

Que peut-on faire sans recourir
à la simulation par agents ?

Une introduction...

Un point d'étape...

Plan de la présentation

- Du concret vers l'abstrait
 - Réflexions méthodologiques sur l'étude Immigrés
 - Présentation d'une approche statistique applicable à cette étude
 - Réflexions préliminaires pour une généralisation
- Les déterminants du choix d'une approche *Simul or Not Simul*
 - Les objectifs poursuivis
 - Les connaissances accessibles sur le système
 - Les moyens disponibles
- Les modélisations possibles
 - Classification et caractéristiques des différents types de modélisation
 - Méthodes applicables
- Conclusion
- Tableau de synthèse

Réflexions théoriques sur l'étude « Immigrés »

- Cette étude comporte deux types de modélisation

1. Un modèle Agents / Interactions

- Les agents sont essentiellement les immigrants et leurs attributs, ainsi que les groupes sociaux de leur environnement
- Les interactions sont les influences d'agent à agent, ou de groupe social à agent
- Une telle modélisation comporte des hypothèses difficiles à établir et à vérifier

2. Une enquête sur un échantillon (non encore effectuée)

- Cette enquête se traduira par un *tableau des attributs* des immigrants interrogés
- Pour chaque immigrant, le tableau comportera ses divers attributs *quantifiés* (ou 0/1)
- Certains de ces attributs caractériseront son *degré d'intégration* (M. Safi)
- Parmi les autres attributs, *les plus influents* sur l'intégration seront recherchés
 - Ce sont les Facteurs-clés

Les objectifs poursuivis

- Dans l'étude Immigrés, la recherche porte sur les « *facteurs-clés* »
- Ce sont les attributs des immigrants ou de leur environnement qui ont la *plus grande influence* sur leur *niveau d'intégration*
- Outre l'intérêt de ces facteurs dans la *compréhension* du système, certains d'entre eux sont susceptibles d'être *modifiés* par des *interventions* des pouvoirs publics ou locaux
- Ces interventions pourraient alors être utilisées dans le cadre d'une *politique publique* de l'accueil et de la gestion des immigrants

Petit lexique

- Immigrés = Agents = Individus = Points (selon le contexte)
- Attributs (des agents) = Propriétés = Facteurs (-clés ou non) = Variables (à expliquer...), Critères...

Les approches envisagées – 1) la simulation

- Elle serait basée sur la modélisation par *agents / interactions*
- Une telle simulation a toujours un caractère *dynamique, diachronique*
- Elle permettrait, *en principe*, de tester l'influence, après un certain temps, de diverses interventions sur l'environnement, et ainsi *d'estimer les actions les plus appropriées* pour améliorer le taux d'intégration, *après un temps suffisant* pour que se manifeste l'effet de ces interventions
- Mais dans l'étude effectuée, *l'impossibilité de simuler ne permet pas* cette approche diachronique
- Une *approche synchronique*, modélisation *sans simulation*, permet de comprendre et d'analyser le système sur le plan statique, mais *pas de prédire son évolution* future, ni de déterminer les *facteurs-clés*
- En effet, les *interactions immédiates ne suffisent pas*, même si elles semblent significatives, du fait des boucles de rétroaction qui peuvent en altérer l'effet (*« effets pervers » par exemple*)
- *Existe-t-il une autre approche plus prometteuse ?*

Les approches envisagées – 2) la régression

- La *régression multiple* est une technique statistique (*résumée ci-après*)
 - Elle porte sur le *tableau de résultats de l'enquête* envisagée auprès d'un échantillon d'immigrés
 - *Pour chaque immigré* interrogé, l'ensemble de ses *attributs significatifs* seront renseignés et codés
 - Certains de ces attributs permettront d'estimer le *degré d'intégration* de l'immigré : c'est la *variable « à expliquer »* (ou dépendante)
 - Les autres attributs seront *automatiquement corrélés* avec cette variable : ces *coefficients de corrélation* constituent les poids des différentes *variables « explicatives »*, ou indépendantes (*ou variantes : ACP, ANOVA*)
- Les attributs *les mieux corrélés* constituent les « *facteurs-clés* » recherchés

Approche statistique des facteurs-clés : Pourquoi ça marche ?

