

Explorer visuellement des données de simulation massives pour analyser le comportement d'un modèle.

Version 2019-05-19

Sommaire

Introduction	2
5.1 Capturer les sorties de SimFeodal	3
5.1.1 Masse des données	3
5.1.2 Répliques	6
5.1.3 Expériences	9
5.1.4 Des données aux indicateurs	10
5.2 Comment explorer les sorties de SimFeodal ?	12
5.2.1 Observation en direct ou <i>a posteriori</i>	12
5.2.2 Générer les indicateurs	16
5.2.3 Organiser les indicateurs en rapports paramétrables	18
5.2.4 Organiser les rapports : Dashboards	23
5.2.5 Interagir avec les rapports : exploration interactive	27
5.2.6 Explorer en comparant : SimEDB	31
5.3 Organiser les données	34
5.3.1 Assurer la capacité d'interrogation des données	34
5.3.2 Structuration des données de SimFeodal	48
5.4 Une plate-forme d'exploration de données de simulations :	
SimEDB	56
5.4.1 Contraintes	56
5.4.2 Construire une plate-forme interactive pour l'exploration de sorties de simulation	66
Conclusion	81

Introduction

A écrire

-> Définir notamment la démarche : contraintes générales -> spécificités Sim-Feodal -> choix méthodes/techniques

5.1 Capter les sorties de SimFeodal

Pour évaluer un modèle, on s'appuie sur plusieurs indicateurs de sortie de simulation, de types divers (indicateurs numériques, graphiques, cartographiques etc., cf. chapitre 3, partie théorique). Quand le nombre d'indicateurs devient important, comme c'est le cas dans le modèle SimFeodal (chap 3, partie présentation des indicateurs), la consultation des indicateurs pendant le déroulement d'une simulation devient difficile. La complexité de ces indicateurs augmente dans le cas d'un modèle stochastique, où il est nécessaire de multiplier les répliques afin d'avoir une idée fiable des tendances simulées par le modèle. Le travail de paramétrage d'un modèle requiert de plus de mener différentes expériences, c'est-à-dire de faire varier les paramètres (chap 4) du modèle, démultipliant encore la masse des sorties, et avec elle, la complexité de leur analyse. Nous détaillons ici les contraintes qu'entraînent ces différentes spécificités des données issues de simulations. Ces contraintes sont transversales à plusieurs types de modèles, et on peut noter que certains autres types de modèles peuvent faire face à d'autres contraintes, propres ou génériques. Dans l'ensemble, les modèles peuvent être amenés à soulever les problèmes génériques à la production de données, quelles qu'en soient la source. Dans ce chapitre, nous n'avons pas l'ambition de dresser le portrait de l'ensemble des contraintes et solutions relatives à l'enregistrement et au stockage de données. Nous nous contenterons donc de soulever les plus fortes limites qui rendent difficile l'enregistrement des données issues d'un modèle de simulation à base d'agents, fortement stochastique, descriptif et exploratoire tel que SimFeodal.

à noter

5.1.1 Masse des données

Dans un premier temps, il convient de noter que l'ensemble des indicateurs observés en sortie de SimFeodal reposent sur des données qu'il est nécessaire de produire et d'enregistrer tout au long de la simulation. Ainsi, pour pouvoir tracer le graphique de l'évolution du nombre d'agrégats au cours du temps, il faut avoir accès à cette information, et dès lors, enregistrer, à chaque pas de temps, cette valeur dans un fichier numérique adapté. Cette information, en tant que telle, est assez faible, aussi bien en valeur sémantique qu'en valeur prise en mémoire. La masse représentée par cette information est toutefois démultipliée par la quantité d'indicateurs de sortie étaient nombreux, et avec eux, la quantité de valeurs à stocker augmente. À chaque pas de temps, il faudra enregistrer les valeurs de plusieurs variables. Cette pratique est habituelle, et un format de données tabulaire se prête bien à un tel enregistrement : une ligne pour chaque pas de temps, et une colonne pour chaque variable à enregistrer. On obtiendrait ainsi en sortie de simulation un tableau contenant 20^1 lignes

