quelles qu'elles soient, ayant pour objectif d'assurer une certaine qualité au modèle, à défaut de lui garantir une validation stricte.

Nous sommes en effet convaincus que même pour des modèles visant à « assister la construction de théories »¹ pour reprendre les termes de LAKE (2014, p. 260), ou encore, selon la classification alternative de l'auteur, pour les modèles à utilité « de développement »², les différents outils d'évaluation permettent d'acquérir une connaissance précieuse sur l'objet modélisé, ne serait-ce que par les effets collatéraux qu'entraîne l'évaluation d'un modèle. Qu'un modèle soit statistique ou à base d'agents, de type descriptif ou explicatif, à visée pédagogique ou prédictive, ou encore constitue un modèle « hybride » entre ces catégories, un modèle de simulation demeure un modèle qu'il convient d'évaluer pour être en mesure d'en tirer des connaissances (SARGENT et BALCI 2017, p. 299-300).

Sans entrer dans les spécificités conceptuelles de ce qu'est l'évaluation d'un modèle ou de l'histoire de ces méthodes³, nous nous contenterons dans la suite de cette partie de donner une vision aussi succincte que possible de ce qu'est l'évaluation, en particulier pour en dégager les méthodes employées usuellement. Cela nous permettra en particulier de défendre et de promouvoir l'une de ces méthodes, la validation visuelle, que nous jugeons très adaptée dans le cadre de co-constructions interdisciplinaires de modèles.

1. « Simulation models to support theory building – so-called heuristic modelling – [...]. »

2. « 'developmental' utility », c'est-à-dire les modèles dont le développement et l'implémentation bénéficient aux chercheurs qui y prennent part plutôt qu'à ceux qui se contentent de les utiliser a posteriori.

3. En particulier parce que ce sujet a été très largement traité dans un travail de thèse récent au sein de notre laboratoire de recherche (REY-COYREHOURCQ 2015, pp. 58-184), travail auquel nous renvoyons vivement pour plus d'approfondissements.

3.1.1 Évaluation, validation, vérification...

Il nous semble important de commencer cette partie par un point de définition et de clarification des concepts mobilisés, non pas par convention, mais parce que les usages en matière d'emploi des termes d'évaluation, de validation (méthodologique, formelle...) ou encore de vérification sont particulièrement diffus et trompeurs dans la littérature relative à la modélisation, y compris dans le champ plus restreint de la simulation à base d'agents en sciences humaines et sociales.

Depuis les travaux fondateurs, dans les années 1960, la logique qui consiste à éprouver un modèle – c'est-à-dire à vérifier qu'il corresponde correctement d'une part (1) au système qu'il décrit, et d'autre part à (2) la manière dont il est décrit – donne lieu à différentes terminologies. On notera en particulier que les deux articles considérés comme pionniers, tous deux parus en 1967, reposent pour l'un sur la notion de vérification (*verification*) (NAYLOR et FINGER 1967), et pour l'autre sur celle de validation (HERMANN 1967), sans pour autant que la distinction entre les deux approches puisse être vue comme consistante.

Quelques décennies plus tard, une fois la pratique de simulation informatique plus développée et mûre, un consensus de pratique a été adopté autour de l'expression englobante de « Validation, Verification and Testing techniques (VV&T) », par l'entremise d'une proposition de BALCI (1994) de vérification et de définition de chacun de ces composants. Pour reprendre ses mots en une distinction devenue courante en simulation à base d'agent, la *validation* consiste à concevoir le bon modèle⁴ – sens (1) exposé plus haut –, alors que la *verification* permet de s'assurer que le modèle est bien construit⁵ – sens (2). Le « *Testing* » correspond aux techniques mises en œuvre, et s'applique donc indistinctement à ces deux termes (*validation* et *verification*).

En dépit de cette définition stricte, les usages persistent dans une absence de distinction formelle entre vérification et validation, le plus souvent en englobant ces pratiques dans le terme plus large et moins défini d'« évaluation ». Il n'est d'ailleurs pas rare que ces trois termes soient employés de manière interchangeable, voir intervertie, comme un recensement rigoureux des usages le démontre (AUGUSIAK, BRINK et GRIMM 2014). Dans cet article, les auteurs mènent une méta-analyse de la littérature sur les usages de chacun des termes liés à l'évaluation⁶, et en particulier de celui de *validation*, et du sens signifié par leurs auteurs respectifs. Ils en tirent le constat qu'une large partie des termes analysés a été employé par plusieurs auteurs leur affectant des sens contradictoires. Par exemple : « the term “validation” has been given virtually any possible meaning in this context » (AUGUSIAK, BRINK et GRIMM 2014, p. 120).

Les auteurs de cette étude puisent dans ce constat les besoins d'une nouvelle terminologie, unique et explicite, permettant de dépasser notamment les suc-

4. « Model validation deals with building the *right* model. » (BALCI 1994, p. 121)

5. « Model verification deals with building the model *right*. » (BALCI 1994, p. 123)

6. *Corroboration, Evaluation, Testing, Validation, Verification et Substantiation*

cessions d'attrait et de rejet que la notion de validation entraîne de par son positivisme. Ils proposent ainsi un nouveau terme, l'« *evaludation* », assorti à une typologie de concepts – explicitement définis (AUGUSIAK, BRINK et GRIMM 2014, table 2, p. 125) – liés à l'évaluation permettant d'identifier l'objet et le sujet de chacune des phases du cycle d'« évaluation » (voir figure 3.1).

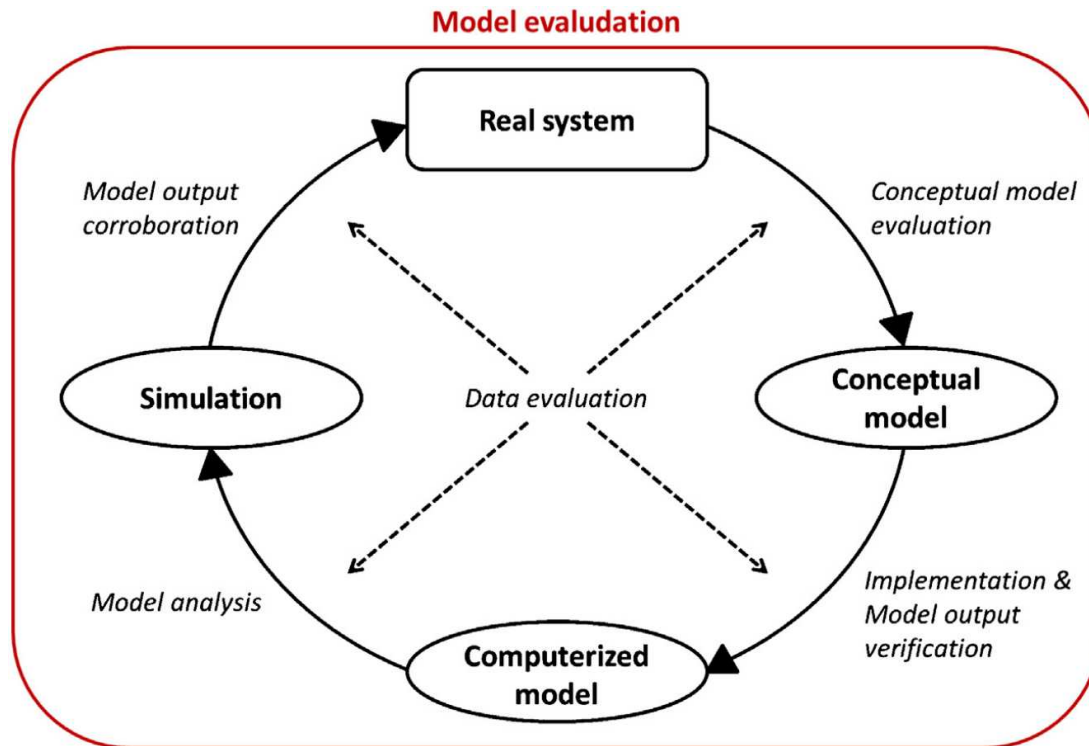


FIGURE 3.1 – Représentation schématique du cycle de modélisation et de la typologie des termes relatifs à l'évaluation de modèle, dans AUGUSIAK, BRINK et GRIMM (2014, Fig. 1, p. 121)

Dans ce schéma, il est intéressant de noter que le terme de « validation » n'est nul part employé (si ne n'est dans le mot-valise *evaludation*), au contraire de celui d'« *evaluation* » assez omniprésent. Les auteurs se détachent de l'usage de « *validation* » en raison de la connotation très positiviste et « binaire » de ce mot : elle peut impliquer qu'un modèle est soit valide, soit invalide, loin du gradient qu'accepte l'évaluation. Notons que l'item central du schéma, l'évaluation des données (*data evaludation*) qui intervient pour l'évaluation de chacun des types de modèle, a une signification plus ouverte qu'on ne pourrait l'entendre au sens premier. AUGUSIAK, BRINK et GRIMM (2014, p. 121) y incluent ainsi des connaissances extraites des données en tant que telles : « The term "data" also refers to patterns (GRIMM et RAILSBACK 2012) or, in economists' terminology, "stylised facts", which are general trends and signals in data, observations, and empirical knowledge. »

Nous partageons absolument le besoin formel, identifié par les auteurs de cette étude, de définir un nouveau terme dans leur proposition, et nous souscrivons à leur approche de définition. Il nous semble toutefois, et nous le déplorons, que ce travail n'ait pas encore suffisamment percolé dans la communauté scientifique, et en particulier dans le monde francophone. Au moment de rédiger ces lignes, très peu d'auteurs en font un usage en français, et surtout sous la forme de recension plus que d'utilisation du concept (par exemple REY-COYREHOURCQ 2015, p. 89,436).

Par soucis d'homogénéité et de compréhension par un plus grand nombre de ce travail, nous nous contenterons donc de nous inscrire dans les choix de AMBLARD, ROUCHIER et BOMMEL (2006, voir encadré 3.1), en particulier parce qu'ils nous semblent assez largement adoptés dans la communauté scientifique francophone de modélisation en sciences humaines et sociales, quand bien même ces concepts nous paraissent moins robustes que ceux présentés auparavant.



Encadré 3.1 : Évaluation, validation interne et externe

Pour AMBLARD, ROUCHIER et BOMMEL (2006), on emploie le concept d'**évaluation** pour définir l'approche d'ensemble, et on distingue alors « **validation interne** » – correspondant à la *verification* définie par BALCI (1994), c'est-à-dire s'assurer de la bonne conception du modèle, et « **validation externe** » – ce que BALCI nomme *validation*, soit l'assurance que le modèle est adapté à ce qu'il cherche à représenter.

« Il est classique de différencier deux étapes dans la validation : interne et externe.

- La phase de vérification ou **validation interne** comprend d'abord une vérification de conformité entre les spécifications et le programme implémenté et pose la question : est-ce que le modèle implémenté est bien celui que je voulais implémenter ? [...] Ensuite, la validation interne concerne la recherche et l'identification des propriétés du modèle. Dans le cas des simulations multi-agents, des preuves logiques ne peuvent être obtenues et se pose alors la question : est-ce que mon modèle possède les propriétés attendues ? Parmi ces bonnes propriétés, on considère par exemple la robustesse ou des études de sensibilité pour vérifier si les réponses sont bien différenciées sur l'espace des paramètres. Cette phase de validation interne concerne de fait une validation dans le contexte ou la logique propre du modèle.
- La deuxième phase de validation, la **validation externe**, correspond à l'évaluation de l'adéquation entre le modèle et le phénomène réel dont il est censé rendre compte. Pour cette dernière phase, la comparaison aux données empiriques ou le fait que le modèle soit capable d'exhiber des faits stylisés identifiés sur le système modélisé sont des critères clés.

Ainsi, ce qui est étudié au travers des simulations, ce sont tout d'abord les propriétés systémiques (structurelles et dynamiques) du modèle, les formes qui peuvent apparaître du fait des hypothèses posées (validation interne); ensuite est évaluée la pertinence du modèle vis-à-vis de situations que l'on souhaite représenter ou prévoir (validation externe). »

AMBLARD, ROUCHIER et BOMMEL 2006, p. 110-111

3.1.2 Les étapes de l'évaluation d'un modèle

Parmi les nombreuses techniques disponibles pour l'évaluation, il est courant de privilégier telle ou telle méthode en fonction de la phase d'avancement d'un modèle. Traditionnellement, l'usage veut ainsi que le modélisateur tende vers des méthodes de plus en plus formelles à mesure que l'évaluation progresse⁷. Les schémas des étapes d'évaluation de KLÜGL (figure 3.2) et de NGO et SEE (figure 3.3) constituent un bon résumé de cette progression – représentée de manière itérative quand bien même chaque auteur insiste sur le fait que ces étapes doivent être menées en multipliant les allers-retours entre elles – que l'on peut brièvement décrire plus avant.