- Déterminer les facteurs-clés : pourquoi ça ne marcherait pas ?
 - Car cette approche statistique n'est *pas diachronique*
 - De plus, l'effet des *rétroactions* n'est *pas pris en compte*
- Pourquoi ça doit marcher (quand même) ?
 - La simulation par agents est une approche *bottom-up* : on part des interactions de détails et *on construit, pas à pas*, les résultats globaux (*émergents*) qui s'en déduisent, dont le niveau d'intégration, en tenant *nécessairement compte* des interactions et rétroactions
 - Mais le questionnaire, lui, *fournit déjà ces résultats globaux*, et en particulier le *degré d'intégration* émergent
 - Le degré d'intégration, qui était l'*inconnue* du premier modèle (*agents/interactions*) devient une *donnée* du nouveau modèle (*le questionnaire*)
 - Les effets des *interactions* et de la *diachronie* sont *intégrés d'emblée* dans les résultats du questionnaire et sont donc *implicitement pris en compte*
 - Les attributs corrélés sont donc jugés par leur *effet réel, à long terme* : il s'agit d'une approche *top-down*, ne nécessitant pas d'analyse « microscopique »

La Régression Multiple - Principes

L'objectif général de la *régression multiple* (*le terme a été utilisé initialement par Pearson, 1908*) est d'en savoir plus sur la relation entre plusieurs variables *indépendantes ou explicatives* et une variable *dépendante ou à expliquer*.

Par exemple, un agent immobilier peut enregistrer pour chaque programme, la taille de la maison (en mètres carrés), le nombre de pièces, le revenu moyen du voisinage respectif, et un taux subjectif de "charme" de la maison.

Une fois cette information *collectée pour diverses maisons*, il peut être intéressant de voir si ces mesures sont liées au *prix de vente* de la maison.

Par exemple, on peut apprendre que le nombre de pièces est un *meilleur indicateur* de prévision du prix de vente d'une maison dans un quartier, plutôt que son aspect extérieur.

(Texte extrait d'un cours sur Internet)

La Régression Multiple - Exemples

- En sciences naturelles et sociales, des procédures de régression multiple sont largement utilisées en recherche
- En général, la régression multiple permet au chercheur de se poser la question générale (et avec un peu de chance, d'y répondre)
« quel est le meilleur indicateur permettant de prévoir ... »
- Par exemple, des chercheurs du domaine éducatif peuvent souhaiter en savoir plus sur les indicateurs qui permettent *le mieux* de prévoir la réussite au baccalauréat
- Les *psychologues* peuvent vouloir déterminer les variables de personnalité qui prévoient *le mieux* la mobilité sociale
- Des *sociologues* peuvent s'intéresser aux multiples *indicateurs sociaux* permettant de prévoir *au mieux* si tel ou tel groupe d'immigrants *risque de bien s'adapter, et donc de s'intégrer dans la société*

Principes de la régression linéaire

- Une variable quantitative Y dite *à expliquer* est mise en relation avec p variables quantitatives $X^1, \dots, X^j, \dots, X^p$ dites *explicatives*.
- Les données sont supposées provenir de *l'observation d'un échantillon statistique* de taille n ($n > p + 1$) de $p + 1$ dimensions quantitatives :

$$(x_i^1, \dots, x_i^j, \dots, x_i^p, y_i) \quad i = 1, 2, \dots, n$$

- L'hypothèse d'un *modèle linéaire* dans cette situation conduit à supposer que :

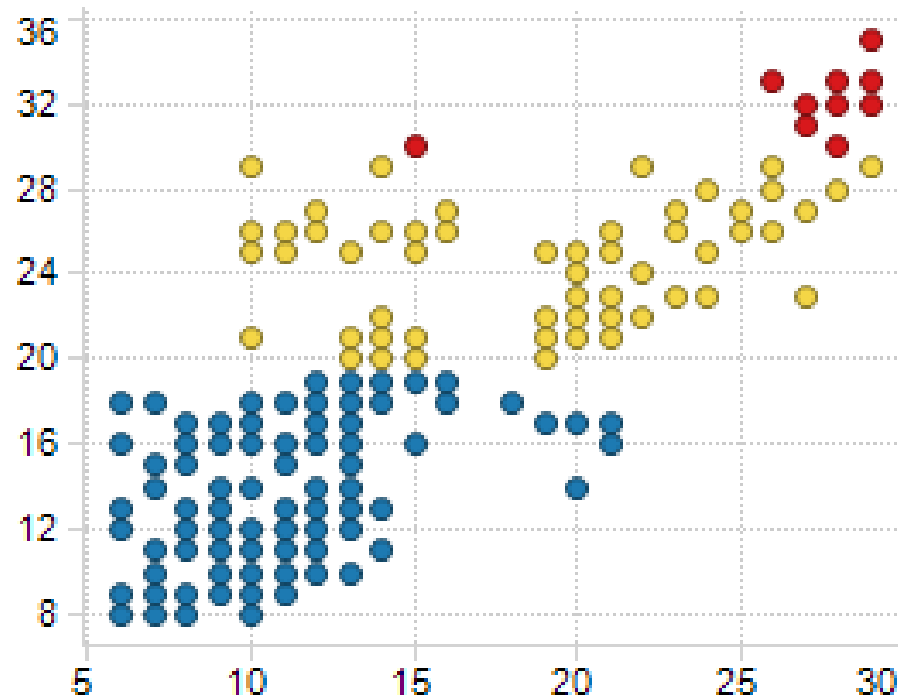
$$y_i = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + \dots + \beta_p x_i^p + \varepsilon_i \quad i = 1, 2, \dots, n$$