1. Il s'agit ici du nombre de pas de temps de SimFeodal. On notera que ce nombre est particulièrement faible au regard de très nombreux modèles de simulation, en particulier vis-à-vis de ceux qui visent à provoquer l'émergence d'un phénomène. Ces modèles sont en général théoriques, et n'ont qu'une faible correspondance entre pas de temps et durée réelle du phénomène modélisé. Dans le cas de SimFeodal, où le temps est un élément crucial du modèle, la résolution temporelle du modèle ne peut être diminuée artificiellement (voir chap2, section 2.2.2.2), et l'on se satisfait donc de ce nombre d'itérations relativement faible.

et une cinquantaine de colonnes², ce qui serait assez raisonnable pour une unique simulation.

que? Cette solution doit être écartée en ce qu'il est nécessaire de prendre en compte un aspect important de l'exploration de données issues de simulations : le coût temporel. L'exécution d'une simulation requiert un certain temps de calcul (3 à 4 minutes pour une exécution du modèle SimFeodal dans la version présentée dans le chapitre 2). Ce temps de calcul ne peut être optimisé que dans des proportions faibles sans avoir à bouleverser l'implémentation des mécanismes, ce qui représenterait un coût temporel encore plus important (voir la succession d'étapes du chapitre 4). En l'état actuel du modèle, la production des données a donc un coût temporel élevé.

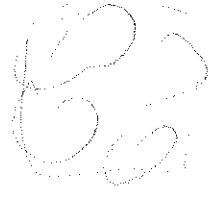
Qui plus est, ce coût est fortement dépendant du nombre de simulations exécutées : si le modèle est exploratoire, donc sujet à de nombreux changements, et notamment à l'ajout ou à la modification des indicateurs observés, il peut donner lieu à d'encore plus nombreuses simulations. Avec une structure de données fixe et agrégée, on ne pourrait introduire de nouveaux indicateurs, et la mise en comparaison des simulations précédentes impliquerait leur adaptation et re-production, c'est-à-dire la ré-exécution systématique de l'ensemble des simulations précédentes à chaque changement dans les indicateurs. L'introduction de nouveaux indicateurs est pourtant très fréquente, en particulier dans le cas d'un modèle exploratoire où l'on affine au fur et à mesure de l'évolution du modèle ce que l'on peut y observer.

Par exemple, parmi les indicateurs de sortie de SimFeodal, on s'intéresse notamment à la composition des pôles, que l'on qualifie assez simplement d'une part avec le nombre d'attracteurs qui les composent, et d'autre part avec l'attractivité globale qui en résulte. À mesure que la calibration du modèle progresse, et si tant est que les indicateurs choisis auparavant ne permettent plus de discriminer les effets de certaines variations fines dans les valeurs de paramètres, une étude plus fine du type d'attracteurs composant chaque pôle et de leurs propriétés spécifiques peut aider à discerner des différences entre ces jeux de paramètres et donc à mieux les comparer. Cet exemple illustre le cas des ajouts d'indicateurs de sorties, mais on peut aussi être confronté à des modifications des indicateurs existants. Le nombre de paroissiens moyen à chaque pas de temps peut être un indicateur utile au départ, mais on peut être amené à faire évoluer cet indicateur en une étude de la médiane si les nouvelles étapes de paramétrage du modèle en augmentent la variabilité. Les indicateurs peuvent évoluer au cours du temps de vie du modèle (cf. encadré chap 3), ou plus simplement, on peut être amené à réaliser une observation plus fine des sorties du modèle au fur et à mesure de la calibration de ce dernier. On se trouverait alors dans une situation impossible requérant de ré-exécuter les simulations après avoir adapté ou mis en place l'indicateur voulu.

