Lena, rendu le 29/06/2018

Chapitre 5

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

Version 2018-06-22
Corrections de la partie 5.1 non prises en compte.

Sommaire

5.2 Comment explorer les données de SimFeodal ?

Pour évaluer une expérience de SimFeodal, on doit passer en revue une trentaine d'indicateurs de sortie de simulation. Cette évaluation ne vise pas à aboutir à une note unique, mais plutôt à une idée de la capacité de l'expérimentation à reproduire les dynamiques modélisées. Il ne s'agit donc pas a proprement parler d'une évaluation du modèle, mais plutôt d'une exploration de son comportement en fonction des mécanismes et valeurs de paramètres choisis. 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. 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 et les choix devant permettre de les dépasser nous paraissent très largement génériques. Nous justifions donc ici la succession de choix d'outils d'explorations au prisme des verrous dans l'exploration que chacun a permis de débloquer, ce qui dresse par là-même un portrait des solutions méthodologies d'exploration de données de simulations dont on peut faire usage selon les contraintes générales des modèles.Pas clair, à reprendre, mais l'idée est là.

5.2.1 Observation en direct vs 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 environnements (ref blogs Gama, NetLogo 3D, GeoMASON, Repast). Visualiser le déroulement d'un modèle "en direct" offre ainsi de nombreux avantages (cf. HDR Arnaud, en sortir une citation et/ou un listing).

Dans l'exploration de SimFeodal, la création en direct de quelques graphiques correspondant à des indicateurs étudiés permet de vérifier, avant le lancement d'une série de simulations, que le déroulement de la simulation ne présente pas de forte incohérence et qu'un *bug* n'a pas été oublié. Pourtant, deux contraintes limitent fortement le recours à ce type de visualisation en direct des simulations.

La première contrainte, déjà évoquée plus haut, est que le modèle SimFeodal est fortement stochastique. Dès lors, la visualisation des indicateurs d'une simulation particulière ne suffit pas à estimer le comportement du modèle. C'est bien pour cela que les indicateurs choisis pour l'évaluation de SimFeodal prennent presque tous en compte la variabilité des résultats induite par l'exécution de réplications.

Certains environnements techniques (ref à multisim dans Gama) permettent toutefois de mener concomitamment plusieurs réplications d'un même modèle et de visualiser directement pendant l'exécution les résultats des réplications agrégés. La première contrainte, liée à la nécessaire étude des réplications du modèle, peut donc être dépassée en adaptant l'implémentation du modèle pour faire usage de ces capacités de multi-simulation.

La seconde contrainte est plus cruciale dans le cas de SimFeodal et invalide l'usage des méthodes de visualisation en direct. On l'a vu, l'exploration des sorties de simulation du modèle repose sur la consultation d'une trentaine d'indicateurs, parfois très spécifiques au cas d'étude de SimFeodal. Outre le fait qu'il serait

concrètement difficile de représenter tous ces indicateurs au sein de l'interface graphique d'une plate-forme de simulation agent, la temporalité de l'exécution d'une simulation (ou même des réplications nécessaires) est bien plus courte que celle requise pour la compréhension des résultats produits. Les indicateurs de sortie de simulation demandent ainsi un examen approfondi avant d'être en mesure de juger de leur adéquation aux attentes thématiques. Cet examen ne peut que difficilement être réalisé en direct, qui plus est quand il demande de faire appel, dans le cadre de co-construction inhérent à SimFeodal, à plusieurs points de vue. Les modalités mêmes de l'exploration des sorties de SimFeodal requièrent donc que les indicateurs soient visibles et explorables à des temporalités différentes, par des chercheurs différents, depuis des lieux différents. Pour un même chercheur, l'évaluation n'étant pas une étape unique et finie, il est utile de pouvoir revenir sur les résultats à différents moments, ne serait-ce que pour comparer les nouveaux résultats produits à ceux générés par des expérimentations précédentes.