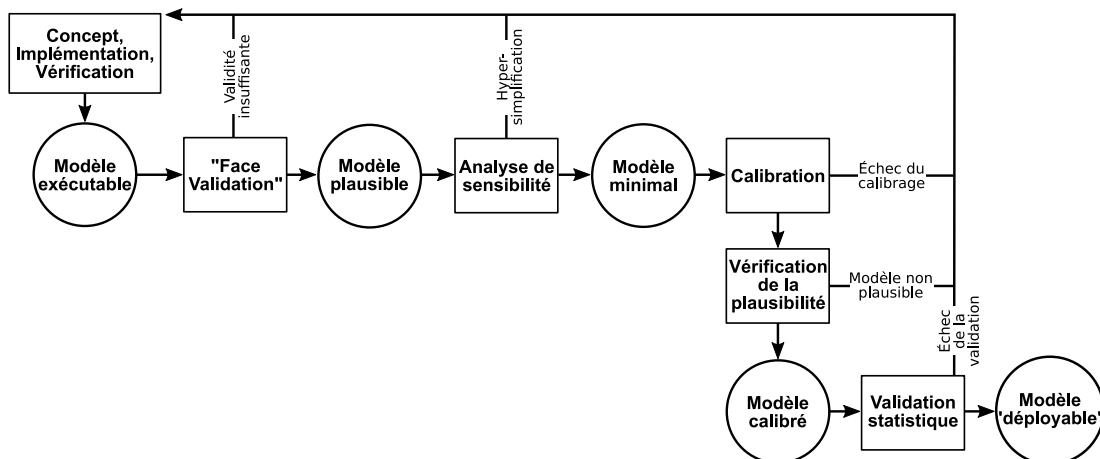


FIGURE 3.2 – Une esquisse de procédure générale de validation de modèles de simulation à base d'agents, traduit d'après KLÜGL (2008, fig. 1 p. 42)

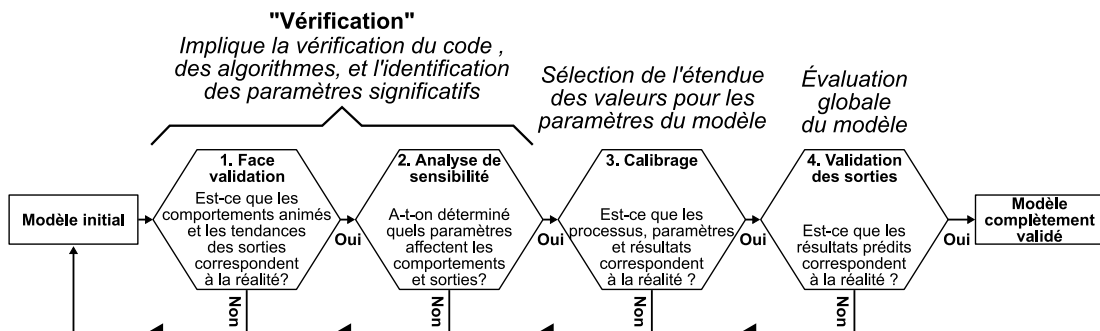


FIGURE 3.3 – Une procédure générale de validation d'un modèle à base d'agents, traduit d'après NGO et SEE (2012, fig. 10.1 p. 183)

« **Face validation** ». La première étape, de « *face validation* », consiste ainsi à vérifier visuellement, en se confortant à des intuitions sur le comportement attendu, la plausibilité du modèle. On entend par plausibilité la potentielle adéquation entre le déroulement (en termes de dynamiques observés) et l'issu (au travers des données produites) d'une simulation, et les connaissances expertes que l'on possède sur le système modélisé. Cette étape est souvent considérée comme une mesure préalable relevant du bon sens plus que de l'évaluation strictement dite, et relève autant de la validation interne que de la validation

7. Voir par exemple la typologie des méthodes d'évaluation, des plus « informelles » aux méthodes « formelles » chez BALCI (1994, figure 3, p. 131)

externe. Nous reviendrons plus longuement dans les pages suivantes (voir section 3.1.3) sur cette étape essentielle d'après l'avis partagé dans la littérature mais néanmoins largement sous-exploité et méjugée à notre avis.

Analyse de sensibilité. Quand une première version du modèle a été implémentée, il est recommandé de procéder à l'analyse de sa « sensibilité », entendue à l'égard des paramètres du modèle : en faisant varier, selon des méthodes plus ou moins complexes⁸, les valeurs des différents paramètres du modèle, on peut observer l'influence de chaque paramètre sur le déroulement du modèle. Cette procédure, relative à la validation interne, intervient tôt, et doit être répétée lors de chaque modification majeure dans les mécanismes du modèle. L'analyse de sensibilité permet en effet de simplifier le modèle conceptuel et son pendant implémenté : si l'analyse révèle qu'un paramètre, quelles que soient les valeurs qui lui sont attribuées, n'a qu'un effet minime voir négligeable sur les sorties du modèles, alors il peut être judicieux de supprimer ce paramètre ou le mécanisme qui le mobilise. Réduire le nombre de paramètres ou de mécanismes d'un modèle peut sensiblement l'améliorer, selon le principe de parcimonie qui voudrait qu'un modèle plus simple soit meilleur⁹.
Même sans aller jusqu'à ce type de découvertes sur l'inutilité de certains paramètres, l'analyse de sensibilité permet de gagner en connaissance sur le fonctionnement d'un modèle complexe et/ou non-déterministe, ne serait-ce que parce qu'elle aide souvent le modélisateur à trouver une « polarité » à l'effet des paramètres : si tel indicateur de sortie (voir section 3.2.1) croît quand on diminue les valeurs d'un paramètre et décroît quand on les augmente, alors on peut prévoir l'effet d'une modification de ce paramètre, ce qui peut éclairer le fonctionnement thématique du modèle.

Calibrage. Une fois le modèle mieux connu et surtout réduit à ses composantes nécessaires et suffisantes, on cherche à en améliorer la qualité de représentation, c'est-à-dire à faire en sorte, en jouant sur les valeurs de paramètres, que le modèle reproduise plus précisément le système qu'il décrit (validation externe) et qui correspond à l'observation empirique ou aux connaissances thématiques. Cette démarche, nommée calibrage, soulève l'enjeu d'isoler, pour chaque paramètre, une étendue de valeurs acceptables et optimales. La complexité – au sens figuré – de cette étape réside dans la complexité – au sens propre – du modèle qu'il convient de calibrer : dans un modèle complexe, où chaque mécanisme peut influencer chacun des autres mécanismes de manière non linéaire, la modification des valeurs d'un paramètre doit certes modifier l'état du modèle en lui-même, mais a le plus souvent tendance à modifier par là même l'optimalité des valeurs des autres paramètres. Le problème ressemble à celui des vases communicants : pour que le modèle soit calibré, il faut que chaque valeur de paramètre soit optimale, mais la modification de chacun des paramètres peut dérégler l'effet des autres paramètres, et par la même les valeurs qu'ils doivent se voir attribuer. On ne peut donc procéder paramètre par

8. Un point plus précis y est consacré dans le **chapitre 6**

9. Les avis divergent nettement sur ce point, voir par exemple la définition de la *simplicité* dans AMBLARD, ROUCHIER et BOMMEL (2006, p. 120)

paramètre, en les réglant un par un, au risque d'entrer dans une boucle infinie de calibrage, mais au contraire, il est nécessaire de considérer l'ensemble – ou un sous-ensemble – des paramètres et de tester des valeurs qui iraient vers une optimisation du comportement du modèle. Le calibrage en lui-même n'est pas à proprement parler une véritable procédure d'évaluation d'un modèle : il vise ainsi plutôt à « améliorer » le modèle plus qu'à juger de sa validité. Il s'agit ainsi plutôt d'une méthode et d'un problème d'optimisation que d'évaluation. Les différents auteurs mettent toutefois en avant son intérêt dans l'évaluation de modèle en ce qu'il permet de garantir une meilleure validation externe du modèle puisqu'il aboutit à l'isolation d'étendues optimales de valeurs de paramètres : en menant de nouveaux tests d'évaluation (analyse de sensibilité, *face validation*, etc.) (KLÜGL 2008, p. 43) sur les valeurs optimales identifiées, on peut évaluer si elles sont porteuses de sens d'un point de vue empirique ou au moins vis-à-vis de la connaissance experte du système modélisé.

Validation statistique. La validation statistique (« *output validation* » dans la figure 3.3) est sans doute la méthode d'évaluation la plus évidente pour quiconque a été amené à concevoir un modèle. Il s'agit de confronter les données produites par le modèle – les *outputs* – aux données empiriques – ou observées – qu'ils cherchent à reproduire. Autrement dit, en termes statistiques, à s'assurer de la qualité de l'ajustement – la *goodness of fit* – des données simulées. On en mesure l'écart avec les données observées, quand de telles données sont disponibles, en cherchant à minimiser cet écart : plus l'écart est faible, alors plus le modèle parvient à reproduire les observations qui ont servi de support à sa conception et construction. La validation statistique est donc une méthode de validation externe. Les différents auteurs du champ de l'évaluation recommandent de ne mener cette étape qu'à la fin du processus d'évaluation, quand le réflexe en pratique est souvent de s'appuyer sur les données empiriques dès le début de la conception du modèle. A défaut de suivre cette recommandation, le modélisateur risque d'emmener le modèle vers du « sur-ajustement » (*overfitting*), et d'inscrire alors celui-ci dans une forme de tautologie, le modèle étant alors construit précisément pour produire ce qu'il devrait plutôt faire émerger. En conservant la validation statistique comme l'une des dernières étapes du cycle d'évaluation, c'est-à-dire en s'empêchant d'essayer de faire coller le modèle aux données qu'il doit reproduire, on s'assure de l'indépendance des mesures de l'ajustement, et on peut donc garantir une certaine objectivité quant à l'évaluation du modèle.

Validations formelles. Absentes des deux figures (3.2 et 3.3), les méthodes de validation formelles sont toutefois porteuses d'un intérêt assez prégnant quand elles sont applicables. Ces méthodes visent à résoudre de manière analytique un modèle complexe, c'est-à-dire à mettre en équations les comportements du modèle, leurs effets d'interaction, et à résoudre ces équations pour en proposer les ensembles finis de solutions ou d'états. Cela requiert d'être en mesure de convertir un modèle exprimé dans un formalisme quelconque en un système d'équations dynamiques, et de parvenir en outre à résoudre l'ensemble de ce système. Dans l'évaluation de modèles au sens large, cette étape peut se révéler indispensable et assez directe, par exemple quand il apparaît

nécessaire d'évaluer un modèle basé sur la théorie des jeux, que l'on traite alors sous forme d'analyse de graphes.

Dans le cas plus spécifique des modèles à base d'agent, cas dans lequel nous nous inscrivons ici, la situation est plus difficile. On emploie généralement la modélisation à base d'agents parce qu'elle encourage une approche anthropomorphique, plus aisément compréhensible et requérant moins de connaissances mathématiques que d'autres approches, mais aussi car il est terriblement complexe d'exprimer des systèmes dotés de multiples interactions, qui plus est multi-scalaires, sous forme de réseaux d'équations. En un sens, on pourrait presque considérer qu'on fait appel à de la modélisation à base d'agents quand on ne peut mobiliser des modèles formels. Le processus qui tendrait à formaliser, mathématiquement, des modèles agents est alors intrinsèquement contre-intuitif et difficile, quand bien même certains pensent que ce n'est pas une fatalité mais une question de temps¹⁰.