En tenant compte de ces deux éléments, on a tout intérêt à se prémunir de ré-exécutions du modèle, et donc à enregistrer l'état de variables qui ne se-

2. Ce qui correspondrait par exemple à environ une colonne par indicateur, en plus des quelques colonnes de bases relatives à l'état d'ensemble de la simulation.

raient pas encore mobilisées pour la production d'indicateurs. Dans l'exemple du nombre de paroissiens, il faudrait en enregistrer au minimum les moyennes, médianes, et sans doute quelques paramètres de dispersions en plus, voir les quantiles, afin d'adapter les indicateurs de sortie de la manière la plus adéquate aux sorties des différentes versions du modèle. Dans le cas contraire, pour chaque changement ou ajout d'indicateur, il faudrait relancer des exécutions du modèle sur l'ensemble des jeux de paramètres précédents afin d'être en mesure d'avoir des indicateurs comparables entre les versions.



Enregistrer l'ensemble des variables d'un modèle est aisé dans le cas d'un modèle théorique simple, par exemple dans le cas d'un modèle comme celui de Schelling (SCHELLING 1971). Cela se complique quand il s'agit d'enregistrer les variables d'un modèle plus complexe comme SimFeodal. Celui-ci requiert en effet bien plus de variables globales (en parler dans chapitre 4, dans distinctions variables, paramètres etc.), pour représenter l'état du système dans son ensemble à chaque instant. Surtout, SimFeodal est un modèle qui voit interagir plusieurs sortes d'entités, chacune relatives à différents niveaux de granularité spatiale et sociale. Afin d'avoir tous les éléments en main une fois la simulation achevée, il est donc nécessaire d'enregistrer l'ensemble des variables non seulement globales, mais aussi afférentes à chacun des types d'agents. D'un unique tableau de données exhaustif en sortie du modèle de Schelling, on passe donc à plusieurs tableaux, dont les variables respectives seront propres à chaque type d'agent.

A ce niveau, l'information en sortie est encore relativement contenue : SimFeodal mobilise cinq types d'agents, chacun étant caractérisé par une douzaine d'attributs. On pourrait donc se contenter de ces cinq tableaux contenant 20 lignes (les pas de temps) et la douzaine d'attributs propres, comme c'est classiquement le cas dans ?³.

Un dernier point invalide cette solution d'enregistrement : une part importante des indicateurs s'appuie sur des données individuelles et non agrégées. Ainsi, on peut, à chaque pas de temps, enregistrer le nombre de paroisses, leur superficie moyenne ou encore le nombre moyen de paroissiens que chacune dessert. Mais cela ne permet en aucun cas d'en dresser une cartographie, c'est-à-dire de réaliser une carte de la localisation et des aires d'attraction des paroisses. Cela demanderait, par définition, d'enregistrer la géométrie de chaque paroisse à chaque pas de temps, les configurations spatiales (localisation de chacune et donc distribution spatiale de l'ensemble) variant à chaque simulation.

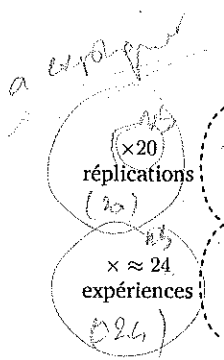
A
peut être à introduire
aussi dans une
liste avec autre
pour
variables
- indicateurs
- niveau
- d'aggrégation

Pour faire face à cette situation, on a donc fait le choix, dans SimFeodal, d'enregistrer les états des variables à des niveaux d'agrégation multiples, y compris au niveau de l'agent, à chaque pas de simulation. Dans le cas des paroisses, le volume de données résultant reste contenu : on obtient un tableau

3. RC : Trouver exemple de modèle SMA avec plusieurs types d'agents qui ont toutes un intérêt à être examinées spécifiquement, plutôt qu'au moyen d'un indicateur résumé classique (proie-prédateur, nb proies, nb prédateurs par exemple).

d'environ 2000 lignes⁴ et une dizaine de colonnes⁵. L'enregistrement systématique de l'état de chaque agent à chaque pas de temps est toutefois bien plus gênant dans le cas d'autres agents, par exemple les foyers paysans. Il est en effet nécessaire d'enregistrer les attributs de chacun d'entre eux pour être en mesure d'étudier les liens entre les valeurs de satisfactions et les choix de déplacement, ou encore d'observer la composition précise de la distribution des satisfactions. Avec 4000 foyers paysans au minimum à chaque pas de temps, les données changent d'ordre de grandeur⁶ : chaque simulation requiert de générer un fichier contenant des dizaines de milliers de lignes, pour un total, pour cet unique fichier, d'une dizaine de mégaoctets occupés.

