



**HAL**  
open science

**Méthodes socio-statistiques pour l'aide à la décision en milieu industriel: Application à la gestion des capacités d'un système d'information en industrie micro-électronique**

Michel Lutz

► **To cite this version:**

Michel Lutz. Méthodes socio-statistiques pour l'aide à la décision en milieu industriel: Application à la gestion des capacités d'un système d'information en industrie micro-électronique. Autre. Ecole Nationale Supérieure des Mines de Saint-Etienne, 2013. Français. NNT: 2013EMSE0688. tel-00849870

**HAL Id: tel-00849870**

**<https://theses.hal.science/tel-00849870>**

Submitted on 1 Aug 2013

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

NNT : 2013 EMSE 0688

## THÈSE

présentée par

Michel Lutz

pour obtenir le grade de  
Docteur de l'École Nationale Supérieure des Mines de Saint-Étienne

Spécialité : Génie Industriel

### METHODES SOCIO-STATISTIQUES POUR L'AIDE A LA DECISION EN MILIEU INDUSTRIEL

*Application à la gestion des capacités d'un système d'information en  
industrie micro-électronique*

soutenue à Saint-Etienne, le 14 mai 2013

#### Membres du jury

Rapporteurs :	A. BERNARD	Professeur, Ecole Centrale, Nantes
	A. BOURAS	Professeur, Université Lumière, Lyon 2
Examineurs :	J.H. CHAUCHAT	Professeur Emérite, Université Lumière, Lyon 2
	H. PANETTO	Professeur, Université Poincaré, Nancy 1
Directeur de thèse :	X. BOUCHER	Professeur, Ecole Nationale Supérieure des Mines, Saint-Etienne
Co-Directeurs :	M.A. GIRARD	Maître-Assistant, Ecole Nationale Supérieure des Mines, Saint-Etienne
	O. ROUSTANT	Maître-Assistant, HDR, Ecole Nationale Supérieure des Mines, Saint-Etienne
Invité :	X. AMBROSIONI	Encadrant industriel, STMicroelectronics, Crolles

**Spécialités doctorales :**  
 SCIENCES ET GENIE DES MATERIAUX  
 MECANIQUE ET INGENIERIE  
 GENIE DES PROCÉDES  
 SCIENCES DE LA TERRE  
 SCIENCES ET GENIE DE L'ENVIRONNEMENT  
 MATHEMATIQUES APPLIQUEES  
 INFORMATIQUE  
 IMAGE, VISION, SIGNAL  
 GENIE INDUSTRIEL  
 MICROELECTRONIQUE

**Responsables :**  
 K. Wolski Directeur de recherche  
 S. Drapier, professeur  
 F. Gruy, Maître de recherche  
 B. Guy, Directeur de recherche  
 D. Graillot, Directeur de recherche  
 O. Roustant, Maître-assistant  
 O. Boissier, Professeur  
 J.C. Pinoli, Professeur  
 A. Dolgui, Professeur

**EMSE : Enseignants-chercheurs et chercheurs autorisés à diriger des thèses de doctorat (titulaires d'un doctorat d'État ou d'une HDR)**

AVRIL	Stéphane	PR2	Mécanique et ingénierie	CIS
BATTON-HUBERT	Mireille	PR2	Sciences et génie de l'environnement	FAYOL
BENABEN	Patrick	PR1	Sciences et génie des matériaux	CMP
BERNACHE-ASSOLLANT	Didier	PR0	Génie des Procédés	CIS
BIGOT	Jean Pierre	MR(DR2)	Génie des Procédés	SPIN
BILAL	Essaid	DR	Sciences de la Terre	SPIN
BOISSIER	Olivier	PR1	Informatique	FAYOL
BORBELY	Andras	MR(DR2)	Sciences et génie de l'environnement	SMS
BOUCHER	Xavier	PR2	Génie Industriel	FAYOL
BRODHAG	Christian	DR	Sciences et génie de l'environnement	FAYOL
BURLAT	Patrick	PR2	Génie Industriel	FAYOL
COLLOT	Philippe	PR0	Microélectronique	CMP
COURNIL	Michel	PR0	Génie des Procédés	DIR
DARRIEULAT	Michel	IGM	Sciences et génie des matériaux	SMS
DAUZERE-PERES	Stéphane	PR1	Génie Industriel	CMP
DEBAYLE	Johan	CR	Image Vision Signal	CIS
DELAFOSSÉ	David	PR1	Sciences et génie des matériaux	SMS
DESRAYAUD	Christophe	PR2	Mécanique et ingénierie	SMS
DOLGUI	Alexandre	PR0	Génie Industriel	FAYOL
DRAPIER	Sylvain	PR1	Mécanique et ingénierie	SMS
FEILLET	Dominique	PR2	Génie Industriel	CMP
FOREST	Bernard	PR1	Sciences et génie des matériaux	CIS
FORMISYN	Pascal	PR0	Sciences et génie de l'environnement	DIR
FRACZKIEWICZ	Anna	DR	Sciences et génie des matériaux	SMS
GARCIA	Daniel	MR(DR2)	Génie des Procédés	SPIN
GERINGER	Jean	MA(MDC)	Sciences et génie des matériaux	CIS
GIRARDOT	Jean-jacques	MR(DR2)	Informatique	FAYOL
GOEURIOT	Dominique	DR	Sciences et génie des matériaux	SMS
GRAILLOT	Didier	DR	Sciences et génie de l'environnement	SPIN
GROSSEAU	Philippe	DR	Génie des Procédés	SPIN
GRUY	Frédéric	PR1	Génie des Procédés	SPIN
GUY	Bernard	DR	Sciences de la Terre	SPIN
GUYONNET	René	DR	Génie des Procédés	SPIN
HAN	Woo-Suck	CR	Mécanique et ingénierie	SMS
HERRI	Jean Michel	PR1	Génie des Procédés	SPIN
INAL	Karim	PR2	Microélectronique	CMP
KLOCKER	Helmut	DR	Sciences et génie des matériaux	SMS
LAFOREST	Valérie	MR(DR2)	Sciences et génie de l'environnement	FAYOL
LERICHE	Rodolphe	CR	Mécanique et ingénierie	FAYOL
LI	Jean Michel		Microélectronique	CMP
MALLIARAS	Georges	PR1	Microélectronique	CMP
MOLIMARD	Jérôme	PR2	Mécanique et ingénierie	CIS
MONTHEILLET	Franck	DR	Sciences et génie des matériaux	SMS
PERIER-CAMBY	Laurent	PR2	Génie des Procédés	DFG
PIJOLAT	Christophe	PR0	Génie des Procédés	SPIN
PIJOLAT	Michèle	PR1	Génie des Procédés	SPIN
PINOLI	Jean Charles	PR0	Image Vision Signal	CIS
POURCHEZ	Jérémy	CR	Génie des Procédés	CIS
ROUSTANT	Olivier	MA(MDC)		FAYOL
STOLARZ	Jacques	CR	Sciences et génie des matériaux	SMS
SZAFNICKI	Konrad	MR(DR2)	Sciences et génie de l'environnement	CMP
TRIA	Assia		Microélectronique	CMP
VALDIVIESO	François	MA(MDC)	Sciences et génie des matériaux	SMS
VIRICELLE	Jean Paul	MR(DR2)	Génie des Procédés	SPIN
WOLSKI	Krzystof	DR	Sciences et génie des matériaux	SMS
XIE	Xiaolan	PR1	Informatique	CIS

**ENISE : Enseignants-chercheurs et chercheurs autorisés à diriger des thèses de doctorat (titulaires d'un doctorat d'État ou d'une HDR)**

FORTUNIER	Roland	PR	Sciences et Génie des matériaux	ENISE
BERGHEAU	Jean-Michel	PU	Mécanique et Ingénierie	ENISE
DUBUJET	Philippe	PU	Mécanique et Ingénierie	ENISE
LYONNET	Patrick	PU	Mécanique et Ingénierie	ENISE
SMUROV	Igor	PU	Mécanique et Ingénierie	ENISE
ZAHOUANI	Hassan	PU	Mécanique et Ingénierie	ENISE
BERTRAND	Philippe	MCF	Génie des procédés	ENISE
HAMDI	Hédi	MCF	Mécanique et Ingénierie	ENISE
KERMOUCHE	Guillaume	MCF	Mécanique et Ingénierie	ENISE
RECH	Joël	MCF	Mécanique et Ingénierie	ENISE
TOSCANO	Rosario	MCF	Mécanique et Ingénierie	ENISE
GUSSAROV	Andrey	Enseignant contractuel	Génie des procédés	ENISE

PR 0	Professeur classe exceptionnelle	Ing.	Ingénieur
PR 1	Professeur 1 <sup>ère</sup> classe	MCF	Maître de conférences
PR 2	Professeur 2 <sup>ème</sup> classe	MR (DR2)	Maître de recherche
PU	Professeur des Universités	CR	Chargé de recherche
MA (MDC)	Maître assistant	EC	Enseignant-chercheur
DR	Directeur de recherche	IGM	Ingénieur général des mines

SMS	Sciences des Matériaux et des Structures
SPIN	Sciences des Processus Industriels et Naturels
FAYOL	Institut Henri Fayol
CMP	Centre de Microélectronique de Provence
CIS	Centre Ingénierie et Santé

*« Durant toutes ces années, et ceci depuis le début, chaque moment d'étude, chaque lecture, chaque recherche sur un auteur m'ont apporté quelque chose : selon les jours, selon les cas, c'était tout à coup une idée qui s'éclairait, un raisonnement que je découvrais et auquel je n'avais pas pensé, une expression qui me frappait, un sentiment humain qui m'atteignait, et, à chaque fois, c'était une impression d'enrichissement intellectuel et moral. La présence de ces textes a été pour moi source d'un étonnement, d'une admiration et d'une émotion constamment renouvelés. »*

Jacqueline de Romilly  
*de l'Académie française*

## **Remerciements**

A l'heure où j'écris ces lignes, un vif sentiment de joie m'anime. En effet, la rédaction de ces remerciements marque le point final de trois années entièrement consacrées aux recherches présentées dans cette thèse. Toutefois, je ne veux pas oublier que ces années furent souvent difficiles. Nombreuses furent les périodes de doutes, d'hésitations et de remises en question. C'est en me remémorant ces moments délicats que je souhaite adresser mes remerciements ; car si j'ai réussi à les surmonter et que mon projet de thèse arrive enfin à son terme, c'est grâce à toutes les personnes qui m'ont entouré, porté, conseillé, rassuré... et parfois un peu brusqué lorsque cela était nécessaire. Je leur serai toujours infiniment reconnaissant, jamais je n'oublierai tout ce qu'elles firent pour moi.

Mes parents tout d'abord. Les plus formidables que l'on puisse rêver. Dans tous mes choix de vie, ils m'ont suivi. Toujours ils ont cru en mes projets, même lorsque moi-même je doutais des directions prises. Merci Maman, merci Papa... merci de m'avoir fait confiance lorsque j'ai choisi de m'engager dans cette aventure et merci pour votre soutien inconditionnel lors de ces trois années de thèse. Plus généralement, merci pour tout ce que vous êtes et pour tout l'amour que vous portez à votre famille. Je ferai tout pour en être digne et continuer à porter vos valeurs morales. Je vous aime de tout mon cœur. Merci aussi à la mère Fouras – élément central et volubile de la famille – et son aveugle admiration pour l'ensemble de mes frasques et poncifs... Héloïse, ma p'tite sœur, on s'est pas beaucoup vu ces derniers temps... ça me coûte de le dire, mais tu m'as manqué !

Un mot pour ma Audrey, que je retrouve en cette fin de thèse. Ma princesse, tu m'as fait comprendre que parfois, ce qu'on cherche avec acharnement et aveuglement est juste sous nos yeux. Drôle d'histoire que la nôtre, peut-être ; mais belle histoire, c'est sûr ! Et crois-moi, ça n'est qu'un début : désormais mon Amour, on ne se lâche plus, on a toute une vie à construire.

Ensuite, je remercie très sincèrement l'Ecole des Mines de Saint-Etienne et STMicroelectronics, qui m'ont fait confiance pour ce projet de recherche, alors que je n'étais pas vraiment l'idéal-type du doctorant en Génie Industriel. Merci d'avoir su voir mes qualités et de m'avoir aidé à les développer, merci de m'avoir fait confiance, merci de m'avoir donné les moyens de mener à bien ces recherches. Particulièrement, je remercie Xavier Boucher qui fut, par ses talents de canalisateur et maïeuticien d'idées, par sa rigueur et son ouverture d'esprit, par sa gentillesse et sa profonde humanité, un Directeur de Recherches exceptionnel ; Olivier Roustant, pour tout le temps consacré à m'initier aux plaisirs et subtilités des séries temporelles, des changements structurels et autres statistiques robustes ; Marie-Agnès Girard, qui a su prêter une oreille patiente et attentive à mes interrogations métaphysiques et multidimensionnelles ; Xavier Ambrosioni, qui fut un manager exemplaire et attentif, malgré un emploi du temps débordant ; Patrick Burlat et Olivier Demeure, qui m'ont accueilli dans leurs Départements respectifs. Au-delà de nos relations professionnelles, sachez que j'ai vraiment apprécié vos qualités humaines et que j'ai été enchanté de faire ce bout de chemin à vos côtés. Je remercie aussi tous les autres membres de mon jury de thèse, pour leur intérêt pour mes travaux et leurs conseils perspicaces : Alain Bernard, Abdelaziz Bouras, Jean-Hugues Chauchat et Hervé Panetto.

De façon générale, merci à l'ensemble des personnes de l'ENSMSE et de ST avec qui j'ai partagé tant de bons moments de discussion et réflexion, de convivialité, de rigolade et/ou de sport : Gabrielle Bruyas, Zahia Mazer, Rodolphe Le Riche, Jean-François

Tchebanoff, Chris Yukna, Espéran Padonou, l'ensemble de l'ITT Crolles 300 et en particulier toute la bande l'archi - Gaël Bergeret, Jérôme Rey, Pascal Palazon, Charlène Guegnon, Haikel Mraïhi, Jean-Claude Favre, Georges Riera et Patrick Farrouch -, mais aussi Nico Blaszczyc, Stéphane Hubac, et tous les autres.

Un grand merci aussi à tous mes amis, qui m'ont soutenu et ont su rester présents lors de ces années de recherche. Vraiment, merci pour votre fidélité et votre compréhension... car je sais que je n'ai pas toujours été aussi disponible que j'aurais dû l'être... Vincent, en premier lieu, mon ami de toujours et pour toujours. En plus de la complicité exceptionnelle qui nous lie, tu m'as toujours aidé à m'élever et à me dépasser, en me faisant réaliser qu'il nous appartient de vivre nos rêves. Je te dois bien plus que tu ne l'imagines. Daniel, mon deuxième plus vieil ami, mon éminence grease, toujours partant pour explorer de nouveaux espaces. Gabrielle et Jean-Louis, ma famille adoptive lozérienne : que de merveilleux souvenirs de nos heures de galop et de tölt, suivis des fameux banana split du Petit Mazelet ! Les potes de la musique aussi (Bast, Lio, Béa) – longue vie à Mimic Octopus ! –, avec une pensée spéciale pour Titi et Laurence et leur incroyable sens de la convivialité : j'ai vraiment kiffé tous nos moments bouffe-musique-sport-discussion littéraire. Et bien sûr, tous les copains d'Alsace et d'ailleurs : Chua et Déborah, JB, Jo (encore merci pour les pictos), Diane et Guy (RDV le 19 octobre !), Jules et Céline (bienvenu à Victor !), les musiciens d'OpenNight (bravo pour Foolish Games), les deux Agnès et Laurent, Lipe de Lorraine, Thomas et Gaël de Paris (Los Angeles bientôt), et tous les autres que je ne peux mentionner ici, faute de place. Bougez pas, j'arrive, on a plein de trucs à faire ensemble !

