Problèmes méthodologiques posés par la simulation de processus de production de services - *version provisoire*

L'exemple d'un service d'urgences

Savas Balin* — Vincent Giard*

Lamsade, Université Paris-Dauphine
Place de Lattre de Tassigny 75016 Paris
Savas.Balin@dauphine.fr, Vincent.Giard@dauphine.fr

RÉSUMÉ : *L'amélioration de l'efficacité et de l'efficience des processus de production de services passe par une mobilisation des techniques de simulation. Dans ce cadre, la modélisation d'un processus dépend des objectifs poursuivis et des propriétés de l'approche de la modélisation mobilisée (simulation basée sur les objets, simulation orientée objet, simulation multi-agents). On discutera ces problèmes méthodologiques avant de les illustrer avec une simulation détaillée des processus d'un service d'urgences.*

ABSTRACT : *The use of the simulation to improve the efficiency and the effectiveness of the service production is indispensable. In addition, the process modeling depends on the objectives and the properties of the approach used (object based simulation, object oriented simulation, agent based simulation). We will discuss these methodological problems and give a service simulation example of an emergency department.*

MOTS-CLÉS : *Approches de modélisation et de simulations, processus de production de services, amélioration de performance, efficience et efficacité.*

KEY WORDS : *modeling and simulation approaches, service production process, performance improvement, efficency and effectiveness.*

1. Introduction

Le discours tenu par les spécialistes des services procède souvent d'une généralisation abusive liée à une insuffisance de vision processus. Les questions fondamentales posées par l'amélioration de la production de services tournent toujours autour de l'efficience, qui n'implique normalement que le producteur, et de l'efficacité qui implique le producteur mais aussi, presque toujours et pas en permanence, le client, ce qui pose des problèmes spécifiques, notamment en matière d'indicateurs de performance et de qualité perçue. Les leviers d'amélioration de la performance de la production de services sont liés à des modifications de ressources, de structures et de procédures. La prédictibilité de l'impact de l'usage de ces leviers sur l'efficacité et l'efficience est difficile. Il est de moins en moins possible de l'aborder de manière empirique par une méthode de type essai/erreur dont la pertinence est liée à un hypothétique régime de croisière et la possibilité, pour l'entité concernée, de survivre quelles que soient les conséquences des actions décidées. Résoudre ce problème de manière analytique n'est pas imaginable en raison de la complexité des processus. Reste l'approche simulatoire qui permet de mieux comprendre le comportement d'un système productif complexe sollicité d'une certaine manière et évite de s'attacher à tort à des problèmes induits, dès lors que leurs origines sont identifiées. La simulation permet alors l'analyse de scénarios alternatifs en terme de structures, ressources et procédures, et d'en évaluer la pertinence et l'intérêt.

Dans une première partie on examinera les problèmes méthodologiques posés lors de la création et de l'utilisation de modèles de simulation relevant de l'une des trois approches de simulation disponibles. On soulignera, pour la plus connue des trois, mise au point pour décrire les processus de production de biens, les problèmes posés par son usage dans la description de processus de production de services, en les illustrant par l'exemple d'un service d'urgences. La seconde partie est consacrée à la simulation de ce service, ce qui permettra de revenir sur des considérations développées en première partie et d'illustrer le caractère incontournable de cette approche pour détecter l'importance de relations causales indirectes peu intuitives, permettant d'améliorer les processus.

2. Analyse des concepts et outils de simulation de processus de production de services

Le pilotage de processus de production implique une mise en relation de trois sous-systèmes. Le premier correspond à une réalité tangible, tandis que les deux autres relèvent d'une modélisation de ce que le processus peut et doit faire.

- Le *sous-système opérant* correspond à l'ensemble des ressources matérielles (machines, outillage, aires de stockage, composants et matières premières…) et humaines (opérateurs), disponibles dans le système sur la période considérée.

- Le *sous-système d'information* comporte un ensemble d'informations techniques (gammes et nomenclatures), un ensemble d'informations de gestion décrivant l'état du système (utilisation des ressources, commandes à exécuter, avancement des commandes en cours d'exécution…) à des instants donnés ou en continu, et un ensemble de procédures mobilisées dans la prise de décision.

- Le *sous-système de conduite* utilise le système d'information pour piloter le système opérant. Un certain nombre de décisions relevant des décisions programmables (au sens de H. Simon) sont prises automatiquement par des applications informatiques (ERP…) ou par des opérateurs appliquant la procédure définie pour résoudre une classe de problèmes décisionnels, sans réelle marge de manœuvre. Les autres décisions impliquent une intervention humaine pour diagnostiquer le problème décisionnel posé, choisir la procédure à appliquer si plusieurs sont mobilisables (une procédure pouvant laisser une marge d'interprétation importante), ou prendre directement une décision pour résoudre un problème d'un type nouveau, ce qui revient à créer une procédure.

« La simulation consiste à faire évoluer une abstraction d'un système au cours du temps afin d'aider à comprendre le fonctionnement et le comportement de ce système et à appréhender certaines de ses caractéristiques dynamiques dans l'objectif d'évaluer différentes décisions » (Hill, 1993). En s'intéressant à la simulation de processus de production on pourrait penser qu'il suffit de créer une représentation du sous-système opérant et de reprendre les deux autres sous-systèmes qui relèvent déjà d'une codification d'informations factuelles ou procédurales. Il n'en est rien. Pour décrire les concepts et approches disponibles, on commencera par présenter l'approche la plus répandue, la « simulation basée sur les objets »; ce faisant, on mettra en évidence certaines limites de cette approche, en particulier dans la modélisation de processus de production de services, ce qui amènera à examiner trois approches complémentaires, la « simulation basée sur les objets », la « simulation orientée objet » et la « simulation multi-agents ». On terminera cette section par l'examen des principaux problèmes posés par le choix d'une approche et son implémentation.

2.1. Simulation basée sur les objets (SBO)

La simulation vise à reproduire le fonctionnement d'un système réel, à travers un ensemble interdépendant de composants; examinons les principes habituellement retenus par les simulateurs de processus de production basés sur les objets pour construire un modèle de simulation.

- Le système opérant est généralement décrit par: des processeurs, des stocks, des ressources en personnels (ressources-personnes) ou en outillage partagées par plusieurs processeurs et des items à traiter, disponibles au début de la simulation et/ou rentrant au cours de la simulation par un (ou plusieurs) point(s) d'entrée pour y subir des traitements, avant de quitter le système par un (ou plusieurs) point(s) de sortie. Ces items peuvent être inanimés (composants…) ou non (personnes…); le second cas pose le problème d'une possible rétroaction de l'item sur le processus de production. Une cartographie du système productif est établie pour positionner dans l'espace les points

d'entrée ou de sortie des items, les processeurs et les stocks. Les flèches reliant des composants de cette carte visualisent les déplacements possibles des items dans le système ; l'ensemble constitue un graphe orienté. Dans cette modélisation du système opérant, les seuls composants actifs sont les processeurs : ils prélèvent des items dans des stocks d'approvisionnement auxquels ils sont reliés, effectuent le traitement requis, puis les expédient dans l'un des stocks de destination auxquels ils sont rattachés. Les ressources sont généralement visualisées à un endroit de la carte et se déplacent à la demande, au cours de la simulation, auprès des processeurs qui les mobilisent pour permettre l'exécution de certains traitements.

- Les simulateurs reprennent les trois parties du système d'information identifiées ci-dessus. L'organisation de ces données est cependant assez différente de celle que l'on observe dans le monde réel.
  - Le sous-système d'information de gestion est assez simple. Un item est caractérisé par les valeurs de l'ensemble de variables quantitatives ou qualitatives (qualifiés aussi d'attributs) qui lui sont assignées ; ces informations « embarquées » peuvent être modifiées au cours de la simulation. Un processeur inoccupé ou se libérant « connaît le contenu » des stocks avec lesquels il est en relation, ainsi que la disponibilité des ressources qu'il peut mobiliser. La connaissance de ce qui se passe au-delà de ce périmètre ne pose pas de difficulté mais implique l'usage du langage de programmation du simulateur.
  - Le sous-système d'information technique est réparti et sa vision d'ensemble, pas toujours aisée. Une partie de la nomenclature, celle relative à la liste des références que peut traiter un système productif, se décrit par l'une des variables attribuées à chaque item ; ce concept de nomenclature, classique en production de biens, est implicitement utilisé dans la production de nombreux services (pathologie d'un patient, par exemple). L'autre partie de la nomenclature, celle qui décrit les relations « composant-composés » lorsque ces relations ont un sens, est fusionnée avec la gamme. Une gamme se définit par la mobilisation d'une séquence de processeurs pour traiter un item rentrant dans le système productif à cette fin et, pour chaque processeur mobilisé, un temps de traitement et une taille de lot (unitaire par défaut). Dans le cas simple où les processeurs ne traitent qu'un item à la fois, le cheminement d'un item résulte de la concaténation des séquences « prélèvement de l'item dans un stock par un processeur lié à ce stock — traitement de l'item dans le processeur — envoi de l'item dans un stock de destination », observées entre son introduction dans le système et son départ du système. Ce cheminement, contraint par le graphe orienté de la modélisation du processus, résulte de l'exploitation par les processeurs des valeurs des variables caractérisant un item. Ces valeurs ne suffisent pas pour décrire ce cheminement, pas plus que les règles de sélection en entrée et en sortie qu'utilise un processeur. Certaines opérations d'une gamme impliquent un traitement par lot sur certains processeurs ; si la taille du lot est la même pour tous les items que peut traiter un processeur, cette information est rattachée au processeur ; dans le cas contraire, cette information est rattachée, d'une manière ou d'une autre, à la référence. La durée du traitement d'un item sur un processeur est la

dernière composante d'une gamme et son traitement s'inspire des mêmes principes que ceux que l'on vient d'évoquer pour les lots. La définition d'une gamme est plus délicate lorsque le traitement d'un processeur implique l'utilisation simultanée de plusieurs items de nature différente, cas de figure pour lequel les relations « composant-composés » de la nomenclature ont un sens. Lorsque le graphe de circulation des items est un arbre, la définition simultanée de la gamme et de la nomenclature est immédiate lorsque chaque stock ne porte que sur un même type d'item (référence) lequel ne peut figurer que dans un stock et que le traitement dans un processeur porte sur un ensemble d'items prélevés dans tous les stocks qui l'alimentent (logique d'assemblage). Dans les autres cas, il faut faire appel au langage de programmation du simulateur et la « reconstitution » de la nomenclature et des gammes est moins immédiate.

- Le sous-système d'informations procédurales est potentiellement intégré dans les options offertes dans le choix d'items à traiter par un processeur et celui du stock de destination après traitement. La panoplie des règles locales proposées permet facilement de simuler le fonctionnement d'un système décentralisé. La simulation de règles de pilotage plus complexes, pouvant aller jusqu'à l'émulation d'un pilotage centralisé utilisant toutes les informations disponibles au moment de la prise de décision, est toujours possible par l'exécution de procédures, déclenchée à cadence régulière ou par certains événements au cours de la simulation.