Au final, l'enregistrement d'un état représentatif d'une simulation est difficile. Cela requiert de disposer de suffisamment d'éléments numériques pour pouvoir générer les indicateurs de sortie et rendre compte d'une partie de leur évolution. La masse de données produite est de ce fait nécessairement conséquente, comme indiquée dans la ligne « simulation » du tableau 5.1.



	Données		Indicateurs	
Intitulé	Quantité	Poids	Type	Quantité
Une simulation	≈ 10 ⁵ lignes	≈ 10 Mo	Visualisations en direct	≈ 10 indicateurs
Une expérience	≈ 10 ⁶ lignes	≈ 200 Mo	Indicateurs de sortie	≈ 30 indicateurs (variabilité des répliques)
Huit étapes de paramétrage	≈ 10 ⁷ lignes	≈ 5 Go	Indicateurs de sortie	≈ 700 indicateurs (à comparer entre les expériences)

TABLE 5.1 – Synthèse de la multiplication des données et indicateurs selon la hiérarchie des simulations.

5.1.2 Répliques

Comme on l'a vu dans le chapitre 3, une simulation ne suffit pas à évaluer le modèle. SimFeodal est ainsi un modèle stochastique, c'est-à-dire qu'une large partie des mécanismes qui l'animent sont basés sur des tirages aléatoires. Cet aléa est évident dans les mécanismes faisant appel à un tirage aléatoire explicite, par exemple le choix de déplacement ou non d'un foyer paysan (cf. chap2, mécanisme déplacement). Dans le cas de ce mécanisme, un foyer paysan mobile se déplacera selon une probabilité dépendant de sa satisfaction. Et s'il y a probabilité, il y a donc aléa. Même avec une forte satisfaction — 99% par

4. Avec une moyenne de 150 paroisses, cela représente $20_{[\text{pas de temps}]} \times 150_{[\text{paroisses}]} \approx 3000$ lignes pour chaque simulation.

5. Les identifiants de la simulation (nom, graine aléatoire), le pas de temps, l'identifiant de la paroisse, puis les différents attributs et enfin la géométrie stockée dans une colonne textuelle.

6. $20_{[\text{pas de temps}]} \times 4000_{[\text{foyers paysans}]} \approx 80\,000$ lignes pour une exécution du modèle.

exemple —, il reste 1% de chance qu'un foyer se déplace, ce qui, sur un grand nombre de tirages (chaque foyer paysan, à chaque pas de temps), aboutit à une probabilité de réalisation non négligeable. Et cette probabilité de réalisation sera encore supérieure pour des foyers paysans ayant des niveaux de satisfaction légèrement moindre mais cependant globalement très élevés, supérieurs à 90% par exemple. En analysant les sorties du modèle, on aura donc la présence d'*outliers*, qu'il sera important d'isoler. Ils présenteront en effet des comportements contre-intuitifs puisque résultant d'une probabilité extrêmement faible. L'aléa a un poids important dans ce type de mécanisme.

Même dans le cas de mécanismes plus anodins, l'aléa est fortement présent, puisqu'il est au cœur de la conception de SimFeodal. L'ordre d'exécution, c'est-à-dire l'ordre aléatoire dans lequel les agents sont appelés pour exécuter leurs mécanismes, aura donc un impact important sur les indicateurs de sortie de simulation, sans que cet impact ne puisse être caractérisé au moyen d'indicateurs agrégés. Par exemple, les seigneurs peuvent créer des châteaux, sous condition de puissance (cf. chapitre 2, section 2.3.12). Pour créer ces châteaux, il faut que des agrégats soient « disponibles », c'est-à-dire ne comportent pas de château pré-existant à une certaine distance. Cette contrainte devient rapidement le facteur principal de la limitation de l'apparition de châteaux. Si un seigneur est plus souvent que les autres « appelé » en premier pour exécuter ce mécanisme, il pourra profiter des nouveaux agrégats disponibles pour créer ses châteaux. À force de création de châteaux, il sera relativement plus puissant, et pourra donc créer d'autant plus de châteaux relativement aux autres seigneurs. Il y aura donc une hiérarchie forte dans le nombre de châteaux possédés par seigneur.