Je n'oublie pas, bien sûr, tous les membres de la famille Lutz-Haeffelé-Staub. A eux aussi, je tiens à adresser tous mes remerciements : ce n'est pas toujours facile d'avoir un apprenti-chercheur préoccupé aux repas de famille, aussi je vous remercie d'avoir été compréhensifs, à l'écoute... et parfois même intéressés par mes propos ! Hervé, tu as su éclairer mon chemin lorsque j'en avais besoin ; Claude, merci pour tes bredelas ; Antoine, chauffe la bécane, sors tes baskets et dépoussière le VTT, je suis en pleine forme après trois années d'entraînement dans les Alpes !

Enfin, je veux terminer par une pensée très émue pour mon grand frère, qui n'est physiquement plus là, mais qui guide chacun des jours de ma vie, chacun de mes choix. Médéric, j'aurais tant aimé te voir à ma soutenance.

## Sommaire



<b>Introduction générale .....</b>	<b>- 1 -</b>
<b>Première partie. Etat de l'art, problématique et présentation du processus socio-statistique d'aide à la décision. ....</b>	<b>- 8 -</b>
<b>Chapitre 1. Etat de l'art et positionnement général de la problématique : nécessité d'un processus socio-statistique d'aide à la décision.....</b>	<b>- 12 -</b>
Introduction.....	- 13 -
Section 1. Etat de l'art : gestion de l'infrastructure des technologies de l'information et des capacités.....	- 15 -
Section 2. Formulation de la problématique scientifique de thèse : vers un processus socio-statistique d'aide à la décision.....	- 37 -
Conclusion du chapitre 1.....	- 41 -
<b>Chapitre 2. Définition et spécification du processus socio-statistique d'aide à la décision.....</b>	<b>- 43 -</b>
Introduction.....	- 44 -
Section 1. Eléments de définition du processus socio-statistique d'aide à la décision.....	- 46 -
Section 2. Spécification du processus socio-statistique d'aide à la décision.....	- 63 -
Conclusion du chapitre 2.....	- 75 -
<b>Deuxième partie. Application du processus socio-statistique d'aide à la décision : analyse du contexte organisationnel et structuration d'une démarche de modélisation.....</b>	<b>- 77 -</b>
<b>Chapitre 3. Introduction : présentation générale du contexte d'application STMicroelectronics. ....</b>	<b>- 82 -</b>
Introduction.....	- 83 -
Section 1. Gestion de l'infrastructure des technologies de l'information à Crolles 2 et enjeux de gestion.....	- 84 -
Section 2. Limites de mise en œuvre de la gestion des capacités des technologies de l'information à Crolles 2.....	- 88 -
Conclusion du chapitre 3.....	- 91 -
<b>Chapitre 4. Analyse du contexte organisationnel : étude de cas. ....</b>	<b>- 92 -</b>
Introduction.....	- 93 -
Section 1. Réalisation d'une étude de cas.....	- 95 -
Section 2. Situation problématique et problèmes de décision : présentation du périmètre industriel d'application des recherches.....	- 101 -
Conclusion du chapitre 4.....	- 107 -
<b>Chapitre 5. Structuration d'une démarche de modélisation destinée à répondre aux problèmes. ....</b>	<b>- 108 -</b>
Introduction.....	- 109 -
Section 1. Structuration d'une démarche de modélisation pour la constitution d'un modèle de prévision ( $II_1$ : dimensionnement des composants d'infrastructure).....	- 111 -
Section 2. Contributions à la détection de comportements anormaux ( $II_2$ : identification du risque de sous-capacité).....	- 124 -
Conclusion du chapitre 5.....	- 135 -

<b>Troisième partie. Application du processus socio-statistique d'aide à la décision : application de la démarche de modélisation, usage de la connaissance créée et évolution.</b> .....	<b>- 136 -</b>
Commentaire préliminaire concernant les traitements quantitatifs .....	- 141 -
<b>Chapitre 6. Application de la démarche de modélisation : création et formalisation de connaissances.</b> .....	<b>- 142 -</b>
Introduction .....	- 143 -
Section 1. Application à la migration MES : constitution d'un modèle de prévision.....	- 145 -
Section 2. Quelle généralité de la démarche de modélisation ? .....	- 166 -
Conclusion du chapitre 6 .....	- 173 -
<b>Chapitre 7. Usage du modèle de prévision pour l'aide à la décision.</b> .....	<b>- 175 -</b>
Introduction .....	- 176 -
Section 1. Construction du modèle d'évaluation : mobilisation de la connaissance pour l'aide à la décision. ....	- 177 -
Section 2. Utilisation du modèle d'évaluation : résultat et recommandation pour l'aide à la décision.-	180 -
Conclusion du chapitre 7 .....	- 182 -
<b>Chapitre 8. Evolution de la connaissance organisationnelle.</b> .....	<b>- 183 -</b>
Introduction .....	- 184 -
Section 1. Stratégie et architecture technique de maintien à jour de la connaissance du système.-	185 -
Section 2. Illustration de l'application du maintien à niveau de la connaissance .....	- 188 -
Conclusion du chapitre 8 .....	- 192 -
<b>Quatrième partie. Perspectives et conclusion générale.</b> .....	<b>- 193 -</b>
Section 1. Perspectives.....	- 196 -
Section 2. Conclusion générale.....	- 207 -
<b>Articles et communications issus des travaux présentés dans cette thèse.....</b>	<b>- 209 -</b>
<b>Références</b> .....	<b>- 212 -</b>
<b>Annexes</b> .....	<b>- 227 -</b>
<b>Annexe 1</b> .....	<b>- 228 -</b>
<b>Annexe 2</b> .....	<b>- 231 -</b>
<b>Annexe 3</b> .....	<b>- 235 -</b>
<b>Annexe 4</b> .....	<b>- 239 -</b>
<b>Annexe 5</b> .....	<b>- 241 -</b>
<b>Annexe 6</b> .....	<b>- 243 -</b>

## **Introduction générale**

## 1. Contexte

### 1.1. *Contexte scientifique général : système d'information et gestion des capacités des technologies de l'information*

Les travaux de recherches présentés dans cette thèse s'inscrivent dans la continuité de collaborations scientifiques antérieures, développées par la société STMicroelectronics et l'Ecole Nationale Supérieure des Mines de Saint-Etienne. Leur objectif général est toujours d'améliorer les pratiques de gestion des systèmes d'information, de façon à contribuer à la performance organisationnelle. On citera pour mémoire :

- *La thèse de Said Izza (Izza, 2006)*, qui porte sur l'interopérabilité des applications informatiques, par l'utilisation des ontologies et de l'approche service ;
- *La thèse de Julie Chapron (Chapron, 2006)*, qui s'intéresse à la modélisation du système d'information en vue d'en piloter l'évolution.

Cette thèse s'intéresse plus spécifiquement à une « strate » plus basse du système d'information : ses composants techniques, qui constituent l'infrastructure des technologies de l'information. Nous précisons ces termes de « système d'information » et d'« infrastructure des technologies de l'information » lors du premier chapitre. Toutefois, retenons dès à présent que cette fondation, constituée de technologies de l'information, permet d'assurer le bon fonctionnement d'un système d'information.

Dans le secteur de la microélectronique, les transformations fréquentes des environnements technologique et économique imposent une grande réactivité organisationnelle et une forte capacité à gérer le changement. Ces changements impactent fortement les processus de fabrication. Ils se répercutent largement sur le système d'information et les technologies de l'information, supports de nombreux processus industriels. L'agilité de l'entreprise repose ainsi sur une gestion prévoyante du système d'information et des technologies de l'information.

Dans ce contexte, les managers en charge des technologies de l'information doivent développer leur aptitude à gérer avec efficacité les capacités du système informatique dont ils ont la charge. Effectivement, une sous-capacité des composants de l'infrastructure des technologies de l'information risquerait de nuire à l'activité du système d'information et d'entraver le bon fonctionnement de l'usine.

Ainsi, l'objectif des recherches présentées dans cette thèse est de proposer des méthodes et des outils décisionnels, qui doivent contribuer à la bonne gestion des capacités des technologies de l'information. Plus spécifiquement, on se situera dans le contexte d'une usine de production de semi-conducteurs de STMicroelectronics. Il sera nécessaire de tenir compte des étroites interactions entre les activités industrielles et les technologies de l'information, de manière à favoriser la performance organisationnelle.

## 1.2. Contexte industriel : STMicroelectronics et le site de Crolles

Précisons l'environnement industriel des travaux présentés dans cette thèse. STMicroelectronics est une société multinationale franco-italienne. En 2010, elle est le septième fabricant mondial de composants microélectroniques, derrière des concurrents tels qu'Intel, Samsung Electronics, Toshiba, Texas Instrument... Présente en Europe, Afrique, Amérique et Asie, STMicroelectronics est active dans le développement et la fabrication de produits semi-conducteurs de type :

- *Composants discrets* : ce sont des systèmes à semi-conducteurs n'exécutant qu'une seule fonction électronique élémentaire (exemple : diode, transistor) ;
- *Circuits intégrés* : ce sont des circuits électroniques fabriqués sur un substrat semi-conducteur, hébergeant une quantité variable de transistors, diodes, résistances, condensateurs. Ils permettent de réaliser des fonctions complexes. Ils sont composés de deux parties : la puce en silicium et le boîtier, qui protège la puce de son environnement et facilite le montage dans les systèmes électroniques.

Dans le monde, STMicroelectronics comprend quinze sites de production et seize centres de recherche et développement, de nombreux centres de conceptions et d'applications, ainsi que des bureaux de vente. Son effectif mondial s'élève à 53000 employés environ en 2010.

Nos recherches se sont déroulées sur le site français de Crolles (Département 38, Isère). Ce site est dévolu aux activités de recherche et développement et de production *front-end*. Les activités *front-end* correspondent à la première partie du processus de production : travail des plaques de silicium vierges, en vue de la réalisation des puces. On distingue deux étapes de *front-end* : la fabrication des composants (FEOL : *Front End Of the Line*), puis des interconnexions (BEOL : *Back End Of the Line*). Deux types de tests accompagnent le travail des plaques de silicium : tests paramétriques (validation du procédé de fabrication) et fonctionnels (validation de la fonctionnalité de chaque puce sur la plaquette). Les activités *back-end* correspondent à la deuxième partie du processus de production : découpage des puces à partir des plaquettes de silicium terminées, puis montage, connexion, moulage et test des circuits intégrés finals.

Le site de Crolles réunit deux usines : Crolles 1 et Crolles 2. Elles traitent des plaquettes de silicium de dimensions différentes : 200mm de diamètre pour Crolles 1 et 300mm de diamètre pour Crolles 2 (pour une épaisseur similaire d'environ 0,5mm). Dans le secteur de la microélectronique, la tendance est d'utiliser des plaquettes les plus grandes possibles, afin de pouvoir y graver davantage de puces simultanément et de limiter les pertes sur le bord de plaque, d'où une production accrue à moindre coût.

Nos travaux ont été effectués au sein de l'usine Crolles 2, dont le chantier a débuté en 2000 (à titre indicatif : 1990 pour Crolles 1) et l'activité de fabrication en 2002 (1992 pour Crolles 1). En 2006, Crolles 2 employait 1721 personnes (2200 pour Crolles 1). Très automatisée, l'activité de Crolles 2 repose fortement sur son système d'information et les technologies de l'information. Afin d'en assurer une gestion appropriée, la structure organisationnelle présentée ci-après était en vigueur au moment où nos recherches étaient réalisées.

1.3. *Périmètre de recherche : la gestion du système d'information et des technologies de l'information à Crolles 2*

Le département MSG Crolles 300 (*Manufacturing Solution Group Crolles 300*) de Crolles 2 est le commanditaire de nos travaux de recherche. Il a pour missions de :

- *Maintenir le système d'information de Crolles 2*, afin de garantir le processus de production pour l'ensemble des lots de fabrication, la haute disponibilité des applications, un accès aisé et rapide aux bonnes informations au bon moment, ainsi qu'une solution à la pointe de la technologie ;
- *Gérer le système d'information de Crolles 2*, en assurant le développement d'une relation de proximité avec les utilisateurs du système d'information, des délais de livraison en adéquation avec les contraintes de la production, une architecture technologique robuste<sup>1</sup> et redondante<sup>2</sup>, un niveau de connaissance permettant de remplir les missions du département, ainsi que sa compétitivité technologique et économique.

Outre sa supervision globale et certaines fonctions transversales de gestion de projets, le département MSG Crolles 300 est divisé en quatre services. Trois services, aux spécialités fonctionnelles différentes, gèrent les applications informatiques du système d'information de Crolles 2 :

- *Le service Decisional Solutions* : il gère et exploite à des fins décisionnelles des données issues de l'activité de l'usine ;
- *Le service Manufacturing Execution System* : il assure la gestion informatisée du processus de production ;
- *Le service Process Control and Automation* : il garantit le contrôle et l'automatisation du processus de production.

Leur mission générale est de garantir une bonne gestion des applications dont ils ont la charge, afin d'assurer les missions du département MSG Crolles 300.

Le quatrième service est responsable de l'architecture de l'infrastructure des technologies de l'information, c'est-à-dire de la sous-couche technique permettant de faire correctement fonctionner et communiquer l'ensemble des applications, dans le respect des missions du département MSG Crolles 300. C'est au sein de ce service « Architecture de l'infrastructure des technologies de l'information » de Crolles 2<sup>3</sup> qu'ont été développés nos travaux de recherche.

---

<sup>1</sup> Une architecture des TI robuste doit permettre d'assurer la haute disponibilité des applications du système d'information, à un niveau de performance satisfaisant.

<sup>2</sup> La redondance d'une architecture est un moyen d'assurer sa robustesse, par une duplication partielle ou totale de la solution technologique.

<sup>3</sup> Pour faciliter la lecture, nous y ferons référence sous le terme « service infrastructure TI-C2 ». De même, nous emploierons le terme « managers de l'infrastructure TI-C2 » pour évoquer l'équipe en ayant la responsabilité.

## 2. Présentation des travaux de recherche

### 2.1. Approche retenue

Dans le cadre de ses missions, le service infrastructure TI-C2 est responsable de la gestion des capacités de l'infrastructure informatique de l'usine de Crolles 2. Pour cela, il cherche à développer un ensemble de supports décisionnels, destinés à faciliter cette activité. Notre objectif est d'apporter un soutien scientifique et applicable industriellement à cette tâche.