D'une manière générale, les concepteurs de simulateurs arbitrent entre la facilité de création d'un modèle de simulation, liée à l'usage de quelques principes de base facilitant un usage intuitif des composants et la largeur du spectre fonctionnel de chacun d'entre eux. Cet arbitrage se fonde sur une idée préconçue des besoins du marché. Lorsque les fonctionnalités de base d'un composant ne sont pas suffisantes, le résultat cherché est obtenu par l'exécution de procédures déclenchée par certains événements et/ou par l'utilisation de processeurs fictifs ; le niveau d'expertise requis pour cette modélisation est alors plus élevé.

Ces modèles de simulation de processus ont été initialement conçus pour décrire et simuler des processus de production de biens. Ce contexte de production jouit de trois caractéristiques : le traitement d'un item ne peut s'effectuer que dans un processeur occupant un endroit dédié à ce seul processeur ; les processeurs ou les ressources partageant les mêmes caractéristiques sont rigoureusement interchangeables ; enfin, ni les ressources, ni les items n'interagissent. Dans les processus de production de services lorsque l'item traité est un être humain, ces caractéristiques peuvent être plus difficilement acceptables.

- Le lieu d'exécution de certains traitements ne nécessitant pas d'environnement matériel spécifique peut ne pas être circonscrit, ce qui rend la définition spatiale du processeur moins précise ou peut conduire à ne pas lier obligatoirement un traitement à un processeur. Par exemple, les informations échangées entre un patient et l'infirmière qui l'accompagne à son box dans un service d'urgences sont une partie d'un « traitement ».

- Un lieu donné peut jouer successivement plusieurs rôles et devoir être considéré conventionnellement comme la répétition d'une séquence de « stock — processeur », chaque processeur effectuant des traitements différents. Ce cas de figure est rare dans la production de biens, sauf dans celle d'items très volumineux (hélicoptères…) où les problèmes de manutention et/ou d'espace disponible jouent en faveur d'une solution de déplacement des hommes et des équipements (et donc des processeurs) pour effectuer en un même lieu une séquence d'opérations de nature très différente. Dans la production de services, ce cas est fréquent lorsque l'item est un client. Le box d'un service d'urgences joue plusieurs fois le rôle de stock lorsque le patient est seul et de processeur, lorsque celui-ci se trouve en présence d'un médecin et/ou d'une infirmière pour l'un des traitements (diagnostic, soins, prélèvements sanguins…) qu'il doit recevoir. Le graphe du processus combine alors une dimension spatiale, lorsque les distances entre composants ont un sens, et une dimension « organigramme de traitements », lorsqu'elles n'en ont pas. L'adoption de conventions graphiques additionnelles est alors souhaitable, pour faciliter la lisibilité de cette représentation qui, en plus d'être un support de conception du modèle de simulation, est un support de communication.

- La banalisation des ressources peut être remise en cause. Par exemple, si la première affectation d'un médecin à un patient arrivant aux urgences est aléatoire, on s'efforcera de garder ce médecin pour les autres actes médicaux effectués sur ce patient aux urgences impliquant un médecin de même spécialité, ces prestations pouvant ne pas s'enchaîner directement.

- L'hypothèse d'une passivité des clients est parfois contestable. Les individus n'acceptent pas nécessairement certaines contraintes découlant du fonctionnement du système productif lorsqu'elles apparaissent difficilement supportables, ce qui conduit à ne pas rentrer dans un système saturé, à ne pas respecter la discipline imposée dans une file d'attente… L'hypothèse d'une absence d'interaction entre un client et certaines ressources (opérateur) dans certaines prestations de services est contestable mais peut ne pas avoir d'incidence dans une évaluation globale de la performance d'un processus à partir d'indicateurs objectifs (par exemple, temps moyen de séjour dans le système productif). Par contre, cette interaction peut avoir une influence sur les évaluations individuelles (qualité perçue).

Certaines de ces hypothèses restrictives peuvent être levées avec la simulation orientée item ou la simulation multi-agents.

2.2. Simulation orientée objet (SOO) et simulation multi-agents (SMA)

Les principes de la programmation orientée objet, proposés dans les années soixante avec SIMULA sont repris dans Smalltalk d'Alan Kay (1993), qui introduit la communication par messages entre les objets, et dans le « C++ » de Bjarne Stroustrup (1986). L'une des versions les plus évoluées de cette approche est JAVA, créée au début des années quatre-vingt-dix sur la base de travaux de Sun Microsystems. On s'accorde à dire que les quatre caractéristiques définissant une approche orientée objet sont : l'identité, l'existence de classes, l'héritage et le polymorphisme (Blaha *et al.*, 2005).

Vers la fin des années soixante-dix, les approches multi-agents voient le jour, d'abord pour faciliter l'obtention de solutions numériques de problèmes complexes dans la lignée des approches de décomposition-coordination imaginées en programmation mathématique. Très rapidement, ces approches sont généralisées à l'émulation de comportements de personnes dans des systèmes de production. Des simulateurs commerciaux, basés sur l'approche multi-agents, voient le jour dans les années quatre-vingt-dix. Différentes approches sont désormais disponibles pour simuler des processus complexes.

De nos jours, la plupart des logiciels – et donc ceux de simulation – s'appuient sur une programmation orientée objet. On s'intéresse ici à la manière dont le concepteur d'un modèle de simulation peut tirer partie des possibilités offertes par un simulateur et comment il peut exploiter les résultats de la simulation. Ces simulateurs disposent tous d'une boîte à outils plus ou moins fournie et performante (composants, langage de programmation, interface graphique). Selon le type de simulation retenue, le composant qui fait l'objet d'un traitement est appelé item, objet ou agent, l'objet étant un item doté de caractéristiques additionnelles ou plus larges par rapport à celles de l'item et l'agent étant un objet doté de caractéristiques supplémentaires par rapport à celles de l'objet (figure 1).

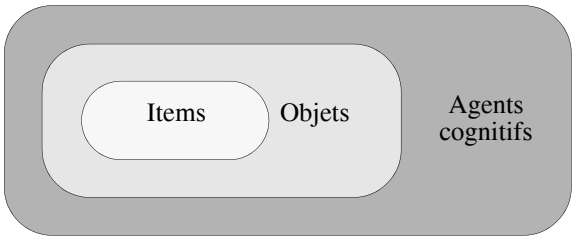

Figure 1. *Inclusion des ensembles item, objet et agent cognitif*

L'approche orientée objet propose une nouvelle manière de penser qui conduit à un nouveau processus de décomposition des problèmes. Avec cette approche, les problèmes abordés peuvent être modélisés par une collection d'objets qui prennent chacun en charge une tâche spécifique. La résolution du problème est conduite par la manière dont interagissent les différents objets. Une opération est simplement une abstraction d'un comportement analogue partagé par différents types d'objets. Chaque objet sait comment exécuter ses propres opérations (Hill, 1993).

Une simulation orientée objet (SOO) modélise le comportement des objets en interaction dans le temps. Chaque objet sait comment exécuter ses propres opérations. Les classes d'objets encapsulent les caractéristiques et les fonctions communes des objets. L'interaction entre les objets est établie par les messages (Joines *et al.*, 1999). Une telle simulation possède donc les avantages d'une conception orientée objet com-

me les propriétés d'« héritage », de « polymorphisme » et la création et la disparition de nouvelles ressources par l'intervention de l'utilisateur au cours de la simulation. La plupart des SOO permettent un large accès à leur langage de programmation de base ce qui donne une assez grande autonomie à l'utilisateur. (Blaha *et al.*, 2005). Les avantages offerts par l'utilisation de l'objet par rapport à celle de l'item sont représentés au tableau 1.

De l'item à l'objet	De l'objet à l'agent
- Types d'objets non limités - Possibilité d'implémentation d'un comportement individuel - Interaction directe entre les individus - Communication par messages - Possibilité d'occuper un espace - Possibilité d'accès aux codes - Polymorphisme - Héritage	- Perception - Adaptation - Raisonnement - Pro-activité (autonomie) - Croyances, émotions, désirs, intentions - Langage de communication plus élaboré

Tableau 1. *De l'item à l'agent, la logique d'enrichissement*

Pour les propriétés d'un agent qui le distinguent d'un objet, Davidson (2000) propose six :

- *Pro-activité.* Les objets sont réactifs tandis que les agents peuvent être autonomes.

- *Langage de communication.* Les objets utilisent une communication restreinte à des simples messages tandis que les agents peuvent utiliser un langage de communication entier.

- *Adaptation.* Les objets sont stimulés tandis que les agents peuvent apprendre tout seul.

- *Concepts mentaux.* Un agent, contrairement à un objet, est capable d'utiliser des concepts mentaux comme les croyances, les désirs et les intentions.

- *Notion de localisation.* Contrairement à un objet, les agents peuvent être affectés à des lieux.

- *Mobilité.* Les objets ne peuvent se déplacer que parmi les processeurs, tandis que les agents peuvent utiliser tout l'espace de la simulation.

Pour Ferber (1995), un objet, contrairement à un agent, est défini par un ensemble de services qu'il offre (ses méthodes) et qu'il ne peut refuser spontanément d'exécuter si un autre objet lui demande. Il le définit comme un « agent obéissant ». Les agents disposent d'objectifs qui leur donnent une certaine autonomie décisionnelle vis-à-vis des messages qu'ils reçoivent. Un agent ne doit pas être contrôlé par une règle de pro-

duction développée par le concepteur du système ; au contraire il est mû par un ensemble de tendances qui peuvent de surcroît évoluer dans le temps.

Dans une SMA, les agents sont capables d'intéragir en dehors des processeurs. Cette interaction, qui résulte de l'échange de messages entre agents, peut conduire à l'exécution de traitements ailleurs que dans un processeur. Certaines simulations multi-agents ne mobilisent pas de ressources et visent à décrire l'évolution d'un système d'agents en interaction représentatif du monde réel (écosystème, marché…) ou créé pour faciliter la résolution de problèmes numériquement difficiles à résoudre. Dans ce cas, on est toujours en présence d'une simulation de processus mais il ne s'agit pas de processus de production et donc de simulation de processus de production multi-agents, auxquels on s'intéresse ici.

La simulation basée sur les objets (SBO) n'est pas précisément définie dans le milieu académique. On peut le définir comme « une variante restrictive de la simulation orientée objet (SOO) qui ne possède pas les propriétés d'héritage et de polymorphisme, qui donne un accès limité à son langage de programmation et qui offre à l'utilisateur un nombre limité de composants de base, permettant la construction modulaire d'une simulation ». La plupart des simulateurs commercialisés, utilisés souvent pour réaliser des simulations de processus de production de biens (comme Arena, Promodel, Simul 8 etc.), ne permettent en effet aux utilisateurs que la réalisation de ces simulations.

