Cena, le 18/06/2018

Chapitre 5

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

Version 2018-06-15

Sommaire

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, elle-même, dans le cas d'un modèle stochastique comme SimFeodal, où il est nécessaire de multiplier les réplications afin d'avoir une idée fiable de la tendance du 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) d'un modèle, démultipliant encore la masse des sorties, et avec elle, la complexité nécessaire à leur étude. Nous détaillons ici les contraintes qu'entraînent ces différentes spécificités des données issues de simulation de SimFeodal.

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 une courbe de l'évolution du nombre d'agrégats au cours du temps, encore faut-il 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. Pour autant, on a montré que les indicateurs de sortie étaient nombreux, et avec eux, la quantité de valeurs à stocker augmente. Ainsi, à chaque pas de temps, il faudra enregistrer les valeurs de plusieurs variables. Ce comportement est habituel, 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 18 lignes (le nombre de pas de temps de SimFeodal) et une cinquantaine de colonnes, ce qui serait assez raisonnable.

Il faut toutefois considérer un aspect important de l'exploration de données issues de simulations : la production de ces données a un coût temporel important, c'est-à-dire que l'exécution d'une simulation requiert un certain temps (3 à 4 minutes pour une exécution du modèle SimFeodal dans la version présentée dans le chapitre 2). Si l'on considère de plus que les indicateurs peuvent évoluer au cours des étapes de construction et de paramétrage d'un modèle (cf. encadré chap 3), on a alors tout intérêt à prévoir ce cas, et donc à enregistrer l'état de variables qui ne seraient pas encore mobilisées pour la production d'indicateurs. 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 (ref). Cela se complique quand il s'agit d'enregistrer les variables d'un modèle plus complexe comme SimFeodal. Celui-ci comprend ainsi bien plus de variables globales, représentant l'état du système dans son ensemble à chaque instant. Surtout, SimFeodal est un modèle qui voit interagir plusieurs types d'agents, chacun situés 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 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 : il y a cinq types d'agents, ayant chacun une douzaine d'attributs, dans SimFeodal. On pourrait donc se contenter de ces cinq tableaux contenant 18 lignes (les pas de temps) et la douzaine d'attributs propres, comme c'est classiquement le cas dans (trouver exemple).

Reste encore un obstacle majeur à un enregistrement suffisant du déroulement d'une simulation : une part importante des indicateurs s'appuie sur des données individuelles et non agrégées. Ainsi, enregistrer une situation globale des paroisses à chaque pas de temps permet par exemple d'en examiner le nombre, la superficie moyenne ou encore le nombre moyen de paroissiens desservis. Mais cela ne permet en aucun cas d'en dresser une cartographie, ce qui nécessite, par définition, d'enregistrer la géométrie de chaque paroisse à chaque pas de temps. De même, on a vu que les indicateurs pouvaient être amenés à évoluer : si l'on enregistre un nombre de paroissiens moyen à chaque pas de temps, puis que l'on décide après coup de comparer non plus cette moyenne, mais une médiane, ou encore d'observer la forme de cette distribution, on se trouve alors dans une situation impossible requérant de ré-exécuter les simulations après avoir adapté l'indicateur voulu.

Pour se prémunir de 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 individuel. Dans le cas des paroisses, le volume de données résultant reste contenu : on obtient un tableau d'environ 2000 lignes [1] et une dizaine de colonnes [2]. L'enregistrement systématique de chaque agent est toutefois bien plus gênant dans le cas d'autres agents, par exemple les foyers paysans. Pour ceux-là, et parce qu'on doit être en mesure d'étudier les liens entre satisfactions et choix de déplacements, ou encore d'observer la composition précise de la distribution des satisfactions, il est aussi nécessaire d'enregistrer les attributs de chacun. Avec 4000 foyers paysans à chaque tour, les données changent ainsi d'ordre de grandeur [3] : 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.

A terme, pour enregistrer un état représentatif d'une simulation, c'est-à-dire avoir suffisamment d'éléments numériques pour pouvoir générer les indicateurs de sortie et prévoir une partie de leur évolution, la masse de données produite est assez conséquente.

5.1.2 Réplications

Comme on l'a vu dans le chapitre 3, une simulation ne suffit toutefois 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, 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

1. Avec une moyenne de 120 paroisses, cela représente $18_{\text{[pas de temps]}} \times 120_{\text{[paroisses]}} \approx 2000$ lignes pour chaque simulation.

2. Les identifiants de la simulation (nom, graine aléatoire), le pas de temps, l'identifiant de la paroisse, puis les différents attributs et la géométrie.