Pour cela, il est décidé de faire usage de méthodes statistiques, dans l'objectif de relever deux enjeux de gestion majeurs : la nécessité de gérer des systèmes technologiques informatiques complexes, tout en tenant pleinement compte de l'évolution de l'activité industrielle de l'usine. Le but de nos travaux est d'intégrer ces méthodes au cœur des processus décisionnels des managers de l'infrastructure TI-C2, en développant une approche scientifiquement structurée. Ce travail doit aboutir à la proposition d'un processus général et générique, permettant d'inscrire l'usage des statistiques dans les processus d'aide à la décision. Ce processus doit être applicable à des cas décisionnels particuliers, afin d'en formaliser les besoins et d'assurer le développement de solutions d'aide à la décision *ad hoc*. Pour cela, cette thèse définira et spécifiera ce que nous nommons un « processus socio-statistique d'aide à la décision ». Impliquant méthodes statistiques, mais également experts<sup>4</sup> de STMicroelectronics, nous montrerons notamment comment il doit permettre de contribuer à la gestion des connaissances organisationnelles.

Par ailleurs, nous cherchons à situer nos travaux dans la perspective d'un référentiel de gestion reconnu dans le milieu de l'informatique industrielle : ITIL (*Information Technology Infrastructure Library*). Ce référentiel fournit une trame, permettant de bien délimiter les besoins décisionnels des managers de l'infrastructure TI-C2. En particulier, cette thèse s'intéressera à deux activités du processus ITIL de gestion des capacités : d'une part, la modélisation et l'analyse des tendances, considérées comme le cœur des travaux de recherches présentés ; d'autre part, la détection de comportements anormaux, dans le cadre de l'activité de contrôle, traitée comme un travail annexe, complémentaire au cœur des recherches. La sélection, la hiérarchisation de ces besoins, et le développement de méthodes statistiques décisionnelles appropriées, est le résultat de l'application du processus socio-statistique d'aide à la décision, dans un contexte particulier de STMicroelectronics. Néanmoins, précisons que le caractère générique de ce processus aurait pu nous amener à développer des solutions toutes autres dans des contextes différents.

---

<sup>4</sup> Nous entendons par experts ou expertise, les acteurs de l'organisation qui possèdent des connaissances plus ou moins formalisées relatives aux systèmes étudiés.

## 2.2. Présentation du plan de thèse général

Afin de développer cette approche basée sur la définition, la spécification, puis l'application d'un processus socio-statistique d'aide à la décision, le plan de thèse général présenté dans la Figure 1 a été retenu.

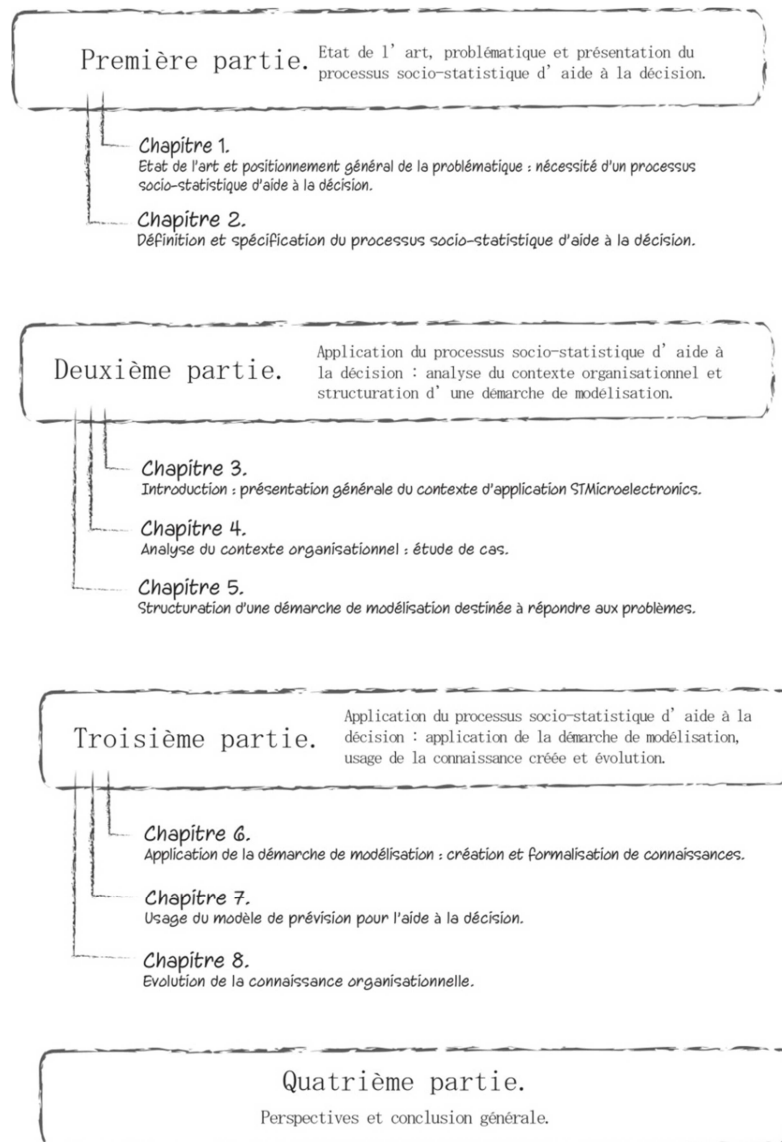


Figure 1. Plan de thèse général.

La première partie permet de positionner le cadre bibliographique de nos travaux. A partir d'un état de l'art relatif à la gestion des technologies de l'information, le chapitre 1 fait émerger le besoin d'un processus socio-statistique d'aide à la décision. La description générale de ce processus sert à formuler la problématique générale de nos travaux. Le chapitre 2 apporte alors une proposition de réponse, en définissant et spécifiant un processus socio-statistique d'aide à la décision, à partir d'éléments de la littérature.

La deuxième partie concerne l'application de la première phase du processus d'aide à la décision défini. C'est une phase de contextualisation, destinée à préciser le besoin décisionnel. Après une présentation générale du terrain de recherche STMicroelectronics



(chapitre 3), le processus est directement appliqué. En résulte la définition d'une étude de cas, permettant de préciser le contexte décisionnel (chapitre 4), puis la structuration d'une démarche de modélisation adaptée (chapitre 5).

La troisième partie traite de l'application de la deuxième phase du processus d'aide à la décision. Elle se base sur les conclusions de la partie précédente. Tout d'abord, la démarche de modélisation définie dans le chapitre 5 est désormais appliquée au cas décisionnel STMicroelectronics (chapitre 6). En résulte un modèle statistique, qui est ensuite mis en usage, pour l'aide à la décision, dans le chapitre 7. Finalement, le chapitre 8 montre comment un tel modèle peut être placé dans une perspective dynamique d'évolution des connaissances organisationnelles.

Cette thèse s'achève sur une quatrième partie, qui propose un certain nombre de perspectives pour les travaux présentés, ainsi qu'une conclusion générale.

### *2.3. Contributions scientifiques attendues*

Outre l'amélioration locale des processus décisionnels et de la gestion des connaissances chez STMicroelectronics, par la définition et l'application du processus socio-statistique d'aide à la décision, un ensemble de contributions scientifiques peut être extrait des travaux présentés dans cette thèse :

- *Le chapitre 4 propose une analyse rigoureuse et structurée d'une situation décisionnelle : scientifiquement, nous montrons comment des concepts existant peuvent être mobilisés pour analyser, formaliser une situation décisionnelle concrète, et être capable d'identifier et de décrire des besoins décisionnels. Le résultat attendu est une description conceptuelle des problématiques décisionnelles traitées dans la thèse ;*
- *Le chapitre 5 montre comment construire une démarche de modélisation contextualisée : d'un point de vue scientifique, ce chapitre illustre comment des démarches de modélisation raisonnées peuvent être mises au point, en accord avec des besoins décisionnels industriels et la problématique générale de cette thèse. Ce chapitre doit donc avoir pour résultat de spécifier une démarche de modélisation, répondant aux problématiques décisionnelles spécifiquement traitées ;*
- *Les chapitres 6, 7, 8 montrent comment démarches de modélisation et méthodes statistiques peuvent être utilisées pour créer, utiliser et maintenir à niveau des connaissances organisationnelles, dans un but d'aide à la décision : on démontre ici l'applicabilité des méthodes proposées, ainsi que leurs contributions à l'amélioration des pratiques, en accord avec la problématique générale de la thèse. En termes de résultats, ces chapitres permettent d'illustrer comment des modèles peuvent être produits, tenus à jour et mis en usage dans un objectif décisionnel.*

Pris dans leur ensemble, les chapitres 2, 4, 5, 6, 7 et 8 constituent également une contribution scientifique à part entière. Elle consiste à proposer une démarche générale de conduite d'un processus d'aide à la décision avec utilisation de modèles quantitatifs, facilitant par là même la gestion des connaissances de l'organisation.

**Première partie.** Etat de l'art, problématique et présentation du processus socio-statistique d'aide à la décision.

## Plan détaillé de la première partie.

### Chapitre 1. Etat de l'art et positionnement général de la problématique : nécessité d'un processus socio-statistique d'aide à la décision.

#### Introduction

#### Section 1. Etat de l'art : gestion de l'infrastructure des technologies de l'information et des capacités.

1.1. Contexte général : gestion de l'infrastructure des technologies de l'information.

*1.1.1. Rôle et place de l'architecture de l'infrastructure des technologies de l'information dans les organisations.*

*1.1.2. Enjeux de gestion posés aux managers en charge de l'architecture de l'infrastructure des technologies de l'information.*

*i. Enjeu n°1 : gérer un système complexe.*

*ii. Enjeu n°2 : soutenir pleinement l'activité de l'organisation.*

*1.1.3. Réponses des bonnes pratiques de gestion.*

*i. Cadre documentaire.*

*ii. Cadre document et enjeux de gestion.*

1.2. Gestion des capacités des technologies de l'information et périmètre industriel d'application des recherches.

*1.2.1. Problème décisionnel général.*

*1.2.2. Introduction au processus ITIL de gestion des capacités et identification du périmètre industriel d'application des recherches.*

*1.2.3. Mise en œuvre de l'activité (Q) Modélisation et analyse des tendances : apports et limitations.*

*i. Généralités.*

*ii. Simulation et modélisation analytique : apports et limites de l'analyse des réseaux de files d'attentes.*

*iii. Analyse des tendances : apports et limites des méthodes d'exploitation des données.*

*1.2.4. Mise en œuvre de la détection de comportements anormaux dans le cadre de l'activité (K) Contrôle : apports et limitations.*

*i. Généralités.*

*ii. Détection de comportements anormaux : apports et limitations.*

#### Section 2. Formulation de la problématique scientifique de thèse : vers un processus socio-statistique d'aide à la décision.

#### Conclusion du chapitre 1.

### Chapitre 2. Définition et spécification du processus socio-statistique d'aide à la décision.

#### Introduction.

#### Section 1. Eléments de définition du processus socio-statistique d'aide à la décision.

1.1. Utilisation de modèles d'aide à la décision dans le cadre d'une approche rationnelle et constructive.

*1.1.1. Positionnement des recherches.*

*i. La décision dans les organisations : diversité des approches.*

*ii. Recherche menée : modèles d'aide à la décision et approche rationnelle constructive.*

*1.1.2. Tsoukiàs et l'utilisation de modèles d'aide à la décision : une approche rationnelle constructive.*

*i. Présentation générale des travaux de Tsoukiàs.*

- ii. Contextualisation de l'aide à la décision : représentation de la situation problématique et formulation du problème.*
- iii. Aide à la décision basée sur l'utilisation de modèles : modèle d'évaluation et recommandation finale.*

#### *1.1.3. Positionnement dans le processus socio-statistique d'aide à la décision.*

### 1.2. Processus socio-statistique d'aide à la décision et gestion des connaissances organisationnelles.

#### *1.2.1. Objectifs.*

#### *1.2.2. Définitions, repères théoriques et positionnement dans le processus socio-statistique d'aide à la décision.*

- i. Artefacts et connaissances organisationnelles.*
- ii. Création de connaissances organisationnelles.*

#### *1.2.3. Mise en œuvre : gestion du système de connaissances et triptyque données-information-connaissance.*

- i. Gestion du système de connaissances organisationnelles et processus socio-statistique d'aide à la décision.*
- ii. Triptyque données-information-connaissance et processus socio-statistique d'aide à la décision.*

## **Section 2. Spécification du processus socio-statistique d'aide à la décision.**

### 2.1. Présentation générale.

### 2.2. Prérequis : analyse du contexte organisation et construction de connaissances organisationnelles.

#### *2.2.1. Généralités.*

#### *2.2.2. Construction des connaissances par la réalisation d'une étude de cas.*

- i. Récolte des données (aspect syntaxique)*
- ii. Structuration des données – informations (aspect sémantique)*
- iii. Interprétation des informations – connaissances (aspect pragmatique)*

#### *2.2.3. Synthèse.*

### 2.3. Questionnement central : structuration d'une démarche de modélisation et construction/évolution des connaissances organisationnelles.

#### *2.3.1. Généralités.*

#### *2.3.2. Construction et évolution des connaissances par l'utilisation des statistiques.*

- i. Récolte des données (aspect syntaxique).*
- ii. Structuration des données – informations (aspect sémantique).*
- iii. Interprétation des informations – connaissances (aspect pragmatique).*

#### *2.3.3. Synthèse.*

## **Conclusion du chapitre 2.**

\*  
\* \*

Les deux chapitres de cette première partie posent les fondements de nos travaux de recherches.

Le chapitre 1 positionne un état de l'art relatif aux technologies de l'information et à la gestion de leurs capacités, thématique de nos travaux. Un certain nombre de difficultés décisionnelles, liées à cette thématique et à l'utilisation des statistiques sont mises en avant (section 1). En découle la nécessité de développer ce que nous dénommons un « processus socio-statistique d'aide à la décision », visant à favoriser l'utilisation des statistiques dans des situations industrielles décisionnelles concrètes. La description générale de ce besoin correspond à la problématique générale de cette thèse (section 2).

Principalement, on s'intéressera à deux questions spécifiques : la réalisation de modèles prospectifs (cœur des travaux de recherche) et la détection de comportements anormaux (travaux annexes). L'identification et la hiérarchisation de ces besoins est un premier résultat de l'application du processus socio-statistique d'aide à la décision chez STMicroelectronics, tel que nous le détaillerons dans la deuxième partie de la thèse. Néanmoins, nous en présenterons les limitations dès ce chapitre d'état de l'art, pour faciliter la lecture et mettre en avant au plus tôt la valeur-ajoutée de l'application du processus socio-statistique d'aide à la décision.

Le chapitre 2 donne les détails de définition et de spécification de ce processus socio-statistique d'aide à la décision. Tout d'abord, la section 1 définit un certain nombre d'éléments théoriques sous-jacents au processus. Sur cette base, la section 2 spécifie la trame générique du processus socio-statistique d'aide à la décision, qui pourra être appliqué concrètement dans les parties 2 et 3 de cette thèse.