Si la nature même de cet exercice implique vraisemblablement que peu s'y essayent, on notera tout de même que quelques auteurs (ZHANG 2011 ; GRAUWIN, GOFFETTE-NAGOT et JENSEN 2012)¹¹ sont parvenus à résoudre de manière analytique un modèle foncièrement pensé comme un modèle agent – en automate cellulaire en l'occurrence –, le modèle de Schelling (voir encadré 3.6). En dehors de l'intérêt que cela peut représenter pour la connaissance de ce modèle en particulier, rappelons tout de même que le modèle de Schelling a été énoncé à la fin des années 1960, que c'est un modèle particulièrement parcimonieux, et qu'il a tout de même fallu attendre le tournant des années 2010 afin d'y trouver une solution formelle. Notons enfin que pour certains auteurs, dont l'un des pères de la modélisation à base d'agents informatiques EPSTEIN (2006), la résolution analytique de modèles de simulation à base d'agents n'est pas véritablement un enjeu, l'objectif étant d'utiliser le paradigme le plus « éclairant » pour un problème donné :

« The oft-claimed distinction between computational agent models, and equation-based models is illusory. Every agent model is, after all, a computer program (typically coded in a structured or object-

10. Par exemple Alain FRANC, mathématicien dans le projet TransMonDyn : « L'une des difficultés de l'acceptation des SMA comme modèles est que ces comportements sont très mal compris mathématiquement. Il existe peu de résultats qui permettent de relier un type de règles avec un type de comportement, alors que de tels liens sont à la base du succès des systèmes dynamiques, où l'on connaît (parfois...) les gammes de paramètres qui mènent à un comportement d'équilibre, cyclique ou chaotique, et l'on sait qu'il ne peut y en avoir d'autres. [...] Il existe donc une tension entre, d'un côté, les systèmes dynamiques qui forment une théorie riche et solide de modélisation mathématique, mais pour un nombre assez restreint de situations (bien des difficultés apparaissent dans le cadre non linéaire, que l'on peut lire dans la richesse des travaux sur la modélisation de la turbulence par exemple) et, d'un autre côté, les SMA qui permettent des simulations à partir de règles plus riches et diversifiées, mais pour des résultats dont la compréhension mathématique très souvent nous échappe (il y a peu de théorèmes). On peut donc dire en résumé que les SMA sont « en avance » sur la compréhension mathématique des systèmes dynamiques et peuvent proposer des cas d'études aux mathématiciens. » (OURIACHI et al. 2018, Annexe 2, « Retour sur les SMA comme outil et cadre conceptuel de modélisation. », pp. 479-482)

11. Lena : regarder Axelrod et Axtell, dans leur article sur les 4 modes d'utilisation d'un SMA, article très maths. -> Je ne trouve pas cette référence.

oriented programming language). As such, each is clearly Turing computable (computable by a Turing machine). But, for every Turing machine, there is a unique corresponding and equivalent partial recursive function [see Hodel (1995)].

[...]

So, in principle, one could cast any agent-based computational model as an explicit set of mathematical formulas (recursive functions). In practice, these formulas might be extremely complex and difficult to interpret. But, speaking technically, they surely exist.[...]

In any case, the issue is not whether equivalent equations exist, but which representation (equations or programs) is most illuminating. »

EPSTEIN (2006, p. 1590-1591)

Notons tout de même une dernière piste, intermédiaire, qui permet d'approcher de l'analyse formelle de modèles à base d'agents. Un collectif de chercheurs a conçu un modèle descriptif et complexe de la participation électorale en partant des comportements individuels des électeurs (EDMONDS, LESSARD-PHILLIPS et FIELDHOUSE 2015 ; FIELDHOUSE, LESSARD-PHILLIPS et EDMONDS 2016, « Modèle 1 »). En accord avec les principes de la modélisation KIDS (EDMONDS et MOSS 2005), proposé par l'un des membres de ce collectif, les chercheurs ont ensuite procédé à une large simplification du modèle, notamment en réduisant une partie de ses aspects les plus spécifiques (réseaux sociaux des électeurs et différenciation des partis politiques entre autre) (LAFUERZA et al. 2016b, « Modèle 2 »). En repartant de cette version parcimonieuse du modèle, les chercheurs ont alors re-construit une nouvelle version du modèle, encore plus parcimonieuse et analysable de manière formelle (LAFUERZA et al. 2016a, « Modèle 3 »). Après avoir démontré les liens entre le modèle 1 et le modèle 2, puis entre le modèle 2 et le modèle 3, les auteurs ont pu montrer à l'aide de l'évaluation formelle du modèle 3 que certains mécanismes du modèle 1, empiriquement pensés importants, n'avaient en fait qu'un effet très modéré. LAFUERZA et al. (2016a) s'appuient sur cette expérience pour démontrer l'utilité de la méthode KIDS vis-à-vis du respect qu'elle permet de conserver vis-à-vis des connaissances expertes :

« One of the most compelling [advantages of this method] is that it combines the best of two worlds : the simplicity appreciated by those trained in the physical sciences, but having an input from the many effects included in complex models. A central point is that, although the models constructed through this procedure are 'simple', in the sense that they have far fewer parameters than the models they are derived from and are more amenable to analysis, they will typically have features that would not have been guessed at if one started from simple models and then added further complexity. This is the strength of the approach : Model 1 contains within it a large amount of social science data and expertise, and a diluted form of this is retained in Model 3. »

LAFUERZA et al. (2016a, p. 6)

Dans l'absolu, une telle expérience demande une énorme quantité de travail,

et sans aller jusque là, on peut chercher à approcher d'une validation formelle sur un modèle complexe de manière directe, c'est-à-dire sans passer par des modèles intermédiaires. Il est ainsi possible d'analyser le comportement d'un modèle de manière globale, c'est-à-dire en explorant l'ensemble de ses comportements possibles. On peut pour cela user de méthodes basées sur du calcul intensif qui visent à cartographier « l'espace des sorties » d'un modèle. C'est par exemple l'un des enjeux principaux, en matière de recherche, d'une plate-forme telle qu'OpenMOLE (REUILLON, LECLAIRE et REY-COYREHOURCQ 2013), dont une partie des algorithmes (par exemple CHÉREL, COTTINEAU et REUILLON 2015) cherche à traverser l'espace des sorties de la manière la plus efficiente possible, c'est-à-dire en cherchant à réduire le nombre de combinaisons de paramètres possible – gigantesque en raison de l'explosion combinatoire – via des solutions d'optimisation.

Quelle évaluation pour quels modèles ? Les étapes d'évaluation énumérées ci-dessus consistent autant en une approche chronologique – relative aux phases successives de la construction d'un modèle – qu'en un gradient de qualité de l'évaluation, souvent considéré en fonction de la difficulté et du coût temporel nécessaire à chacune de ces méthodes¹². Il est évident à la lecture des auteurs de références du champ (par exemple ceux référencés en section 3.1) que pour eux, « plus » le modèle est évalué, c'est-à-dire se confronte aux étapes d'évaluation de plus en plus formelles, plus il sera digne de confiance et donc capable d'apporter des connaissances sur les objets qu'il tend à représenter. Robert SARGENT par exemple différencie les méthodes d'évaluation selon que le système modélisé est observable ou non, c'est-à-dire « s'il est possible ou non de collecter des données sur le comportement opérationnel de l'entité »¹³. Pour autant, ces auteurs soulignent aussi que selon les choix de modélisation et les caractéristiques du système modélisé, toutes ces étapes ne sont pas nécessairement accessibles ou possibles.

Nous pensons qu'un autre facteur peut affecter plus fortement l'éventail des méthodes possibles d'évaluation : la parcimonie du modèle réalisé. Ainsi, avec un modèle très parcimonieux, qui s'inscrirait dans un certain purisme des méthodes « KISS », doté d'un nombre minime d'*inputs* et d'*outputs*, il nous semble que toutes les méthodes, y compris les plus formelles, sont assez simplement – si ce n'est pour la résolution analytique, on l'a vu plus haut – applicables. A contrario, un modèle très descriptif, ancré dans une approche « KIDS », fourmillant d'*inputs*, de paramètres et d'*outputs* sera bien plus complexe à évaluer de manière quantitative, ou « objective » selon les mots des pionniers de l'évaluation.

Pour illustrer l'écart entre ces approches en matière de possibilités de quantification de l'évaluation, prenons l'exemple d'une analyse de sensibilité : cette

12. « [One] should start with cheap tests that allow fast rejection of the model and continue investing more and more effort when the model becomes more and more valid. », KLÜGL (2008, p. 42), par exemple.

13. « The major attribute affecting operational validity is whether the problem entity (or system) is observable, where observable means it is possible to collect data on the operational behavior of the program entity. », SARGENT (2009, p. 6).

technique consiste à faire co-varier les valeurs des paramètres afin d'observer les effets que, chacun ou conjoints, ils produisent sur les sorties du modèle. Avec un modèle de Schelling, dans lequel on identifie en général trois paramètres (cf. encadré 3.6), que l'on peut faire varier chacun selon une granularité de dix valeurs, et tenir compte de l'aléa en menant dix répliques, on peut mener une analyse de sensibilité basique au moyen de $(10^3 \times 10)$ 10 000 simulations. Dans le cas d'un modèle doté d'une dizaine de paramètres, et avec le même type d'analyse basique, le nombre de simulations nécessaire dépasserait déjà le milliard...

Pour de tels modèles, malgré tout assez peu complexes au regard de certains des tenants du genre KIDS, une analyse de sensibilité rigoureuse ou un calibrage fin ne sont en aucun cas envisageables selon les canons méthodologiques de l'évaluation. Dans ce cas, les théoriciens de l'évaluation recommandent, à défaut de mieux, de tout de même mener les premières étapes d'un cycle d'évaluation (PETTY 2010, p. 342) : « While moving beyond face validation to more objective and quantitative methods should always be a goal, face validation is clearly preferable to no validation at all. »

Nous ne partageons pas la réticence associée à cette recommandation. Au contraire, nous considérons que dans ce type de cas, des méthodes de « *face validation* » peuvent être utiles et suffisantes pour évaluer un modèle de simulation. Une des conditions est que ces méthodes soient menées de façon systématique et en suivant un protocole précis. Après avoir défini de manière plus approfondie ce qu'est la face validation, nous formulerons ensuite une proposition d'un tel protocole, intitulé « évaluation visuelle ».

3.1.3 Une évaluation de la plausibilité d'un modèle : la « *face validation* »

Avant d'aller plus avant dans la justification de l'utilité des méthodes de *face validation*, il convient de définir plus précisément ce à quoi la littérature réfère quand elle préconise cette méthode d'analyse de plausibilité d'un modèle.

3.1.3.1 Définition

Le terme semble avoir émergé dans les années 1940, en particulier dans le champ scientifique de la psychologie et des **études pédagogiques** (NEVO 1985). Concept discuté et disputé dans ces domaines (MOSIER 1947), on y attribue un besoin pour les modèles, statistiques dans ce cas, de présenter à la fois une validité à l'épreuve des données, mais aussi de présenter une apparence de validité, c'est-à-dire de sembler plausibles¹⁴.

14. « In this usage, the term “face validity” implies that a test which is to be used in a practical situation should, in addition to having pragmatic or statistical validity, appear practical, pertinent and related to the purpose of the test as well; i.e., it should not only be valid but it should also appear valid. This usage of the term assumes that “face validity” is not validity in any usual sense of the word but merely an additional attribute of the test which is highly desirable in certain situations. » MOSIER (1947, p. 192)

Pour illustrer ce besoin de « plausibilité », on peut prendre l'exemple des problèmes de corrélations fallacieuses (ou « *spurious correlations* »). À la suite d'un article (SHAW 2017) liant utilisation de glyphosate et nombre d'enfants diagnostiqués autistes (figure 3.4), plusieurs chercheurs ont renvoyé à l'analyse menée par un membre de l'espace de discussion *reddit*¹⁵. Celui-ci qui proposait en effet – de manière ironique – une explication opposée, liant prévalence de l'autisme et vente de produits de l'agriculture biologique (figure 3.5).

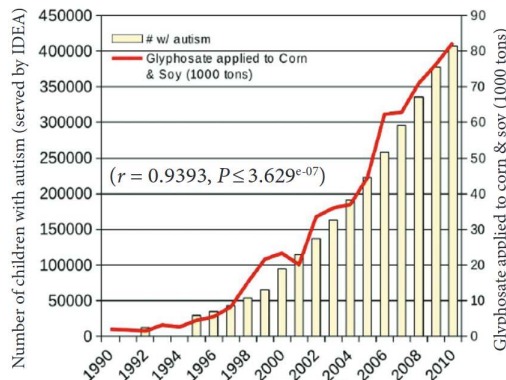


FIGURE 3.4 – Relation entre autisme et utilisation de glyphosate, d'après SHAW 2017, Figure 2, p. 51

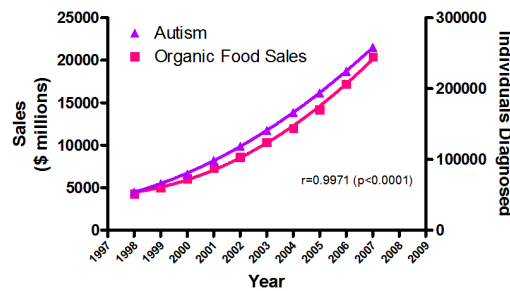


FIGURE 3.5 – Relation entre autisme et vente d'aliments « bio », d'après « jasonp55 », 2012.