Il est donc indispensable que les indicateurs soient enregistrés et consultables simplement à tout moment, ce qui élimine de fait la visualisation des indicateurs en direct pendant l'exécution des simulations comme unique méthode d'exploration du comportement de SimFeodal.

La visualisation en direct n'est donc pas mobilisable en tant que telle, mais elle peut tout de même, comme dans un usage très classique, être utilisée comme un outil de validation interne pour tester chaque modification dans les valeurs de paramètres. Visualiser une seule simulation, avant d'en exécuter les réplications nécessaires, permet ainsi déjà de vérifier que les modifications apportées dans les valeurs de paramètre ou dans les mécanismes n'ont pas entraîné l'apparition de *bugs* ou d'incohérences immédiatement visibles.

Nous avons donc choisi de doter SimFeodal d'une interface graphique, très sommaire, mais permettant des allers-retours rapides entre l'implémentation et l'exécution. Cette interface n'affiche qu'un nombre réduit d'indicateurs (Figure 5.1), ainsi qu'une représentation cartographique (Figure 5.2) utile à une analyse rapide du comportement d'ensemble du modèle.

FIGURE 5.1 – Visualisations intégrées à l'interface graphique de SimFeodal : indicateurs liés aux foyers paysans et aux seigneurs.

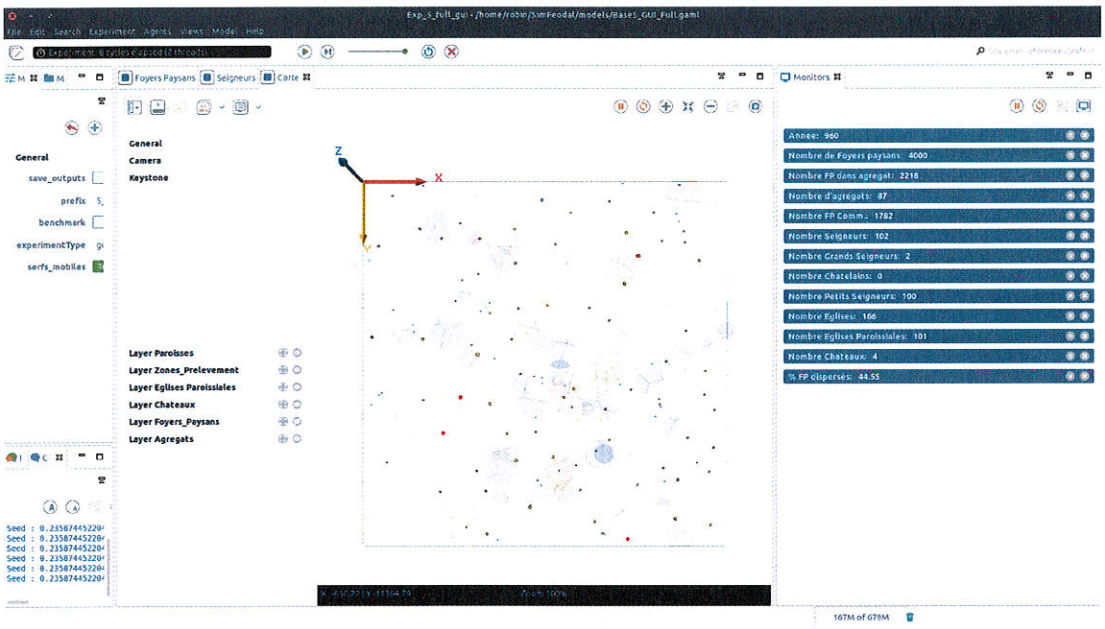

FIGURE 5.2 – Visualisations intégrées à l'interface graphique de SimFeodal : cartographie synthétique de l'espace modélisé.

Ces représentations ne suffisent pas et ne remettent aucunement en cause les constats de leur inadéquation aux contraintes d'ensemble de l'évaluation de SimFeodal. Elles complètent cependant la démarche d'évaluation du modèle par l'ajout d'une étape de contrôle préalable au lancement d'expériences,ce qui ne peut être qu'avantageux.

