

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

Version 2019-01-18

- N.B. : Commentaires de Lena (Mars et Avril 2018) sur les parties 3.2 et 3.3 non pris en compte
- Décembre 2018 / Janvier 2019 : Ajout de la partie 3.1

Sommaire

3.1	Comment évaluer un modèle ?	2
3.1.1	Évaluation, validation, vérification... : désambiguïsation	3
3.1.2	Les étapes de l'évaluation de modèle	5
3.1.3	Une évaluation de la plausibilité d'un modèle : la « <i>face validation</i> »	9
3.1.4	Des critères pour l'évaluation : les indicateurs	13
3.1.5	L'importance de la réPLICATION	13
3.2	Évaluer le modèle SimFeodal	14
3.2.1	Indices et indicateurs	14
3.2.2	Hiérarchiser et catégoriser les indicateurs	20
3.3	Les indicateurs et dimensions de SimFeodal	23
3.3.1	Évaluer la polarisation des foyers paysans	23
3.3.2	Évaluer la hiérarchisation du système de peuplement .	31
3.3.3	Évaluer la fixation et la dissémination du peuplement	38

Although the construction and analysis of a simulation model, the validity of which has not been ascertained by empirical observation, may prove to be of interest for expository or pedagogical purposes (e.g., to illustrate particular simulation techniques), such a model contributes nothing to the understanding of the system being simulated. [...]

Unless the construction of simulation models is viewed as a game with no purpose other than the formulation of a model, it is hard to escape the conclusion that the purpose of a simulation experiment is to predict some aspect of reality.

NAYLOR et FINGER 1967, B-92-93;96

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, VAN DEN 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 les plus anciens), 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 inscrivons SimFeodal comme un modèle résolument pédagogique, et l'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 de d'en assurer une certaine qualité, à défaut d'en garantir une validation stricte.

Nous sommes en effet convaincu 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 usage « 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. Soit-il à visée pédagogique, à base d'agents, de type descriptif ou encore vu comme « hybride », 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 de 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

considérations

utile

soit forme à horizon
d'utilisation

¹ 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.

méthode, la validation visuelle, que nous jugeons très adaptée dans le cadre de co-constructions interdisciplinaires de modèles.

3.1.1 Évaluation, validation, vérification... : désambiguïsation

Il nous semble important de commencer cette partie par un point de définition et de clarification des concepts mobilisés, non pas par convention datée, 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 au système qu'il décrit, et d'autre part à 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 (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 d'Osman BALCI (BALCI 1994) de clarification 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 alors que la *vérification* permet de s'assurer que le modèle est bien construit⁵. Le « *Testing* » correspond aux techniques mises en œuvre, et s'applique donc indistinctement à ces deux composantes. *du quoi ?*

En dépit de cette définition stricte, les usages persistent à ne pas formellement différencier 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 intervertis, comme un recensement rigoureux des usages le démontre (AUGUSIAK, VAN DEN BRINK et GRIMM 2014, Table 1, p. 120).

Nous partageons le besoin formel – identifié par les auteurs de cette étude – de définir un nouveau terme – « *evaludation* » dans leur proposition – et nous souscrivons à leur approche de définition (figure 3.1). Nous nous contenterons toutefois, dans ce travail, de nous inscrire dans les choix de AMBLARD, ROUCHIER et BOMMEL 2006, en particulier parce qu'ils nous semblent assez largement adoptés dans la communauté scientifique de modélisation en sciences humaines et sociales francophone. Pour ces auteurs, et donc dans le présent ouvrage, 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, 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.

4. « Model validation deals with building the *right* model. »

5. « Model verification deals with building the model *right*. »