- Les coefficients β_j , *communs à tout l'échantillon*, représentent les *poids relatifs* des X^j dans la variable à expliquer
 - Les X^j dont les poids sont les plus élevés sont les *facteurs-critiques* (ou *-clés*)
- Il existe de nombreux *logiciels statistiques* permettant de calculer ces régressions

(Texte adapté de cours sur Internet)

Une variante : l'analyse en composantes principales

- Cette approche permet de *visualiser* les positions des immigrés comme un *nuage de points* représentés sur 2 (ou 3) *axes « principaux »*, *explicatifs* de leur *degré d'intégration*
- Chaque axe est lui-même une *combinaison linéaire de facteurs-clés*
 - *Il convient donc d'« interpréter » intuitivement chaque axe*
- Un *codage couleurs* peut représenter le *degré d'intégration*



Recherche d'une méthode d'analyse d'un système complexe

Simulation : applications et limitations

Autres approches : « NoSimul »

Les déterminants de la méthode à appliquer

- Le choix d'une méthode d'analyse repose sur trois déterminants principaux :
 1. Les objectifs poursuivis (*en grandes catégories*)
 - *Connaître* les mécanismes, *ajuster* les paramètres
 - *Comprendre* la nature et le fonctionnement du système
 - *Prédire* (*en faisant varier certaines hypothèses*)
 - *Agir, optimiser*
 2. Ce qu'il est possible de savoir en observant / étudiant le système ; selon les cas :
 - Les *mécanismes intimes* (*agents, interactions*), *toujours existants mais pas toujours accessibles*
 - Connaissance des *attributs des agents* mais ignorance de leurs *interactions* mutuelles (*Immigrés...*)
 - Observation des *comportements des agents* (*Big Data, épidémiologie, sociologie...*)
 - Ignorance des agents mais observation du *comportement du système* (« *boîte noire* » : *Stimuli → Réponses : Psychologie, robot, véhicule autonome, phénomène physique...*)
 3. Les ressources disponibles : budget, temps, compétences, moyens matériels...
- La modélisation dépend de ce qu'on peut connaître en pratique du système
 - *Tout système* comporte *en général* des agents en interactions mutuelles, mais si ces éléments *ne sont pas observables*, il faudra recourir à un autre type de modélisation
 - En revanche, pour un même système, *plusieurs modélisations* sont parfois possibles

Place de la simulation

- La simulation par ordinateur est, lorsqu'elle est possible, *la plus puissante* des méthodes d'analyse et de prédiction d'un système
- Elle permet de réaliser pratiquement *tous les objectifs précédents*, pourvu que le modèle soit *suffisamment précis*
- Mais son *domaine d'application* est très limité :
 - Elle exige une *modélisation quantitative* par agents et interactions
 - Même si une telle modélisation est possible en théorie, elle peut exiger un effort en équipes et en moyens *disproportionnée* aux objectifs poursuivis et à l'ambition du projet
 - Même si l'on dispose d'un tel modèle, un projet de simulation peut être très long, complexe et coûteux
- Il faut alors recourir à d'autres approches, abordées ci-après

Les différents modèles et les
méthodes applicables

Classification des modèles en deux classes

1. Modèles « microscopiques » : Agents et interactions

- Les seuls à permettre une simulation
 - Ne demandent pas de connaître, mais permettent d'établir, les *propriétés émergentes* du système
 - Mais si l'on peut observer ces propriétés, on peut les utiliser pour *calibrer le modèle* après simulation
 - La simulation permet d'évaluer *l'effet global, cumulatif*, des interactions (*approche bottom-up*)
 - La simulation permet d'estimer les *propriétés émergentes* sous *différentes hypothèses*
 - *Propriété fondamentale pour préparer et évaluer toute action sur le système*
 - La *qualité des prévisions* dépend de plusieurs facteurs
 - *Précision* et *fiabilité* de la modélisation
 - Influence plus ou moins importante des *aléas* du modèle
 - Mais des approches statistiques (*Monte-Carlo*) permettent de traiter les modèles aléatoires de manière *probabiliste*
 - *Stabilité intrinsèque du système* dans le domaine étudié
 - Instabilité = *grandes variations* des résultats pour de *petites altérations* des paramètres
 - Le degré de stabilité peut être estimé par une *étude de sensibilité* (impliquant *plusieurs simulations*)
- La simulation est *nécessaire* pour tenir compte des interactions cumulées
- Hors simulation, on se « contentera » de résultats qualitatifs ou observationnels