L'articulation générale de cette partie 1 est illustrée dans la Figure 2.

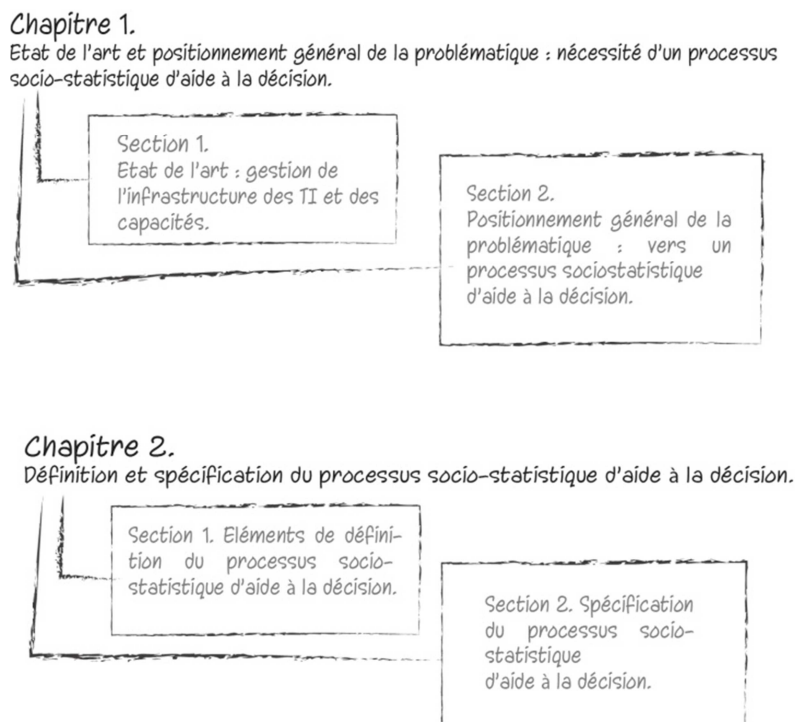


Figure 2. Plan général de la partie 1.

\*  
\* \*

**Chapitre 1. Etat de l'art et positionnement général de la  
problématique : nécessité d'un processus socio-statistique d'aide à  
la décision.**

## Introduction

Ce chapitre expose le cadre et les motivations de nos recherches, afin d'en démontrer l'intérêt et la valeur-ajoutée. Sa structure est résumée dans la Figure 3.

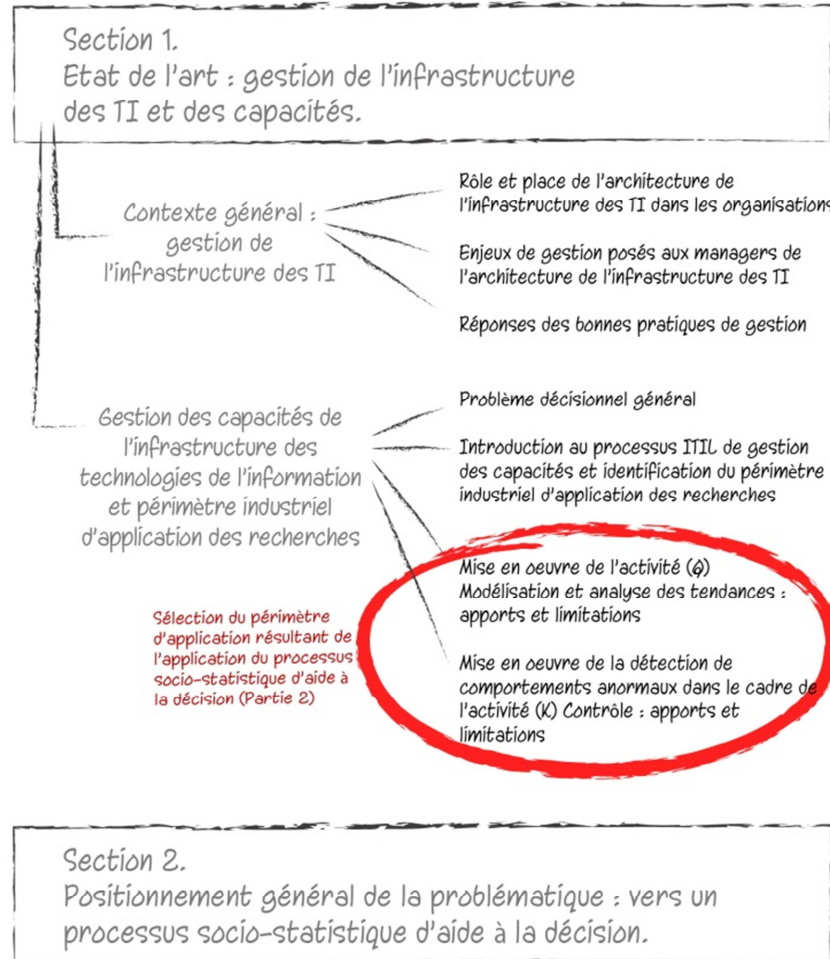


Figure 3. Plan du chapitre 1.

La section 1 propose un état de l'art. D'une part, il cadre le contexte général de nos travaux : la gestion de l'infrastructure des technologies de l'information. Leur importance dans l'activité des organisations est précisée. Sont également indiqués deux enjeux de gestion auxquels sont confrontés les managers responsables de ces technologies de l'information : la gestion d'un système technologique complexe et la prise en considération de l'activité métier de l'organisation. Nous expliquons que des référentiels et guides de bonnes pratiques professionnelles existent pour guider les managers dans cette voie, et en présentons les grands principes.

D'autre part, la section 1 précise le périmètre direct des recherches présentées dans cette thèse : la gestion des capacités. Après avoir décrit le problème décisionnel général, une présentation du processus de gestion des capacités du référentiel ITIL (*Information Technology Infrastructure Library*) est proposée. Ce référentiel comporte un ensemble d'activités de gestion des technologies de l'information. Elles peuvent être utilisées par les managers, pour faciliter leurs prises de décisions. Toutefois, leur mise en œuvre peut

s'avérer problématique. En particulier, nous discutons certains apports et limitations de deux activités de gestion au sein de ce référentiel : la modélisation et l'analyse des tendances, puis la détection de comportements anormaux, dans le cadre de l'activité de contrôle. Comme nous l'avons déjà mentionné, notre objectif de vouloir améliorer spécifiquement ces deux activités correspond à une application industrielle directe de nos travaux, même si leur identification et leur sélection s'inscrit dans une volonté scientifique plus générale de renforcement des processus décisionnels. Au sein de ce périmètre d'application, la modélisation et l'analyse des tendances représentent le cœur de nos recherches ; la détection de comportements anormaux est considérée comme un apport complémentaire.

Ensuite, les limitations suggérées permettent de préciser la problématique générale de la thèse (section 2). Pour ce faire, nous expliquons pourquoi il semble nécessaire de développer ce que nous nommons un « processus socio-statistique d'aide à la décision ». Ce processus permet d'englober l'amélioration des deux activités ITIL mentionnées, dans un cadre plus exhaustif d'aide à la décision, dont les objectifs généraux seront précisés.



## **Section 1. Etat de l'art : gestion de l'infrastructure des technologies de l'information et des capacités.**

### **1.1. Contexte général : gestion de l'infrastructure des technologies de l'information**

#### *1.1.1. Rôle et place de l'architecture de l'infrastructure des technologies de l'information dans les organisations*

En fournissant un ensemble de fonctions relatives à la gestion des informations (acquérir, traiter, stocker, communiquer des informations), les Systèmes d'Information (SI) contribuent à la performance et à la compétitivité des organisations (Earl, 1989 ; Cardinali, 1992 ; Reix et *al.*, 2011). C'est aussi vrai pour certains cas particuliers d'organisations, en rapport avec les recherches présentées dans cette thèse : systèmes industriels de production (Gowan et Mathieu, 1996 ; Briffaut, 2000), systèmes industriels de production de semi-conducteurs (Leachman et Hodges, 1996). Un système d'information est un objet multidimensionnel, pouvant être caractérisé selon trois dimensions principales (Reix et *al.*, 2011) :

- *La dimension organisationnelle*, en tant qu'élément des processus et de la structure de l'organisation ;
- *La dimension informationnelle*, en tant que producteur de représentations du monde réel ;
- *La dimension technologique*, en tant qu'assemblage d'équipements et de dispositifs techniques.

Relativement à notre sujet de recherche, on s'intéressera principalement à la gestion de la dimension technologique du système d'information. Cette dimension technologique correspond à la mise en œuvre opérationnelle d'un système d'information informatisé<sup>5</sup>, *via* des Technologies de l'Information (TI). En conséquence, ces technologies de l'information jouent un rôle important dans le fonctionnement des processus et la performance de l'organisation. Ce rôle est souligné par de nombreuses publications (Bharadwaj, 2000 ; Aerts et *al.*, 2004 ; Smith et *al.*, 2007 ; Berghout et *al.*, 2011 pour les dimensions organisationnelles en générales, et Deuel, 1994 ; Kumar et *al.*, 2005 ; Samad et *al.*, 2007, pour un focus sur les activités industrielles).

Les activités et processus métiers d'une organisation dépendent donc d'une définition stratégique de l'architecture des TI (Earl, 1989 ; Weill et Ross, 2004 ; Ross et *al.*, 2006).

---

<sup>5</sup> La notion de système d'information n'implique pas nécessairement informatisation (Reix et *al.*, 2011), mais c'est exclusivement de SI informatisés dont nous parlerons dans cette thèse.

On peut distinguer (Aerts et al., 2004) :

- *L'architecture applicative*, qui organise les logiciels et les données permettant la mise en œuvre de la dimension informationnelle du système d'information ;
- *L'architecture de l'infrastructure des technologies de l'information*<sup>6</sup>, qui organise l'ensemble des moyens techniques permettant de faire fonctionner l'architecture applicative (ordinateurs, réseau, base de données, systèmes d'exploitation, *middleware*<sup>7</sup>, etc.).

Nos travaux se sont déroulés au sein d'un service responsable de cette architecture des TI. La Figure 4 résume le positionnement d'un tel service dans une organisation. Ces concepts seront mobilisés dans la suite de cette thèse.

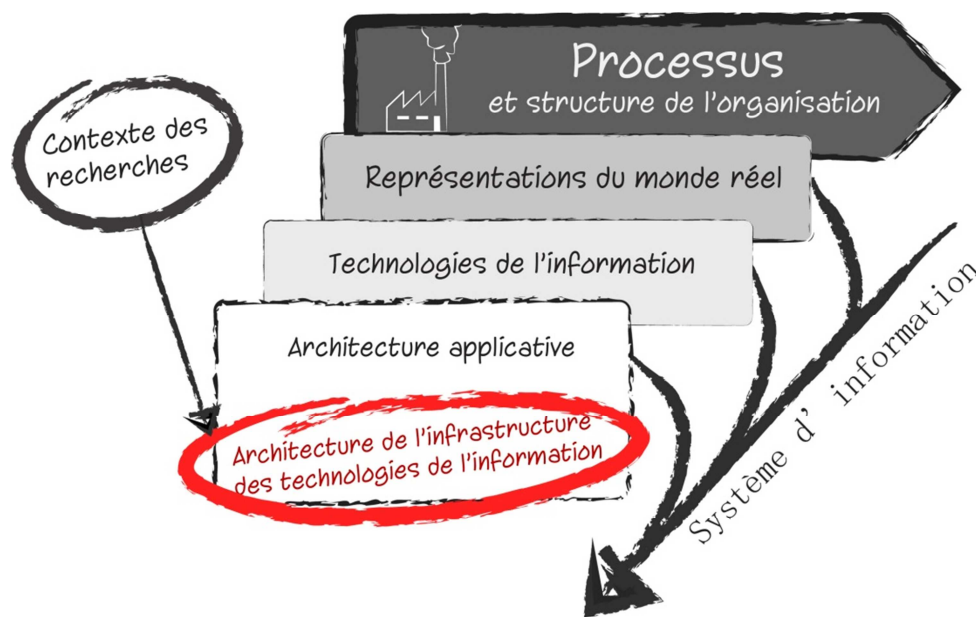


Figure 4. Rôle et place de l'architecture de l'infrastructure des technologies de l'information dans les organisations.

Ainsi, les managers en charge de l'architecture de l'infrastructure des TI doivent, par la bonne gestion d'un ensemble de composants techniques, contribuer à la performance et à la compétitivité d'une organisation en soutenant ses processus et sa structure. Cette mission n'est pas tâche facile : l'analyse de la littérature permet d'identifier deux enjeux principaux auxquels sont confrontés ces managers.

#### 1.1.2. Enjeux de gestion posés aux managers en charge de l'architecture de l'infrastructure des technologies de l'information

L'identification de ces deux enjeux est une proposition propre à nos recherches, fruit d'une analyse de la littérature liée à la gestion des technologies de l'information.

<sup>6</sup> Les auteurs parlent d' « *information et communication technologies platform architecture* ». Nous utiliserons le terme infrastructure plutôt que plate-forme.

<sup>7</sup> Le *middleware*, essentiel dans les architectures modernes, assure la communication entre différents composants applicatifs et techniques hétérogènes (Lefebvre, 1994).

*i. Enjeu n°1 : gérer un système complexe*

Selon Mitleton-Kelly et Land (2011), un système d'information est un système complexe. Sa dimension technologique<sup>8</sup> est caractérisée par des composants fortement interconnectés, interdépendants et entremêlés. Elle est à la source de propriétés émergentes spécifiques au système technique, qui constituent l'une des dimensions de la complexité du SI.

Les évolutions des techniques en usage dans les infrastructures des technologies de l'information confirment cette proposition. De 1960 à 2000, les innovations en informatique et réseau ont contribué au développement de technologies de plus en plus décentralisées et communicantes (Buyya, 2002). On constate la transition d'architectures centralisées, fondées sur des *mainframes* monolithiques et isolés, vers des systèmes en réseaux fortement distribués : modèle client/serveur (Lefebvre, 1994 ; Englander, 2010), systèmes multi-ordinateurs de type *clusters* (Englander, 2010), etc. Ces architectures sont caractérisées par une multitude de composants et répondent tout à fait à la description de Mitleton-Kelly et Land (2011) d'un système technologique complexe.

Ces architectures complexes des TI sont aujourd'hui largement présentes dans les entreprises (Aerts et *al.*, 2003 ; Schiesser, 2010). Or, la gestion de systèmes complexes implique des postures spécifiques (Von Bertalanffy, 1993 ; Le Moigne, 1999 ; Morin, 2005), que les managers en charge de l'architecture de l'infrastructure des technologies de l'information doivent être capables d'intégrer.