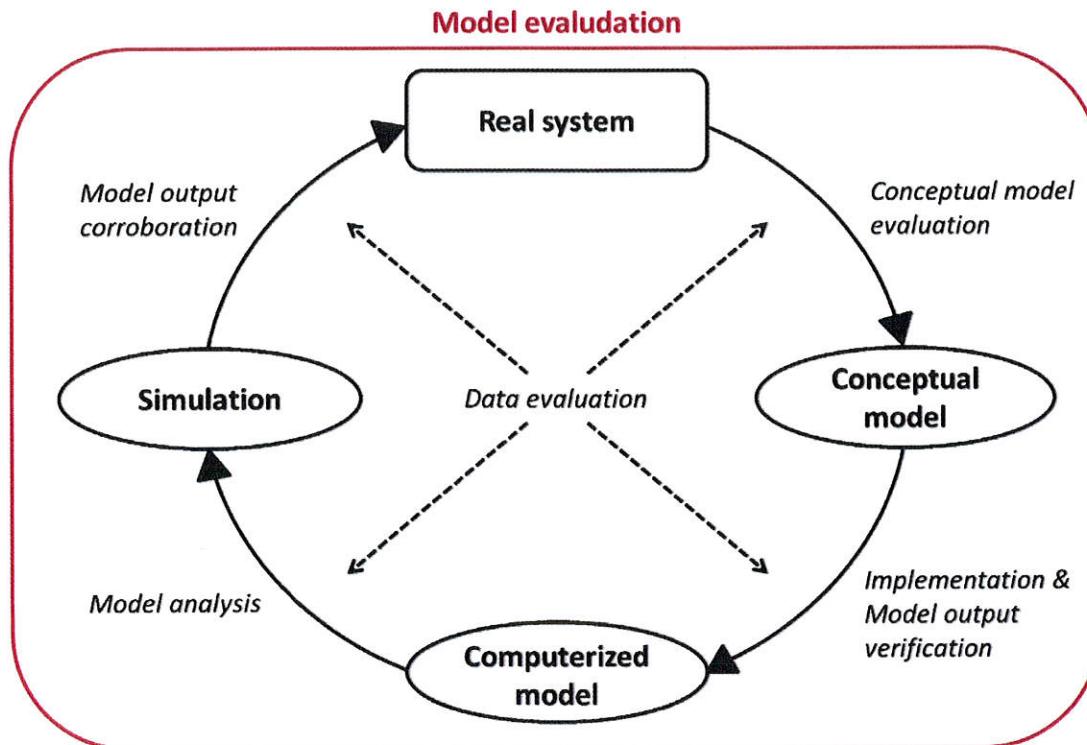


FIGURE 3.1 – Représentation schématique du cycle de modélisation et de la typologie des termes relatifs à l'évaluation de modèle, dans AUGUSIAK, VAN DEN BRINK et GRIMM 2014, Fig. 1, p. 121

« 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

disputer
un peu +
dans leur ur
dim pourrai
tin ~
l'utile p's
peu m'a
intre l-
seut
(per p'manda)

encadré ?

3.1.2 Les étapes de l'évaluation de 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 d'aller vers des méthodes de plus en plus formelles à mesure que le modèle mûrit⁶. Les schéma des étapes d'évaluation de KLÜGL (figure 3.2) et de NGO et SEE (figure 3.3) constituent un bon résumé de cette progression, que l'on peut brièvement décrire plus avant.

d'un ?
|| expliqu +

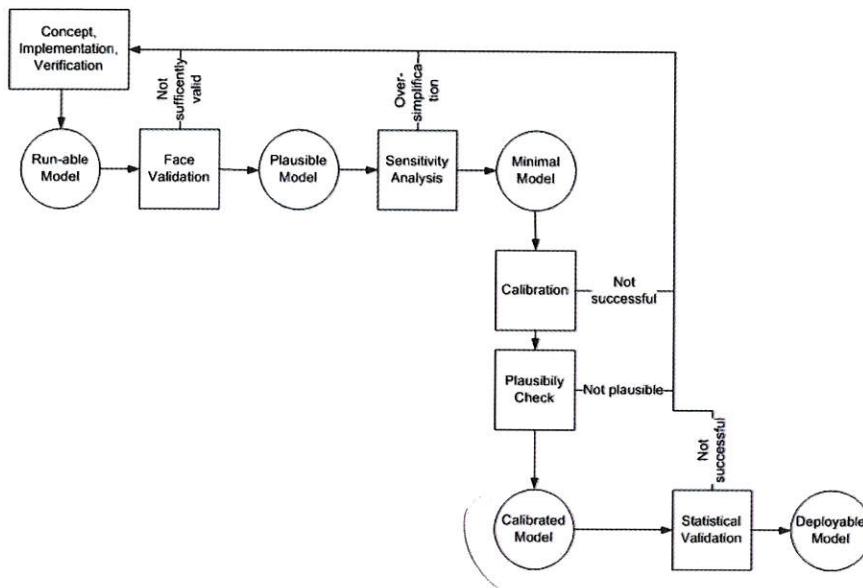


FIGURE 3.2 – « Sketch of a general procedure for validating an agent-based simulation », KLÜGL 2008, fig. 1 p. 42