Technique de simulation \ Caractéristiques de la technique	**SBO**	**SOO**	**SMA**
Disponibilité de composants de base (stocks, processeurs…) de description d'un processus	Oui	Oui	Oui
Capacité de prendre en compte les relations directes entre les clients et/ou ressources-personnes	Non	Oui	Oui
Capacité de modéliser une communication directe entre les clients et/ou ressources-personnes	Non	Oui	Oui
Capacité de modéliser des ressources portables mobiles	Non	Oui	Oui
Existence de la notion de localisation	Non	Oui	Oui
Capacité de modéliser des clients réactifs	Non	Oui	Oui
Capacité de modéliser les comportements individuels	Non	Oui	Oui
Capacité de réaliser des évaluations individuelles	Non	Oui	Oui
Capacité de simuler les scénarios d'urgence	Non	Oui	Oui
Capacité de modéliser une communication élaborée entre individus	Non	Non	Oui
Capacité de prendre en compte les agents cognitifs	Non	Non	Oui

Tableau 2. *Comparaison des logiciels de SBO, SOO et SMA*

2.3. Choix préalables à l'élaboration d'un modèle de simulation de processus de production

Très souvent, un modèle de simulation est établi pour apporter des éléments de réponse à des questions portant sur l'efficacité d'un processus, c'est-à-dire sa capacité à répondre à des objectifs quantifiés, et sur son efficience, c'est-à-dire le niveau minimal de ressources à mobiliser pour atteindre une efficacité désirée. Dans ce contexte, certaines hypothèses simplificatrices jouent un rôle mineur et une simulation basée sur les objets peut suffire. Il convient cependant de s'interroger sur leur pertinence lorsque l'on utilise certains critères d'efficacité spécifiques à la production de services et sans objet dans celle de biens, comme ceux tournant autour de la qualité perçue. La question de la supériorité de l'approche orientée objet ou de l'approche multi-agents se pose.

La pertinence d'un modèle de simulation est une affaire de jugement parce que les relations postulées entre facteurs sous contrôle dans le modèle et les indicateurs d'évaluation associés aux objectifs qui lui sont assignés doivent paraître suffisamment plausibles et précises aux yeux du concepteur du modèle et de ses utilisateurs. C'est aussi une affaire de validation empirique, lorsque cela est possible, par le biais de comparaisons entre les observations du terrain et les résultats d'une simulation s'appuyant sur des hypothèses de fonctionnement comparables. Dans les deux cas, la disponibilité des données et le coût de mise au point d'un modèle de simulation[1], au regard des avantages escomptés du modèle de simulation à créer, conduisent nécessairement à des compromis.

Dans ce contexte, le niveau de détail retenu dans un modèle de simulation joue un rôle important et se traduit par des choix relatifs à la granularité temporelle, qui joue sur le niveau de détail des activités élémentaires dont les durées se mesurent en un multiple de l'unité de temps choisie, et au niveau de discrimination entre entités et entre ressources de même nature. Dans une modélisation basée sur les objets, quelques principes simples peuvent être mobilisés, comme : lorsqu'un processeur réquisitionne en permanence un même ensemble de ressources non partagées avec d'autres processeurs, il est inutile d'isoler ces ressources[2] ; la décomposition d'un traitement en un ensemble de traitements élémentaires à réaliser sur le même objet et mobilisant tous le même ensemble de ressources ne présente pas d'intérêt[3] ; il est inutile de prendre en compte des ressources surabondantes lorsqu'une action sur leur niveau n'est pas envisageable[4]… Dans une modélisation orientée objet ou multi-agents, certains de ces

1. Le « ticket d'entrée » est généralement plus élevé dans la simulation orientée objet ou la simulation multi-agents que dans la simulation basée sur les objets, en raison d'un travail de programmation plus important.

2. L'accueil d'un service d'urgences est un poste de travail comportant de manière stable lorsqu'il est « actif » : du mobilier, un téléphone, un ordinateur et une infirmière. Cet ensemble peut être vu comme un processeur.

3. La séquence d'opérations conduisant à un prélèvement sanguin sur un patient peut, sans problème, être ignorée au profit d'une opération globale.

4. Si le nombre de brancards disponibles dans un service d'urgences est toujours supérieur à la demande, cette ressource peut être ignorée.

principes peuvent être invalidés en raison du rejet de certaines limitations de l'approche basée objet (substituabilité parfaite de certaines ressources, absence d'interactions entre objets et/ou ressources…).

La prise en compte du comportement individuel des objets ou de certaines ressources, permettant des interactions directes, est une composante importante du niveau de détail d'un modèle de simulation qui conduit à privilégier une simulation orientée objet ou une simulation multi-agents. Le réalisme de la modélisation est, a priori, plus grand mais son intérêt, au regard des objectifs du modèle de simulation, n'est pas toujours évident. Il convient sans doute de distinguer la prise en compte du comportement individuel lié aux déplacements, de celui lié aux interactions directes entre agents ou objets. Si l'on décide de simuler à un niveau individuel le déplacement d'une personne, par exemple de la porte d'entrée d'un aéroport au guichet d'enregistrement de son vol, il faudra prendre en compte non seulement les interactions possibles entre la personne et celles qu'elle croise mais aussi la plus ou moins bonne connaissance de l'endroit où se rendre ; avec une simulation parfaite de ces comportements dans des conditions normales (pas d'incident grave ni d'afflux massif de clients), on doit pouvoir retrouver, à partir de l'analyse des temps individuels de déplacement tirés de la simulation, la distribution statistiquement observée dans la réalité ; les deux modélisations fourniront des enseignements similaires si l'on ne s'intéresse qu'à des problèmes d'efficience ou d'efficacité « collective ». La prise en compte d'interactions entre agents ou objets peut avoir une meilleure plus-value potentielle qu'il s'agit d'apprécier avant de se lancer dans cette direction. Elle peut jouer au niveau de l'efficience lorsqu'elle récuse l'hypothèse de banalisation des certaines ressources ; par exemple, faire en sorte que le même médecin suive un patient tout au long de son séjour dans un service d'urgences conduit nécessairement à une moins bonne utilisation de ce type de ressource et le gain potentiel de qualité perçue lié à ce suivi peut être contrebalancé, toujours au niveau perceptuel, par l'allongement de séjour dans le service d'urgences. Par ailleurs, la construction d'un indicateur de qualité perçue est difficile et l'évaluation finale est tributaire des règles de comportement encapsulées dans l'agent ou l'objet ; le risque de conclusions inscrites en grande partie dans les prémices n'est pas nul.

Une fois retenu un niveau de détail jugé pertinent, le modèle de simulation peut s'avérer difficile à appréhender dans sa globalité. Il est alors possible de réunir un ensemble de composants interconnectés pour créer un méta-composant et de répéter ce mécanisme d'agrégation autant de fois que nécessaire. L'objectif est de faciliter la vision globale du processus complexe étudié car la simulation s'effectue toujours au niveau de détail initial et le fonctionnement interne d'un méta-composant reste toujours accessible. Les mécanismes d'agrégation et de désagrégation n'ont pas d'impact sur la simulation. On peut ajouter qu'une fois établi un modèle à un certain niveau de détail, on peut souhaiter établir un modèle à un niveau plus détaillé, par exemple en décomposant un processeur en processeurs élémentaires. Les liens logiques entre les deux niveaux de modélisation sont généralement assez clairs, car ils résultent normalement de mécanismes de décomposition. L'utilisation du mécanisme de création de méta-composants à partir de composants du modèle détaillé pour établir un modèle similaire à celui d'origine redonne une cartographie des flux similaire mais conduit à

des résultats de simulation différents, parce que la distribution du temps de traitement d'un méta-composant, combinant des distributions des composants élémentaires, ne peut être la même que celle du composant correspondant dans le modèle d'origine.

3. L'exemple de simulation d'un service d'urgences

Dans plusieurs articles ciblant l'« amélioration des processus de production des systèmes de santé », le besoin de la simulation est mis en évidence et son utilisation est considérée comme indispensable. Ces remarques laissent penser que les recherches sur cet axe sont multiples et largement répandues. Mais il n'en est pas ainsi, nos recherches dans la base de données EBSCOhost, non exhaustives mais sûrement indicatives, où sont cotés la plupart des meilleurs journaux de la gestion de production, nous a montrés le contraire. La recherche d'articles publiés depuis 1990 sur ce sujet en a donné 55[5], que l'on peut regrouper sur trois catégories. Certains de ces articles (22/55) ne traitent pas de « simulation de système de production » mais d'essais réalisés dans les simulateurs, d'interventions à distance etc. Donc, le nombre d'articles traitant notre sujet n'est en réalité que 33. Certains des articles (26/33) ne traitent pas la « modélisation d'un système par la simulation ». Le reste des articles (7/33) présente quelques similitudes, car ils traitent de la « modélisation d'un système de production de soins de santé par la simulation du flux de patients ».

Une analyse approfondie de ces 7 articles montre que, la plupart sont trop simplistes (4/33) (Badri, Hollingsworth, 1993 ; Shim, Kumar, 2005 ; Swisher *et al.,* 2001, Swisher, Jacobson, 2002) et les reproches méthodologiques que l'on peut leur adresser sont les suivants :

- absence de rétroactions (boucles), le patient suit un chemin à sens unique ;
- niveau de détail insuffisant, agrégation excessive des tâches ;
- choix inapproprié des variables et des indicateurs, utilisation des temps d'attente en tant que données ;
- hypothèses difficilement acceptables (ne plus accepter de patients si les box sont occupés).

Très peu d'articles (3/33) (Altinel, Ulas, 1999 ; Blake, Carter, 1996 ; Rytilä, Spens, 2006) comportent une modélisation élaborée de système de santé. La même remarque a été faite par Jun *et al.* (1999) dans leur revue de littérature qui repose sur 20 ans : « curieusement, il existe peu d'articles qui fournissent la simulation des systèmes complexes ». Néanmoins, ces travaux amènent les remarques suivantes qui en limitent la portée :

5.Selon la présence du mot « simulation » dans le titre et des mots « soin de santé (healthcare ou health care) » ou « service d'urgence (emergency department) » dans le titre ou résumé ou parmi les mots-clés de l'article.

- du point de vue « efficience », c'est seulement le « niveau de ressources » qui est visé ;

- du point de vue « efficacité », pour diminuer le « temps de séjour » ou le « temps d'attente », ce sont toujours et seulement le « nombre de postes » et le « nombre de ressources humaines affectées à ces postes » qui sont mises en question ;

- des transformations de processus ne sont presque jamais envisagées.