*ii. Enjeu n°2 : soutenir pleinement l'activité de l'organisation*

Le rôle important des TI dans le bon fonctionnement dans l'activité de l'organisation a déjà été mentionné. Il est en conséquence essentiel que les managers des technologies de l'information ne se limitent pas à un rôle d'experts techniques des systèmes dont ils ont la charge. Ils doivent au contraire s'engager pleinement en tant que contributeurs à la performance de l'organisation. Néanmoins, le rôle fondamental du SI et des TI dans les organisations n'est pas une évidence. Il est le fruit d'une évolution progressive des pratiques (Mason *in* McFarlan, 1984 ; Bounfour, 2010). Il convient alors de s'assurer que les approches des managers des technologies de l'information soient en phase avec cette évolution, afin qu'ils soient capables d'intégrer les enjeux des activités métier de l'organisation, dans l'exercice de leur mission. Pour autant, même lorsque l'importance du soutien aux activités de l'organisation est intégrée par les managers de l'infrastructure des TI, un risque de manque de maîtrise de ces activités existe.

---

<sup>8</sup> Les auteurs suggèrent d'autres sources de complexité, tels que les aspects sociaux et organisationnels des systèmes d'information ; mais elles sont hors du périmètre des thématiques abordées dans cette thèse.

Nous suggérons deux explications :

- *Le mode de fonctionnement décentralisé* des TI modernes tend à déconnecter la gestion de l'architecture de l'infrastructure des TI de l'architecture applicative (Aerts et al., 2004 ; Porra et al., 2006 ; voir aussi Lefebvre, 1994 pour une description du modèle client/serveur). En conséquence, les managers en charge l'architecture de l'infrastructure des TI risquent de perdre la maîtrise des usages applicatifs et fonctionnels des systèmes qu'ils gèrent ;
- *Le déploiement des SI/TI dans les organisations entraîne une augmentation de la taille des systèmes techniques*, mais pas nécessairement une hausse du nombre d'experts gérant les infrastructures (Fronheiser, 2004, 2006 ; Grummitt, 2009). Ces experts ont donc moins de temps à consacrer à la compréhension des usages applicatifs et fonctionnels des systèmes dont ils ont la charge.

C'est pourquoi les gestionnaires responsables de l'architecture de l'infrastructure des TI doivent rester vigilants lors de l'exercice de leur mission, afin d'assurer leur capacité à soutenir les activités de leur organisation.

La Figure 5 synthétise les enjeux ainsi identifiés.

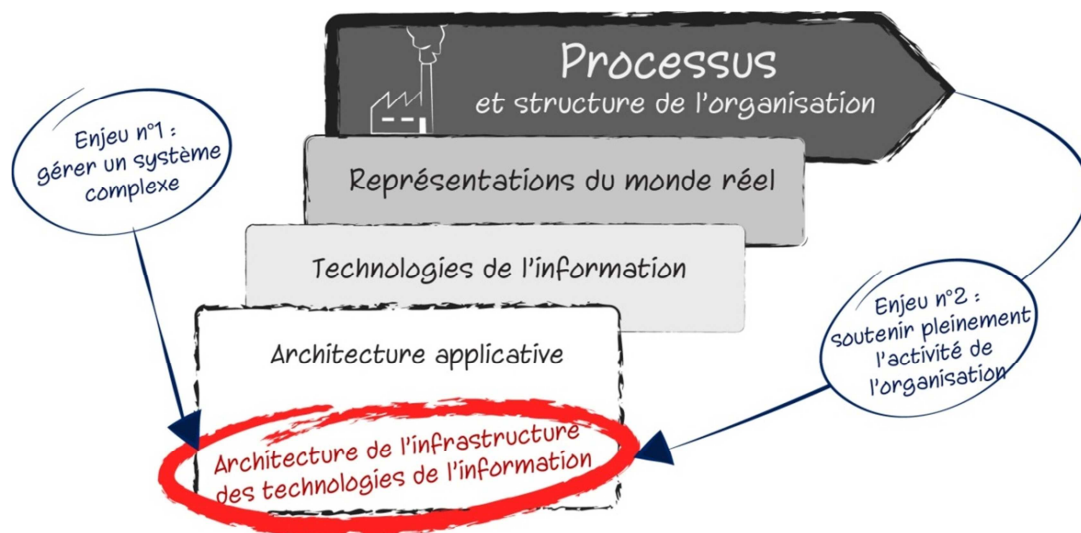


Figure 5. Les enjeux de gestion posés aux managers en charge de l'infrastructure des technologies de l'information.

Face à ces enjeux, les managers peuvent avoir recours à des guides de bonnes pratiques.

### 1.1.3. Réponses des bonnes pratiques de gestion

#### i. Cadre documentaire

Les managers en charge de l'infrastructure des TI peuvent s'appuyer sur de nombreux référentiels et normes, mettant à disposition un ensemble de bonnes pratiques professionnelles : Alford et al. (2006) en recensent 22 principaux. Nous nous

concentrerons sur deux d'entre eux : la norme ISO 20000 (ISO-1, 2005 ; ISO-2, 2005) et le référentiel ITIL<sup>9</sup> (ITIL, 2012). En effet, ils représentent bien les activités de gestion des technologies de l'information généralement recommandées et comptent parmi les cadres documentaires les plus déployés en entreprise :

- *La norme ISO 20000* est un standard international pouvant donner lieu à certifications ;
- *Le référentiel ITIL* tend à s'imposer comme un standard industriel *de facto* (McNaughton et al., 2010 ; Valiente et al., 2012).

ISO 20000 et ITIL sont complémentaires (Dugmore et Taylor, 2008) : la norme décrit les grands principes de gestion (spécifications et code de bonnes pratiques), tandis que le référentiel précise des processus aidant à leur mise en œuvre<sup>10</sup>. Charge aux organisations de définir ensuite leurs propres modes d'opérationnalisation *in situ* des processus ITIL (ISO-3, 2012). Ces relations sont résumées dans la Figure 6.

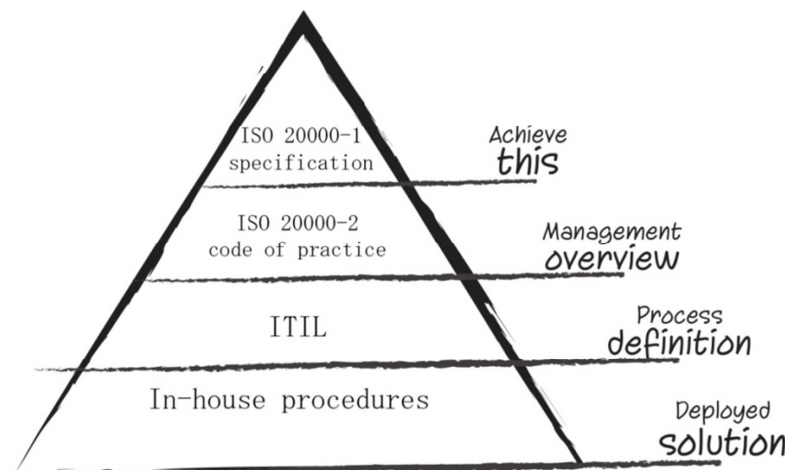


Figure 6. Bonnes pratiques de gestion : relations entre ISO 20000, processus ITIL et implémentation *in situ* (ISO-3).

## ii. Cadre documentaire et enjeux de gestion

Deux concepts fondamentaux du couple ISO 20000/ITIL montrent que ce cadre documentaire vise à prendre en compte les enjeux identifiés précédemment. Le livre ITIL *Service Delivery* indique :

« *A holistic approach should be adopted [...] to adopt consistency and integration within all activities and processes across the entire IT technology [...]* » (Lloyd et Rudd, 2007)

Cette dimension holistique préconisée par ITIL permet de prendre en compte la complexité du système technologique à gérer. Effectivement, les démarches holistiques sont à la base des approches systémiques, de leurs fondements positivistes (Von

<sup>9</sup> *Information Technology Infrastructure Library*

<sup>10</sup> La dernière version de ITIL (ITIL V3) est organisée autour de cinq livres : *Service strategy, Service Design, Service Transition, Service Operation, Continuous Service Improvement*.

Bertalanffy, 1993) aux ramifications ultérieures basées sur la notion de complexité (Le Moigne, 1999 ; Morin, 2005)<sup>11</sup>. L'enjeu n°1 n'est donc pas étranger aux processus ITIL. Ceci implique que les managers de l'infrastructure des technologies de l'information doivent avoir une approche globale du système qu'ils gèrent. Leur mission est de maintenir une solution complète et non fragmentée, finalisée sur sa capacité à soutenir l'activité organisationnelle, comme l'indique la suite de la définition du livre *Service Delivery* :

« [...], *providing end-to-end business-related functionality and quality.* »

Cette finalité est formalisée dans le couple ISO 20000/ITIL par le concept central de service. Les managers des TI doivent en effet être guidés par leur aptitude à fournir un service complet, conforme aux exigences des utilisateurs (les « clients »), dans le cadre de leurs participations aux processus de l'organisation (Wan et *al.*, 2007 ; McNaughton et *al.*, 2010 ; Wan et Jones, 2011 ; Valiente et *al.*, 2012). On parle d'approche ITSM (*Information Technology Service Management*). Ainsi, les managers de l'infrastructure des TI deviennent des fournisseurs de services, pour les clients de l'organisation. Cette approche par la gestion des services est fondamentale : elle joue un rôle de « glue » établissant une connexion organique entre les activités de l'organisation et les technologies de l'information (Wan et *al.*, 2007). L'enjeu n°2 est donc pris en considération dans les processus ITIL. La Figure 7 résume le positionnement du couple ISO 20000/ITIL par rapport aux enjeux de gestion identifiés. Elle illustre également que, correctement mise en œuvre, l'application des processus ITIL favorise la qualité de la gestion de l'infrastructure des technologies de l'information, et contribue ainsi pleinement aux processus et à la structure de l'organisation, donc à sa performance et à sa compétitivité.

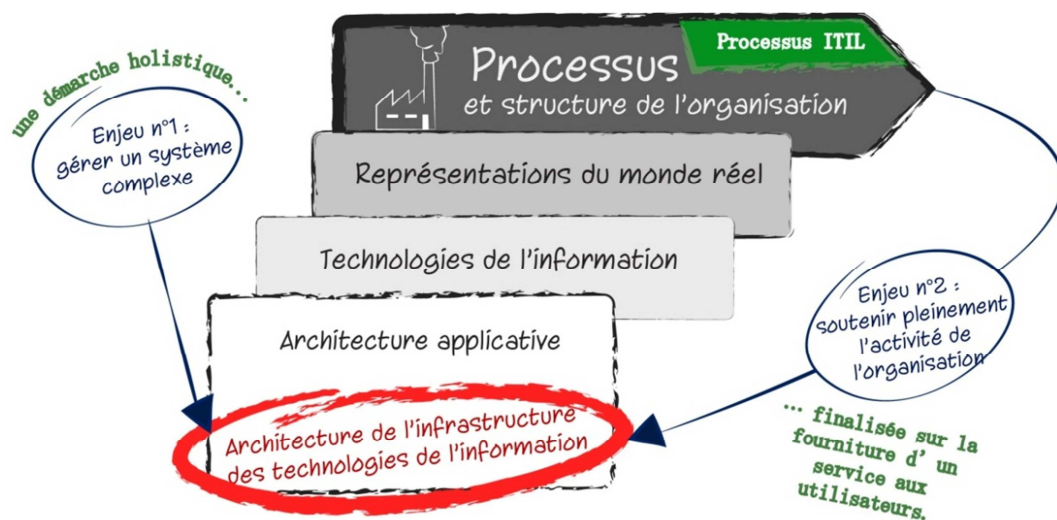


Figure 7. ISO 20000/ITIL : une réponse aux enjeux posés aux managers de l'infrastructure des TI.

<sup>11</sup> L'approche systémique vise à appréhender les comportements de systèmes, pris dans leur globalité. Cette approche, telle qu'on la connaît aujourd'hui, est le résultat d'une évolution épistémologique. Initiée sur la base de travaux positivistes, tels que la théorie générale des systèmes de Von Bertalanffy, on parlait surtout de sciences des systèmes. Ensuite, pour mieux répondre à la complexité des systèmes à comprendre, une ramification du paradigme systémique s'est progressivement tournée vers une épistémologie constructiviste, intégrant les concepts d'artificialité et de modélisation interactionniste. On parle alors parfois de sciences de la complexité (Le Moigne, 2003, 2007 ; Avenier, 2011).

Cette situation générale de gestion étant posée, nous nous intéressons désormais aux problèmes spécifiques considérés dans nos travaux de recherches et nous montrons la valeur-ajoutée qu'ils peuvent apporter, lorsqu'on les situe dans un référentiel tel que ITIL.

## **1.2. Gestion des capacités des technologies de l'information et périmètre industriel d'application des recherches**

### *1.2.1. Problème décisionnel général*

La gestion des capacités d'un site industriel est un problème bien connu (Forgionne, 1991) : comment garantir une infrastructure globale (« *facility* »), apte à soutenir l'activité de production ? Dans les sites de productions de semi-conducteurs, les investissements sont tels que la rentabilité d'une usine n'est garantie qu'à la condition d'une maîtrise excellente du flux de production. En effet, un arrêt de production inopportun causé par un manque de capacité, aura un impact grave sur la performance de l'organisation. Le pilotage de la capacité des équipements de production de semi-conducteurs est par exemple une problématique bien connue, faisant l'objet d'une recherche active (cf. Geng et Jiang, 2009, pour un état de l'art récent).

Les technologies de l'information sont également concernées. Elles jouent d'ailleurs un rôle de plus en plus important, dans des usines dont l'automatisation, donc le recours au système d'information, va croissant. La sous-capacité de l'infrastructure des TI risque en effet d'entraîner la dégradation, voir l'arrêt d'un ou plusieurs services du système d'information et, ainsi, de nuire au bon fonctionnement de l'usine et à la rentabilité du site de production. La norme ISO 20000 définit la gestion des capacités comme le fait de :

*« s'assurer que le fournisseur de services dispose à tout moment d'une capacité suffisante pour répondre aux exigences acceptées, actuelles et futures, des clients de l'entreprise. »* (ISO-1, 2005)

Plus précisément, pour les managers de l'infrastructure des TI, la question est de garantir que le système technique dont ils ont la charge soit suffisamment dimensionné, pour fournir sans interruption un service permettant de soutenir les activités de l'organisation<sup>12</sup> présentes et futures (Dugmore et Lacy, 2005 ; Grummitt, 2009).

Cette responsabilité sous-entend un ensemble de prises de positions. Les managers des TI doivent en effet chercher des compromis entre le choix des composants de l'architecture des TI (les ressources techniques) et leur coût (les ressources financières), afin d'ajuster au mieux la qualité de service fournie par le système, à un niveau de service satisfaisant pour les clients, dans le cadre de leurs activités présentes et futures. Le problème décisionnel général posé aux managers est de continuellement parvenir à assurer un juste équilibre entre ces différents éléments. Comme l'indique la Figure 8, cet équilibre requiert la prise en compte des deux enjeux de gestion proposés ci-dessus.

---

<sup>12</sup> La littérature relative à la gestion des capacités des TI parle généralement d'activité « *business* » ou « métier ».