5.2.2 Générer les indicateurs

Si la production des indicateurs doit donc nécessairement être réalisée en aval de l'execution des simulations, encore faut-il disposer d'outils adaptés au traitement des données produites, c'est-à-dire répondant aux contraintes identifiées auparavant (sous-section 5.1.4). La contrainte principale est d'être en mesure de gérer la masse de données produites. On l'a vu, cela élimine d'office les outils de type tableurs, ou encore les outils de manipulation graphique de données les plus courants. Pour les raisons évoquées dans le chapitre 1 (Positionnement : pourquoi utiliser des outils libres ?), seules les solutions techniques non-propriétaires étaient envisageables.

Certains outils graphiques, basés sur des logiciels libres en arrière-plan (PSPP, R Commander, Orange), sont extrêmement aisés à prendre en main et auraient pu constituer un bon choix. Pourtant, avec une trentaine d'indicateurs à produire pour chaque expérience, donc de manière régulière, nous avons préféré nous tourner vers des outils plus orientés vers une interface en ligne de commande (*Command Line Interface*, abrégés *CLI*).

L'utilisation de CLI a plusieurs interêts gravitant autour de la reproductibilité des traitements. En premier lieu, ils permettent une adaptation aisée et rapide aux différents jeux de données. Ainsi, partant du principe que les données générées par les réplications et expérimentations sont de même structures, il suffit généralement de modifier le chemin d'entrée des fichiers résultants pour reproduire à l'identique une analyse sur un nouveau jeu de données.

De manière plus technique et interne à la génération des indicateurs, on peut remarquer que les différents indicateurs de sortie de simulation choisis présentent souvent des caractéristiques communes, aussi bien dans le traitement nécessaire que dans les formats (graphiques) produits.

Par exemple, la grande majorité des indicateurs reposent sur une première agrégation des données par réplication et pas de temps simulé, puis par une seconde agrégation montrant la variabilité des situations générées, au niveau de l'expérimentation donc. En terme de manipulation de données, seul l'indicateur statistique final, et éventuellement l'agent caractérisé, est ainsi modifié dans ces nombreux indicateurs de sortie. Le recours à des traitements en CLI permet ainsi un simple copier/coller, voir la création de fonctions dédiées, pour effectuer ces traitements très récurrents.

Au niveau des sorties graphiques, et donc des indicateurs multi-dimensionnels (cf. chap 3 et schéma des indicateurs), on peut aussi remarquer que la forme est assez largement identique : on représente les pas des temps (les années simulées) en abscisse, un indicateur statistique en ordonnée, et les figurés sont sous forme de *box-plot* minimalistes (« *minimal boxplot* », promus par Edward Tufte pour minimiser le ratio données-encre.Mettre les réfs de Tufte). Là aussi, en disposant d'un environnement de type CLI, et qui plus est en faisant usage de solutions graphiques construites sur une syntaxe régulière et générique[7], il devient très confortable de n'avoir qu'a adapter un premier graphique conçu aux autres indicateurs souhaités.

Je ne sais pas si il faut argumenter ici le choix d'utiliser R, mais pour l'instant, je laisse ça pour ailleurs/plus tard.

Avec ces solutions, il est facile de concevoir et d'implémenter les codes informatiques nécessaires à la génération des indicateurs de sortie de simulation. Cela est de plus, dans l'exécution de ces programmes, extrêmement rapides, les différents fichiers de sortie de simulation étant lus et parcourus un unique fois pour en tirer toutes les variables nécessaires à l'établissement des indicateurs.

En sortie, on obtient une table synthétique (indicateurs numériques, cf. schéma chap. 3) et de nombreux graphiques dans des formats vectoriels, donc aisément modifiables, par exemple pour publication, et surtout, lisibles et transférables très simplement.

7. On utilise pour tous ces graphiques le *package* R `ggplot2`, qui repose sur la grammaire graphique conçue par Leland Wilkinson (ref).

5.2.3 Organiser les indicateurs en rapports paramétrables

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