La simulation d'un service d'urgences fictif proposée par notre travail, s'appuyant sur un recueil approfondi de données venant de plusieurs sources[6], permet d'illustrer les remarques méthodologiques faites dans les paragraphes précédents. Le point de vue privilégié est celui du service d'urgences ; la qualité perçue des prestations fournies est essentiellement liée au temps de séjour du patient dans le système (efficacité), la qualité des soins étant considérée comme indépendante de la durée du séjour. L'objectif de la simulation est de permettre une analyse simultanée de l'efficience des processus mobilisés et de leur efficacité dans le cadre d'un scénario initial et d'une recherche d'amélioration des processus, rendue possible par la mise en évidence de certains chemins critiques dans la simulation qui définissent des relations de causalité indirectes entre certaines variables de contrôle et certaines variables d'état. Dans ce cadre, une simulation basée sur les objets a été retenue.

3.1. Le modèle de simulation d'un service d'urgences

L'organigramme de la figure 2 illustre tous les chemins que peut emprunter un patient dans un service d'urgences et explicite certaines conditions sous-jacentes à certains aiguillages ou à la mobilisation de certaines ressources. La simulation utilise des gammes où les cheminements, temps opératoires et ressources mobilisées sont définis en probabilité. La figure 3 reproduit deux copies d'écran : la première est une vue générale de la circulation des patients dans le service et visualise le calibrage initial des ressources, la seconde est un zoom sur le méta-composant « box », illustrant le mécanisme d'agrégation associé à une simulation effectuée au niveau de détail le plus fin. Ce box est une illustration des remarques faites sur la multiplicité des rôles (stocks et processeurs) joués en un même lieu dans la production de certains services et sur le nécessaire usage de processeurs fictifs pour pouvoir implémenter des gammes sophistiquées (par exemple, l'ordonnancement des traitements et analyse dans un box relève d'une approche d'*open shop*).

Le patient commence à son parcours par l'accueil (lois d'arrivées variables au cours d'une journée de 24 heures et du type d'admission). S'il n'est pas réorienté ou conduit immédiatement au centre de déchoquage dans les cas les plus graves, il passe

6. Pour construire notre modèle, nous nous sommes appuyés sur les travaux que nous avons menés pendant 3 mois avec les étudiants du master 2 « droit et management de la santé » de la Faculté Jean Monnet, dont certains avaient des expériences aux CHU « St Antoine-AP-HP » et « Henri Mondor de Créteil ». Pour les données statistiques, nous nous sommes également appuyés sur des documents publiés par la MEAH et la DREES.

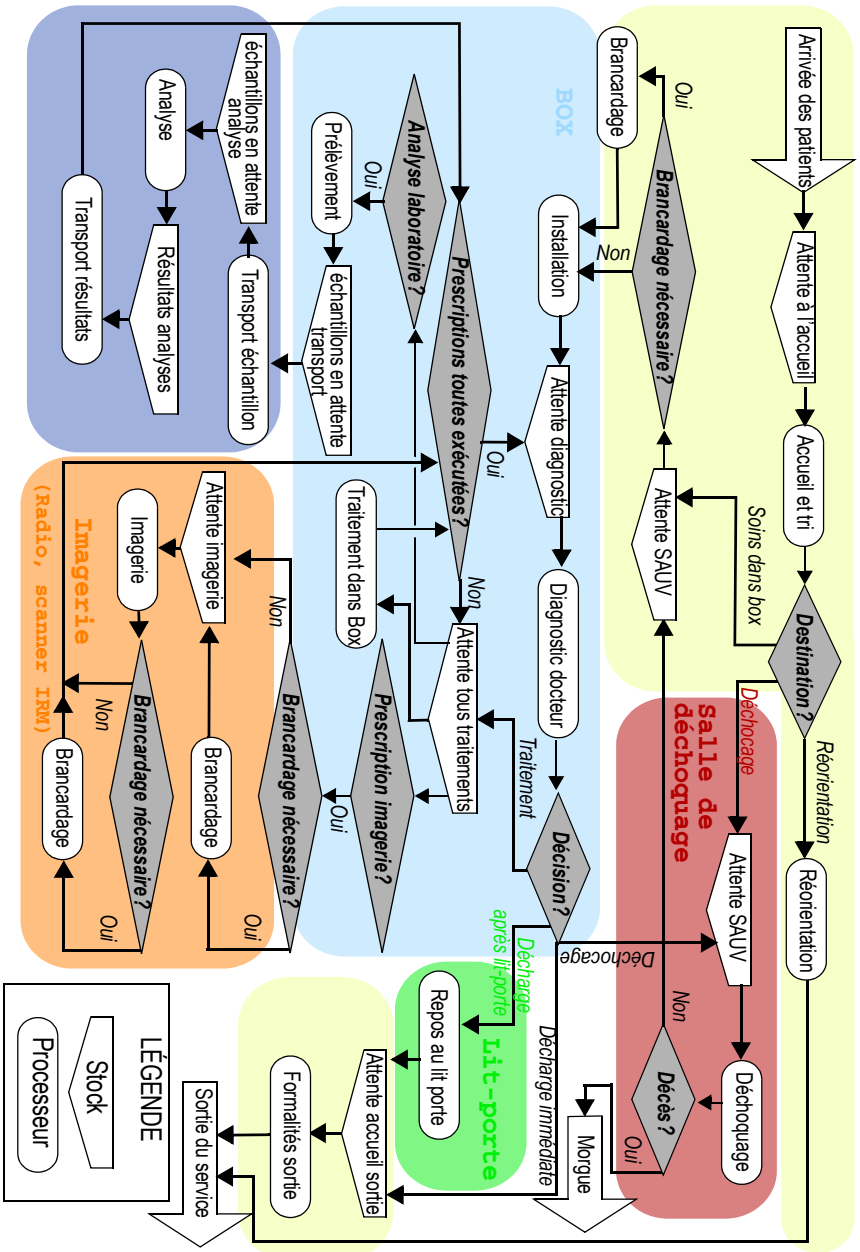

Figure 2. *Diagramme de flux du modèle*

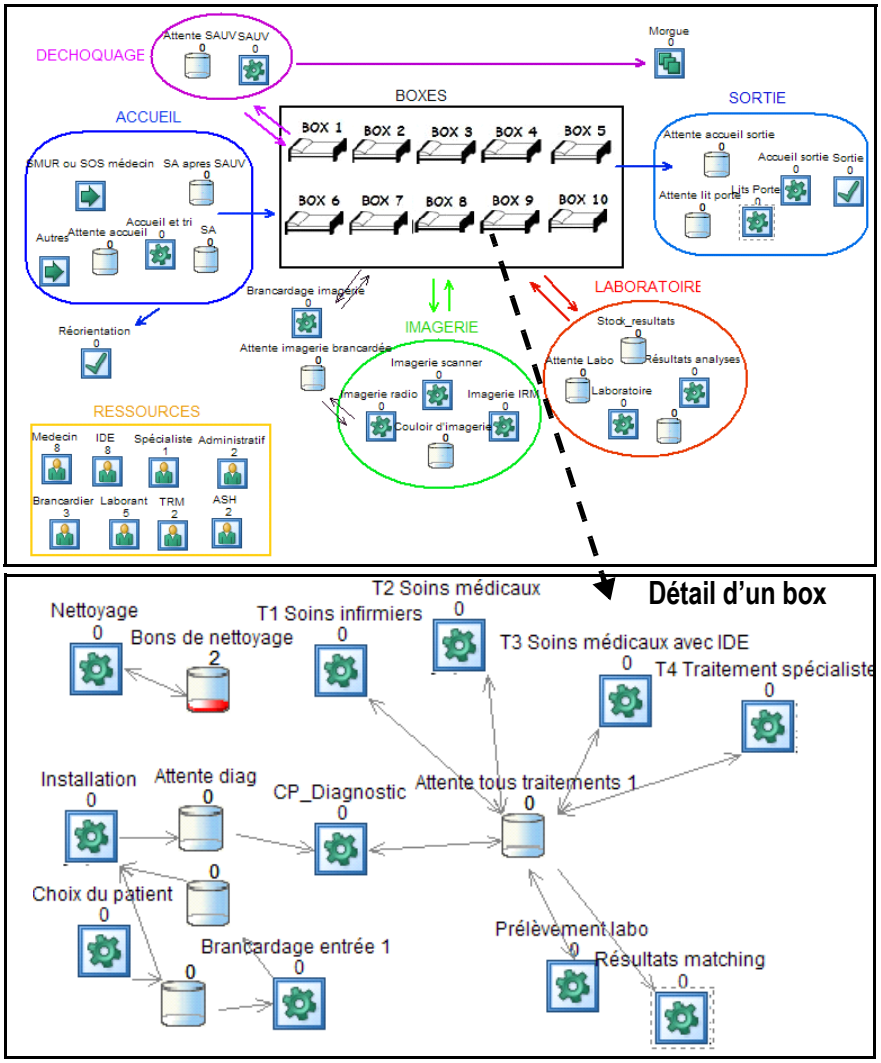

Figure 3. *Simulation d'un service d'urgence*

en salle d'attente avant d'être installé dans un box par une infirmière où il attend qu'un médecin l'examine, ce premier diagnostic conduisant à un ensemble de prescriptions (traitements, prélèvements, imagerie). Une fois ces prescriptions exécutées, le patient est revu par un médecin (pas forcément le même) qui peut à nouveau prescrire un nouvel ensemble de traitements et/ou d'examens. Dans la simulation, ce nombre d'itérations est limité à trois et le box est réservé à un même patient, tant qu'il n'est pas définitivement sorti (« déchargement » immédiatement ou après un repos au « lit-porte », envoi en salle de déchoquage).

- Quatre types de traitements, notés T1 à T4, peuvent être prescrits ; ils diffèrent par les ressources qu'ils mobilisent. Les probabilités des choix de traitements dépendent du niveau de gravité de l'état du patient (CCMU), attribué aux patients entrants ; faute d'informations fiables, la même distribution est utilisée, que le patient arrive avec le SMUR ou par ses propres moyens.

- Les besoins d'analyses sont déterminés en probabilité par le CCMU du patient ; faute d'informations fiables, l'indépendance est postulée entre le type de traitement et les besoins d'analyses. Les différentes prescriptions peuvent être exécutées dans un ordre quelconque (*open shop*), en fonction de la disponibilité des ressources mobilisées ; en réalité, l'importance du temps d'analyse d'un prélèvement conduit à commencer par le prélèvement, si cette prescription a été faite, afin de limiter le temps de séjour d'un patient dans le box (voir figure 5). L'un des traitements implique la présence d'un médecin spécialiste de l'hôpital ; on suppose ici qu'une demande du service d'urgences préempte celle de l'hôpital ; seul l'élargissement du périmètre à l'ensemble consolidé peut permettre une prise en compte plus réaliste de la disponibilité du spécialiste.

- En cas de prescription d'analyse sanguine, une infirmière prélève un échantillon de sang dans le box, après quoi le patient est disponible, si nécessaire, pour d'autres analyses ou traitement ; cet échantillon de sang est transmis au laboratoire ; les résultats de l'analyse sont ensuite acheminés directement au box.