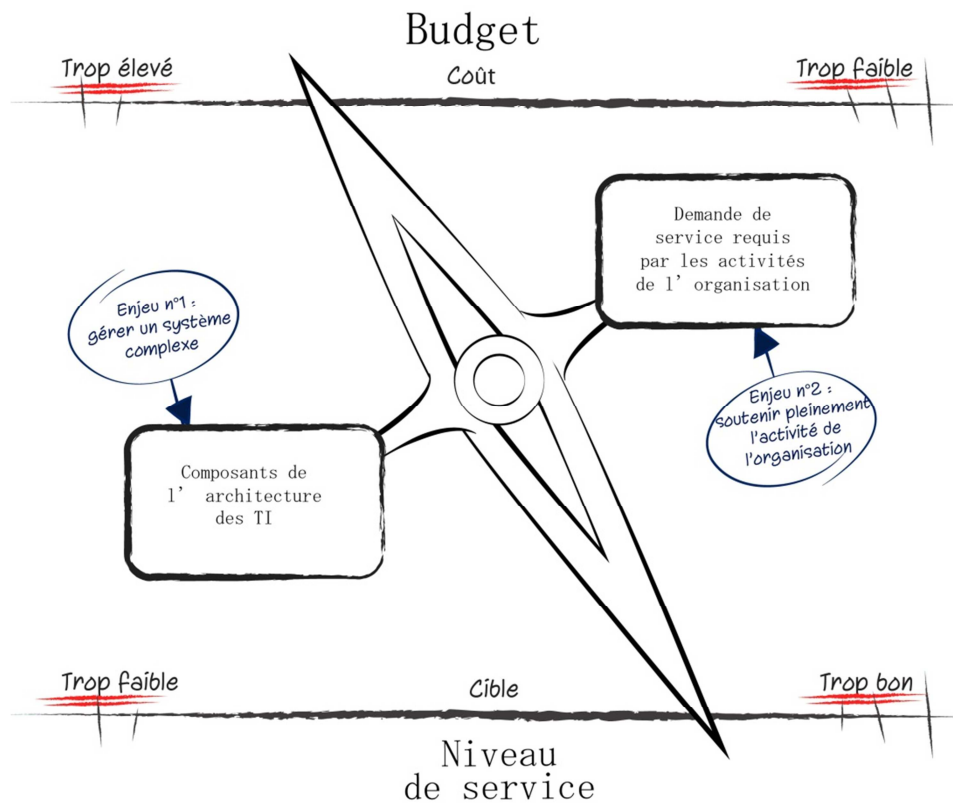


Figure 8. Gestion des capacités : un problème décisionnel pour les managers de l'architecture des TI (inspiré de Grummitt, 2009).

Le référentiel ITIL propose un processus de gestion des capacités, pouvant aider les managers de l'architecture des TI dans leurs prises de décisions.

#### 1.2.2. Introduction au processus ITIL de gestion des capacités et identification du périmètre industriel d'application des recherches

Par souci de concision, seules les idées essentielles du processus ITIL de gestion des capacités sont évoquées. Pour plus de détail, on pourra se référer au chapitre 4.3. *Capacity Management* du livre ITIL *Service Delivery* (Lloyd et Rudd, 2007) et à plusieurs références spécifiques : Fronheiser (2006), Grummitt (2009), Kloesterboer (2011). Sauf mention expresse, la rédaction de cette section est basée sur notre interprétation de ces sources. En accord avec le problème décisionnel général précisé ci-dessus, le processus ITIL de gestion des capacités vise à fournir des recommandations managériales, telles que :

*« cost-justifiable IT capacity in all areas of IT always exist and is matched to the current and future agreed needs of the business, in a timely manner »* (Lloyd et Rudd, 2007)



Comme le montre la Figure 9, le processus est divisé en quatre activités-clés :

- (A) *Passer en revue l'activité et les performances actuelles (Review current activity and performance) ;*
- (B) *Améliorer le service et la capacité du système actuels (Improve current service and system capacity) ;*
- (C) *Evaluer, valider et documenter les nouveaux besoins et capacité (Assess, agree and document new requirements and capacity) ;*
- (D) *Planifier la nouvelle capacité (Plan new capacity).*

Trois types de livrables sont attendus, afin de documenter la réalisation de ces activités-clés :

- (E) *Rapports de capacité et performance (Capacity and performance reports) ;*
- (F) *Prévisions (Forecasts) ;*
- (G) *Plan de capacité (Capacity plan).*

La réalisation des activités-clés doit être menée à trois niveaux, liés les uns aux autres :

- (I) *Niveau « service » : la qualité du service fourni par les TI est évaluée sur la base de contrats de services, négociés avec les clients, représentants de l'activité métier (exemples : temps de réponse applicatif, temps de rupture de service autorisé, ...). On fait généralement référence à ces contrats par leur acronyme anglais : SLA pour *Service Level Agreements* (cf. Paschke et Bichler (2008), Goo et Huang (2008) ou Edgington et al. (2010) pour un aperçu de recherches récentes sur cette notion) ;*
- (H) *Niveau « métier » : pour intégrer au mieux l'activité organisationnelle, les managers des TI doivent explicitement incorporer l'activité métier dans leur gestion des capacités ;*
- (J) *Niveau « ressource » ou « composant » : le service effectif sera fonction de la gestion des divers éléments constituant l'architecture de l'infrastructure des TI.*

Afin de mettre en œuvre ce processus de gestion des capacités, le référentiel ITIL décrit un ensemble d'activités support. D'une part, les activités itératives, qui constituent une boucle d'amélioration continue du système des TI :

- (K) *Contrôle (Monitoring) ;*
- (L) *Analyse ;*
- (M) *Ajustement (Tuning) ;*
- (N) *Implémentation.*

D'autre part, plusieurs activités *ad hoc* sont également proposées :

- (O) *Evaluation des besoins applicatifs (application sizing)* ;
- (P) *Gestion de la demande* ;
- (Q) *Modélisation et analyse des tendances (Modelling and trending)* ;
- (R) *Gestion et contrôle des seuils*.

On distingue un sous-ensemble « quantitatif » parmi ces activités : le *capacity planning*. Son objectif est de décrire, sur la base d'éléments quantifiables, le comportement présent de l'infrastructure des TI, d'anticiper son évolution future et d'être capable de fournir des propositions d'évolutions en conséquence (Dugmore et Lacy, 2005 ; Lloyd et Rudd, 2007 ; Grummitt, 2009 ; Kloesterboer, 2011).

Sont généralement considérées comme activités du *capacity planning* : (K) Contrôle, (Q) Modélisation et analyse des tendances, (R) Gestion et contrôle des seuils, et éventuellement (L) Analyse et (O) Evaluation des besoins applicatifs<sup>13</sup> (Bronner, 1980 ; Domanski 1999b, Gunther, 2007 ; Allspaw, 2008). Ce sous-ensemble constitue le périmètre général de nos recherches : on s'interroge sur leur rapport aux enjeux décisionnels identifiés précédemment.

Ce référentiel ITIL nous permet donc de clarifier le positionnement applicatif de nos travaux. Dans la réalité du déroulement de la recherche, ce positionnement n'existait pas *a priori*. Il résulte à la fois de l'analyse de la problématique industrielle et de l'identification des verrous scientifiques associés. C'est donc par soucis de clarté que nous utilisons dès maintenant ITIL pour ce positionnement applicatif. Nous justifierons ces choix plus loin (chapitre 2), lors de l'application du processus socio-statistique d'aide à la décision à un cas opérationnel de STMicroelectronics.

Ainsi, en nous situant par rapport au cadre ITIL, le périmètre d'application de nos recherches concerne principalement l'activité (Q). Nous avons par ailleurs été amenés à nous intéresser à l'un des aspects de l'activité (K), concernant la détection de comportements anormaux, en tant que complément à l'activité (Q). La Figure 9 résume l'ensemble des éléments du processus ITIL et le périmètre de recherche identifié.

---

<sup>13</sup> Discipline souvent référencée dans la littérature sous le terme *Software Performance Engineering* (Smith et Williams, 1993, 2002 ; Smith *in* Bernardo et Hillston, 2007).

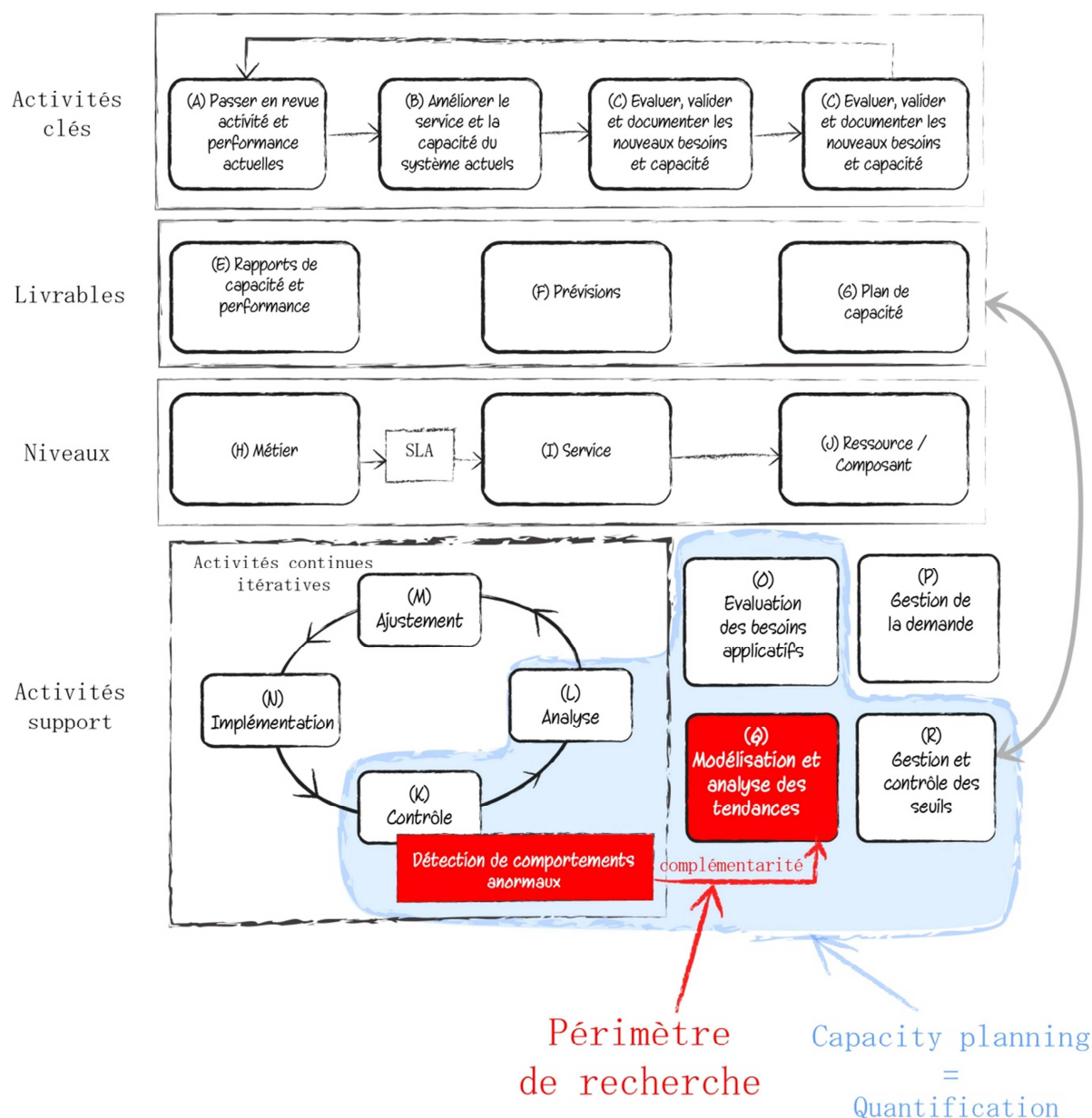


Figure 9. Vue d'ensemble du processus ITIL de gestion des capacités et périmètre industriel d'application des recherches (inspiré de Fronheiser, 2006 ; Lloyd et Rudd, 2007 ; Grummitt, 2009 ; Kloesterboer, 2011).

Pour mettre en œuvre ce processus, le référentiel ITIL préconise la mise en place d'un système d'information de gestion de la capacité, le CMIS (*Capacity Management Information System*). Son but est de mettre à disposition et d'assurer l'exploitation de toutes les données nécessaires pour la gestion des capacités, quels que soient les niveaux (ressource, service et métier). Recourir à une telle approche statistique pour soutenir les activités de *capacity planning* est une approche quasi-généralisée chez les praticiens depuis les années 2000 (Domanski, 1999a, 1999b ; Cockroft et Walker, 2001 ; Gunther, 2007 ; Allspaw, 2008)<sup>14</sup>. S'appuyer ainsi sur l'utilisation des données disponibles favorise la prise en compte d'éléments quantitatifs dans les prises de décision.

<sup>14</sup> Quelques citations caractéristiques : « Assume that anything that is not measured is out of control [...] Measure things that you have no intention of tuning or controlling » (Cockroft et Walker, 2001) ; « Measure, measure, measure » (Allspaw, 2008).

Comme indiqué en Figure 6, les processus ITIL se sont pas des solutions « clés en main », directement applicables dans une organisation. Le déploiement sur le terrain relève de la responsabilité des organisations elles-mêmes, et peut poser des difficultés aux managers concernés (Wan et *al.*, 2007 ; Valiente et *al.*, 2011 ; Wan et Jones, 2011). C'est souvent le cas des activités de *capacity planning*. Plusieurs références montrent des limitations dans leurs applications, au regard des enjeux de gestion proposés. Jain (1991), Lacity et *al.* (1997), Gunther (2007) montrent que la complexification des architectures des infrastructures des TI modernes rend les études de *capacity planning* de plus en plus difficiles (Jain, 1991 ; 1997 ; Gunther, 2007). Porra et *al.* (2006), Kloesterboer (2011) indiquent que l'activité métier y reste insuffisamment intégrée. Nous précisons cette constatation dans le cadre des deux activités faisant l'objet de nos recherches.

### *1.2.3. Mise en œuvre de l'activité (Q) Modélisation et analyse des tendances : apports et limitations*

#### *i. Généralités*

La modélisation prédictive est l'une des activités centrale de la gestion des capacités et fait l'objet d'une littérature abondante. Elle doit permettre de constituer des modèles quantitatifs caractérisant le système des TI en place, pour tester des scénarios d'anticipation qualifiés de « *what if?* », afin de disposer d'éléments chiffrés pour prendre des décisions à visée prospective. Le référentiel ITIL retient trois approches qui peuvent être utilisées pour établir un modèle (Lloyd et Rudd, 2007) : simulation, modélisation analytique (*analytical and simulation modelling*), et analyse des tendances (*trend analysis*). Certains auteurs (Bronner, 1980 ; Cortada, 1983 ; Grummitt, 2009) proposent une typologie un peu plus détaillée, selon une classification allant des méthodes les plus simples (instinct, voir « laissez-faire ») aux plus complexes à mettre en œuvre (protocoles expérimentaux). On trouve également des approches plus rares : par exemple Chakravarty et Jain (1990) utilisent des méthodes de programmation linéaire ; Norton (1998) propose une méthode de modélisation intégrant une approche de type système dynamique. Néanmoins, la typologie ITIL est représentative des principaux courants rencontrés dans la littérature.