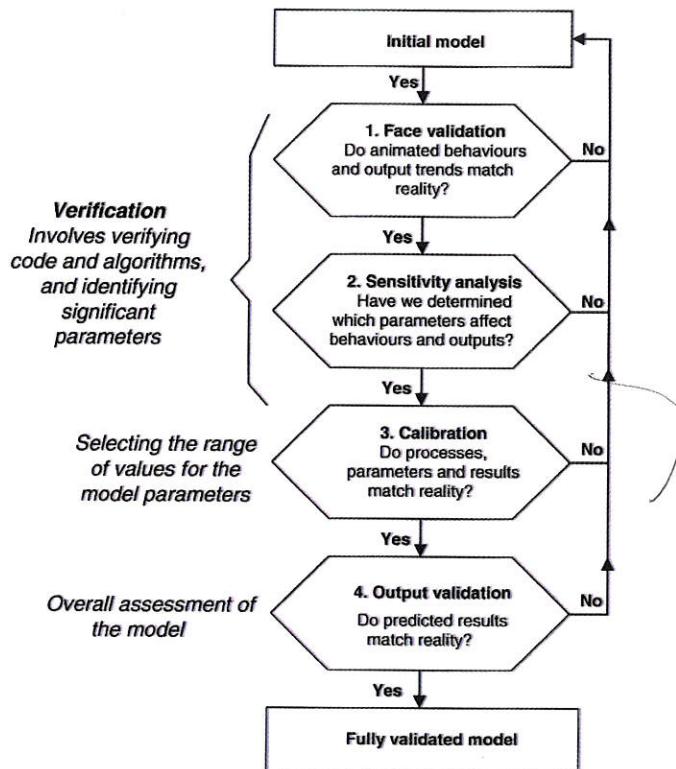


FIGURE 3.3 – « General validation process of an ABM », NGO et SEE 2012, fig. 10.1 p. 183

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

Face validation La première étape, de « *face validation* », consiste ainsi à vérifier visuellement, rapidement, la plausibilité du modèle. Cette étape est souvent considérée comme une mesure préalable relevant du bon sens plus que de l'évaluation strictement dite. Nous y reviendrons plus longuement par la suite.

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 complexes⁷, les valeurs des différents paramètres du modèle, on peut observer l'influence de chaque paramètre sur le déroulement du modèle. Cette mesure intervient tôt, et doit être répétée lors de chaque modifications majeures dans les mécanismes du modèle. L'analyse de sensibilité permet en effet de simplifier le modèle conceptuel et son pendant implanté : si l'analyse révèle qu'un paramètre, quelles que soient les valeurs qui lui sont attribuées, n'a qu'un effet minime voire négligeable sur les sorties du modèles, alors il peut être judicieux de supprimer ce paramètre et/ou le mécanisme qui le mobilise. Réduire le nombre de paramètres et/ou de mécanismes d'un modèle peut sensiblement l'améliorer, selon le principe de parcimonie qui voudrait qu'un modèle plus simple soit meilleur⁸. 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, ne serait-ce que parce qu'elle aide souvent le modélisateur à trouver une « *polarité* » à l'effet des paramètres : si tel indicateur 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.

Calibration 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. Cette démarche, nommée calibration, 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, 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 paramètre, en les réglant un par un, au risque d'entrer dans une boucle infinie de calibration, 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. La calibration en elle-même n'est pas à proprement parler une véritable procédure d'évaluation d'un modèle : il s'agit plutôt d'une méthode et d'un problème d'optimisation. Les différents auteurs mettent toutefois en avant son intérêt dans l'évaluation de modèles en ce qu'elle permet de garantir une meilleure validation externe du modèle puisqu'elle

précise
plausibilité du
modèle
et cohérence
(tendant) par
produit et
résultat ?

de sorte
pas en wa
intradict

et qui corr. à
l'obs empirique et
aux connaissances
théoriques

L de manière
non linéaire
à ajouter ?

pas très clair

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

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

aboutit à l'isolation d'étendues optimales de valeurs de paramètres : en menant de nouveaux tests de plausibilité (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.

qif. où ?

= validation externe ?

?

Validation statistique La validation statistique (« *output validation* » dans la figure 3.3) est sans doute la méthode d'évaluation la plus évidente pour qui-conque a été amené à concevoir un modèle. Il s'agit de confronter les données produites par le modèle – les *outputs* – aux données empiriques – ou théoriques – 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 théoriques, 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. 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 théoriques 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 et/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 multis 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 des modèles agents est alors intrinsèquement contre-intuitif et inexorablement ardu. 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) sont parvenus à résoudre de manière analytique un modèle foncièrement pensé comme un modèle agent – en automate

y Ax1 et
ou Ax2

cellulaire en l'occurrence –, le modèle de Schelling⁹. 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 tout de même fallu attendre le tournant des années 2010 afin d'y trouver une solution formelle.