En termes de qualité de l'ajustement, le second modèle est meilleur : son coefficient de corrélation est de 0.997 quand celui du premier ne vaut que 0.939. En dehors du principe statistique qui veut que causalité et corrélation ne soient pas équivalents, cet exemple illustre la nécessité d'apporter un éclairage en termes de plausibilité dans l'évaluation. Tout épidémiologiste pourra en effet rejeter les conclusions du second graphique en se basant sur son intuition face à l'absence de relation empirique, au niveau individuel, entre autisme et alimentation biologique. Cet exemple, trivial, illustre le besoin d'une évaluation basée sur la plausibilité, quand bien même une mesure de la validité aurait déjà été effectuée et jugée concluante.

C'est à cette évaluation de plausibilité que correspond, dans les figures 3.1 et 3.2 notamment, l'étape de « *face validation* ». L'utilisation très polysémique – et donc contradictoire – de ce terme, et les importants débats autour de son usage ayant poussé à sa désuétude¹⁶, on le retrouve pourtant au cœur de l'un des articles fondateurs de l'évaluation de modèles de simulation, où HERMANN le définit ainsi :

« Face validity is a surface or initial impression of a simulation or game's realism. Probably no approach to model validity is reported more frequently than the subjective estimates of experimenters, observers, or human participants as to the correspondence between

15. Message posté par l'utilisateur « jasonp55 » en 2012 : www.reddit.com/r/skeptic/comments/14qbn9/

16. MOSIER 1947, p. 205 recommande même son abandon : « Since the term "face validity" has become overlaid with a high degree of emotional content and since its referents are not only highly ambiguous but lead to widely divergent conclusions, it is recommended that the term be abandoned. »

the model's operation and their perception of the actual phenomena which the game or simulation represent. [...]

Face validity can be a significant part of a validity strategy. A quick impression that "things don't seem right" may be the only validity check possible during the actual operation of a game or simulation. Such validity judgments and their evaluation may also be part of the learning experience provided by operating models designed for instructional purposes. »

HERMANN (1967, p. 221)

Quelques années plus tard, on en trouve une définition plus succincte chez un des fondateurs de l'évaluation de modèles : « Face validity is asking people knowledgeable about the system whether the model is reasonable. » (SARGENT 1979, p. 500). Cette définition introduit un aspect qui nous semble important en matière de *face validation* : il ne s'agit pas de faire évaluer la plausibilité d'un modèle par un quelconque examinateur, mais bel et bien par un expert du sujet modélisé¹⁷. Ce type d'évaluation n'a donc pas uniquement vocation à démasquer des comportements contre-intuitifs, mais bel et bien à faire expertiser, par un thématicien, le déroulement et l'aboutissement d'un modèle de simulation.

Il a fallu attendre la relative démocratisation des plate-formes de modélisation à base d'agents pour qu'une auteure, KLÜGL, se penche véritablement sur l'identification et l'explicitation de la *face validity*, et en donne une définition plus précise, mais englobante car centrée sur les usages plus que sur la méthode en elle-même :

« Face validity can be seen as the result of face validation. Under this paradigm I want to subsume all methods that rely on natural human intelligence. Examples are structured walk-through, expert assessments of descriptions, animations or results. Thus, face validity shows that processes and outcomes are reasonable and plausible within the frame of theoretic basis and implicit knowledge of system experts or stake-holder. Face validation may be applied from the early phases of the simulation study under the umbrella of conceptual validations. It is often also called plausibility checking. [...]

Face validation usually plays an important role during model design. All tests based on reviews, audits, involving presentation and justification of assumptions and model structure are used for reaching this form of plausibility. »

KLÜGL (2008, p. 39–41)

La description des méthodes possibles menant à cette évaluation n'est pas en reste non plus dans cet article, puisque l'auteur identifie trois familles de cette

17. On retrouve l'expression de « people knowledgeable about the system » chez BALCI (1994, p. 130), et KENNEDY et al. (2006, p. 2) parlent de « domain experts ».

face validation, chacune pouvant être menée par des experts différents¹⁸ :

Composantes de la *face validation*.

- **Évaluation du déroulement.** Ce type d'évaluation vise à analyser le déroulement d'une simulation dans son ensemble. Il s'agit ici de juger de la plausibilité des dynamiques (à l'échelle du système dans son ensemble, ou de composantes de celui-ci) reproduites dans la simulation, via une observation en direct de la simulation.
- **Évaluation des sorties.** Cette approche consiste plutôt à une évaluation qualitative des sorties produites par la simulation. Cela peut prendre la forme de vérifications des valeurs (approche que l'on retrouve dans les méthodes d'évaluation plus formelles, via une automatisation de ces types d'évaluation) par un expert, mais aussi d'analyse des covariations et évolutions temporelles de différents indicateurs de sortie. L'évaluation des sorties peut être appliquée sur le système modélisé dans son ensemble, mais aussi au niveau des types d'agents mobilisés.
- **Évaluation « immersive ».** Il s'agit ici d'évaluer le modèle au travers de la vraisemblance des actions et réactions individuelles des agents qui y interagissent. L'accent est donc mis sur la plausibilité du comportement des agents (niveau micro), plus que sur celle des dynamiques macroscopiques résultantes. Les experts de ces deux niveaux d'observation peuvent être différents (un psychologue spécialiste des réactions individuelles en cas d'incident ne peut porter un jugement de même niveau qu'un physicien spécialisé dans les dynamiques de foules par exemple), et il faut donc, à chaque niveau d'observation du modèle, faire intervenir un expert adéquat.

Pour l'auteur, ces trois approches d'évaluation sont complémentaires et s'inscrivent dans des temporalités différentes de la phase de vie du modèle. Elle encourage ainsi plutôt à mener l'évaluation des sorties après les deux autres, puisque ces dernières sont comparativement moins coûteuses en termes de calcul (KLÜGL 2008, p. 42).

Il nous semble que si les deux premières approches sont applicables à tout modèle, l'évaluation immersive comporte un postulat lourd sur la plausibilité des trajectoires individuelles. Cela se prête bien à de nombreux modèles où les agents représentent des humains dotés de comportements rationnels, ou encore des particules dont la trajectoire individuelle est prévisible en dehors des effets d'interaction. Toutefois, tout un pan de la modélisation en géographie repose sur des agents non anthropomorphiques, ou encore sur des entités primaires dont seules les interactions ont vocation à faire émerger un comportement d'ensemble. Dans le cas de SimFeodal par exemple (ref chap2), les comportements individuels des foyers paysans ne reposent pas sur des hypothèses de vraisemblance : le foyer paysan qui se déplace de villes en villes, parfois en faisant des allers-retours, au cours des 300 ans modélisés, ne s'appuie sur aucune connaissance empirique, et tendrait même à contrevenir aux

18. L'énumération qui suit est une traduction libre et une reformulation partielle de KLÜGL (2008, p. 41-42)

connaissances expertes de la mobilité résidentielle des foyers paysans médiévaux. On pourrait y voir une « méta-plausibilité » inter-générationnelle, mais il serait difficile de l'interpréter et surtout de la différencier d'autres comportements de foyers paysans. Le suivi d'un foyer paysan, isolé de ses co-agents, au cours du déroulement du modèle, par un expert thématique, n'est donc pas sujet à évaluation, au contraire du suivi des structures spatiales de niveau macroscopique engendrées par cette accumulation de déplacements.

L'évaluation immersive, bien que peu adaptée à certains types de modèles, peut toutefois s'avérer de manière universelle utile en matière d'évaluation interne, dans un aspect de « débogage ». Même si les réactions et attributs des agents ne reproduisent pas une connaissance experte, leur observation peut toujours servir au modélisateur pour vérifier l'absence de valeurs aberrantes ou encore la juste activation de chacun des mécanismes.

3.1.3.2 Limites

Comme mentionné auparavant (3.1.2), pour de nombreux auteurs (HERMANN 1967 ; BALCI 1994 ; KENNEDY et al. 2006), la *face validation* ne peut qu'être une étape préalable des méthodes d'évaluation plus quantitatives et formelles. Les raisons données sont souvent le manque d'objectivité d'une démarche fondamentalement basée sur l'expertise et l'impression. Parmi ces auteurs, HERMANN est sans doute celui qui se montre le plus méfiant vis-à-vis de la pratique de la *face validation*, en en pointant plusieurs limites :

« Although face validity has value in the early stages of model building or for quick checks during actual operation, its severe limitations should be recognized. Sometimes the experimenter will not know what behaviors are "realistic" because of his limited experience observing the actual phenomena. Participants can become interested and highly motivated in an incorrect representation of the desired environment. If the simulation involves the substitution of one property for another, some features may appear quite unreal and yet replicate the performance of the reference system for which the simulation was designed. The acceptance of face validity as a rough, first approximation might be improved if the simulator explicitly stated in advance what observations would constitute indications that an aspect of the observable universe had been successfully captured. In summary, face validity in its usual form suffers from the lack of explicit validity criteria. »

HERMANN (1967, p. 222)

Ces réserves nous semblent être autant de pistes pour justifier de l'intérêt d'une démarche scientifique de *face validation*. En reprenant les critiques dans l'ordre énoncé par l'auteur, on peut y répondre ainsi :

- **Manque de connaissance experte.** Cette première remarque nous apparaît comme quelque peu biaisée : si l'on confie une évaluation experte à des non experts, naturellement, cela ne peut déboucher sur une évaluation correcte du modèle. Cela est d'ailleurs applicable quelle que soit

la méthode d'évaluation : une expertise ne vaut que par la qualité de l'expert. De manière plus nuancée, on notera d'ailleurs que cette phrase montre ici l'absence d'un élément de définition de la *face validation* partagé par les autres auteurs : HERMANN considère par là que c'est au modélisateur uniquement de mener cette phase d'évaluation, alors que la littérature s'entend quant au fait que ce rôle échoit à des experts. Ce faisant, HERMANN se positionne dans la logique de construction de modèles par des modélisateurs, sans apport des thématiciens, et donc dans l'approche classique de séparation forte entre ces deux acteurs indispensables du modèle (voir **chapitre 1, prestation vs co-construction**).

- **Invraisemblance de certains comportements.** HERMANN met en avant que dans un modèle, tous les mécanismes n'ont pas vocation à être vraisemblables. Ainsi, en mentionnant ces « propriétés de substitutions », il rappelle un élément important d'une évaluation, quelle qu'en soit la méthode. On ne doit et ne peut en effet juger de la plausibilité que des aspects du modèle qui cherchent à reproduire un comportement plausible. Il nous semble qu'ici aussi, la critique de l'auteur revient à ignorer l'importance du dialogue entre modélisateur et évaluateur, tout en présumant que l'évaluateur ne serait pas le modélisateur. Si le modélisateur connaît les « substitutions » opérées dans le modèle, il se gardera donc bien de juger de leur vraisemblance. *A contrario*, un expert thématicien pourrait être étonné par certains comportements micro, dans la mesure où il ne connaîtrait pas les correspondances entre éléments du modèle et éléments du système modélisé. Là encore, cette limite repose donc surtout sur le choix d'un mode de construction isolé, c'est-à-dire n'impliquant pas et le thématicien et le modélisateur. Plus généralement, cette limite est présente s'il n'y a pas d'explicitation des « substitutions » implémentées, c'est-à-dire de la correspondance entre le domaine conceptuel et le domaine du modèle implémenté : le niveau attendu de plausibilité individuelle de chaque comportement doit être décidé pendant la construction du modèle.
- **Explicitation préalable des objectifs.** La dernière remarque de cette citation nous semble, sans conteste, être la plus importante et la plus juste. L'auteur note ainsi que la *face validation* ne peut constituer une méthode d'évaluation adaptée si l'on ne spécifie pas, en amont, les critères qu'elle doit s'attacher à examiner. C'est là encore vrai de toutes les méthodes d'évaluation, mais nous souscrivons aux remarques de HERMANN quant à l'importance primordiale que cela revêt pour la *face validation*. En matière de plausibilité, on pourrait ainsi, comme cela nous semble souvent être le cas, se contenter d'évaluer « à chaud » les différentes dynamiques et sorties d'un modèle, sans s'encombrer d'une démarche, ou feuille de route, spécifique. Le risque est alors d'introduire encore plus de subjectivité dans cette analyse, et en particulier de briser la capacité de reproductibilité ou de justification d'une évaluation : une évaluation peut être subjective tout en étant justifiée, appuyée par des arguments, et dès lors, reproductible si tant est que chacun de ces éléments soit explicités. Quand un modèle est évalué par une seule personne, par exemple

un expert thématique, la nécessité d'une telle démarche est peu visible, chacun étant en capacité d'estimer qu'il sera en mesure de justifier *a posteriori* son évaluation. A contrario, quand un modèle résulte d'un travail collaboratif, qui plus est quand il implique plusieurs évaluateurs, les évaluations d'un même résultat peuvent varier. Il est donc indispensable de les expliciter autant que possible, et pour se prémunir d'un travail gigantesque d'analyse postérieure des résultats tout autant que pour se doter d'un outil de discussion et de débat commun, il apparaît primordial de fixer une grille d'évaluation, ou, en d'autres termes, d'un ensemble de critères à observer. Cela ne limite aucunement la nécessaire subjectivité et complémentarité des évaluateurs experts, mais permet au contraire d'inscrire leurs discours dans un référentiel de comparabilité.