- Plusieurs types d'image peuvent être prescrits (radio, scanner, IRM) avec des probabilités dépendant du CCMU du patient ; faute d'informations fiables, ces prescriptions sont considérées comme indépendantes en probabilité. Chaque centre d'imagerie possède un seul équipement. Les résultats sont remis au patient avant son retour au box.

- Si nécessaire, le patient est transporté par des brancardiers (probabilité variant avec le CCMU du patient).

- Un box peut avoir besoin d'être nettoyé (probabilité variant avec le CCMU du patient) avant de pouvoir être réutilisé.

- Avant de quitter le service, le patient doit passer encore une fois par l'accueil pour les taches administratives. Une fois cette opération terminée il quitte le système et ce qu'il devient (hospitalisation, par exemple) ne nous concerne plus.

- Les temps de déplacement sont négligés et le niveau des ressources est constant. On aurait pu améliorer le modèle en faisant varier les effectifs de personnel dans le cadre de l'organisation en « trois-huit » car les arrivées de l'après-midi sont plus espacées que celles du matin et de la nuit, de caractéristiques proches. Le surdimensionnement des ressources de l'après-midi a été maintenu pour résorber les files d'attente importantes se produisant assez souvent en fin de matinée.

3.2. Résultats

Sept modélisations du service d'urgences ont fait l'objet d'une simulation sur 144 000 minutes (soit un fonctionnement du lundi au vendredi, 24 heures sur 24 heures, pendant 20 semaines), avec les mêmes caractéristiques de patients entrants (environ

10000 patients) : même date d'arrivée, même niveau de CCMU et même besoin de brancardage. Dans une file d'attente, la durée d'attente d'une personne est tributaire de l'attente des personnes qui la précède dans la file, ce qui explique l'auto-corrélation d'ordre 1 observée entre les temps de séjour de deux patients se suivant en salle d'attente ou en quittant les urgences. L'existence de cette corrélation implique qu'il n'est pas possible de considérer l'échantillon des patients (obtenu dans une simulation sur période longue) comme un échantillon aléatoire de réalisations des variables indépendantes. Le jugement par intervalle de confiance du temps de séjour d'un patient n'est donc pas possible, de même que celle du temps moyen de séjour. Ce dernier intervalle de confiance peut être cependant obtenu à partir de plusieurs simulations (Harrell *et al.,* 2000 ; Law, Kelton, 1982, 1984) d'un même processus n'utilisant pas la même séquence de nombres aléatoires suivant la loi uniforme [0,1]. Chaque simulation conduit à une réalisation indépendante de la variable aléatoire (VA) « temps de séjour moyen d'un patient dans le service d'urgences » dans le même système productif. Ces différentes simulations s'appuient exactement sur les mêmes arrivées dans le système (dates et attributs) mais pas sur les mêmes réalisations de processus (temps opératoires,...). La somme de plusieurs VA « temps moyen de séjour » d'une simulation étant une somme de VA indépendantes et de mêmes caractéristiques, on peut faire appel au « théorème de la limite centrale » si cette sommation porte sur au moins une trentaine de réalisations. Dans ce cas, on est en mesure d'estimer un intervalle de confiance par la durée moyenne de séjour d'un patient dans le système étudié. Bien évidemment, il n'est toujours pas possible de déterminer un intervalle de confiance de la durée de séjour d'un patient dans ce système. Le tableau 3 présente les 30 simulations que nous avons réalisées pour chacun des modèles.

Lieu	**Variable mesurée**	Patient suivi obligatoirement par le même médecin						
		Non				Oui		
		Processus initiaux		Processus améliorés[1]		Proc. init.	Processus améliorés[1]	
		modèle 1[2]	modèle 1'[3]	modèle 2[2]	modèle 2'[3]	modèle 3[2]	modèle 3'[3]	modèle 3''[4]
Attente accueil	Taille moyenne de la queue	0,36						
	Temps moyen d'attente	5,05						
Salle d'attente	Taille moyenne de la queue	4,51	6,14	3,52	4,19	8,31	10,84	9,09
	Temps moyen d'attente	65,62	88,39	49,57	60,45	103,80	121,33	106,56
Temps de séjour (T)	Temps moyen de séjour	244,26	269,47	224,33	236,89	276,86	293,86	281,08
	IC[5] (95%), borne inférieure	241,99	264,82	222,61	235,03	273,18	288,07	276,01
	IC (95%), borne supérieure	246,53	274,13	226,06	238,74	280,54	299,65	286,14
	P (T > 480)	7,23%	10,05%	5,20%	6,59%	10,82%	15,19%	10,86%

Tableau 3. *Résultats des 7 modèles de simulation*

1. Amélioration par la diminution de 10' du processus d'acheminement des prélèvements sanguins et de leur analyse.
2. 8 médecins, 8 infirmières, 5 laborantins, 3 brancardiers et 2 ASH.
3. 7 médecins, 6 infirmières, 4 laborantins, 1 brancardier et 1 ASH.
4. 8 médecins, 7 infirmières, 4 laborantins, 1 brancardier et 1 ASH.
5. IC : Intervalle de Confiance

Le modèle 1 est qualifié de modèle de référence. Les autres modèles diffèrent du premier par le calibrage de ressources en personnel, le temps opératoire d'un processeur et la gestion du suivi d'un patient par les médecins. L'indicateur le plus important est le « temps de séjour ». Les temps d'attente en « salle d'accueil » et en « salle d'attente » sont retenus parce que fortement anxiogènes (cf. études de MEAH, Le Spégagne & Cauterman, 2005) ; ils ne varient pas d'un modèle à l'autre car ces stocks sont des points de passage obligés dans tous les cas et sont donc systématiquement sur le chemin critique.

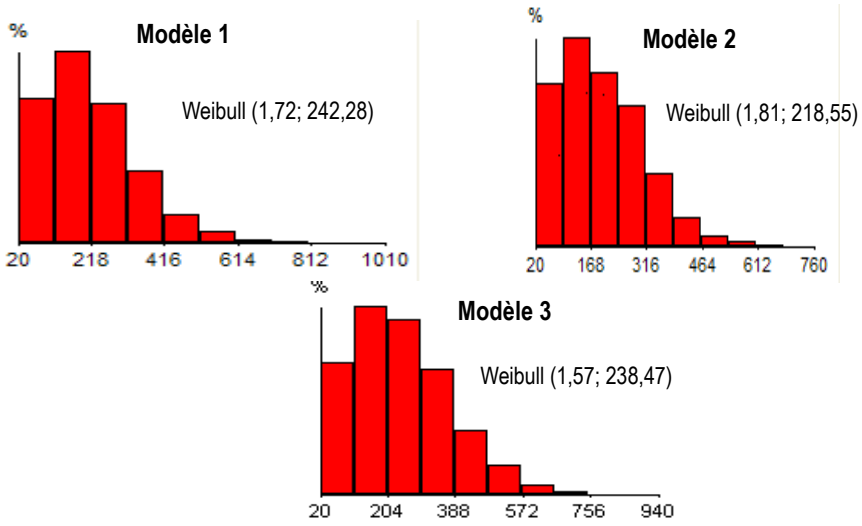

Figure 4. *Temps de séjour des patients dans le modèle de la simulation.*

3.3. Axes d'amélioration des processus du service d'urgences

On s'intéressera à l'amélioration de l'efficacité des processus du service d'urgence, avant de s'intéresser à l'amélioration de son efficience, la durée de séjour d'un patient dans un service d'urgence étant considéré comme un indicateur majeur d'efficacité. On commencera par quelques considérations générales sur la recherche des actions ayant le plus de chances (si elles sont possibles) d'avoir un effet sur la réduction du temps de séjour dans le système productif. Pour ce faire, on évitera la solution onéreuse d'une augmentation du nombre de processeurs, et on privilégiera des actions sur le niveau de ressources partagées par plusieurs processeurs et sur les gammes (temps opératoires et routes empruntables).

3.3.1. Règles d'amélioration de l'efficacité du processus, le temps de séjour

Le temps passé par un item dans un système productif est une somme de temps de traitement (dans les processeurs), de temps d'attente (dans les stocks) et de temps de déplacement. On négligera ici les temps de déplacement parce que la seule modification possible de ce temps total de déplacement résulte d'une variation de la vitesse de déplacement (beaucoup de logiciels de simulation permettent de neutraliser les déplacements). D'une manière générale, un item attend dans un stock pour trois catégories de raisons : le processeur dans lequel il doit se rendre est occupé et n'a pas fini le traitement en cours ; le processeur est occupé et a fini de traiter l'item mais se trouve dans l'impossibilité de l'envoyer dans un stock-aval (stock saturé) ; enfin, certaines conditions requises avant le début du travail ne sont pas remplies, en particulier, la disponibilité des ressources mobilisées par le processeur pour l'exécution de l'opération ou celle d'autres items qu'il faut traiter simultanément par cette opération. Lorsqu'elles surviennent, ces deux dernières catégories de raisons rendent le processus complexe et son comportement plus difficilement prévisible. On commencera par s'intéresser aux processus non complexes.

Dans ce cadre, il convient de souligner la relation de dépendance qui existe entre le temps d'attente en stock et le temps aléatoire de traitement du processeur qui prélève dans ce stock : une modification de la loi du temps de traitement entraînera mécaniquement une modification de la distribution des temps d'attente dans le stock. La théorie des files d'attente donne quelques résultats analytiques pour des problèmes simples. Par exemple, dans un système caractérisé par un intervalle de temps séparant deux arrivées successives dans le stock suivant une loi exponentielle de paramètre λ et par une durée de traitement dans le processeur suivant une loi exponentielle de paramètre μ, on sait (loi de Little) que l'espérance mathématique de la durée d'attente dans le stock est égale à $\mu^2/(\lambda - \mu)$; la diminution de la durée moyenne μ du traitement s'accompagne nécessairement d'une diminution du temps d'attente, le plus souvent[7] proportionnellement plus forte (« effet amplificateur »). On peut ajouter que si l'on remplace le processeur unique par k processeurs fonctionnant en parallèle, dotés des mêmes caractéristiques et prélevant dans le même stock, les relations analytiques sont plus complexes et font intervenir le nombre de processeurs[8]. Dans ce cas, l'effet amplificateur d'une baisse de temps opératoire des processeurs en parallèle sera d'autant plus important que le nombre de processeurs en parallèle est élevé.

Dans le cas général, de telles relations analytiques n'existent pas mais trois relations de cause à effet peuvent être observées :

---

7. Dans les conditions de la loi de Little, avec $\rho = \mu / \lambda < 1$, la dérivée du temps moyen d'attente par rapport à μ, $\rho(2 - \rho)/(1 - \rho)^2$, doit être supérieure à 1 pour que la variation relative du temps passé en file d'attente soit supérieure à celle du temps opératoire. Il s'ensuit que ρ doit être dans l'intervalle $(3 \pm \sqrt{5})/2$; en pratique l'effet amplificateur existe pour $\rho > 0{,}292893$.