Au contraire, si l'ordre d'appel des mécanismes favorise des seigneurs différents à chaque pas de temps, alors plus de seigneurs seront en mesure de créer des châteaux, et la hiérarchie sera alors plus faible.

Ces mécanismes sont sensibles à l'ordre d'appel, et il est ainsi difficile de discerner ce qui relève d'une tendance simulée et ce qui relève de fines variations dues à l'aléa, dans le comportement du modèle.

On pourrait objecter qu'en considérant les agents de manière agrégée, donc globale, les probabilités sont appliquées à suffisamment d'individus pour présenter un résultat cohérent et robuste au niveau de la population dans son ensemble. En corollaire, le comportement de chaque agent serait régulé par tant de variables aléatoires qu'on entrerait dans le cadre d'application de la loi forte des grands nombres, les agents adoptant alors en moyenne un comportement proche de l'espérance (moyenne théorique) de chaque tirage. Avec ces considérations, on pourrait justifier la robustesse probable des différentes exécutions de SimFeodal.

SimFeodal n'est toutefois pas simplement un modèle stochastique, mais avant tout, un modèle complexe, c'est-à-dire s'inscrivant dans le champ des systèmes complexes. Sans vouloir ici entrer dans les détails des implications et raisons de ceci, on peut simplement en retenir qu'un modèle tel que SimFeodal est extrêmement sensible aussi bien aux conditions initiales qu'aux différents

une analyse?

tirages aléatoires. A développer sérieusement ici, ou bien dans les chapitres 1 ou 2. Il faudra de toute façon faire un point quelque part sur les systèmes complexes, l'émergence etc. Pour illustrer, on peut s'appuyer sur un exemple, caricatural mais possible : à l'initialisation, tous les foyers paysans, placés aléatoirement dans l'espace, seraient concentrés dans un espace d'étendue restreinte. Seul un énorme agrégat émergerait donc, et aucun pôle ne serait susceptible dès lors de diviser cet agrégat géant. On atteindrait ainsi une situation très éloignée des configurations spatiales observées empiriquement, et très éloignée aussi des réalisations habituelles du modèle. En présence d'un seul agrégat, les possibilités de développement d'attracteurs (châteaux et paroisses) pourraient tout aussi bien être fortes que faibles. À partir d'une telle configuration initiale, on ne peut savoir si la situation convergerait vers un agrégat « paradisiaque », extrêmement développé et doté de pôles satisfaisants, ou au contraire, vers un agrégat « prison », où aucun des foyers paysans ne serait satisfait, mais n'aurait non plus d'alternative.

Cet exemple fictif, volontairement caricatural, ne s'est pas présenté jusqu'ici, mais le cas échéant il faudrait pouvoir repérer ce type de comportement aberrant. Cela serait par exemple utile pour éventuellement les distinguer des autres simulations et ne pas le laisser influencer l'analyse d'un jeu de paramètres données. De plus, cet exemple concerne uniquement une configuration initiale qui présenterait des caractéristiques tout à fait exceptionnelles. Plus généralement, il existe un grand nombre de situations initiales potentielles éloignées de l'empirique et même du vraisemblable. Les réalisations aberrantes peuvent apparaître à toute étape de la simulation, et déformer les tendances observées dans les indicateurs de sortie du modèle. Au delà de l'initialisation, elles peuvent être issues de tirages aléatoires particulièrement défavorables, ou encore apparaître suite à une succession d'événements improbables qui s'auto-renforceraient. Pour distinguer ces réalisations aberrantes de ce que l'on pourrait caractériser d'une tendance normale, il est nécessaire de multiplier les répliques, afin de constituer un contexte suffisant pour isoler ces événements anormaux.

