Lena 02/07/2018.

# Chapitre 5

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

Version 2018-06-29

Corrections de la partie 5.1 non prises en compte.

## Sommaire

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.

> *Encadré 5.1 : Générer les indicateurs avec R et `ggplot2`*
>
> Ça pourrait être bien de faire un encadré technique ici sur les choix technologiques (R, dplyr, ggplot2) retenus pour SimFeodal

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.

### 5.2.3 Organiser les indicateurs en rapports paramétrables

Du point de vue de la manipulation, la création de fichiers informatiques indépendants correspondant aux différents indicateurs de sortie de simulation est extrêmement pratique : on peut facilement les échanger et les adapter, par exemple pour inclusion dans une publication.

Du point de vue de la comparaison des résultats, cette forme n'est pourtant pas la plus adaptée. Si l'on peut facilement comparer un même indicateur portant sur deux expériences différentes, la tache se complique quand il s'agit d'avoir une vision globale des différences dans les indicateurs entre deux expériences. Pour cela, la démultiplication des fichiers correspondant aux indicateurs se révèle rapidement être un obstacle : on est alors amené à jongler entre de très nombreux fichiers.

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

Pour rendre la comparaison des indicateurs plus aisée, en présence d'une forte diversité d'indicateurs, il convient donc, ~~a minima,~~ d'organiser les indicateurs. Nous entendons ici par organisation, une présentation structurée, suivant un certain ordre, identique selon les expériences, adapté à une évaluation des résultats du modèle SimFeodal. Pour cela, nous avons choisi d'organiser ces « indicateurs » au sein de « rapports ». Cela permet, dès lors que les expériences simulées ont été nombreuses, de rassembler l'ensemble des indicateurs de sortie propres à chacune dans un unique fichier, à la structure toujours similaire.

En dehors du simple archivage des sorties, la production de rapports facilite aussi la comparaison des expériences par le biais de leurs indicateurs de sortie. On peut ainsi, par exemple, placer côte à côte deux rapports rendant compte de deux expériences différentes, et, en les faisant défiler simultanément, comparer point par point, c'est-à-dire indicateur par indicateur, leurs résultats respectifs.

Les formes que peuvent prendre des rapports, tout autant que les modalités de leur production, sont multiples et extrêmement diverses, depuis le document produit manuellement en insérant des bons indicateurs au fur et à mesure, par exemple dans un traitement de texte, jusqu'au rapport entièrement automatisé produisant des commentaires automatiques des indicateurs insérés en fonction d'expressions conditionnelles.

Pour SimFeodal, nous avons choisi de restreindre au maximum la manipulation manuelle, c'est-à-dire de générer un rapport entièrement automatique, ne requérant pas d'action spécifique en dehors du choix des données depuis lesquelles créer les indicateurs ~~et donc le rapport~~. Le rapport produit (figure 5.3 pour un aperçu global, et annexe X pour un exemple de rapport complet) n'intègre ainsi que les indicateurs, ~~c'est-à-dire les indicateurs~~ numériques – sous forme de tableau – et les indicateurs multi-dimensionnels – sous forme de graphiques –. Ces indicateurs sont ~~toutefois~~ organisés par partie, en l'occurrence ~~par~~ le type d'~~agents~~ et de comportement qu'ils décrivent.