3.1.3.3 Intérêts de la *face validation*

En dépit des limites identifiées ci-dessus, qui nous semblent surpassables à condition de définir une grille d'évaluation avec précision en amont, la *face validation* présente de nombreux atouts en dehors de la facilité de sa mise en œuvre traditionnelle.

Là où HERMANN et les auteurs classiques cantonnent la *face validation* à une étape préalable à une véritable évaluation, KLÜGL justifie l'intérêt propre de cette démarche méthodologique, y compris dans les phases plus avancées de la démarche classique d'évaluation :

« One may argue why face validity is need, when statistical validation is successfully done ? Face validation assures that the processes and structures are reasonable for a human expert. Especially, when there is (semi-)automatic calibration of a simulation that is used in combination with statistical validation, a careful check of plausibility is necessary. This is in general true for all kinds of simulation, but it is particularly important for agent-based simulations.[...]

Although face validation may be informal and inconsistent, but it at least results in plausibility of modeled processes. Our experience with modeling and simulation in many interdisciplinary projects showed that even the formulation of a plausible model supports theory building and future empirical research. »

KLÜGL (2008, p. 40;43)

Pour l'auteure, la face validation complète ainsi d'autres méthodes d'évaluations mieux considérées, et nous irons même plus loin en considérant que chacune de ces méthodes peut potentiellement être améliorée en la conjuguant à une analyse de plausibilité visuelle de ses résultats.



On notera d'ailleurs que chez BALCI, dans une analyse de l'applicabilité des méthodes de « VV&T » (*Validation, Verification and Testing techniques*) aux différentes phases du cycle de vie d'un modèle, seules 5 techniques¹⁹ (sur 77

19. Il s'agit systématiquement de méthodes « informelles » : (1) La vérification de la docu-



analysées) présentent la caractéristique d'être mobilisables à chacune de ces phases. Pour obtenir cette information, nous avons croisé plusieurs travaux de BALCI. Dans un premier article (BALCI 1997), cet auteur présente toutes les méthodes de VV&T qu'il a identifiées, et il les catégorise en quatre catégories. Chaque méthode peut ainsi être formelle, informelle, statique ou dynamique. Nous avons tiré d'un second article (BALCI 1998) un tableau présentant l'applicabilité de chacune de ces méthodes à l'une des 18 étapes du « cycle de vie » des modèles de simulation qu'il identifie. En numérisant ces deux informations, relatives aux catégories des méthodes et à leur applicabilité, nous avons pu créer une représentation graphique (figure 3.6) qui les croise.

mentation (*Documentation Checking*); (2) la *Face Validation*; (3) les inspections de code; (4) les « revues » de modèle (*Reviews*) et enfin (5) les « procédures pas à pas » d'évaluation (*walk-through*).

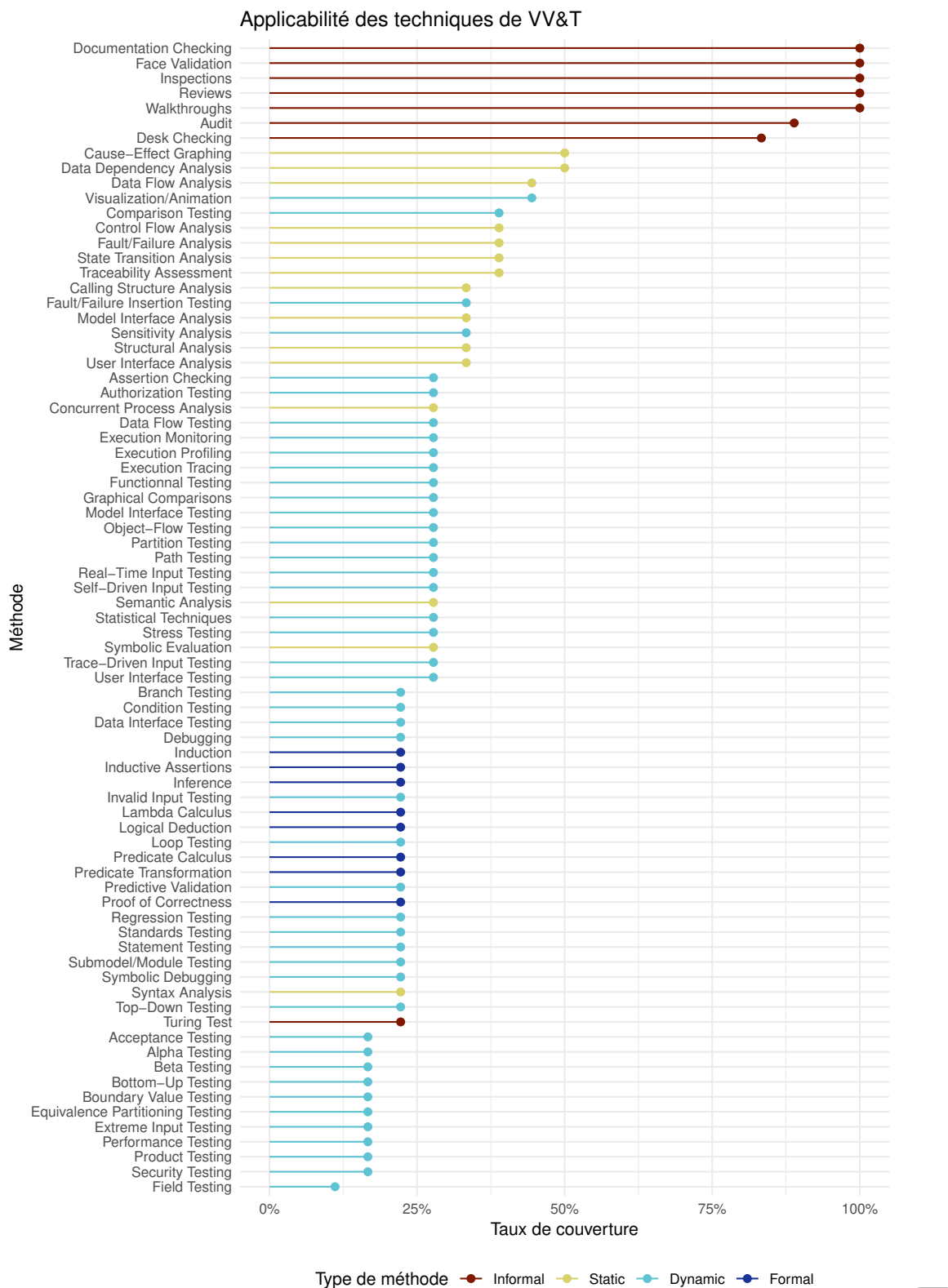


FIGURE 3.6 – Part des étapes de la cycle de vie d'un modèle pour lesquelles différentes méthodes de « VV&T » peuvent être mobilisées, selon le type de méthode (intitulés d'origine).

Taxonomie des méthodes : d'après BALCI (1997, Figure 2, p. 139);

Taux de couverture : d'après BALCI (1998, Table 3, pp. 45-47).

Cette figure fait apparaître très nettement une « hiérarchie » d'applicabilité des méthodes selon leur type : les méthodes informelles sont les plus souvent mobilisables, suivies par les méthodes statiques et certaines méthodes statiques.

3.1.4 Vers une évaluation visuelle

Au regard de ces éléments, il nous semble que l'utilité des approches de *face validation* est assez largement prouvée. C'est le cas que cette approche soit cantonnée à une phase préalable mineure ou au contraire à une bonne pratique plus générale, à mener lors de chacune des phases de construction et d'évaluation d'un modèle.

3.1.4.1 Une démarche comparative

Comme identifié en limites et en intérêts, il nous semble de plus que la *face validation* souffre, en matière de réputation, d'un manque de clarification de la démarche qu'elle met en œuvre. ~~Nous considérons que~~ dotée d'un protocole d'évaluation rigoureux, cette méthode peut constituer une alternative crédible à des méthodes d'évaluation plus répandues, par exemple les méthodes statistiques. Ces dernières sont souvent basées sur l'analyse de l'écart entre des données empiriques et des données simulées, et cherchent à quantifier et à minimiser cet écart. La *face validation* procède à la même démarche comparative :

« Face validation is a validation method that compares simuland behavior to model results. [...] Based on their knowledge of the simuland, the observers subjectively compare the behavior of the simuland as reflected in the simulation results with their knowledge of the behavior of the actual simuland under the same conditions, and judge whether the former is acceptably accurate. Differences between the simulation results and the experts' expectations may indicate model accuracy issues. PETTY (2010, p. 341)

Nous postulons que la démarche comparative que l'on retrouve dans l'évaluation statistique peut donc être appliquée sans recherche de quantification, c'est-à-dire en évaluant ces mêmes écarts de manière subjective et qualitative. Dans les modèles descriptifs dotés de nombreux indicateurs non résumables, tel que SimFeodal (ref chap 2), la comparaison terme à terme entre des valeurs numériques est potentiellement possible, mais non ordonnable et ne peut donc déboucher sur un indicateur unique (voir Quels types d'indicateurs pour SimFeodal?, section 3.2.1.2). De plus, les données empiriques qui permettraient de mener une comparaison sont trop lacunaires et incertaines pour être jugées suffisamment fiables pour évaluer le modèle.

3.1.4.2 Une démarche rigoureuse

Nous pensons toutefois que ces carences quantitatives peuvent être compensées par les connaissances expertes des différents thématiciens impliqués dans la construction et l'évaluation d'un modèle. Ainsi, dès lors que le processus d'évaluation est pensé en amont de son application, et qu'il est possible de parvenir à la création d'une grille d'analyse, c'est-à-dire à un protocole d'évaluation, il nous semble que la différence avec l'évaluation statistique est assez restreinte. Pour évaluer les modèles du type de SimFeodal, nous considérons dès lors qu'il est tout à fait possible de faire reposer cette démarche sur une

évaluation experte, s'inscrivant dans les logiques de *face validation*, et en particulier de sa composante d'évaluation des sorties (voir Composantes de la *face validation*, section 3.1.3.1).

3.1.4.3 Un nouveau terme ?

Dans la suite de cet ouvrage, nous nommerons cette approche « **évaluation visuelle** ». Ce terme n'est, à notre connaissance, que peu employé, et le semble surtout en études environnementales, par exemple pour définir une méthode de comptage d'espèces animales et végétales (par exemple HARMELIN-VIVIEN et al. 1985). Le pendant anglophone, la « *visual evaluation* », semble s'inscrire dans le même champ disciplinaire (par exemple HORST, ENGELKE et MEYERS 1984), et paraît aussi assez faiblement utilisée dans un usage scientifique. Dans l'usage qui en sera fait dans ce manuscrit, ce terme est forgé au regard de la « *face validation* » naturellement, et pourrait être confondu avec. Il s'agit toutefois de s'éloigner de l'aspect « apparence » présent dans le terme – qui insiste donc sur une validité de façade –, pour embrasser au contraire la méthode visuelle. Cette dernière a fait ses preuves – dans de nombreux autres champs disciplinaires – et constitue un pan non négligeable des méthodes d'analyse, et nous pensons donc adéquat d'en faire un usage argumenté dans le domaine de l'évaluation de modèles de simulations. On notera un usage proche de cette dernière, appliquée aux modèles aussi, mais statistiques cette fois-ci, visant à l'évaluation visuelle de modèles de « *Data Science* ». EILERS et al. (2017) rapportent le constat d'une utilité réelle de l'évaluation visuelle (intitulée « *Visual Model Evaluation* » dans leur cas), et mettent une emphase particulière dans l'intérêt des intuitions que peuvent avoir les experts thématiques à la vue des résultats d'un modèle. Pour ce faire, ils insistent sur l'intégration d'experts dans le processus d'évaluation, et sur le besoin de communications, lors de cette phase de travail, entre les modélisateurs et ces experts²⁰.

3.1.4.4 Définir l'évaluation visuelle

Pour définir cette évaluation visuelle, nous repartirons de la définition de la *face validation* sur laquelle cette méthode s'appuie.