Division des modèles en deux classes

2. Modèles « macroscopiques » : comportent des propriétés émergentes

- Si les agents *ou* leurs interactions sont inconnus, ces modèles *ne permettent pas la simulation*
- Mais il est possible d'*exploiter les propriétés émergentes* du système *si elles sont observables*
- Ils peuvent aider à préciser les *rôles des différents facteurs* sur ce comportement global
- Cela permet *d'anticiper l'effet* de diverses *interventions* sur le comportement global
 - *Propriété fondamentale pour préparer et évaluer toute action sur le système*
- La modélisation tient compte *implicitement* de l'effet global, résultant, des *interactions* et de la *diachronie*, car les propriétés émergentes observées *intègrent ces effets*
- La *qualité des prévisions* dépend de plusieurs facteurs
 - *Précision* et *fiabilité* de la modélisation
 - *Stabilité intrinsèque* du système dans le domaine étudié (*définie précédemment*)
 - La stabilité peut être estimée par la *variance* des tableaux statistiques

➤ Il n'est *plus nécessaire* de simuler pour tenir compte des interactions ou du temps

Méthodes « douces » pour modèles microscopiques

- Un *modèle microscopique (Agents / Interactions)* permet *en principe* la simulation par agents sur ordinateur
- Mais dans de nombreux cas celle-ci est *contre-indiquée* :
 - Modèle trop imprécis ou trop aléatoire
 - Moyens insuffisants (*temps, équipes, compétences, budget...*)
- On peut alors tenter *par différentes approches* « douces » de *tirer le maximum* de la modélisation agents/interactions obtenue :
 - *Simulation « mentale », expérimentation*
 - Utilisation d'outils de *visualisation* pour rendre le modèle plus « intuitif »
 - *Par exemple, transformation de données en images de visages humains*
 - Exploiter les *interactions directes (sans itérer)* en espérant que les rétroactions, ainsi ignorées, n'auront pas d'effets trop importants dans le cas du système étudié

Méthodes Sans Simul pour modèles macroscopiques :

- Modèles par agents / attributs, *interactions ignorées*
 - *Tableaux statistiques « multivariées »* : propriétés d'un échantillon de taille limitée
 - Permettent des *traitements statistiques* (*Régression, ACP... Exemples: études sociologiques, médicales...*)
 - *Big data, data lakes* : saisie ± structurée du comportement d'un grand nombre d'agents
 - Relèvent de méthodes beaucoup plus *sophistiquées* (*mise en forme, approches Big Data, stats, IA...*)
- Modèles dont les agents et les interactions sont *inconnus* ou volontairement *ignorés*
 - Modèles « *boîte noire* » : on a seulement accès au comportement global *stimuli → réponses*
 - A) Un programme peut *enregistrer* et *structurer* le modèle puis, en présence d'un nouveau stimulus, *interpoler* entre les divers stimuli appris (*IA, apprentissage profond...*)
 - *Programmation de véhicules autonomes...*
 - B) Ou on peut rechercher un modèle par agents/interactions *à la base explicative du système*
 - On peut alors *tester* et *valider* ce modèle *par simulation* et comparaison avec la boîte noire
 - *Recherche scientifique* (Exemples : *cosmologie, recherche médicale... : élaboration et tests de modèles causaux*)
 - *Recherche SHS* (*Elaboration et tests de modèles économiques, sociologiques...*)
- Modèles spécifiques à certaines méthodologies
 - Ex : *modèles « évolutionnaires »* utilisés pour l'*optimisation* (*Génotypes, phénotypes, reproduction*)
 - Mise au point de designs optimaux (*antennes, ailes d'avion...*)
 - Optimisation de parcours (*« voyageur de commerce » ...*)

Conclusion

- Les méthodes que l'on peut employer pour analyser un système complexe dépendent de *trois facteurs* :
 1. Les objectifs poursuivis
 - Mieux *connaître* le système ; *prévoir* ; *agir*...
 2. Les connaissances accessibles sur le système
 - Données microscopique (*Agents, interactions*), ou macro (*comportements globaux observés*)
 - A chaque observable correspond un *type de modélisation*
 - *Agents/interactions, tableaux d'attributs, boîtes noires, analogies...*
 - Les *méthodes* utilisables dépendent du *type de modélisation* disponible
 3. Les moyens techniques, financiers et humains disponibles
- Lorsque c'est possible, on peut *combiner plusieurs approches* sur le même système pour mieux en appréhender ou préciser les *divers aspects*