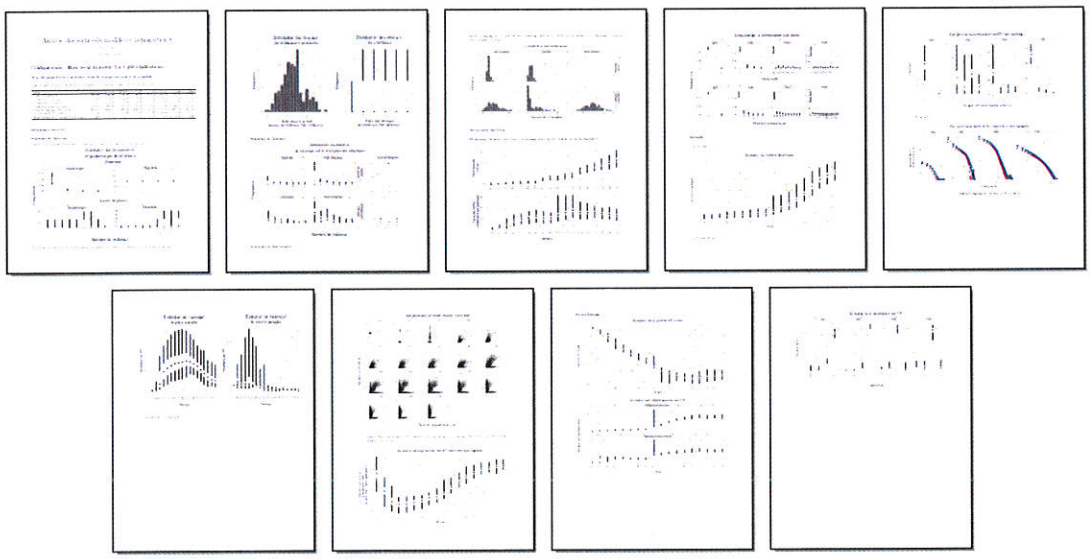

FIGURE 5.3 – Un exemple de rapport automatique généré pour une expérimentation (étape 0) de SimFeodal. La version en taille réelle est reproduite en annexe X.

Les raisons de ce choix sont multiples, mais ont en commun une recherche de reproductibilité des résultats et des analyses menées. Reproductibilité certes théorique (encore une ref au positionnement), les résultats de simulation devant être en mesure d'être analysés et reproduis par de potentiels intéressés, mais

reproductibilité aussi rendue nécessaire par la pratique de la modélisation, tel qu'explicité auparavant (ref. à partie 1 du chap, dans réplication -> expériences). La quantité d'expériences requises pour arriver à un état satisfaisant de SimFeodal a ainsi été importante, et le nombre de rapports tout autant. La création d'un rapport automatisé garantissait d'une part une génération rapide des indicateurs sur les nouvelles données, permettant donc un examen des sorties de simulation presque immédiatement après leur execution. D'autre part, la structure assez fixe d'un rapport automatisé, c'est-à-dire se basant une structure de données figée (des fichiers tabulés dotés d'en-têtes constantes), ou encore sur une masse suffisante et nécessaire de réplications (pour que les indicateurs soient comparables, ils doivent être réalisés sur le même nombre de réplications, 20 dans notre cas, ce qui est introduit comme contrainte dans la génération du rapport), permet une première évaluation du bon déroulement « interne »[8] de la simulation : si le rapport ne peut être généré, c'est que le modèle a été modifié d'une manière qui le rend non rétro-compatible, du moins en ce qui concerne ses sorties.

Un autre interêt majeur des rapports, déjà pointé en avantage des outils de type *CLI* est leur adaptabilité. On a vu (chap. 3) que les indicateurs à examiner sont nombreux et surtout, évolutifs, dans le sens où ces indicateurs ont fortement été modifiés, remplacés, affinés, au cours des étapes de paramétrage de SimFeodal. L'utilisation de rapports automatique permet de changer le code source, qui permet de générer les indicateurs, en un unique endroit. À partir de là, pour mettre à jour l'ensemble des rapports déjà produits, c'est-à-dire regroupant les indicateurs de chacune des expériences passées, il suffit de ré-exécuter la routine de production des rapports, ce qui représente un gain de temps et d'efficacité conséquent dans les situations de changements fréquents d'indicateurs, comme cela a été le cas pour SimFeodal.

*Encadré 5.2 : Générer les rapports avec `R` et `knitr`*

Encadré sur les rapports automatiques, le *« litterate programming »*, le choix et le paramétrage de knitr + lien vers le code-source des rapports.