3. $18_{\text{[pas de temps]}} \times 4000_{\text{[foyers paysans]}} \approx 70\,000$ lignes pour une exécution du modèle.

une probabilité dépendant de sa satisfaction. Et s'il y a probabilité, il y a donc aléa. Même avec une forte satisfaction — 0.99 par exemple —, il reste donc 1% de chance qu'un foyer se déplace, ce qui, sur un grand nombre de tirages (chaque foyers paysans, à chaque pas de temps), a une forte probabilité de réalisation. L'aléa a donc un poids important dans ce type de mécanisme.

Pour autant, même dans des mécanismes plus anodins, l'aléa est fortement présent, étant au cœur de la conception de SimFeodal. Ainsi, le simple ordre d'exécution du mécanisme de déplacement des foyers paysans peut avoir une importance considérable.Cet exemple est faux, en trouver un juste où l'ordre d'exécution importe dans SimFeodal et le développer.

On pourrait objecter qu'en considérant les agents de manière agrégée, donc globale, les probabilités s'effectuent sur suffisamment d'individus pour présenter un résultat cohérent et déterministe 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 le déterminisme 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 champs 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 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 — fictif jusqu'ici dans les différentes exécutions du modèle — possible : à l'initialisation, tous les foyers paysans, placés aléatoirement dans l'espace, seraient ici concentrés dans un espace faible. 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 de l'empirie, 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 alors tout aussi bien être fortes que faibles. À partir de cette 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 jamais produit jusqu'ici, mais le cas échéant, encore faudrait-il pouvoir le repérer, pour éventuellement l'isoler d'autres simulations. On ne peut donc pas raisonner sur une unique simulation pour évaluer un jeu de paramètres (cf. chap 3), mais on ne peut pas non plus se contenter de récupérer des différentes réplications et d'en tirer une moyenne (selon qu'on s'intéresse par exemple à la tendance générale) ou un écart-type (si l'on cherche justement 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 réplications 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 réplications. Pour l'exploration de SimFeodal, après différents tests, ce nombre a été fixé à 20 réplications (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 mégaoctet 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 1 400 000[4].

5.1.3 Expériences

Comme décrit dans le chapitre 4, le paramétrage de SimFeodal a demandé plusieurs étapes. Chacune de ces étapes représente, ~~qui plus est,~~ plusieurs sous-étapes, 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 ~~donc~~ été nécessaire de tester des dizaines de configurations de paramètres. L'objectif étant de comparer, à chaque fois, les résultats en sortie de simulation d'un nouveau jeu de paramètres testé, il était indispensable de conserver, au minimum, l'ensemble des jeux de données de la version précédemment testée du modèle.

Cela n'est pourtant pas suffisant, pour plusieurs raisons aussi bien éthiques que méthodologiques. En premier lieu, pour des impératifs de reproductibilité (Ça aussi faudrait quand même en parler quelque part. Chap. 7 ?) de la démarche engagée, aussi bien que pour la simple capacité à restituer honnêtement et rigoureusement les étapes suivies, il était nécessaire de conserver l'ensemble des indicateurs de sortie de simulations correspondant à chaque étape ou sous-étape. Cette démarche de paramétrage s'inscrivant ainsi sur une durée assez étalée, et suivant ~~surtout~~ un avancement non linéaire fait d'allers-retours, il était indispensable de documenter autant que possible chaque avancée, et pour cela, de conserver l'ensemble des résultats en résultant.

Une autre contrainte, méthodologique, déjà évoquée (cf. encadré incrémentalité des indicateurs dans chap. 3), complexifie toutefois encore la tâche. Ainsi, les indicateurs jugés utiles évoluent tout au long des étapes de paramétrage. Or, pour pouvoir comparer les résultats de simulations issues de jeux de paramètres différents, ~~encore~~ faut-~~il~~ disposer d'indicateurs comparables, et donc, identiques dans leur définition. Si l'on ne conserve que les indicateurs de chaque simulation, on ne peut donc les ajuster sans avoir accès aux données sources ayant permis de les produire.

Pour ces raisons, il n'était pas possible de mener un travail de paramétrage de SimFeodal sans conserver l'ensemble des données produites, c'est-à-dire l'ensemble des attributs de l'ensemble des agents de chacune des réplications, tout cela pour chacune des expérimentations.

En supposant que les 8 étapes présentées dans le chapitre précédent (ref chap2, étapes) soient ne serait-ce que constituées de 3 sous-étapes chacune — ce qui est bien en deçà de la réalité —, on obtient ~~donc~~ 24 jeux de paramètres à stocker, puis à devoir mobiliser. Cela représente ~~alors~~ une somme considérable de données, qui se chiffrent en dizaines de millions d'enregistrement[5]. Si cela ne représente jamais que quelques gigaoctets de données, ce que quiconque a désormais l'habitude de manipuler dans un cadre personnel, en terme de traitement, cette masse de données est à la limite de ce que l'on peut traiter sur un ordinateur individuel.