APPROCHES D'ETUDES DE SYSTÈMES AVEC OU SANS SIMULATION								
MODELISATION				OBJECTIFS POSSIBLES (plus ou moins précis)				
Type de MODELISATION (dépend des observables du système)	Exemple(s)	SIMULATION PAR AGENTS théoriq ^t possible ?	APPROCHES de MODELISATION	COMPRENDRE la structure et les mécanismes du système	VALIDER le système, AJUSTER les paramètres	ESTIMER les propriétés globales (émergentes)	PREDIRE le comportement sous diverses hypothèses	Comment AGIR au mieux sur le système ou sur son environnement
Agents et interactions, modèle QUANTITATIF (Réaliste ou schématique)	Exemples Vadémécum Systèmes techniques, scientifiques, économiques ou sociologiques	OUI EN THEORIE NON si le projet est trop lourd pour l'équipe	Etude Documentation Expérimentations Enquête(s)	OUI Résulte de la modélisation	OUI si observations globales Par comparaison entre simulation et observations (NON si NoSimul)	OUI Résulte de la simulation (NON si NoSimul)	OUI Résulte de la simulation (NON si NoSimul)	OUI Résulte de la simulation (NON si NoSimul)
Agents et interactions, modèle qualitatif FORTES RETROACTIONS	Modèle Immigrés initial (après restructuration éventuelle)	NON (OUI sous pseudo- quantification)	Etude Documentation Expérimentations Enquête(s)	OUI Résulte de la modélisation	NON (Sauf à pousser plus loin l'étude)	NON	NON Les rétroactions peuvent s'opposer aux interactions directes	NON (ou avec prudence) Les rétroactions négatives peuvent provoquer des "effets pervers"
Agents et interactions Modèle qualitatif FAIBLES RETROACTIONS	Modèle Handicap	NON (OUI sous pseudo- quantification)	Etude Documentation Expérimentations Enquête(s)	OUI Résulte de la modélisation	NON (Sauf à pousser plus loin l'étude)	PLUS OU MOINS Du fait des faibles rétroactions	OUI Du fait des faibles rétroactions	OUI En appliquant les interactions directes
Connaissance d'un SYSTÈME ANALOGUE	Chiffrage d'un chantier complexe (Centrale nucléaire, aéroport...)	NON	Comparaison Evaluation des différences Facteur d'échelle ?	OUI Résulte de l'analogie	NON	PLUS OU MOINS Selon la précision de l'analogie	PLUS OU MOINS Selon la précision de l'analogie	PLUS OU MOINS Selon la précision de l'analogie
TABLEAU Agents et leurs propriétés	Enquête Immigrés (tableau de réponses) Autres études sociologiques	NON	Etude Documentation Enquête(s)	OUI Régression/corrélation ACP (Analyse en composantes principales)	Système valide par hypothèse	Elles sont incluses dans la modélisation	OUI Régression, ACP	OUI Régression, ACP (Mise en évidence des "facteurs principaux")
BIG DATA Amas de données Agents/Comportements	Données épidémiologiques Comportements humains "capturés" sur ordinateurs	NON	Capture systématique informatisée de comportements + Prétraitements éventuels	OUI (Approche BIG DATA) Relevé de régularités et de corrélations (mais pas toujours causales)	Système valide par hypothèse	Elles sont incluses dans la modélisation	OUI Dans la mesure de la nature causale des corrélations relevées	OUI Dans la mesure de la nature causale des corrélations relevées
BOITE NOIRE Observation des comportements Stimulus → Réponse	Comportements humains Psychologie Sociologie	NON	Observations Expérimentations Enquêtes	NON Mais on peut proposer des modèles explicatifs	Raffinements progressifs par tâtonnements (essais/erreurs)	Elles sont incluses dans la modélisation	OUI IA ou interpolation entre comportements connus	OUI IA ou interpolation entre comportements connus
Modèle EVOLUTIONNAIRE Génotype, phénotype, reproduction	Solutions optimales (chemins, structures : voyageur de commerce, design sous contraintes)	OUI (Etape de reconstruction du phénotype)	Recherches de génotypes efficaces	On peut interpréter un génotype efficace	Ajustement d'un génotype efficace	NON	NON	OUI par construction d'une solution optimale