Naturellement, la reproductibilité des rapports constitue plus un objectif qu'une réalisation effective. Ainsi, comme explicité dans l'encadré sur l'incrémentalité des indicateurs (ref encadré incrémentalité chap 3), une contrainte forte empêchent une absolue reproductibilité des analyses du comportement des différentes versions de SimFeodal : les données générées par les différentes versions du modèle ne sont pas systématiquement « compatibles », c'est-à-dire qu'elles ne présentent pas toute exactement la même structure, à commencer par les variables enregistrées.

Dès lors, les différents rapports peuvent être considérés comme reproductibles et automatiques au sein de « générations » de SimFeodal, c'est-à-dire pour les versions créées et paramétrées dans le cadre d'une même phase de co-construction, avant donc d'avoir eu à adapter les indicateurs[9].

A l'issue de la conception et de l'implémentation de ces rapports automatiques, on dispose donc, pour chaque expérience, d'un document aisément partageable et lisible. Ce pourrait être la dernière étape de la création d'outils d'évaluation de SimFeodal si le nombre de versions ou d'étapes de SimFeodal était plus restreint. Pourtant, comme vu dans la partie 1 du chapitre, le paramétrage de SimFeodal a été caractérisé par une forte quantité d'allers-retours entre le modèle et ses résultats, entraînant à chaque fois de nouvelles expérimentations. De la

8. Au sens de l'évaluation interne, c'est-à-dire du bon fonctionnement, exempt de *bugs*, du modèle de simulation implémenté.

9. Pas clair du tout, à reprendre !

même manière qu'au regard du nombre d'indicateurs à évaluer il n'est donc pas possible de manipuler les indicateurs un par un dans des fichiers individuels, la masse d'expériences rend partiellement caduque l'utilisation unique des rapports automatiques. Il est facile de comparer, sur un même écran d'ordinateur, deux ou trois rapports, mais dès lors qu'il faut en comparer plus que cela, la manipulation conjointe des rapports devient complexe, tout autant que d'avoir une vision globale des résultats principaux de chaque expérience.

### 5.2.4 Organiser les rapports : Dashboards

Pour être en mesure de comparer de nombreux éléments, il est nécessaire de passer d'une exploration linéaire, voyant défiler les indicateurs les uns après les autres, à une exploration globale et interactive. C'est-à-dire que plutôt que de voir se succéder des pages d'indicateurs, mieux vaut une interface présentant les points clefs de l'évaluation et permettant d'entrer dans le détail de chacun des indicateurs après-coup.

Cette logique, assez universelle désormais, est celle qui préside à la création des nombreux « tableaux de bord », ou « *dashboards* » que l'ont voit émerger depuis la fin des années 1990. Très répandus dans le monde de l'informatique décisionnelle (*Business Intelligence, BI*), ces outils permettent d'explorer des données d'entreprises, par exemple des résultats financiers. Pour ce faire, ils mettent en avant, dans une interface unique, des indicateurs clés (*Key Performance Indicators, KPI*), qu'il est ensuite possible de filtrer et d'affiner, par exemple par sélection de différents intervalles temporels.

L'avènement des données massives et de leur prise en compte pour la gestion des villes (*smart cities*) a amené à une utilisation de plus en plus fréquente de ce type de *dashboards* en géographie (Batty 2015 ; Kitchin, Lauriault et McArdle 2015 ; Roumpani, O'Brien et Hudson-Smith 2013), tant les villes sont des objets complexes à observer et à gérer. Au regard des problématiques de gestion urbaine autant que de celles de supervision des performances d'entreprises, on comprend bien le parallèle qui peut être fait avec une utilisation appliquée à un modèle complexe à évaluer tel que SimFeodal.

L'analogie entre les indicateurs clés (KPI) et les indicateurs numériques de SimFeodal d'une part, et entre les indices issus de fouille de données du monde de l'informatique décisionnelle et les indicateurs multi-dimensionnels d'autre part, est assez explicite.

### 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