On ne peut donc pas raisonner sur une unique simulation pour évaluer un jeu de paramètres (cf. chap 3). On ne peut pas non plus se contenter d'une agrégation des résultats des différentes répliques, sous la forme de moyennes ou d'écarts-types, selon qu'on s'intéresse par exemple à la tendance générale ou qu'on cherche à observer les variations que peut entraîner l'aléa.

Pour ces raisons, et pour être en mesure d'embrasser l'entière diversité des sorties de simulations issues de variation de la graine aléatoire, il est donc nécessaire de mener plusieurs répliques de chaque simulation, et d'enregistrer l'entièreté des sorties de simulations dans chacun des cas. Le jeu de données produit par une simulation, contenant quelques dizaines de milliers de lignes, est ainsi obligatoirement multiplié par le nombre de répliques. Pour l'exploration de SimFeodal, après différents tests, ce nombre a été fixé à 20 répliques (J'en aurais sans doute parlé dans le chapitre 3 (évaluation), mais à laisser ici jusqu'à ce que ce soit certain.). La dizaine de mégaoctets issue d'une simulation devient donc approximativement 200 mégaoctets, et le

nombre de lignes contenues, par exemple pour les foyers paysans, passe d'à peu près 70 000 à 1 400 000⁷ (voir la ligne « expérience » du tableau 5.1).

5.1.3 Expériences

Comme décrit dans le chapitre 4, le paramétrage de SimFeodal a demandé plusieurs étapes. De plus, chacune de ces étapes représente plusieurs sous-étapes – les expériences – faites d'essais et d'erreurs, en faisant varier à chaque fois les valeurs de paramètres de SimFeodal. Afin de construire le modèle, puis de l'explorer de manière plus systématique, il a été nécessaire de tester des dizaines de configurations de paramètres. Pour comparer, à chaque nouvelle version du modèle, les résultats produits par rapport aux résultats de la version précédente, il est indispensable de conserver, au minimum, l'ensemble des jeux de données de cette version précédente.

Cet archivage des résultats immédiatement précédents n'est pourtant pas suffisant, pour des raisons tenant à la reproductibilité et à traçabilité du modèle obtenu au final. On serait en effet tenté, à chaque nouvelle version « majeure » du modèle, de ne conserver que les indicateurs de sorties des versions précédentes en considérant que le modèle a atteint une phase de maturité supérieure à chaque fois. Les étapes intermédiaires, reléguées au rang de brouillons ou d'esquisses, deviendraient alors inutiles. Le processus de conception et de paramétrage d'un modèle n'est pourtant pas linéaire (cf. chapitre 4), et on peut avoir besoin de comparer une version intermédiaire « actuelle » à une version intermédiaire précédente, quitte à réaliser qu'une modification erronée a été ajoutée au modèle.

La conservation des résultats de chacune des expériences joue donc à nouveau un rôle multiplicatif dans la masse de données à conserver (voir le tableau 5.1).

En supposant que les 8 étapes présentées dans le chapitre précédent (ref chap4, étapes) soient ne serait-ce que constituées de 3 sous-étapes chacune — ce qui est bien en deçà de la réalité —, on obtient 24 jeux de paramètres à stocker, puis à devoir mobiliser. Cela représente une somme considérable de données (voir tableau 5.1), qui se chiffrent en dizaines de millions d'enregistrement⁸. En matière de stockage, il ne s'agit jamais que de quelques gigaoctets de données, pourtant à la limite de ce que l'on peut traiter sur un ordinateur individuel⁹.

7. Si cette quantité de données semble tout à fait raisonnable et peut largement être traitée sur un ordinateur classique, on peut toutefois noter qu'elle dépasse déjà le maximum de lignes (2²⁰, ≈ 1 000 000) que les tableurs classiques — LibreOffice ou Microsoft Excel dans leurs dernières versions en 2018 — sont en capacité de gérer.

8. $20_{\text{[pas de temps]}} \times 4000_{\text{[foyers paysans]}} \times 20_{\text{[répliques]}} \times 24_{\text{[jeux de paramètres]}} \approx 40\,000\,000$ de lignes enregistrées pour les seuls foyers paysans.