Avant d'aborder la littérature relative à ce sujet, il est nécessaire de rappeler le cadre de cette thèse : elle n'est pas une thèse en sciences de l'informatique, ni en mathématiques appliquées. L'objectif n'est donc pas d'entrer en détail dans toutes les subtilités des recherches dans ces domaines, mais de donner un aperçu des méthodes applicables dans un contexte industriel, ainsi que des limites à leur mise en œuvre. Comme l'indique Jain (1991) :

*« Sophisticated queueing theoretic models and statistical techniques are of interest to performance specialists [nous précisons ces termes ci-dessous], but in the majority of cases simple analysis yields sufficiently useful results. Complex analyses are not feasible, for example, in an industrial environment where there are considerable time restraints and deadline »*

ii. *Simulation et modélisation analytique : apports et limites de l'analyse des réseaux de files d'attente*

Simulation et modélisation analytique sont les méthodes traditionnelles, largement répandues dans la littérature de *capacity planning* depuis les années 1970. Leur principal objectif est avant tout la construction de modèles de performance (cf. le terme « *performance specialists* » de la citation précédente), caractérisant les files d'attente au sein des systèmes informatiques (« *queueing theoretic models* »). La théorie des files d'attente est un cas classique de recherche opérationnelle (Cohen, 1995). Nous en présentons les grands principes. Cette théorie s'intéresse à des systèmes décrits par des variables d'états discrètes, dont on étudie les changements d'états sous occurrence d'un ensemble d'événements (Baynat, 2000). Les événements peuvent avoir lieu de façon continue dans le temps, mais les changements des variables d'états sont discrets.

De façon générale, la constitution d'un modèle de performance se déroule en deux étapes. Tout d'abord, le système considéré est formalisé par un ensemble d'équations. Différents formalismes peuvent être employés : réseaux de files d'attente, réseaux de Pétri, automates stochastiques, etc. La littérature informatique se concentre sur ces premiers. Dans ce cadre d'un système informatique, les variables d'état peuvent être : nombre de processus en cours de traitement sur l'unité centrale, état des processus (en attente, prêt, actif), nombre de requêtes pour un accès en écriture sur l'unité de stockage, état des ressources partagées ; les événements associés étant alors : début ou fin de traitement d'un processus par l'unité centrale, demande d'écriture ou fin d'écriture sur l'unité de stockage, demande d'accès à une ressource ou libération de cette ressource. Un système de files d'attente peut être composé d'une file simple ou d'un réseau de files d'attente, et peut être défini de façon très générale comme :

*« une boîte noire, dans lequel les clients arrivent suivant un processus quelconque, séjournent pour recevoir un ou plusieurs services et finalement quittent le système »* (Baynat, 2000)

Un tel système est défini par un certain nombre de paramètres (généralement stochastiques) :

- *Débit d'entrée ( $X_e$ )* : nombre de clients arrivant dans le système, c'est-à-dire changements d'état à opérer (par unité de temps sur une période d'observation donnée) ;
- *Débit de sortie ( $X_s$ )* : nombre de clients quittant le système, c'est-à-dire changements d'état opérés ;
- *Nombre de clients présents dans le système ( $Q$ )* : changements d'état en cours ;
- *Temps de réponse ( $R$ )* : temps de changement d'état.

Ces paramètres sont illustrés par la Figure 10.

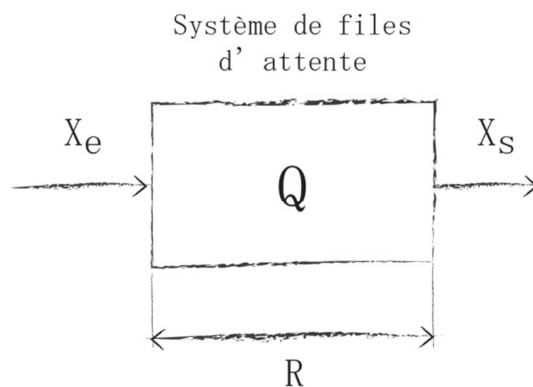


Figure 10. Système de files d'attente général (Baynat, 2000).

On associe un cinquième paramètre à ce système,  $U$ , qui représente le taux d'utilisation du système, c'est-à-dire la proportion de temps pendant laquelle il est occupé sur la période d'observation donnée (pour opérer les changements d'états). Une vision formelle du système peut donc être constituée par la mise en relation de ces variables (par exemple, en utilisant la loi générale de Little (1961), très connue en sciences de l'informatique). Ce cadre général permet d'appréhender une multitude de formes de systèmes de files d'attente (Baynat, 2000), des plus simples au plus compliquées (réseaux monoclases ouverts ou fermés, réseaux multiclases purement ouverts ou purement fermés, avec ou sans changement de classes, etc.). Lorsqu'on considère un réseau de files d'attente, on peut considérer les paramètres de performances pour l'ensemble du réseau, ou pour différents sous-ensembles du réseau. Par exemple, dans un système informatique, on peut considérer un *mainframe* dans sa globalité, ou on peut s'intéresser à l'activité du processeur, aux accès disques, etc.

La deuxième étape consiste à analyser le modèle construit dans le formalisme choisi (Baynat, 2000). Tout d'abord, une analyse qualitative est destinée à définir les propriétés structurelles et comportementales du système (absence de blocage, comportement fini ou borné, etc.). Ensuite, une analyse quantitative calcule les paramètres du système. Deux types de méthodes principales existent :

- *La simulation* : le modèle est programmé informatiquement et son évolution est reproduite pas à pas. Plusieurs réalisations particulières du modèle sont étudiées ;
- *Les méthodes analytiques* : les paramètres sont calculés mathématiquement en résolvant les équations du modèle.

Nous retrouvons les deux approches ITIL, largement développées dans la littérature informatique propre au *capacity planning* (exemples d'utilisation généraux : Ferrari, 1978 ; Lavenberg, 1983 ; Lazowska et al., 1984 ; Cady et Howarth, 1990 ; Jain, 1991), qui a également parfois recours à un troisième type d'évaluation des files d'attentes : la simulation par mise en œuvre de protocoles expérimentaux.

Ces modèles sont principalement utilisés pour analyser les performances d'un système informatique, soit le paramètre  $R$  (Menascé et al., 1994), mais on y a également recours pour considérer l'utilisation du système. Dans ce dernier cas, on considère plus particulièrement le paramètre  $U$ , tout en sachant que  $U$  et  $R$  sont liés. Ces modèles caractérisent suffisamment finement un système pour pouvoir tester des scénarios « *what-if* » à différents plans : on peut évaluer l'impact de changements au niveau du

système et au niveau de ses entrants. Dans le cas du *Software Performance Engineering*, ces entrants peuvent également être découpés séquentiellement, selon les différentes étapes d'exécution d'un programme (Smith et Williams, 1993, 2002 ; Smith in Bernardo et Hillston, 2007), ou selon l'utilisation faite par un programme des ressources systèmes (Sia et Ho, 1997). Néanmoins, nous identifions deux limitations dans le cadre d'un usage industriel de ce type de modèles.

La première limitation concerne la constitution de ces modèles, que la complexification des architectures des TI (enjeu de gestion n°1) rend de plus en plus compliqués : voir par exemple Menascé et *al.* (1994), Foxon et *al.* (1996), pour les systèmes clients-serveurs, Gunther (1998) pour les architectures distribuées en général, ou Menascé et Almeida (1998, 2000, 2002) pour les infrastructures *web*. Leur constitution nécessite des compétences fortes en théorie des files d'attente et beaucoup de temps pour l'analyse des modèles et de leurs paramètres. Or ces deux facteurs ne sont pas toujours réunis dans un contexte industriel. De plus, cette constatation est accentuée par un changement d'état d'esprit dans la considération du matériel informatique. En effet, dans les années 1970-1980, la question était d'exploiter au mieux des investissements technologiques coûteux (Cortada, 1983). La baisse constante du coût du matériel informatique a rendu cette hypothèse caduque (Lefebvre, 1994 ; Domanski, 1999a, 1999b ; Gunther, 2007) : la tendance est plus à l'achat *ad hoc* de nouveaux composants informatique qu'à l'optimisation du matériel disponible. En conséquence, les infrastructures des TI deviennent très changeantes. De fait, il n'y a souvent plus grand intérêt à modéliser finement le comportement d'une ressource informatique, qui de toute manière sera rapidement obsolète et remplacée. Comme l'indique Jain (1991) :

*« Initial computer systems consisted of only a few components. Each component was expensive enough to justify the cost of accurately modeling its behavior. Also, the number of users on the system was large. Thus, even though each user's behavior is highly variable, the aggregate performance of all users did not vary that much and could be modeled accurately. In distributed environments of today, the system consists of a large number of semi-autonomous clients, servers, network links, and I/O [inputs-outputs] devices. Workstation usage is very different from others, and the interactions are also rather complex. Also, the cost of individual components is not high enough to justify accurate modeling. »*

Compte-tenu des évolutions des technologies de l'information observées lors des vingt dernière années, cette citation n'est que plus vraie en 2012, année de rédaction de cette thèse.

L'autre limitation concerne l'utilisation de ces modèles pour effectuer des prévisions concernant l'état du système en considérant l'activité métier (enjeu de gestion n°2, niveau (H) du processus ITIL de gestion des capacités). En effet, à partir des années 1990, la capacité à quantifier l'activité métier et d'en tenir compte lors de l'utilisation des modèles de performance devient une préoccupation de plus en plus importante. On parle de *natural forecasting unit*, ou de *key volume indicator*, ou encore de *forecasting business unit*. Pour pouvoir utiliser ces variables métier dans le cadre de modèles de performance, il faut pouvoir les rattacher à des variables d'états des modèles de file d'attente. Une telle configuration arborescente n'est pas aisée, particulièrement au sein des systèmes distribués, dans lesquels les responsables des systèmes informatiques peuvent perdre la maîtrise des usages des activités applicatives et métier (voir plus haut).

Comme le mentionne Kelly (2005) :

*« Today's commercial production applications are composed of numerous opaque software components [...] To make matters worse, applications are increasingly distributed across both geographical and organizational boundaries. Merely to collect in one place sufficient measurement data and knowledge of system design to support a detailed performance analysis is often very difficult in practice. »*

C'est pourquoi de plus en plus d'ouvrages, orientés vers la pratique professionnelle, considèrent que le recours à la théorie des files d'attente est de moins en moins adapté à un usage industriel. Lourds à concevoir et à maintenir, ces modèles semblent peu appropriés à la gestion des infrastructures informatiques industrielles modernes, complexes, décentralisées et changeantes (Jain, 1991 ; Browning, 1994 ; Cockroft et Walker, 2001)<sup>15</sup>.

### *iii. Analyse des tendances : apports et limites des méthodes d'exploitation des données*

Il est déjà fait référence aux méthodes statistiques exploratoires et prévisionnelles dans les ouvrages les plus anciens. Toutefois, on constate que le rôle de l'analyse des données dans la gestion des capacités a fortement évolué au cours de ces 40 dernières années. Le premier usage de ces méthodes est complémentaire aux modèles de performance, pour les calibrer à partir de données issues du système réel (Ferrari, 1978 ; Jain, 1991 ; Browning, 1994 ; Menascé et al., 1994 ; Menascé et Almeida, 1998, 2000, 2002). Des structures au sein des variables métiers ou des variables d'état peuvent être identifiées par l'emploi de méthodes de *clustering*. L'activité de ces variables peut également être anticipée par l'emploi de modèles de régression ou modélisation des séries temporelles.

Cependant, les statistiques sont progressivement sorties de ce rôle secondaire. Elles sont devenues un outil de pilotage des capacités et de modélisation à part entière, largement employé depuis les années 2000, notamment comme une réponse aux limitations des modèles de performance dans le cadre d'usages industriels, évoquées ci-dessus. L'objectif est de s'appuyer sur des modèles plus opérationnels, construits à partir de l'observation des données issues de l'activité réelle du système en fonctionnement (Domanski, 1999a, 1999b ; Cockroft et Walker, 2001 ; Gunther, 2007 ; Allspaw, 2008). Ces méthodes semblent en effet une bonne réponse à une problématique de gestion des systèmes complexes (enjeu de gestion n°1), sous contrainte de temps. Ahituv et Igbaria (1988) indiquent que le recours à la statistique est la meilleure approche face à un système déjà en place, car elle permet d'analyser son comportement à partir de données collectées à partir de son activité réelle. C'est une approche holiste, qui permet d'observer d'éventuelles propriétés émergentes du système informatique complexe.

---

<sup>15</sup> Nous précisons toutefois que les modèles de performance encore aujourd'hui totalement adaptés et pertinents dans le cadre de périmètres bien identifiés. Simplement, ils sont hors du sujet de cette thèse.



Gunther (2007) ajoute :

*« for complex systems where planning time is limited, statistical models can offer a powerful approach »*

Dans ce contexte, les concepts de performance et d'utilisation tendent à se dissocier. On ne cherche donc plus à constituer un modèle global du système, mais des modèles *ad hoc*, constitués en fonction d'objectifs précis, orientés soit vers la gestion des performances, soit vers l'anticipation de l'utilisation des ressources du système (Gunther, 2007 ; Allspaw, 2008). Les modèles d'anticipation ne visent donc plus à prédire simultanément un niveau de performance, en fonction de l'utilisation du système informatique. Néanmoins, on associe généralement à l'utilisation du système des valeurs-seuils empiriques, au-delà desquelles on sait que les performances se dégraderont, ou que le système se bloquera (selon les auteurs, on parle de « *threshold* », « *ceiling* » ou « *upgrade indicator* », en lien avec l'activité ITIL (R) Gestion et contrôle des seuils).

Que ce soit pour l'analyse des performances ou de l'utilisation du système, une multitude de méthodes statistiques est envisageable. Potentiellement, toute méthode d'exploitation de données pourrait être appliquée à la gestion des capacités et il serait impossible d'en donner un aperçu général ici. On peut renvoyer aux articles du *Computer Measurement Group* (CMG, 2012) pour un aperçu régulièrement mis à jour de l'évolution de ces pratiques. De façon générale, on observe le recours à deux grandes catégories de méthodes : d'une part, l'application directe et « manuelle » de méthodes statistiques usuelles (régression, modélisation des séries temporelles) ; d'autre part, le recours à des approches plus automatisées, basées sur le développement d'algorithmes. La Figure 11 résume cette typologie de la littérature que nous proposons.