8. Cette classe de problèmes est identifiée dans la littérature des files d'attente comme étant du type M/M/k, celui étudié précédemment étant du type M/M/1.

- La baisse du temps moyen passé dans un processeur s'accompagne nécessairement d'une baisse du temps d'attente moyen du stock qui l'approvisionne (s'il n'est pas négligeable). En effet, dans tous les cas, l'anticipation de la sortie d'un item qui résulte de la libération plus rapide du processeur, diminue mécaniquement d'autant le temps d'attente de chacun des items présents dans le stock.

- Pour des raisons similaires, dans le cas de systèmes complexes tels que définis ci-dessus, la diminution de l'attente par un processeur soit d'une ressource partagée, soit d'autres items requis par l'opération, s'accompagne d'une baisse du temps moyen d'attente dans le stock-amont. Il en est de même pour la diminution du temps d'attente, par le processeur, de la libération d'une place dans un stock-aval saturé.

- Enfin, pour un même taux moyen des arrivées en stock, l'attente moyenne dans le stock varie dans le même sens que l'écart-type de l'intervalle séparant deux arrivées successives dans le stock. En effet, une plus grande régularité des arrivées en stock permet une meilleure utilisation du processeur dont la probabilité d'attente de travail baisse. Cette relation de cause à effet est difficile à utiliser, la distribution de l'intervalle de temps entre deux sorties successives d'un processeur étant la conséquence de tout ce qui se passe en amont.

Lorsque tous les items suivent le même processus[9], la diminution du temps moyen que passe un item dans le système productif est obtenue en utilisant les implications de ce qui vient d'être évoqué. Si on laisse momentanément de côté le cas de systèmes complexes, la question se pose de savoir sur quel poste il est a priori le plus intéressant de diminuer le temps opératoire. Il semble judicieux de s'intéresser d'abord au processeur approvisionné par le stock ayant la durée d'attente moyenne la plus élevée (*règle 1*), en raison de l'effet d'amplification évoqué ci-dessus; dans cette configuration de processeurs en série, il s'agira du processeur ayant le temps opératoire moyen le plus élevé. Ce couple « processeur, stock-amont » apporte la plus forte contribution à la durée moyenne de séjour d'un item dans le système productif.

Dans les processus complexes, les processeurs ne sont pas en série et les items ne suivent pas tous le même chemin. Il convient donc d'identifier préalablement chacun de ces chemins et de déterminer, si possible, la probabilité de chacun d'entre eux. Trois cas de figure peuvent être identifiés.

- Un premier cas de figure se caractérise par le fait que ces chemins ne partagent aucun stock et aucun processeur. Dans ce cas, en l'absence d'interaction (imputable par exemple au partage de certaines ressources), la durée moyenne de séjour des items dans le système est la somme des espérances mathématiques des durées de chacun de ces chemins, pondérée par la probabilité d'emprunt de ces chemins. L'adaptation du raisonnement conduit au paragraphe précédent est immédiate, on comparera les temps d'attente moyens pondérés par la probabilité de passage par les stocks (*règle 1 bis*).

- Ces chemins peuvent comporter des parties communes, ce qui rend seulement plus compliqué le calcul de la probabilité du passage d'un item dans un stock utilisé

9. Processeurs en série, ce qui n'exclut pas la possibilité d'avoir des processeurs identiques travaillant en parallèle pour exécuter la même opération.

par un sous-ensemble d'items. Si ce calcul est compliqué, notamment pour des raisons combinatoires, on peut s'appuyer sur le taux d'utilisation des processeurs utilisant ces stocks, à l'issue d'une simulation suffisante du régime de croisière, cette information étant fournie par tous les simulateurs (*règle 1 ter*).

- Plusieurs processeurs en parallèle peuvent prélever les items dans un même et unique stock. Ces items emprunteront des chemins différents. On a vu que la relation liant la distribution du temps d'attente dans le stock, aux distributions des arrivées et du temps de traitement, existe dans des cas très simples. En tout état de cause, la diminution du temps d'opératoire de chacun des processeurs identiques permet d'augmenter le nombre d'items traité par chacun d'eux au cours d'une journée, ce qui provoque une diminution plus forte du temps d'attente du stock en amont. Il s'ensuit qu'entre deux stocks ayant la même durée moyenne d'attente, il semble judicieux de s'intéresser à celui qui alimente le plus de processeurs et donc de chercher à diminuer son temps opératoire ; l'effet amplificateur sera d'autant plus important que le nombre de processeurs en parallèle est grand (*règle 1 quatro*).

- Le troisième cas de figure reprend le précédent en intégrant des relations d'interdépendance entre les processeurs pour trois types de raisons : partage de ressources ; opérations combinant des items venant de sous-processus différents ; limitation du nombre d'items simultanément présents dans certains stocks ou dans certains groupes de stocks et de processeurs. Dans tous ces cas, la prédictibilité du comportement du système productif devient difficile et l'incidence de l'application de la règle proposée dans les cas 1 et 2 n'est plus aussi simple. Il est alors judicieux de partir des effets considérés comme indésirables – ici partir des temps d'attente les plus importants – pour agir sur les processeurs puisant dans ces stocks en cherchant à tirer le meilleur parti de l'effet amplificateur mis en évidence ci-dessus.

  - En cas de ressources en personnel partagées, il est intéressant de vérifier si une augmentation de la ressource permet une diminution significative de la durée moyenne d'attente. Si l'impact est négligeable ou si l'augmentation de la ressource est difficilement envisageable[10], on peut alors examiner s'il est intéressant de modifier les règles d'affectation des ressources appelées par plusieurs processeurs (*règle 2*), ce qui implique que l'on puisse avoir plusieurs processeurs simultanément en attente de ressources (s'il s'agit de processeurs parallèles interchangeables, cette solution est sans intérêt). Si l'application des *règles 1 bis* ou *1 ter* conduit à s'intéresser à un processeur à ressource partagée, une diminution de son temps opératoire a non seulement l'effet induit sur le temps d'attente du stock-amont, évoqué ci-dessus, mais aussi un second effet induit. En permettant une libération plus rapide de la ressource partagée, celle-ci est alors en mesure d'exécuter plus d'opérations sur l'ensemble des processeurs qui la mobilise. Cette propagation conduit à un effet de levier plus important de la diminution du temps opératoire du processeur sur le temps moyen de

10. Le gain d'efficacité, essentiellement lié à la diminution du temps passé dans le stock-amont, s'accompagne d'une perte d'efficience, le coût de fonctionnement du processus productif étant plus grand.

séjour. Il s'ensuit qu'en cas de ressource partagée, l'application des *règles 1 bis* ou *1 ter* n'est pas forcément la plus judicieuse ; elle ne l'est que si une simulation préalable levant la contrainte de disponibilité de la ressource partagée a un impact négligeable.

- En cas de limitation de capacité d'un stock, il faut d'abord s'assurer par simulation que la levée de cette contrainte permet de réduire sensiblement le temps moyen de séjour dans le système productif, sinon, on peut la négliger. Si cette contrainte joue un rôle, il faut chercher à diminuer le temps opératoire du ou des processeurs qui s'approvisionnent dans ce stock (*règle 3*). On peut ajouter que le lissage des arrivées dans ce stock réduit sa probabilité d'être saturé mais qu'il est difficile d'agir directement sur ce point sauf si le processus étudié implique des traitements par lot en amont de ce stock, auquel cas une diminution de la taille des lots peut avoir un impact significatif (*règle 4*).
- Le dernier cas est celui de processeurs réalisant des opérations d'assemblage. Derrière le problème apparent de synchronisation insuffisante, peut se trouver un problème de chemin critique, en raison d'une charge de travail différente entre les sous-processus convergeant vers le processeur considéré. Il convient alors d'identifier les chemins critiques (*règle 5*) et de ne pas appliquer les *règles 1 bis* ou *1 ter* aux processeurs des chemins considérés comme non critiques.

Ces quelques règles, fruits d'une expérience nécessairement limitée, ne prétendent pas définir une stratégie générale mais seulement de guider une réflexion dans la recherche des bons leviers d'action.

3.3.2. Amélioration de l'efficacité du processus, le temps de séjour

L'efficacité du processus d'un service d'urgence, comporte deux volets : la qualité médicale des soins, sur laquelle la simulation ne peut guère apporter d'information (sauf si l'on est en droit de penser qu'un niveau excessif d'utilisation de certaines ressources médicales a un impact sur la qualité des prestations effectuées) et le temps de séjour aux urgences, que l'on peut supposer indépendante de la qualité des prestations fournies. Ce dernier indicateur est cependant peu pertinent pour les patients arrivant dans un état grave.

Le séjour des patients dans le système peut se décomposer en une phase d'admission, une phase de traitement en box et une phase de sortie. La première est la dernière phase sont des cas simples de processus et des actions d'améliorations du temps de séjour sont facilement envisageables via les règles 1. Néanmoins, la réduction des durées des phases d'admission et de sortie, suivies par tous les patients, est difficilement envisageable, sauf à multiplier les postes de travail, piste que l'on n'explorera pas, ces durées étant faibles. C'est donc sur la phase intermédiaire qu'il convient d'agir ; le processus correspondant peut être qualifié de complexe :

- il existe des relations d'interdépendance entre les processeurs en raison d'un partage de ressources (les ressources « médecins », « IDE », « spécialiste », « brancardiers », « ASH » interviennent dans le fonctionnement de plusieurs postes de travail),

- certaines opérations combinent des items venant de sous-processus différents (patient + résultat d'analyse de laboratoire avant de passer au diagnostic suivant),

- enfin, il existe une limitation du nombre d'items simultanément présents dans un groupe de stocks et de processeurs ; la modélisation réaliste d'un box n'est possible qu'en faisant appel à un groupe de stocks et de processeurs, ce groupe ne pouvant accueillir qu'un seul patient.

Nous nous trouvons donc dans le troisième cas de figure identifié au paragraphe précédent et des améliorations peuvent être apportées en appliquant les règles 2, 3, 4 ou 5. On a indiqué qu'il est souhaitable d'agir en priorité sur ce qui est à l'origine du temps d'attente élevé dans certains stocks.

- La règle 2 (augmentation du nombre de ressources partagées : infirmières, médecins et brancardiers) s'avère inintéressante, la simulation montrant que l'accroissement du niveau de ces ressources n'induit pas de baisse significative du temps de séjour susceptible de justifier la perte d'efficience induite par ces augmentations. La règle 4 (impact du lotissement) est sans objet.

- L'application de la règle 3 conduit à se focaliser sur la salle d'attente où le temps de séjour est le plus élevé. Les box sont des processeurs parallèles qui prélèvent des patients dans ce stock. L'augmentation du nombre de box n'est pas possible et, quand bien même elle le serait, il n'est pas évident que cette solution serait économiquement la meilleure. L'amélioration ne peut donc venir que d'une diminution du temps passé dans un box.