9. Selon une approximation courante, on ne peut charger en mémoire des données d'une taille supérieure à la moitié de la mémoire vive. Approximation qui approche du tiers quand on prend en compte les autres processus en cours, et éventuellement des modifications à l'échelle de l'ensemble du jeu de données plutôt que sur des extraits. Pour pouvoir traiter ces 5 Go de données (tableau 5.1), l'ordinateur utilisé doit donc disposer d'au moins 16 gigaoctets de mémoire vive, et encore, au prix d'un traitement potentiellement lent et bloquant.

Biblio?

Figure
étapes?

à rendre
finir
pour ton
modèle

On ne mentionne ici que les expérimentations issues des étapes de paramétrage. Les phases suivantes, visant à l'exploration du comportement du modèle (analyse de sensibilité, calibration...), demandent ainsi d'exécuter, et donc d'enregistrer, une masse bien plus importante de simulations.

5.1.4 Des données aux indicateurs

Dans l'ensemble, l'enregistrement et la sauvegarde des données issues de simulations constituent, pour les modèles de simulations basés sur de nombreux agents et mécanismes, une contrainte importante vis-à-vis de l'exploration du comportement de ces modèles.

C'est particulièrement le cas pour SimFeodal, où l'on ne peut se contenter de produire à la volée les indicateurs pour des raisons de reproductibilité¹⁰.

Analyser une masse de données La masse de données en sortie est impressionnante et requiert dès lors, d'un point de vue technique, d'utiliser des outils adaptés à la manipulation de grands jeux de données. Cela exclut de fait l'outillage le plus simple de la géographie quantitative, ne laissant par exemple pas la possibilité d'utiliser les outils à interface graphique classiques. Au contraire, face à des données de cet ordre, seules des solutions statistiques, basées sur des analyses en ligne de commande, peuvent être mobilisées. Ces solutions doivent en plus être appuyées par des capacités de calculs importantes, sans toutefois justifier encore l'usage de technologies de calcul intensif¹¹. . Cela pose une contrainte dans l'accessibilité aux analyses : le traitement des données requiert des compétences spécifiques en analyse de données volumineuses. Dans un contexte interdisciplinaire caractérisé par une large hétérogénéité en matière de pratiques quantitatives, il n'est pas possible de se contenter d'envoyer les jeux de données produits aux thématiciens – qui ne disposent le plus souvent pas de ces compétences – : ils seraient alors en difficulté pour en tirer les analyses nécessaire à leur interprétation.

Analyser une masse d'indicateurs D'un point de vue thématique, et c'est là l'objectif, cette masse de données doit servir à la production d'indicateurs, nombreux et divers aussi bien dans leur forme que dans les caractéristiques des processus qu'ils décrivent (ref. chap. 3, indicateurs). Les mêmes raisonnements que pour les données s'appliquent ainsi aux indicateurs. On peut prendre en compte la variabilité des réplifications directement dans les indicateurs produits (par exemple avec des représentations graphiques de type *box-plot*, utilisés ici une large partie des indicateurs). La production de tels indicateurs au niveau de la variabilité inter-expérience est pourtant difficile, si tant est qu'elle soit souhaitable. De fait, chaque expérience doit pouvoir être comparée aux précédentes sur la base de leurs seules réplifications respectives. Dès lors, la raison d'être des indicateurs de sortie est de rendre possible une

10. La reproductibilité sera abordée « longuement » dans le chapitre 1 (positionnement).

11. Le « *High-Performance Computing* » (HPC) par exemple, mobilisé pour l'étude de données plus massives, c'est-à-dire trop importantes pour être analysées sur un unique ordinateur ou serveur. Faire ref à la thèse de Seb là où il en parle, dans sa partie 1

?

cote que
99 peut être
bien modifier
objectif ?
- modèle
- modèle
- modèle
- modèle

comparaison, indicateur par indicateur, entre chacune des expériences. Il est donc indispensable de générer, pour chaque expérience, l'ensemble des indicateurs. En ne considérant ici encore que 24 expériences, cela fait donc déjà plusieurs centaines¹² d'indicateurs (tableau 5.1).