Définition. Il s'agit d'évaluer, visuellement, la plausibilité du comportement d'un modèle à partir des données qu'il produit. Cette plausibilité peut être entendue comme la correspondance entre le système modélisé et le modèle du système, correspondance s'exprimant en comparant les données en sortie de simulation – et en les agrégeant au besoin pour tenir compte de la nécessaire

20. « Integrating these expert groups [data scientists and domain experts] to follow a common objective is still a major challenge today for a successful data science project in the industry and therefore a suitable field for information systems research. A collaborative analysis system addressing this issue should therefore focus on both aspects. It is important to most efficiently support human decision-makers with data-driven expert systems, and much research has been carried out in this area (Shim et al. 2002; Power 2008). But it is equally important that domain experts are also part of the system itself, e.g. by supporting data scientists with their domain knowledge when constructing the underlying models. A key success factor for this purpose is communication between different groups. » (EILERS et al. 2017, p. 2)

réplication²¹ – au comportement du système modélisé. Cette correspondance doit être qualifiée avant de mener cette phase d'évaluation, c'est-à-dire qu'il est nécessaire de spécifier les critères d'observation et les réponses attendues. Ces éléments, les critères d'évaluation, ne peuvent être formulés par n'importe qui : si le modélisateur autant qu'un expert externe peuvent les spécifier, il convient de s'assurer de l'expertise – thématique et de la connaissance du système tel que modélisé – de l'évaluateur. On obtient ainsi un système à évaluer au filtre d'une grille d'analyse qualitative et basée sur le visuel. Il devient alors possible d'apprécier l'écart le modèle et le système qu'il représente, sans chercher pour autant à quantifier ou à mesurer cet écart. Il s'agit en effet plutôt d'ordonner différentes versions ou paramétrages d'un modèle de simulation afin de juger de ceux qui semblent minimiser cet écart.

Un dernier point nous semble particulièrement appréciable dans ce recours à l'appréciation visuelle plutôt qu'à une mesure stricte de l'écart entre une situation estimée parfaite et une situation simulée. C'est la capacité, humaine, à estimer semblable des configurations spatiales qui seraient jugées très différentes par une méthode de calcul quelconque. Si l'on cherche par exemple à estimer l'écart entre un semis de point cible et un semis simulé, on peut mesurer l'écart comme, par exemple, une somme des écarts de chacun des points à celui qu'ils sont sensés représenter. Un décalage minime, par exemple obtenu par translation horizontales de quelques dizaines de mètre sur un semis régional, engendrera alors une démultiplication des erreurs, alors même qu'un œil humain aurait jugé ces deux semis quasi-identiques. En terme de configuration spatiale, la quantification peut ainsi amener à des contres interprétations dans l'analyse. Dans le cadre d'un modèle comme SimFeodal où l'espace est important et qui plus est, théorique et aléatoire, et donc difficilement agrégeable (cf. chapitre 7), le recours systématique à l'évaluation visuelle nous paraît alors indispensable.

Cette méthode, contrairement à d'autres, plus quantitatives, permet donc au final de tirer avantage des méthodes qualitatives telles que la *face validation* – par exemple la capacité d'évaluer un modèle qui ne reposerait que sur peu de données empiriques ou encore sur des données incertaines –, tout en se confortant à une démarche d'évaluation rigoureuse, loin de l'estimation « au doigt mouillé » à laquelle peuvent donner lieu certaines méthodes reposant sur la plausibilité et l'estimation.

21. Cet aspect est discuté dans les chapitres 1, 2 et 7. En matière d'évaluation, tel que pointé par la majorité des auteurs cités dans cette sous-partie de chapitre, il est ainsi nécessaire de tenir compte de la variabilité d'un modèle, variabilité intrinsèque dans un modèle stochastique. Il n'est donc pas possible d'évaluer, visuellement ou non, un modèle stochastique sur la base d'une seule exécution. Au contraire, seule l'exécution d'un certain nombre de **réplications** (voir chapitre 1) permet de s'assurer que le comportement évalué correspond bien au comportement habituel, ou tendanciel, du modèle.

3.1.4.5 Des critères pour l'évaluation visuelle : construire des indicateurs de sortie de simulation

Dans la présentation de la démarche d'évaluation visuelle, nous précisions que pour que cette évaluation qualitative et experte soit rigoureuse, il était nécessaire de fixer des objectifs de manière préalable, et d'en expliciter la teneur autant que possible.



Pour l'évaluation de SimFeodal, face à la multiplicité des attentes thématiques vis-à-vis du modèle, nous avons choisi de mobiliser à cet effet des « indicateurs de sortie ». Ceux-ci relèvent du domaine de la simulation et leur évaluation doit être guidée par les connaissances empiriques, formalisées au sein d'« indices empiriques » qui correspondent à ces indicateurs.

Pour finir de décrire la démarche d'évaluation de SimFeodal, il reste donc à définir plus précisément ces composantes de l'évaluation, ainsi qu'à expliciter les objectifs fixés pour chacun des indicateurs que l'on va construire.

3.2 Une grille d'analyse composée d'indicateurs de sortie

Le modèle SimFeodal présenté dans le **chapitre 2** correspond à la « version 6.3 » du modèle, c'est-à-dire qu'il en constitue une version qui n'est ni la première, ni sans doute la dernière dans cette expérience de co-construction interdisciplinaire de modèle qui s'inscrit résolument dans le temps long. L'ensemble des mécanismes figurant dans le modèle conceptuel ont été implémentés mais l'ensemble des liens, interactions et valeurs de paramètres ne sont pas encore stabilisés. De ce fait les résultats des simulations ne répondent pas toujours complètement aux attentes définies dans le **chapitre 2**.

Si l'on a déjà décrit le principal objectif du modèle dans le chapitre précédent (celui de comprendre les mécanismes sous-jacents au processus de polarisation qui s'est déroulé entre 800 et 1100), il convient ici d'explicitier comment les résultats d'un tel processus peuvent être saisis. Ceux-ci sont en effet nombreux et hétérogènes, concernant aussi bien des concentrations de foyers paysans que l'émergence de pôles. Certains sont centraux, d'autres secondaires, et le modélisateur a des attentes relativement aux résultats qui devraient être obtenus en fin de simulation. La description précise de ces attentes se révèle importante dans le cadre du paramétrage – et de l'ensemble des étapes de la vie du modèle – de SimFeodal.

Une telle description repose sur la construction d'« indices empiriques » et d'« indicateurs de sortie de simulation » qui vont permettre de rendre compte des résultats des simulations. Il s'agira d'exprimer les attentes sous formes de critères relatifs à ces indicateurs de sortie de simulation.

Dans cette partie, on explicitera d'abord le sens que l'on prête à ces « indices empiriques » et « d'indicateurs de sortie de simulation ». Ces indices et indicateurs sont nombreux, certains sont multivariés, et il s'agira donc de présenter des méthodes visant à réduire la complexité de ces indicateurs de sortie, en adoptant une démarche proche de ce qui se fait en statistiques : réduction de dimensionnalité et/ou catégorisation et hiérarchisation de ces indicateurs. L'utilisation de ces méthodes permettra, seule, de décrire et qualifier le comportement du modèle SimFeodal tel qu'il a été décrit dans le chapitre précédent, avant d'en analyser les résultats par ce biais dans le **chapitre 6**.

3.2.1 Indices et indicateurs

~~On attend d'un modèle, sans entrer encore dans le détail, qu'il reproduise au moins les grands traits de l'élément empirique dont il est sensé rendre compte. Ces grands traits peuvent s'entendre de multiples manières, et se formaliser avec encore plus d'approches. Ici, nous avons souhaité proposer une dichotomie simple entre le domaine de l'empirique et celui de la simulation, en systématisant l'usage d'un vocabulaire qui est souvent employé de manière plurielle.~~ Pour être en mesure d'évaluer la vraisemblance du comportement reproduit par le modèle sur le plan empirique, il est nécessaire de mettre en

correspondance des éléments empiriques et des éléments issus de la simulation. Nous caractérisons ces éléments en deux grands ensembles : (1) **les indices empiriques**, éléments quantifiables ou au moins descriptibles émanant du domaine empirique, et (2) **les indicateurs de sortie**, variables informatiques produites par le modèle de simulation et devant pouvoir être comparés à chacun des indices empiriques. La figure 3.7 reprend, sous forme de schéma ontologique synthétique, ces deux ensembles de mesures, explicitant le vocabulaire mobilisé dans cette partie.

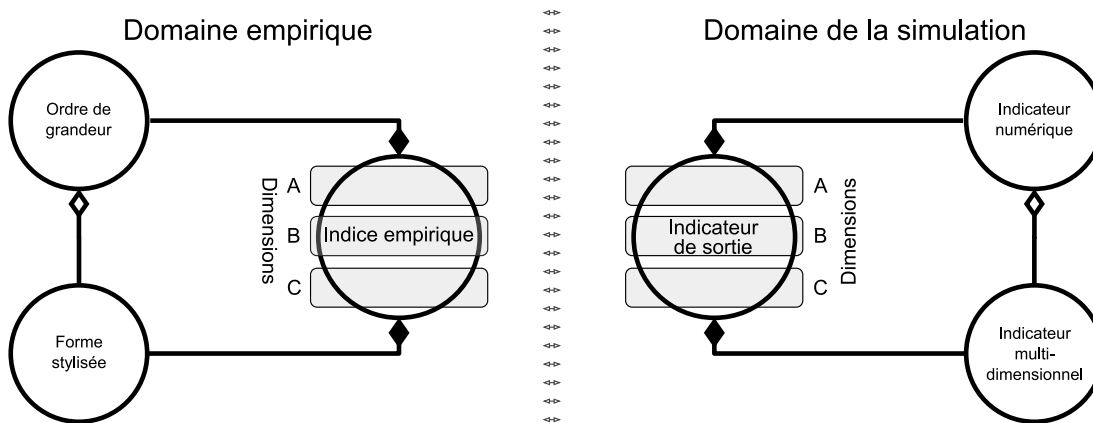


FIGURE 3.7 – Schéma de synthèse des correspondances entre mesures relevant du domaine empirique et mesures issues des simulations pour l'évaluation du modèle SimFeodal

Dans cette figure, on représente au moyen d'une symétrie axiale la correspondance entre le **domaine empirique** et le **domaine de la simulation**. Les **indicateurs de sortie** sont donc la correspondance simulée des **indices** du domaine empirique. Indices et indicateurs de sortie sont catégorisés selon des **dimensions** (A, B, C) qui correspondent aux processus que l'on cherche à modéliser : la polarisation du système de peuplement, sa hiérarchisation, sa fixation, etc. Les dimensions sont donc composées²² d'un ensemble d'indices, pour la partie empirique, et des indicateurs de sortie de simulation correspondants.

Les indices empiriques peuvent être de deux formes. Nous nommons la forme la plus simple « **ordre de grandeur** ». Ces indices expriment donc une valeur observée, connue avec plus ou moins de certitude (d'où le terme d'ordre de grandeur), qui trouvera une correspondance dans le domaine simulé sous la forme d'un **indicateur numérique**. D'autres indices ont une forme plus complexe, reposant sur l'agrégation²³ de plusieurs ordres de grandeur, et on les nomme « formes stylisés », par analogie aux faits stylisés déjà décrits. Dans le domaine de la simulation, les indicateurs de sortie correspondant sont appelés « **indicateurs multidimensionnels** », pour appuyer l'idée qu'ils sont constitués de plusieurs indicateurs numériques.

On peut illustrer cela avec un exemple tiré de SimFeodal. Dans le domaine empirique, on sait que la part de l'habitat rural dispersé tend à diminuer au cours

22. Les losanges noirs marquent une relation de composition dans le formalisme UML, repris dans ce schéma. La position du losange indique le sens de la relation : les indices sont composés d'ordres de grandeur et de formes stylisées.

23. En UML, un losange vide symbolise les relations d'agrégation : plusieurs ordres de grandeurs sont nécessaires pour définir une forme stylisée.

de la période, depuis environ 95% jusqu'à environ 20%. C'est notamment l'un des indices du processus de polarisation, qui constitue une dimension d'analyse. Ces deux valeurs sont des indices empiriques, de type « ordre de grandeur ». Dans le modèle de simulation, on a un indicateur de sortie correspondant qui correspond au taux de foyers paysans isolés. Le taux de foyers paysans isolés en début de simulation est un indicateur numérique, de même que celui en fin de simulation. La forme de la courbe d'évolution du taux de foyers paysans isolés au cours de la simulation est un indicateur multi-dimensionnel : elle repose sur une succession d'indicateurs numériques que sont le relevé du taux à chaque pas de temps. Cette forme de courbe peut être évalué au regard de la forme stylisée, dans le domaine empirique, sur laquelle est s'appuie. On sait que la part de l'habitat rural diminue régulièrement au cours du temps, et on évaluera donc, pour cette dimension (la polarisation), la capacité du modèle à reproduire un processus de polarisation semblable à celui observé par le biais de la forme stylisé « diminution régulière de la part de l'habitat rural dispersé ».