Dans l'absolu, en ce qui concerne les modèles à base d'agents, il est donc difficile d'aller vers une validation formelle totale, mais on peut tout de même gagner en connaissances en mobilisant des méthodes qui cherchent à « résoudre » des modèles de simulation, de manière non formelle, par exemple autour des nombreux travaux actuels (CHÉREL, COTTINEAU et REUILLOU 2015 par exemple) qui visent à l'exploration complète des comportements possibles d'un modèle¹⁰.

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 méthodes¹¹. Il est évident à la lecture des références du champ que pour ces auteurs, « 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 »¹². 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'*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 » dans 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 classique d'une analyse de sensibilité : cette technique consiste à faire co-varier les valeurs des paramètres afin d'observer les effets que, chacun ou conjointement, ils produisent sur les sorties du modèle. Avec un modèle de Schelling, dans lequel on identifie en général trois paramètres (on y reviendra dans le chapitre 4), que l'on peut faire varier chacun selon une granularité de dix valeurs, et tenir compte de l'aléa en menant dix réplications, on peut mener une analyse de sensibilité basique au moyen de $(10^3 \times 10)$ 10 000

9. On reviendra et on présentera ce modèle de manière détaillée dans le chapitre 4. La description du modèle, en particulier, est effectuée dans la description.

10. C'est par exemple l'un des enjeux principaux, en matière de recherche, d'une plate-forme telle qu'OpenMOLE (REUILLOU, LECLAIRE et REY-COYREHOURCQ 2013)

11. « [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.

12. « 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.

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 complexe au regard de certains des tenants du genre KIDS, une analyse de sensibilité rigoureuse ou une calibration fine 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. » *à priori, le rétenu*

Nous ne partageons pas le dépit associé à cette recommandation, et au contraire, considérons que dans ce type de cas, des méthodes de « *face validation* » peuvent être, à condition d'être rigoureuses, extrêmement satisfaisantes et sans doute suffisante pour évaluer un modèle de simulation, ce que nous allons essayer de montrer par la suite.

?

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 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 plausibles¹³. L'utilisation très polysémique – et donc contradictoire – et les importants débats autour de l'usage de ce terme ayant poussé à sa désuétude¹⁴, on le retrouve pourtant au cœur de l'un des articles fondateurs de l'évaluation de modèles de simulation, où Charles 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 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

*de face validity
en dire + ce
donner un ex -
avoir une ex.
la corrélation*

13. « 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

14. 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. »

*= plausible ?
à bien définir*

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 d'évaluer la plausibilité d'un modèle pour un quelconque examinateur, mais bel et bien pour un expert du sujet modélisé. 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, Franziska 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 *face validation*, chacune pouvant être menée par des experts différents¹⁵ :

- **Évaluation du déroulement** Évaluation du déroulement d'une simulation dans son ensemble. Il s'agit donc 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 vérification 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 des d'analyse des co-variations 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.

) est-ce la bonne formulation ?

ici ok

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

- **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, ce qui est donc susceptible de requérir un expert différent de l'analyse de la plausibilité du système résultant des comportements de ces agents.

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 rationalité des comportements individuels qui n'est pas universelle dans les modèles à base d'agents. Cela se prête bien à de nombreux modèles où les agents sont 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 (ni même des seigneurs ou autre d'ailleurs) ne reposent pas sur des hypothèses de vraisemblance : le suivi d'un foyer paysan au cours du déroulement du modèle, par un expert thématique, n'est donc pas sujet à évaluation.

L'évaluation immersive nous paraît donc peu adaptée à l'évaluation de plausibilité de tous les modèles. Notons cependant que même quand elle n'a pas de sens thématique, cette méthode peut toutefois s'avérer utile en matière d'évaluation interne, dans un aspect de « débugage » : quand bien même les réactions et attributs des agents ne reproduisent pas une connaissance experte, l'observation de celles-ci 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 à 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, Charles 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

pourquoi ?
au niv. menu.
quick phas.
menu / menu,

représentent ?

explique !

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 malhonnête : 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 quelque 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 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, tout en assumant cette fois que l'évaluateur ne serait pas le modélisateur : le modélisateur connaît les « substitutions » opérées dans le modèle, et se gardera donc bien de juger de leur vraisemblance. A contrario, un expert thématique pourrait être étonné par certains comportements micro, dans la mesure où il ne connaît 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 le thématique et le modélisateur.

- **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

*- incorrect ?
- limite - biaise*

oui !

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.

FIN DE LA PARTIE RÉDIGÉE

3.1.3.3 Vers une évaluation visuelle

Distinguer Face validity (vraisemblance) et visual validation

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 :

« 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

BALCI 1998, Table 3, p. 46 : Face validity applicable à toutes les étapes du cycle de vie de la simulation

3.1.4 Des critères pour l'évaluation : les indicateurs

3.1.5 L'importance de la réPLICATION