Le choix ayant été fait de mener une comparaison visuelle (ref. dans chapitre 3 : indicateurs uniques vs fonctions objectifs), on imagine dès lors que celle-ci va être difficile en présence de tant d'indicateurs.

En sus de la contrainte de l'enregistrement et de la production des indicateurs, le verrou majeur à la compréhension des phénomènes modélisés dans SimFeodal est donc la simple capacité à visualiser et à explorer l'ensemble des indicateurs de sortie. Ce qui doit de plus être rendu accessible y compris pour un auditoire non habitué à la manipulation de nombreuses données et sorties quantitatives.

cf. remarque précédente

12. En considérant ainsi une trentaine d'indicateurs, on obtient donc $30 \text{ [indicateurs]} \times 24 \text{ [jeux de paramètres]} \approx 700$ indicateurs uniques.

5.2 Comment explorer les sorties de SimFeodal ?

Pour évaluer, de manière approfondie, une expérience (voir tableau 5.1) d'un modèle tel que SimFeodal, il est nécessaire de passer en revue de nombreux indicateurs de sortie de simulation. Cette évaluation ne vise pas à produire une « note » unique et synthétique, mais plutôt à tester la capacité de l'expérience à reproduire les dynamiques que le modèle cherche à reproduire. Il ne s'agit pas, à proprement parler, d'une validation du modèle, au sens quantitatif où on pourrait l'entendre. On vise plutôt à explorer le comportement du modèle en fonction des mécanismes et valeurs de paramètres choisis. Cela aboutit donc sur un jugement qualitatif sur la capacité du modèle à reproduire les dynamiques souhaitées. Pour mener cette exploration, il convient d'utiliser des outils adaptés, c'est-à-dire de disposer de solutions techniques permettant le calcul et l'affichage des indicateurs à partir des données produites par le modèle.

pour compléter
la précédente
- Vario 0 1
- Vario 0 1000

Dans le travail mené autour de SimFeodal, plusieurs solutions ont été utilisées au cours des différentes étapes de construction du modèle. La restitution purement chronologique de ces solutions ne revêt pas d'intérêt propre, mais les contraintes accumulées au cours de la construction du modèle ainsi que les choix devant permettre de les dépasser nous paraissent très largement génériques et généralisables.

La succession de choix d'outils d'explorations se justifie par les verrous dans l'exploration que chacun de ces outils a permis de débloquent. Cela dresse par là-même un portrait des solutions méthodologiques d'exploration de données de simulations dont on peut faire usage selon les contraintes générales des modèles.

5.2.1 Observation en direct ou *a posteriori*

Classiquement, le premier réflexe d'un modélisateur, du moins pour les modèles à base d'agents, est de définir des sorties graphiques pour accompagner son modèle. Les différentes plate-formes de modélisation agent mettent d'ailleurs régulièrement en avant les possibilités de représentations qu'offrent leurs environnements¹³. Visualiser le déroulement d'un modèle « en direct » (« *online* » dans GRIGNARD et DROGOUL 2017), c'est-à-dire au sein de la plate-forme de simulation et au cours l'exécution du modèle, offre ainsi de nombreux avantages : évaluation visuelle du niveau de ségrégation (et de son évolution) dans une implémentation du modèle de Schelling ; visualisation de cohérence du déplacement des nuées d'oiseaux dans un modèle de type « Flocks » (REYNOLDS 1987) ; ou encore suivi d'un indicateur dans le temps – la quantité de ressources collectées – dans un modèle de type « Sugarscape » (EPSTEIN et AXTELL 1996).

~~Double
vérifier
l'axe~~

13. Voir par exemple les collections de visualisations sur les pages d'accueil de Gama (<https://gama-platform.github.io/>), de NetLogo (<https://ccl.northwestern.edu/netlogo/>), de GeoMASON (<https://cs.gmu.edu/~eclab/projects/mason/extensions/geomason/>) ou encore de Repast (<https://repast.github.io/screenshots.html>).