3.2.1.1 Les indices empiriques.


Afin d'évaluer la capacité du modèle à reproduire un phénomène observé, il est nécessaire de disposer dans le domaine empirique, de « points de repère ». Selon les modèles, ceux-ci peuvent revêtir de multiples formes et relever de l'ensemble des échelles spatiales et temporelles que l'on choisit de mettre en scène dans le modèle. Leur point commun est qu'ils doivent pouvoir être mesurés, au sens le plus large, c'est-à-dire être en capacité d'être reproduits et comparables avec d'autres mesures. Dans cette étude, on a décidé de qualifier ces points de repère d'« **indices empiriques** » et de les regrouper en deux catégories basées sur la précision avec laquelle ils peuvent être décrits et non sur la précision de leur connaissance, cf. section 3.2.2.1. La figure 3.7 illustre cette catégorisation entre la première catégorie – les ordres de grandeur – et la seconde – les formes stylisées –.



Ordres de grandeur. La première catégorie est constituée d'**ordres de grandeurs** empiriques estimés – ~~avec une précision plus ou moins importante~~ (cf. tableau 3 p. 317 du chap. TMD, à reproduire dans chap 2). Certaines valeurs empiriques sont ~~ainsi~~ connues, que ce soit d'après des sources primaires ou secondaires, et peuvent ainsi constituer des indices. Par exemple, on connaît avec quasi-certitude le nombre d'églises paroissiales de la région Touraine en 1100. D'autres valeurs empiriques sont en revanche issues d'estimations. Tel est le cas, par exemple, du taux de foyers paysans isolés en fin de période. Celui-ci ne peut être renseigné par des sources primaires et il a donc été nécessaire de l'estimer à partir de sources secondaires et en menant des extrapolations. Il est cependant possible de ~~construire~~ des indicateurs de sortie de simulation offrant une correspondance presque exacte avec ces différents indices observés ou estimés (cf. Correspondance entre indicateurs de sortie de simulation et indices empiriques, p. 31). Il est dès lors possible de mener une comparaison entre données observées/estimées et données simulées. Ces ordres de grandeur peuvent ~~ainsi~~ participer à l'évaluation du comportement du modèle simulé.



 **Faits et formes stylisés.** La seconde catégorie d'indices empiriques est moins précise et ne repose pas sur une valeur observable ou estimable, mais plutôt sur la connaissance experte d'un phénomène. Il s'agit des « **faits stylisés** »²⁴, qui rendent davantage compte d'une tendance dans la forme d'une relation ou d'une organisation que d'un ordre de grandeur. On fait un large usage de ces faits stylisés en économie, mais aussi en géographie, par exemple quand on qualifie la tendance des systèmes de peuplement à se hiérarchiser. La valeur de la pente associée à la courbe rang-taille d'un système de villes a ainsi été décrite comme tendant vers 1 à mesure que le système évolue et se hiérarchise (BERRY et OKULICZ-KOZARYN 2012, in PUMAIN et al. 2015, §9)²⁵. De la même manière, le modèle de transition démographique d'Adolphe Landry est un fait stylisé, énoncé à partir de l'observation de nombreuses récurrences de l'évolution des populations d'un pays en fonction de leurs taux de natalité et de mortalité. Ces exemples montrent qu'au sein des faits stylisés, il y a une certaine diversité quant à la précision de leurs énoncés. On peut ainsi quantifier précisément la courbe d'une relation rang-taille et l'allure de son évolution dans le temps, au moyen de l'évolution du coefficient de sa pente, et ces indices empiriques sont le plus souvent communiqués dans la littérature. Pour la transition démographique, on peut certes l'exprimer sous la forme d'une courbe logistique, mais les paramètres de cette courbe ne sont le plus souvent pas donnés dans les études thématiques. Le fait stylisé « transition démographique » est ainsi communiqué d'une manière moins précise que le fait stylisé « hiérarchisation d'un système de villes ».

Dans notre cas d'étude, les faits stylisés sur lesquels on s'appuiera seront d'une part des « allures »  de courbes, temporelles (par exemple l'évolution dans le temps d'un indicateur tel que le taux de concentration des foyers paysans) ou liées à une composition de valeurs (par exemple la courbe rang-taille correspondant à la hiérarchie des agrégats), et d'autre part des formes de répartition spatiale (densité du semis d'églises paroissiales par exemple). Notons que ces formes stylisées relèvent le plus souvent d'une agrégation ou d'une composition d'ordres de grandeurs (comme figuré dans la figure 3.7) : l'évolution dans le temps de la population, par exemple, correspond à un vecteur d'ordres de grandeur, c'est-à-dire à une succession de mesures de la quantité de population pour chaque date étudiée.

24. Définis ainsi par (LIVET, PHAN et SANDERS 2014) : « Un “fait stylisé” est une présentation simplifiée (i.e. taux, ratio ou écart, structure spatiale) d'une régularité empirique sur l'observation de laquelle il y a un large accord. Le terme a été popularisé en économie par Nicholas Kaldor (1961). [Les] faits stylisés peuvent être construits de la manière suivante : 1) en partant du domaine empirique, on identifie des relations saillantes ; 2) on opère quelques simplifications qui permettent d'inclure formellement ces relations dans des modèles ; 3) une fois admis que ces simplifications ne faussent pas trop les choses, on érige ces relations à la fois simplificatrices et formalisables au rang de “ faits stylisés”, dont les concepts théoriques doivent rendre compte. »

25. Notons que les auteurs de PUMAIN et al. (2015) nuancent cette affirmation, en montrant empiriquement que les pentes convergent vers des valeurs supérieures ou inférieures à 1 en fonction de l'ancienneté de l'intégration des systèmes de ville considérés (CURA et al. 2017a).

3.2.1.2 Les indicateurs de sortie de simulation

Les ordres de grandeur et formes stylisées évoqués ci-dessus relèvent du domaine empirique, c'est-à-dire qu'on dispose de données ou de connaissances d'experts à leur sujet. Afin de pouvoir les mobiliser pour évaluer la capacité du modèle à reproduire le phénomène d'intérêt, il est nécessaire de définir des **indicateurs de sortie** dans le modèle de simulation, c'est-à-dire des variables informatiques que l'on enregistrera durant l'exécution du modèle et que l'on pourra ensuite comparer aux indices empiriques définis.

Définition. Comme pour les indices empiriques qui sont leurs équivalents dans le domaine empirique, on peut définir les indicateurs de sortie de simulation, en distinguant des formes numériques simples (des scalaires), et des indicateurs plus complexes, multidimensionnels. Ces derniers sont en effet nécessaires pour pouvoir confronter les sorties du modèle de simulation avec les formes stylisées identifiées dans le domaine empirique. Chaque indice empirique doit ainsi se voir correspondre un indicateur de sortie (figure 3.7).

Correspondance entre indicateurs de sortie de simulation et indices empiriques. La correspondance entre indicateurs et indices ne correspond pas toujours à une équivalence exacte. En effet, si certains indicateurs peuvent trouver un équivalent strict dans le domaine empirique – le nombre de châteaux connus à chaque date a un sens strictement équivalent au nombre de châteaux simulés par le modèle –, d'autres correspondances sont moins directes.

Il peut s'agir de correspondances ayant trait aux mêmes éléments de base et le passage de l'indicateur à l'indice résulte alors d'une simple conversion. Par exemple, du point de vue empirique, on connaît à peu près les populations de la région étudiée au début et à la fin de la période. Dans SimFeodal cependant, on ne modélise pas des individus en tant que tels, mais des foyers paysans. Le nombre de foyers paysans simulé n'est pas directement comparable à la population estimée, mais en supposant une moyenne de 4 ou 5 habitants par foyer paysan, il est possible d'en déduire un nombre d'habitants.

Dans d'autres cas enfin, le décalage entre indicateurs et indices est plus important. Il s'agit notamment de caractéristiques du système féodal que l'on sait importantes mais pour lesquelles on ne dispose pas de données facilement quantifiables. La puissance militaire des seigneurs, par exemple, est complexe à quantifier. On sait d'après les connaissances expertes que la hiérarchie des puissances était forte à l'époque étudiée, majoritairement dominée par deux seigneurs (les comtes de Tours et de Blois) et assortie d'une grande quantité de petits chevaliers. On sait de plus qu'avec les liens de vassalité, les grands seigneurs disposaient des forces militaires des seigneurs qui leur étaient assujettis. Dans le domaine empirique on ne dispose pas d'éléments plus précis pour quantifier la puissance militaire des seigneurs. Dans le domaine du modèle, en revanche, on a défini un indicateur « proxy » de cette puissance à partir du nombre de foyers paysans s'acquittant de droits à chaque seigneur. De cette manière, on peut observer précisément en sortie de simulation la hiérarchie implicite entre les seigneurs reproduite par le modèle, avec une quantifica-

tion de leurs puissances respectives. Ces éléments peuvent être comparés aux connaissances empiriques sur ces rapports de puissance entre les seigneurs à différents moments de l'époque féodale.



Les correspondances entre indicateurs de sortie et indices empiriques sont ainsi de nature multiple, reflétant différents niveaux de proximité entre le concept mobilisé dans le modèle et ce qui est observable dans le domaine empirique : les châteaux, entités d'intérêt dans le modèle, ont un équivalent direct dans le domaine empirique (il s'agit d'entités facilement observables et des données historiques les concernant sont disponibles) alors que la puissance militaire des seigneurs, élément moteur dans le modèle, a conduit à utiliser une variable dans le modèle pour laquelle on ne dispose pas d'observations empiriques.

La création d'indicateurs de sortie correspondant aux indices empiriques permet donc de quantifier une information qui n'est pas forcément aisément quantifiable dans le domaine empirique.



Indicateur composite. La forme « informatique » (numérique ou multidimensionnelle) des indicateurs de sortie permet de trouver des manières plus simples d'évaluer le modèle que d'observer l'ensemble des indicateurs. Chaque indicateur étant numérique, il devient en effet possible de les combiner au sein d'indicateurs composites, résultant en quelques indicateurs synthétiques permettant une évaluation plus rapide des résultats d'une simulation. Ces indicateurs composites sont très fréquemment utilisés en statistiques, permettant par exemple de résumer une information multidimensionnelle en un indicateur simple. L'Indice de Développement Humain (IDH), par exemple, est un indicateur composite dépendant de l'espérance de vie à la naissance, du niveau d'éducation et du niveau de revenu de chacun des pays caractérisés. On le trouve très souvent utilisé, parce qu'il permet de résumer le niveau de développement d'un pays en agrégeant trois dimensions majeures, l'aspect sanitaire, culturel et économique.

Fonction objectif. En renforçant cette logique de synthèse de plusieurs dimensions, on peut aller plus loin dans la définition d'un unique indicateur, parfois composite et synthétique, qui permet d'évaluer à lui seul la qualité de représentation d'un modèle. On nomme d'ordinaire cet indicateur « fonction objectif » (ou « fonction de *fitness* »). C'est une pratique très fréquente, qui plus est dans le domaine de la simulation informatique en particulier sur des modèles de type « KISS » (ref. chap 1 ou 2). Il s'agit alors de définir une « fonction objectif », parfois composée d'une pondération des quelques indicateurs composites qui auront été identifiés, ou plus simplement, basée sur une unique variable que l'on juge représentative de l'ensemble du modèle.



Être en mesure d'évaluer un modèle à l'aide d'un unique indicateur a des avantages majeurs en pratique. Cela permet par exemple d'explorer et de paramétrer un modèle de simulation de manière entièrement automatique puisqu'on peut alors générer une cartographie simple des résultats du modèle en fonction des valeurs de paramètres utilisés (voir CHÉREL, COTTINEAU et REUILLON 2015, par exemple).

Ces indicateurs composites et synthétiques résultent d'une quantification des autres indicateurs (excluant donc les formes stylisées qui sont plus libres d'interprétation), et apportent un grand confort dans le paramétrage d'un modèle de simulation.

Quels types d'indicateurs pour SimFeodal ? SimFeodal n'est pas adapté à de tels indicateurs, parce qu'une large partie des faits stylisés et ordres de grandeur mobilisés proviennent de connaissances expertes, et les thématiciens qui les ont consolidées rechignent à créer de tels indicateurs composites. Ces derniers demandent en effet de pondérer précisément l'importance de chacun des indicateurs par rapport aux autres. Pour pouvoir pondérer cette importance, il faudrait de plus que les différents indices empiriques mobilisés présentent le même niveau de certitude, et que les indicateurs de sortie aient des variabilités similaires. Cela n'est le cas ni des indices empiriques sur lesquels SimFeodal s'appuie, ni des indicateurs de sortie que le modèle produit.