4. 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 toutefois déjà le maximum de lignes ($2^{20}, \approx 1\ 000\ 000$) que les tableurs classiques — LibreOffice ou Microsoft Excel dans leurs dernières versions en 2018 — sont en capacité de gérer.

5. 18 [pas de temps] × 4000 [foyers paysans] × 20 [réplications] × 24 [jeux de paramètres] ≈ 35 000 000 de lignes enregistrées pour les seuls foyers paysans.

Ainsi, selon une approximation courante, on ne peut charger en mémoire de données d'une taille supérieure à la moitié de la mémoire vive disponible, sans même prendre en compte les autres éventuels processus en cours. Avec 5 Go de données , il faut donc disposer d'un ordinateur personnel possédant au moins une douzaine de gigaoctets de mémoire vive, et encore, au prix d'un traitement extrêmement lent et bloquant.

Et encore, on ne mentionne ici que les expérimentations issues des étapes de paramétrage. Les phases suivantes d'exploration du comportement du modèle, par exemple autour de variations systématiques de valeurs de paramètres en vue de 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, en vue de leur mobilisation pour produire les indicateurs de sortie, se révèle une contrainte importante dans la compréhension du comportement d'un modèle. 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 reproductibilité théorique et pratique.

La masse de données en sortie est impressionnante et requiert donc premièrement d'utiliser des outils adaptés à la manipulation de grands jeux de données. Cela exclue de fait l'outillage traditionnel 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 intensif. Cela pose une première contrainte dans l'universalité de l'analyse, en particulier dans un contexte interdisciplinaire porteur d'une large hétérogénéité en matière de pratiques quantitatives : il n'est pas possible de juste envoyer les jeux de données produits aux thématiciens, qui ne pourraient en l'état pas en tirer les indicateurs nécessaires.

Dans un second temps, et c'est là la contrainte principale, 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 raisonnement que pour les données s'appliquent ainsi aux indicateurs. Si on peut prendre en compte la variabilité des réplications directement dans les indicateurs produits (par exemple avec des représentations graphiques de type *box-plot* qu'adoptent une forte partie des indicateurs), ce n'est ni possible ni souhaitable entre les différentes expériences. De fait, chaque expérience doit pouvoir être comparée aux précédentes sur la base de leurs seules réplications respectives. Dès lors, la raison d'être des indicateurs de sortie est de rendre possible une 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 centaines[6] d'indicateurs (table 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.

6. En considérant ainsi une trentaine d'indicateurs, on obtient donc 30 [indicateurs] × 24 [jeux de paramètres] ≈ 700 indicateurs uniques.

	Données		Indicateurs	
Intitulé	**Quantité**	**Poids**	**Type**	**Quantité**
Une simulation	$\approx 10^5$ lignes	$\approx$ 10 Mo	Visualisations en direct	$\approx$ 10 indicateurs
×20 réplications → Une expérience	$\approx 10^6$ lignes	$\approx$ 200 Mo	Indicateurs de sortie	$\approx$ 30 indicateurs (variabilité des réplications)
× $\approx$ 24 expériences → Huit étapes de paramétrage	$\approx 10^7$ lignes	$\approx$ 5 Go	Indicateurs de sortie	$\approx$ 700 indicateurs (à comparer entre les expériences)

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

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 possible et accessible y compris pour un auditoire non habitué à la manipulation de nombreuses données et sorties quantitatives.

5.2 Comment explorer les données de SimFeodal ?

5.2.1 Observation en direct vs a posteriori

5.2.2 Générer des rapports

5.2.3 Automatiser les rapports

5.2.4 Organiser les rapports : Dashboards

5.2.5 Interagir avec les rapports : exploration interactive

5.2.6 Explorer en comparant

5.3 Organiser les données

5.3.1 Modèle de données

5.3.2 Assurer la capacité d'interrogation des données

5.3.2.1 Interroger de manière universelle

5.3.2.2 Interroger rapidement

5.3.3 Assurer la pérennité et la stabilité des données

5.3.3.1 Stockage fichier vs BDD vs projet de recherche

5.3.4 Présentation de la/des solution(s) adoptée(s)

5.3.4.1 Historique et raisons

5.3.4.2 MapD

5.4 Une plate-forme d'exploration de données de simulations : SimEDB

5.4.1 Contraintes

5.4.1.1 Efficacité

5.4.1.2 Interopérabilité

5.4.1.3 Adaptabilité

5.4.1.4 Généricité / indépendance aux données

5.4.2 SimVADB / SimEDB

5.4.2.1 Choix des technologies

5.4.2.2 Choix de l'organisation

5.4.2.3 Choix des modes d'interactions

5.4.2.4 Présentation générale