Le temps passé dans un box se décompose en une somme de temps opératoires et de temps d'attente. Certaines de ces opérations sont effectuées dans le box tandis que d'autres (analyses d'imagerie et de prélèvements) sont effectués en dehors du box, dans des centres mutualisés, sachant que le box du patient parti en imagerie ne peut être réaffecté en son absence. Le laboratoire analyse l'échantillon prélevé sur le patient par une infirmière dans le box. Le patient est ensuite libre de se déplacer pour les autres analyses (s'il y en a eu de prescrites) mais le diagnostic suivant ne peut se faire que si toutes les prescriptions ont été réalisées, ce qui relève d'une logique d'assemblage. L'application de la règle 5 s'impose donc ; le temps coulé entre deux diagnostics est déterminé par la durée la plus longue des durées d'analyses réalisées en parallèle (chemin critique).

Le temps opératoire du processeur « box » est égal à la somme des durées d'exécution des opérations en amont du premier diagnostic et du temps passé entre le premier et le dernier diagnostic, après quoi le patient quitte le box. Sachant que les opérations en amont du diagnostic permettent difficilement des diminutions dans leur temps de traitement (à savoir l'installation du patient et le diagnostic), nous nous orientons vers l'analyse du temps passé entre les diagnostics. La figure 5 décrit les parcours d'un patient entre les deux premiers et entre les deux derniers diagnostics, en indiquant l'espérance mathématique des temps opératoires et celle des pourcentages de patients empruntant un chemin (sur ceux ayant atteint le nœud précédent). Dans cette représentation, en aval de l'opérateur ET (⋋), les deux chemins doivent être empruntés à la fois. Il s'ensuit que les opérations de ces chemins peuvent être réalisées en pa-

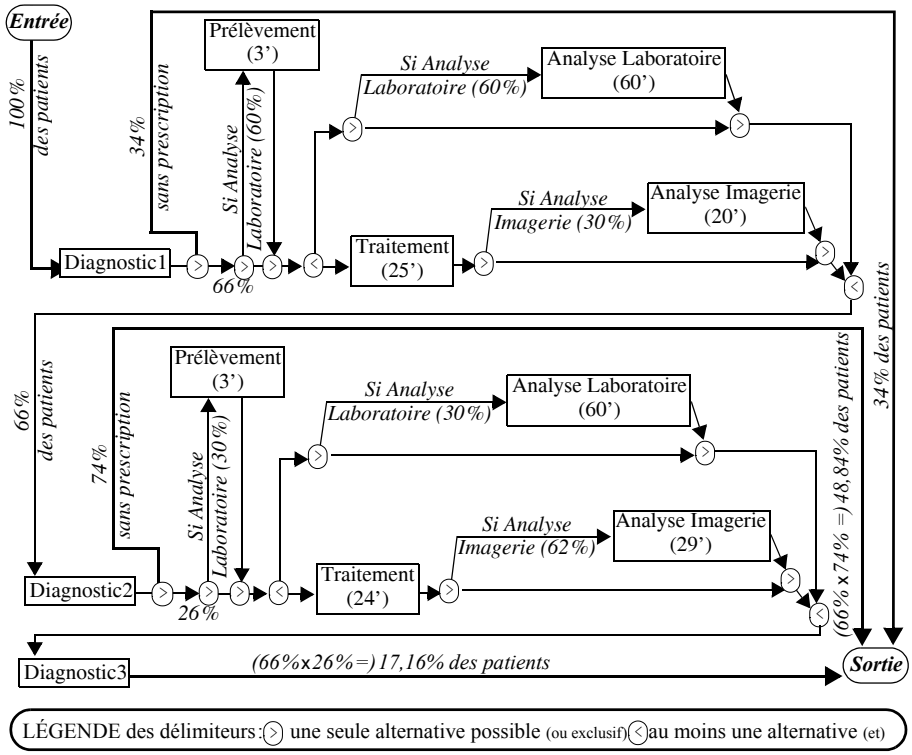

Figure 5. *Réalisation des prescriptions après les diagnostics 1 et 2 (avec durées moyennes et répartitions moyennes — les opérations « Traitement » et « Analyse Imagerie » sont permutables)*

rallèle (d'une part, l'*Analyse Laboratoire* et, d'autre part, le *Traitement* et/ou l'*Analyse Imagerie*). Entre deux diagnostics successifs, 4 groupes exclusifs de prescriptions sont possibles[11]. À partir de ces informations, on peut calculer la probabilité de ces groupes, en application des théorèmes classiques de probabilité (tableau 4[12]).

Un patient qui passe le diagnostic 1 et se voit prescrire au moins un traitement (pour mémoire 66% des patients) peut avoir 20 parcours : après l'un des 4 groupes de prescriptions du diagnostic 1, il passe le diagnostic 2 après quoi, soit il sort du box (4 premiers parcours possibles[13]), soit il continue avec l'un des 4 groupes de prescriptions du diagnostic 2 (ce qui donne 4x4 = 16 autres parcours possibles dont les proba-

11. Avec T pour Traitement, AI pour Analyse Imagerie, et AL pour Analyse Laboratoire, les possibilités sont : T&AI & L ; T&AI & $\overline{L}$; T&$\overline{AI}$ & L ; T&$\overline{AI}$ & $\overline{L}$.

12. Par exemple, à partir des informations de probabilité données dans la figure 5, la probabilité de l'événement T_1&$\overline{AI}_1$ & $\overline{L}_1$, se produisant juste après le diagnostic 1 est : (1 – 0,3).(1 – 0,6) = 0,28 et celle de l'événement T_2&$\overline{AI}_2$ & $\overline{L}_2$, se produisant juste après le diagnostic 2 est : (1 – 0,3).(1 – 0,62) = 0,266.

bilités se calculent facilement[14]). Il est facile ensuite de déterminer la probabilité qu'ont ces prescriptions[15] de se trouver sur le chemin critique qui implique, selon le patient, une, deux ou trois de ces prescriptions. On a présenté le résultat de ces calculs dans la figure 6, sous une forme ensembliste; on constate que c'est la prescription d'analyse de laboratoire qui a la plus forte probabilité d'intervenir sur un chemin critique (63,1%). Sachant que 34% des patients sont orientés vers la sortie ou vers la salle de déchoquage après le premier diagnostic et donc que 66% auront au moins une prescription, nous pouvons dire qu'en moyenne l'analyse de laboratoire a une probabilité de 66%x63,1% = 41,64% d'être sur un chemin critique. C'est donc d'abord sur ce processus que des efforts d'amélioration doivent être envisagés. Dans la suite, on a postulé qu'il était possible de réduire de 10 minutes l'espérance mathématique de la distribution de la durée d'analyse de laboratoire (**modèle 2**).

Prescriptions à effectuer	**Probabilités**	
	Entre diagnostics 1 et 2	***Entre diagnostics 2 et 3***
Traitement (T)	28,0%	26,6%
Analyse de laboratoire (AL) + T	42,0%	11,4%
Analyse d'imagerie (AI) + T	12,0%	43,4%
AL + AI + T	18,0%	18,6%

Tableau 4. *Probabilités de prescriptions entre deux diagnostics successifs*

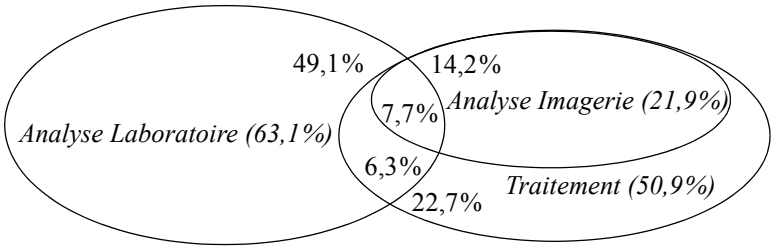

Figure 6. *Répartition des chemins critiques entre le premier et le*

13. Par exemple, la probabilité du parcours « $T_1\&\overline{AI}_1$ & $\overline{L}_1$ » + « Diagnostic 2 » + « Sortie » est: 0,28x1,0x0,74 = 0,2072.

14. Par exemple, la probabilité du parcours « $T_1\&\overline{AI}_1$ & $\overline{L}_1$ » + « Diagnostic 2 » + « $T_2\&\overline{AI}_2$ & $\overline{L}_2$ » + « Diagnostic 3 » + « Sortie » est: 0,28x1,0x0,26x0,266x1 = 0,0194

15. Exemple le calcul de probabilité de l'événement « Traitement seul », qui correspond à l'un des deux événements exclusifs analysés dans les deux notes précédentes: 0,2072 + 0,0194 = 0,2266 ($\Rightarrow$ 22,7% dans la figure 6).

Le raisonnement du chemin critique conduit à supposer que la baisse du temps opératoire du processeur « box » doit être imputable à la diminution du temps d'« attente tous traitements ». La simulation du modèle 1 et du modèle 2 fournit une différence de 6,83 dans le temps opératoire du box contre une diminution de 6,99 du temps d'« attente tous traitements »[16]. On peut donc écrire :

$$\Delta(\text{Temps passé dans le box}) = k_1 \times \Delta(\text{Durée d'analyse labo}) = 0{,}683 \times 10' = 6{,}83$$

La contribution de cette diminution au temps de séjour est supérieure à celle-ci. La relation suivante met en évidence cet effet amplificateur avec le coefficient k_2.

$$\Delta(\text{Temps de séjour}) = k_2 \times \Delta(\text{Durée d'analyse labo}) = k_2 \times 10'$$

Cette baisse du temps de séjour, la somme de la baisse du « temps passé dans le box » et de celle du temps d'attente de son stock en amont (salle d'attente) est impossible à déterminer analytiquement, parce que le temps passé en salle d'attente dépend du nombre de box. On sait que la diminution simultanée du temps opératoire des plusieurs processeurs identiques (6,83 minutes) provoquera un effet amplificateur encore plus grand (*règle 1 quatro*), ce gain peut être calculé en cas des lois connues mais pas possible pour les processeurs box. Ce coefficient k_2 ne peut être calculé que par la simulation laquelle évalue la différence des temps de séjour entre le modèle 1 et le modèle 2, à 19,93 minutes :

$$\Delta(\text{Temps de séjour}) = 19{,}93' = 1{,}993 \times 10' = 1{,}993 \times \Delta(\text{Durée d'analyse labo})$$

On trouve donc k_2=1,993, ce qui justifie notre hypothèse d'effet amplificateur avec un gain plus important que celui de k_1. La différence est essentiellement imputable à une libération de 6,83 minutes plus rapide des box, conduisant à une baisse de 16,05 minutes de l'attente de la libération d'un box[17]. Cette diminution dépend du nombre du box et si le nombre de box est assez grand pour ne pas faire attendre de patient dans la salle d'attente, on aura :

$$\Delta(\text{Temps de séjour}) = \Delta(\text{Temps passé dans le box})$$

Or, $\Delta(\text{Temps de séjour}) = (k_2 \backslash k_1) \times \Delta(\text{Temps passé dans le box})$

Donc, $k_2 = k_1$

On en déduit que k_2 est une fonction monotone et décroissante du nombre de box et tend vers k_1. Cette fonction ne peut pas être obtenue analytiquement mais peut être tracée à l'aide de la simulation (figure 7), ce qui illustre tout l'intérêt d'un modèle de simulation dans la recherche d'amélioration de l'efficacité d'un processus.