On aurait ainsi pu créer quelques indicateurs composites, mais ceux-ci n'auraient pas eu de véritable correspondance dans le champ empirique, les thématiciens ne faisant pas appel à des indices empiriques de telle sorte. Un indicateur composite serait donc nécessairement « hors-sol », et qui plus est, perdrait beaucoup dans la finesse de description du système modélisé.

Par exemple, pour caractériser la polarisation du système de peuplement, il pourrait suffire de définir un indicateur composite fonction du niveau de concentration – le taux de foyers paysans dispersés –, du nombre de pôles et de l'espacement moyen entre les agrégats. Les valeurs de l'indicateur généré pourraient renseigner efficacement sur la capacité d'un ensemble de valeurs de paramètres à reproduire le phénomène de polarisation attendu. Cette information serait cependant grossière, dans la mesure où seraient agrégées dans le groupe des « simulations réussies » des configurations extrêmement diverses. L'information fournie risquerait alors d'être très éloignée des connaissances empiriques des thématiciens : une information multivariée ne peut pas toujours être résumée, en gardant tout son sens, par une seule variable (de manière univariée).

On a donc fait le choix d'évaluer SimFeodal en conservant des indicateurs de sortie « simples », c'est-à-dire ni composites ni exprimés sous forme de fonction objectif. Ce choix a toutefois des implications majeures pour la méthodologie mise en place pour l'analyse des sorties de simulation. Il est en effet bien plus simple d'analyser quelques indicateurs composites plutôt qu'un grand nombre d'indicateurs hétérogènes.

3.2.2 Hiérarchiser et catégoriser les indicateurs

SimFeodal s'appuie sur une dizaine d'indicateurs numériques, ainsi que sur plus d'une trentaine d'indicateurs multidimensionnels. Tous ces indicateurs ne présentent pas le même degré de certitude, la même échelle d'observation, et surtout, le même niveau de précision sur les phénomènes modélisés. À chaque changement dans le modèle, pour une évaluation complète de la capacité de cette version à reproduire les indices empiriques, il faudrait donc observer et

analyser chacun de ces nombreux et divers indicateurs. Dans le contexte du paramétrage d'un modèle s'appuyant sur une logique itérative et incrémentielle (voir Encadré 3.2), on imagine bien que cela n'est pas possible : le nombre d'indicateurs est bien trop élevé pour avoir rapidement une vision globale de la qualité de représentation du modèle. Il faut dès lors, comme pour toute analyse synthétique, concevoir une hiérarchie d'observation et d'utilisation des indicateurs : il ne sera pas nécessaire d'analyser chacun des indicateurs dans la plupart des cas, seuls les indicateurs jugés plus importants pourront être analysés. Les indicateurs de moindre importance ne seront mobilisés que pour départager des situations dont la différence ne serait pas suffisamment explicitée par l'usage des indicateurs principaux.

3.2.2.1 Incertitude

Dans le modèle de simulation, les indicateurs de sortie sont à analyser en tenant compte de la précision des indices qu'ils représentent. Il ne faudra ainsi pas étudier la croissance du nombre d'agrégats au cours de la simulation de manière fine, par exemple en étudiant le coefficient directeur de la courbe, quand les données empiriques ne donnent quasiment aucune information à ce sujet si ce n'est qu'il y a bien plus d'agrégats en fin de période qu'au début. On peut vouloir quantifier la précision de ces données, par exemple à l'aide des méthodes développées dans le champ des observations floues et/ou incertaines (voir par exemple le travail de Cyril de Runz sur les données « imparfaites » (DE RUNZ 2008)).

Cette quantification de l'incertitude pourrait alors servir de base à l'établissement d'une hiérarchie des indicateurs : on analyserait en premier lieu l'écart entre les ordres de grandeurs empiriques bien connus (cf. tableau du niveau de certitude des objectifs) et les indicateurs calculés sur les données simulées. Les ordres de grandeur plus incertains seraient analysés dans un second temps (augmentation de la charge fiscale entre 800 et 1100 par exemple), et les formes stylisées viendraient enfin clore cette hiérarchie d'indicateurs.

SimFeodal se caractérise d'une part par une très forte hétérogénéité dans les niveaux de connaissance des ordres de grandeurs et faits stylisés modélisés, et d'autre part, se voulant un modèle théorique (ref dans chap1), « coller aux données » à tout prix n'est pas la priorité. La vraisemblance d'ensemble du modèle compte en effet bien plus que la précision de chacune de ses composantes. Pour l'évaluation de SimFeodal, nous ne tiendrons compte de l'incertitude des indicateurs qu'au cas par cas, sans la mesurer de manière systématique et donc sans établir de hiérarchie à partir de cette incertitude.

3.2.2.2 Catégoriser les indicateurs : définir des dimensions d'analyse

En présence de plus d'une quarantaine d'indicateurs, il est nécessaire, *a minima*, d'organiser leur analyse. On a vu qu'il n'était pas justifié de mener cet ordonnancement à partir des propriétés intrinsèques des indicateurs du modèle. Au contraire, et cela nous semble plus adapté pour un modèle à forte visée exploratoire et heuristique, la hiérarchisation des sorties du modèle doit

suivre la hiérarchie implicite qui structure les hypothèses et objectifs du modèle en lui-même. Ces hypothèses et objectifs sont multiples dans SimFeodal, et dès lors, une hiérarchie globale ne peut être définie. Il convient donc de catégoriser les indices empiriques – et les indicateurs de sortie de simulation leur correspondant –, avant de chercher à hiérarchiser ces indicateurs de manière globale. La hiérarchisation des indicateurs se fera donc relativement à chacune de ces catégories.



La dynamique du système de peuplement que l'on cherche à reproduire sur la période IX^e-XII^e siècle comprend trois processus (cf. chapitre 2), que nous nommerons dimensions (voir figure 3.7) : (1) polarisation de l'habitat rural, (2) hiérarchisation du système de peuplement et (3) fixation des foyers paysans. On peut s'appuyer sur ces trois dimensions pour caractériser les sorties du modèle, c'est-à-dire mener la confrontation entre indices empiriques et indicateurs de sortie.

On va répartir chacun des indicateurs dans la dimension qu'il sera le mieux en mesure de décrire. Cette répartition n'a pas à être égale, chaque dimension pouvant s'appuyer sur un nombre différent d'indicateurs. De même, chaque dimension sera composée d'indicateurs dotés d'une qualité de représentation ou d'un niveau de certitude hétérogène. Le seul point commun des indicateurs de sortie de chaque dimension doit être thématique. Les trois dimensions choisies – polarisation, hiérarchisation et fixation –, et les indicateurs qui les caractérisent dans le modèle, sont dès lors considérés comme les trois dimensions d'analyse des sorties de SimFeodal.

3.2.2.3 Hiérarchiser les indicateurs dans chaque dimension

Chacune de ces dimensions s'applique à plusieurs types d'agents du modèle. Pour définir la hiérarchie interne aux dimensions, on retiendra les agents les plus impactés par les dynamiques correspondant à ces dimensions : la polarisation, par exemple, peut être observé depuis le point de vue de ce qui polarise (les attracteurs) tout autant que de ce qui est polarisé (les foyers paysans).

Pour trancher, on examinera d'abord un indicateur de sortie numérique, caractéristique de la structure dans son ensemble à son état final. Les indicateurs de sortie représentatifs des dynamiques ayant mené à cette structure finale, par exemple les indicateurs multi-dimensionnels temporels, seront étudiés dans un second temps. Dans cet exemple, on analysera donc dans un premier temps le résultat effectif de la polarisation, c'est-à-dire la concentration des foyers paysans en agrégats, avant d'observer dans un second temps la répartition et la diversité des attracteurs ayant entraîné ce phénomène. On peut dès lors définir des « indicateurs principaux » pour chaque dimension, représentatifs des grands traits structurels auxquels on souhaite aboutir en sortie de simulation, et des « indicateurs secondaires », permettant d'affiner l'évaluation de chacune de ces dimensions.

3.2.2.4 Une hiérarchie mouvante

Notons que l'analyse des indicateurs de sortie suit une hiérarchie parfois mouvante, et en tous les cas, assez peu quantifiable. L'ordre d'observation des indicateurs est plutôt stable, mais l'importance que l'on portera à chacun peut varier. Les indicateurs principaux de chaque dynamique sont ainsi « incontournables », c'est-à-dire qu'un résultat trop loin de celui des indices empiriques est disqualifiant. Parmi les indicateurs secondaires, il n'est pas toujours possible, d'après les connaissances des experts sur le sujet, d'établir une priorité ou une pondération de chaque indicateur.

L'évaluation de la polarisation, par exemple (section 3.2.3.1), se définit principalement par rapport à un indicateur principal – le taux de foyers paysans dispersés –, mais selon les résultats des autres indicateurs de sortie, chacun aura une importance variable. L'étude de la dispersion des agrégats et pôles peut en effet se révéler plus importante que celle de l'évolution du nombre d'agrégats selon les paramètres que l'on souhaite ajuster, ou se montrer tout au moins plus différenciante selon l'état du paramétrage.



Encadré 3.2 : Incrémentalité des indicateurs

De la même manière que les paramètres et mécanismes d'un modèle de simulation tendent à évoluer^a au cours du temps de la construction, souvent afin d'affiner un comportement observé, les indicateurs de sortie sont amenés à évoluer aussi.

Ainsi, en cas de modifications fines du modèle, il est fréquent que les indicateurs initialement choisis ne suffisent plus à départager des versions du modèle quant à un phénomène spécifique. Par exemple, quand on observe le phénomène de polarisation dans les sorties de SimFeodal, l'indicateur du nombre d'agrégats est extrêmement synthétique et informatif jusqu'à ce que l'objectif soit atteint ou que les modifications ne parviennent plus à le faire évoluer. À partir de ce moment, afin d'améliorer la vraisemblance de la situation simulée par le modèle, on peut se focaliser sur la distribution spatiale de ces agrégats, par exemple pour vérifier qu'ils sont bien répartis de manière homogène dans l'espace, et non trop concentrés.

L'observation de la répartition spatiale requiert certes de nouvelles analyses, mais surtout, par exemple, d'enregistrer les positions des agrégats au cours du temps. Si cet indicateur de sortie n'était pas utile avant cela, il n'y avait aucun intérêt à l'enregistrer. Il faut donc adapter l'implémentation du modèle pour générer, faire évoluer et enregistrer une nouvelle variable informatique correspondant à cet indicateur. Dès lors, on pourra composer un nouvel indicateur synthétique, qui, dans cet exemple, pourrait prendre la forme d'un indice de concentration spatiale.

Ce procédé incrémental dans la construction des indicateurs est très fréquent, mais pose toutefois un problème majeur : sauf à adapter chacune des anciennes versions du modèle implémenté pour y ajouter l'enregistrement des nouveaux indicateurs nécessaires, on ne pourra rendre stric-

tement comparable les sorties de toutes les itérations du modèles informatique. Et même alors, il faudrait ré-exécuter des réplifications de chaque version du modèle implémenté à chaque ajout d'indicateur, quand bien même les indicateurs présents initialement étaient jugés suffisants. Un dernier obstacle est plus gênant : certains indicateurs sont spécifiques à des mécanismes, et en cas de changement de ces derniers, ils peuvent ne plus être calculables ou simplement comparables. Par exemple, des versions antérieures du modèle enregistraient les comportements individuels des foyers paysans quant à leur « choix » de déplacement, selon qu'ils étaient à l'origine localisés dans un agrégat ou dispersés. Une simplification du modèle a abouti à la modification des règles différenciant les possibilités de déplacement : on n'observe plus si le foyer paysan est dans un agrégat, mais plutôt s'il est dans un agrégat doté d'un pôle d'attraction. Dès lors, les analyses basées sur les choix de déplacement des foyers paysans selon leur origine ne sont plus comparables avec celles des versions antérieures au changement dans le modèle, quels que soient les détails d'implémentation de ce dernier.

Ces éléments expliquent que dans les résultats de chaque étape du paramétrage du modèle, on ne présente pas systématiquement l'ensemble des indicateurs, y compris quand ceux-ci pourraient être plus pertinents que les indicateurs présentés.

a. De manière incrémentielle et itérative, voir [dans le chapitre x ?](#) et THOMAS 2012, <http://itsadeliverything.com/revisiting-the-iterative-incremental-mona-lisa>