---

16. La diminution du temps opératoire « box » peut être calculée à partir des informations de la figure 5 et du tableau 4. Ces calculs donnent une baisse de 6,49 minutes, un chiffre assez proche de ce qui est observé dans la simulation (l'écart étant lié au fait que l'indisponibilité de certaines ressources peut rendre impossible le parallélisme postulé).

17. On peut noter que la somme de ces deux valeurs est très proche à la diminution du temps de séjour. Cette différence (0,95% du temps de séjour) est d'autant plus petite si les analyses se font par des groupes de patients homogènes (jusqu'à 0,02%).

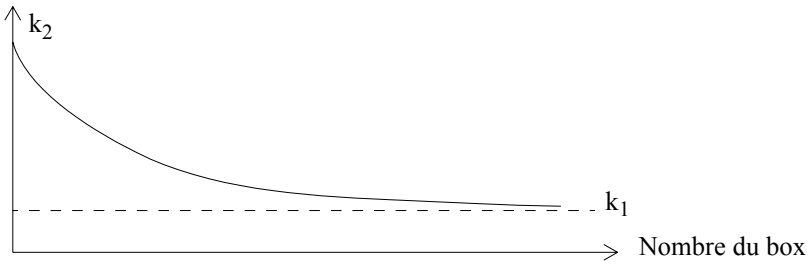

Figure 7. *« k_2 » en fonction du nombre de box*

3.3.3. Amélioration de l'efficacité du processus, la qualité perçue

Pour augmenter la qualité du service, les diminutions des temps de séjour et des temps d'attente sont importantes mais pas suffisantes à elles seules. Certaines dimensions de la qualité, que l'on appelle la « qualité perçue », ne sont pas objectives et très difficiles à évaluer mais doivent être prises en compte dans les recherches d'amélioration de la qualité dans la production de services. À partir du modèle 2, on a créé le modèle 3, dans lequel le même médecin est affecté au même patient durant tout son séjour, ce qui implique l'usage de nouvelles règles d'afffectation des médecins. Deux conséquences en résultent. La continuité du contact médical sécurise le patient et a des chances d'améliorer la qualité perçue. Par contre ce médecin peut ne pas être libre immédiatement quand on en a besoin, ce qui augmente le « temps passé dans le box » et donc le « temps de séjour ». L'intérêt de cette nouvelle règle dépend du compromis retenu (par chaque patient ou collectivement) entre deux dimensions de l'efficacité et donc entre l'amélioration d'un attribut subjectif (le fait d'être soigné par le même médecin) et la dégradation d'un attribut objectif[18] (un temps de séjour). La simulation du service d'urgence avec ces nouvelles règles (modèle 3) conduit à une augmentation moyenne de 52,53 minutes du temps de séjour (par rapport au modèle 2). Le compromis (l'évaluation) entre la « satisfaction du patient d'être soigné par le même médecin » et son « insatisfaction d'un allongement du temps d'attente » ne peut relever des méthodologies mobilisées ici. On peut ajouter qu'ici on a contourné une limitation classique des simulations basées sur les objets, en exploitant certaines possibilités de partage d'informations par les processeurs, combinées à des options de programmation.

3.3.4. Amélioration de l'efficience du processus

Comme indiqué au § 3.3.1, pendant la recherche d'amélioration d'efficacité du processus, il convient parfois d'augmenter le nombre de ressources. Toute analyse d'efficacité doit donc être couplée à une analyse d'efficience. Nous avons donc cher-

18. La perception du temps qui passe est en grande partie subjective, ce qui complique singulièrement la définition d'une qualité perçue.

ché à réduire les niveaux des ressources mobilisées sans compromettre les indicateurs d'efficacité. La situation de départ est la même pour les modèles 1 à 3 (voir tableau 3). Les solutions efficientes sont celles des modèles 1', 2' et 3'' (le modèle 3', utilise le niveau de ressource du modèle 2' mais avec un fort allongement de la durée de séjour). S'il est possible d'améliorer simultanément l'efficacité et l'efficience (le modèle 2' domine le modèle 1[19]) et il est, par contre, impossible de porter un jugement sur le suivi du patient par le même médecin. Cela étant, la simulation fournit à la réflexion quelques éléments utiles : *s'il faut troquer ce suivi contre un allongement de la durée de séjour de moins de 44,19 minutes (comparaison des modèles 2' et 3'')*, le fonctionnement du service d'urgences avec un suivi du patient par le même médecin est dominé tant du point de vue de l'efficacité que de l'efficience par un fonctionnement dans lequel les médecins sont complètement interchangeables.

4. Conclusion

La comparaison des approches de simulation faite dans la première partie (§ 2) montre les avantages d'une approche orientée objet vis-à-vis d'un système de production de services. Cela dit, le but d'une simulation est d'apporter des éléments de réponse à des questions portant sur l'efficacité et l'efficience d'un processus. Le choix de l'approche utilisée est conduit par les données récupérables du terrain, les hypothèses de fonctionnement, les objectifs de l'utilisateur et les indicateurs choisis. Dans ce contexte, le niveau de détail retenu dans le modèle de simulation joue un rôle important dans le choix de l'approche. Dans la deuxième partie (§ 3), nous avons montré par un exemple que l'utilisation d'une approche basée sur objets peut permettre un niveau de détail assez fin de modélisation d'un système de production de services. Nous avons étudié les relations causales indirectes peu intuitives d'un service d'urgences afin d'améliorer l'efficacité (par la diminution du temps de séjour) (§ 3.3.2) et l'efficience de processus (par la diminution du niveau de ressources-personnel) (§ 3.3.4). Nous avons élargi notre analyse d'efficacité avec celle de la qualité perçue (§ 3.3.3); ce qui a permis de mettre en évidence qu'une bonne maîtrise de l'outil de la simulation peut permettre l'inclusion de certaines spécificités d'interaction interpersonnelle d'un système de production de services.

5. Bibliographie

Altinel K. I., Ulas E., « Simulation modeling for emergency bed requirement planning », *Annals of Operations Research,* vol. 67, 1999, p. 183-210.

Badri M.A, Hollingsworth J., « A Simulation Model for Scheduling in the Emergency Room », *International Journal of Operations&Production Management*, vol. 13, n° 3, 1993, p. 13-24.

19. La borne supérieure de l'intervalle de confiance du moyen de «temps de séjour» du modèle 1 étant plus petite que la borne inférieure du celui du modèle 2', cette domination est évidente.

Blaha M., Rumbaugh J., *Modélisation et Conception Orientées Objet avec UML 2*, Paris, Pearson Education, 2005.

Blake J. T, Carter M. W., « An analysis of emergency room wait time issues via computer simulation », *INFOR*, Canadian Operational Research Society, vol. 34, n° 4, 1996, p. 263-273.

Davidson P., « Multi agent simulation : Beyond social simulation », dans *MABS 2000*, Moss S. et Davidson P. Eds., 8-9 juin 2000, p. 97-107.

Ferber J., *Les systèmes multi-agents. Vers une intelligence collective*, Paris, Inter Editions, 1995.

Ferber J., « La modélisation multi-agents : un outil d'aide à l'analyse de phénomènes complexes », dans *Journées du Programme Environnement*, 1997, p. 113-133.

Harrell C., Ghosh B.K., Bowden R., *Simulation using Promodel*, McGraw-Hill, NY, 2000.

Hill R. D. C., *Analyse Orientée Objets & Modélisation par Simulation*, France, Addison-Wesley, 1993.

Joines J., Roberts S., « Simulation in an object-oriented world », dans *Proceedings of the 31st conference on Winter simulation : Simulation-a bridge to the future*, 1999, New York, ACM Press, p. 132-140.

Jun J. B., Jacobson S. H., Swisher J. R., « Application of Discrete-Event Simulation in Health Care Clinics: A Survey », *The Journal of the Operational Research Society*, vol. 50, n° 2, 1999, p. 109-123.

Kay A., « The early history of Smalltalk », *ACM SIGPLAN Notices*, vol. 28, n° 3, 1993, p. 69-95.

Law A.W., Kelton D., *Simulation Modeling and Analysis*, 1st ed., McGraw-Hill, NY, 1982.

Law A.W., Kelton D., « Confidence intervals for steady-state simulations: I. A survey of fixed sample size procedures », *Operations Research*, vol. 32, n° 6, 1984, p. 1221-1239.

Le Spégagne D., Cauterman M. (2005), « *Rapport de fin de mission. Temps d'attente et de passage aux urgences* », Mission Nationale d'Expertise et d'Audit Hospitaliers (MEAH).

Maes P., « Intelligent Software », *Scientific American*, vol. 273, n°3, 1995, p. 84-86.

Savall M., « Une architecture d'agents pour la simulation : Le modèle YAMAM et sa plate-forme Phoenix », Thèse en Informatique, INSA Rouen, 2003.

Millischer L., Modélisation individu centrée des comportements de recherche des navires de pêche, Thèse en Halieutique, ENS Agronomique de Rennes, 2000.

Odell J., Parunak H. V. D., Fleischer M., Sven B., « Modeling Agents and their Environnement », dans *Agent-Oriented Software Enginering*, Giunchiglia F., Odell J. et Weiss G. Eds., Springer, Berlin, 2002, p. 16-31.

Rytilä J. S., Spens K. M., « Using simulation to increase efficiency in blood supply chains », *Management Research News*, vol. 29, n° 12, 2006, p. 801-819.

Shim S. J., Kumar A., « Using computer simulation for surgical care process reengineering in hospitals », *INFOR*, vol. 43, n° 4, 2005, p. 303-319.

Stroustrup B., *The C++ programming language*, Addison-Wesley Longman Publishing Co. Inc., Boston, MA, USA, 1986.

Swisher J. R., Jacobson S. H., « Evaluating the Design of a Family Practice Healthcare Clinic Using Discrete-Event Simulation », *Health Care Management Science*, vol. 5, 2002, p. 75-85.

Swisher J. R., Jacobson S. H., Jun B. J., Balci O., « Modeling and analyzing a physician clinic environment using discrete-event (visual) simulation », Computers & Operations Research, vol. 28, 2001, p.105-125.

Weyns D., Parunak H. V. D., Fabien M., Holvoet T., Ferber J., « Environments for Multiagent Systems : State-of-the-Art and Research Challenges », dans *First International Workshop-E4MAS*, 19 Juillet 2004, New York, Springer, p. 1-47.