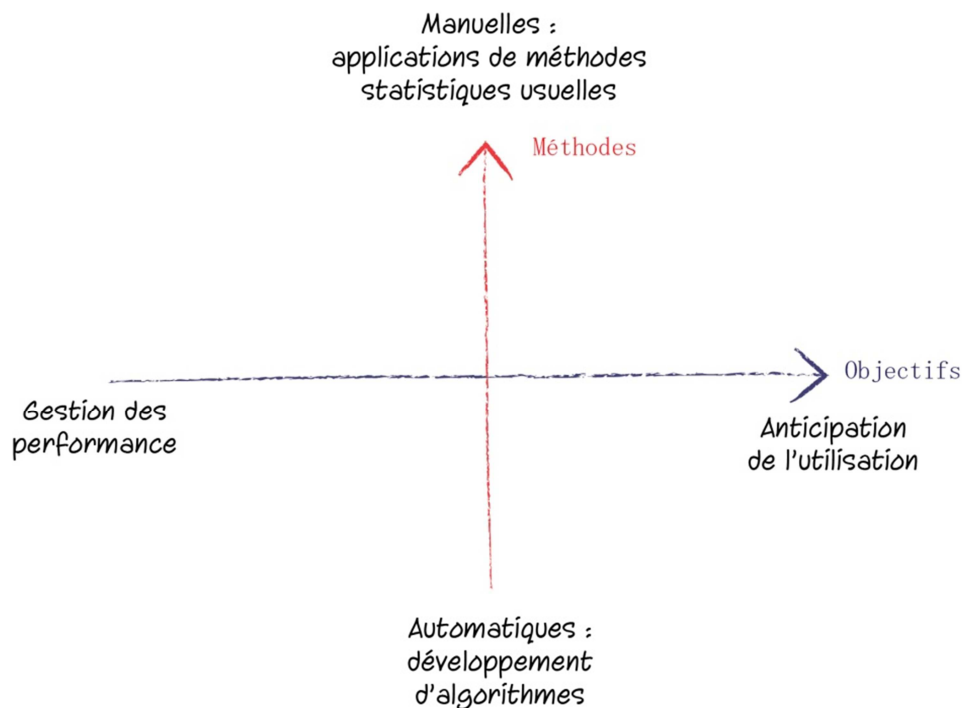


Figure 11. Catégorisation de la littérature relative à l'analyse des données : proposition.

Nous donnons quelques exemples d'applications ci-après :

- *Gestion des performances selon des méthodes « manuelles »* : Kelly (Kelly, 2005 ; Kelly et Zhang, 2006) applique un modèle de régression linéaire multiple pour expliquer la performance d'une application, en fonction du nombre d'appels à certaines transactions. Hendriksen (2011) propose une méthode complète pour analyser les temps de réponses applicatifs mesurés dans des systèmes distribués, grâce à l'emploi des statistiques exploratoires, puis pour construire des modèles de prévision qui mettent ces temps de réponses en relation avec l'activité des ressources de l'infrastructure des TI ;
- *Gestion des performances selon des méthodes automatisées* : Aguilera et al. (2003) développent des approches algorithmiques pour traiter des problèmes de performances en identifiant des chemins causaux au sein d'architectures des TI complexes. Jiang et al. (2007, 2008) proposent des algorithmes génériques pour identifier des invariants au sein de systèmes distribués. Ils pourraient être utilisés à des fins diverses, notamment la gestion des performances ;
- *Anticipation de l'utilisation selon des méthodes manuelles* : on distingue :
  - \* *L'usage de modèles de régression linéaire* : on prédit l'utilisation d'un processeur en fonction d'autres paramètres d'activités de l'infrastructure informatique (Jain, 1991 ; Gunther, 2007), d'activités applicatives (Ahituv et Igarria, 1988), voire d'activités métiers non précisées (Browning, 1994, qui parle d'un « *business factor* », sans préciser),
  - \* *L'extrapolation simple de séries temporelles (droites de tendance)* : Browning (1994), Gunther (2007), Allspaw (2008) ont recours à ces méthodes pour anticiper la taille d'une base de données, le nombre de processus occupés, l'utilisation d'un processeur, etc. ;
- *Anticipation de l'utilisation selon des méthodes automatiques* : on peut à nouveau citer Jiang et al. (2007, 2008). Dans Jiang et al. (2008), les auteurs appliquent notamment une méthode de recherches d'invariants pour analyser des mesures qui pourraient permettre d'anticiper l'utilisation d'un système informatique (nombre de requêtes SQL, utilisation processeur, ...), sans pour autant fournir un modèle prévisionnel *in fine*.

En conséquence, on peut dire que la littérature ne propose pas d'approche générale pour la constitution de modèles de *capacity planning* basés sur l'exploitation des données, mais plutôt différents outils, développés en fonction d'objectifs précis et de contextes locaux d'applications, qui conditionnent la forme et les variables du modèle. C'est une limitation que nous identifions : si le recours aux méthodes statistiques semble être adapté à la gestion d'un système des TI industriel complexe, comment guider la constitution d'un modèle ?

Par ailleurs, nous précisons cette limitation en soulignant que l'intégration de l'activité métier (enjeu de gestion n°2) semble peu abordée dans le choix des variables des modèles statistiques. D'une part, les modèles basés sur des extrapolations de séries temporelles excluent d'emblée la possibilité de faire quelque hypothèse que ce soit quant à un changement de configuration de l'activité métier. D'autre part, les modèles à notre

connaissance, mettant des variables en relations, n'intègrent pas de variables métiers clairement identifiées, qu'on soit dans le cadre d'approches manuelles ou automatiques. Les contextes locaux d'applications évoqués dans les exemples ci-dessus en attestent : on parle de requêtes SQL, d'activités applicatives, ou d'un énigmatique *business factor*, mais jamais de volume de production ou de nature d'activité industrielle en cours. Cette constatation semble revêtir une portée assez générale, puisqu'on peut lire dans une publication récente :

*« You may not be able to make a direct correlation between the business trends and your IT capacity, but you should at least look at whatever your organization publishes and see how it compares your component and service trends »* (Kloesterboer, 2011)

Qui plus est, même si l'on disposait d'un exemple local de constitution d'un modèle orienté métier, on pourrait chercher à avoir une approche plus générique, afin d'apporter une solution industrielle qui ne soit pas que strictement idiosyncratique. La limitation identifiée peut donc être complétée : si le recours aux méthodes statistiques semble être adapté à la gestion d'un système des TI industriel complexe, comment guider la constitution d'un modèle orienté métier ? Considérer l'activité de *capacity planning* (Q), en tenant compte des enjeux de gestion identifiés, nécessite d'apporter une réponse à cette question.

#### *1.2.4. Mise en œuvre de la détection de comportements anormaux dans le cadre de l'activité (K) Contrôle : apports et limitations*

##### *i. Généralités*

La nécessité de contrôler l'activité du système en fonctionnement est une évidence largement partagée dans la littérature (Jain, 1991 ; Browning, 1994 ; Domanski, 1999a, 1999b ; Cockroft et Walker, 2001 ; Fronheiser, 2004 ; Gunther, 2007 ; Lloyd et Rudd, 2007 ; Allspaw, 2008). Correctement mise en œuvre, elle améliore l'observation et la compréhension de l'activité de l'infrastructure des technologies de l'information. En ce sens, elle peut être complémentaire à (Q), en permettant d'identifier des comportements utiles à la constitution des modèles prospectifs de *capacity planning*. C'est cette application de l'activité de contrôle qui nous intéresse dans le cadre de cette thèse.

Remarquons toutefois que d'autres bénéfices peuvent en être attendus. Effectivement, elle doit également favoriser l'identification de dysfonctionnements avérés ou potentiels, tels que de mauvaises performances ou des risques de sous-capacités, et ainsi motiver des prises de décisions ultérieures. Ainsi, si (Q) est une activité définitivement proactive, (K) est à l'origine de comportements proactifs et réactifs<sup>16</sup>. Nous pouvons résumer ces bénéfices attendus par la Figure 12.

---

<sup>16</sup> Neubert (1997) définit ces deux types de comportements de la façon suivante :

*« Réactif : comportement par lequel un agent ou une organisation s'adapte à l'apparition d'un aléa ;*

*Proactif : comportement par lequel un agent, ou une organisation anticipe l'apparition d'aléas et apprend à mieux les gérer »*

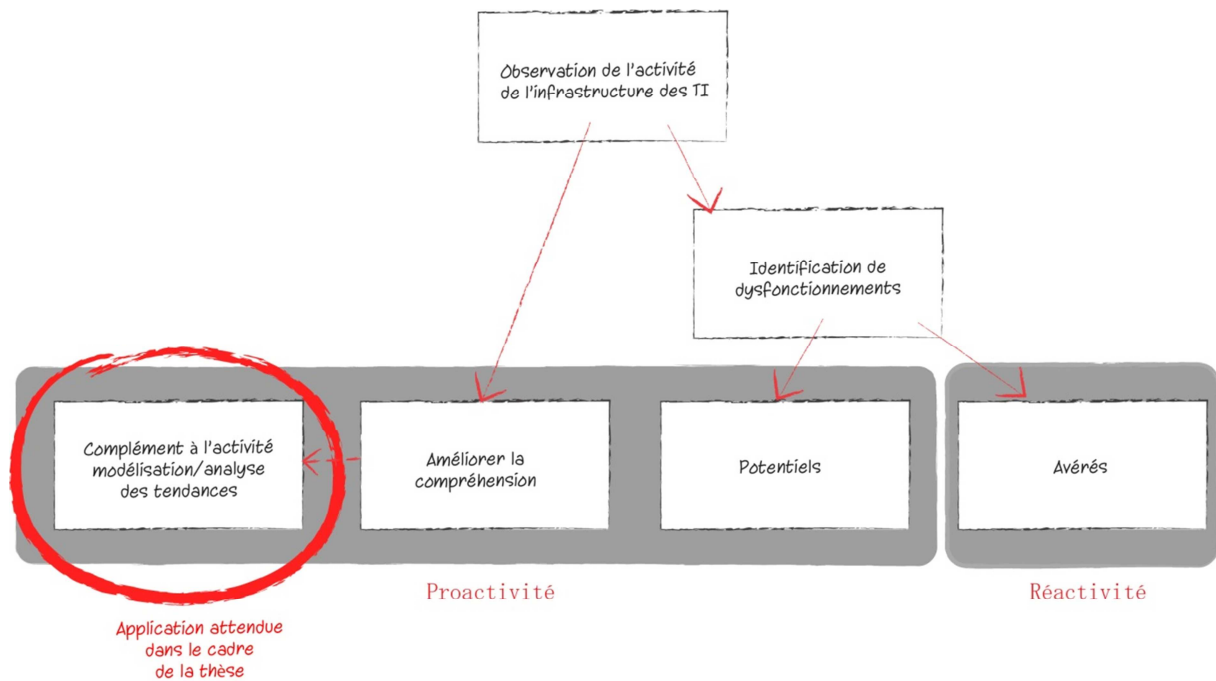


Figure 12. Bénéfices attendus de l'activité (K) : « contrôle ».

C'est donc une activité clé pour inscrire les activités de *capacity planning* dans une perspective d'amélioration continue du système. Comme l'indique Jain (1991) :

« *That which is monitored improves* »

La littérature indique que la complexité des systèmes modernes implique d'avoir recours à des procédures de contrôle automatisées (Fronheiser, 2004 ; Dugmore et Lacy, 2005). En effet, lorsqu'un système est caractérisé par des dizaines, voire des centaines de variables, on ne peut espérer une vérification manuelle de leurs fluctuations. L'objectif est alors de contrôler de façon périodique (par heure, par jour, etc.) un grand nombre de données, sans intervention humaine. L'approche la plus couramment rencontrée dans la littérature pour ces contrôles automatiques est la détection de dépassements de seuils, à rapprocher de l'activité ITIL (R) Gestion et contrôle des seuils. Ces contrôles sont surtout appliqués au niveau ressources (cf. les « *threshold* », « *ceiling* » ou « *upgrade indicator* », dont nous avons parlé précédemment), mais peuvent potentiellement aussi l'être aux autres niveaux du processus ITIL de gestion des capacités. Seulement, ce type de contrôle nécessite une compréhension *a priori* du comportement du système, pour pouvoir définir les valeurs de ces seuils empiriques. Toutefois, comment faire pour mettre sous contrôle des variables pour lesquelles l'on n'est pas capable d'émettre de telles hypothèses *a priori* ?

## ii. Détection de comportements anormaux : apport et limitations

Lorsque l'on ne dispose pas d'hypothèse *a priori* concernant le comportement d'une variable, une intelligence statistique devient nécessaire. Cette problématique particulière, qui consiste à identifier des comportements anormaux dans l'évolution d'une série chronologique, nous intéressera dans le cadre de cette thèse. L'objectif est d'être capable

de définir une méthode de contrôle générale et peu coûteuse, suffisamment générique pour s'appliquer à un grand nombre de données. Ces données sont organisées dans le temps, donc généralement non indépendantes : elles impliquent des contraintes de contrôles particulières. C'est un cas connu en maîtrise statistique des procédés. Sa procédure usuelle de mise en œuvre se décompose en deux étapes (Alwan et Roberts, 1988 ; Montgomery, 2005) :

- *Etape 1* : modélisation de l'évolution de chaque variable en tant que série chronologique ;
- *Etape 2* : modélisation des résidus du modèle obtenu lors de l'étape 1, par carte de contrôle sous hypothèse de normalité.

Diverses méthodes statistiques peuvent être appliquées à chacune des étapes et cette approche peut être définie de façon à être adaptée à un grand nombre de cas de figures.

Cette méthode de détection de comportements anormaux est connue dans le cadre de la gestion des technologies de l'information, avec différentes méthodes statistiques de modélisation des séries temporelles. Par exemple, Hellerstein et *al.* (1998) se basent sur une modélisation ARIMA, Brutlag (2000) et Leikis (2007) sur un lissage exponentiel Holt-Winters (nous reviendrons plus en détail sur ces méthodes dans la deuxième partie de la thèse). On propose ci-après (Figure 13) un exemple de contrôle, appliqué dans Brutlag (2000) pour le contrôle automatique de l'activité d'un routeur (variable contrôlée : nombres de bytes transmis par seconde).

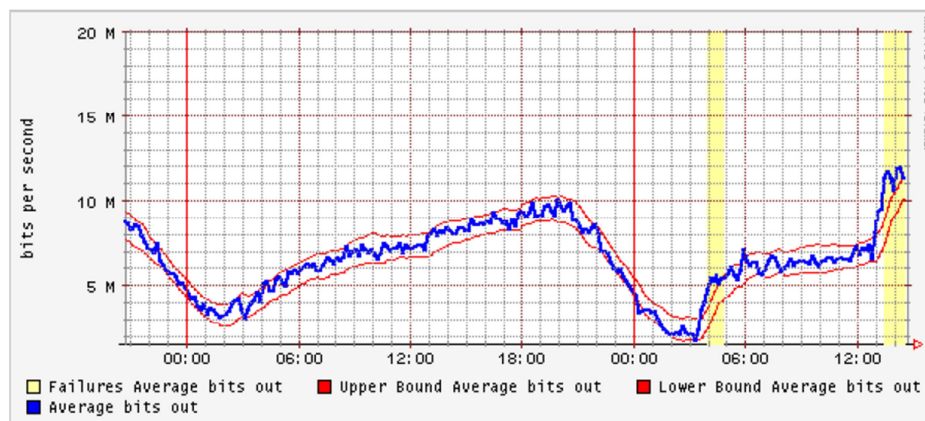


Figure 13. Détection automatique de comportements anormaux : application à l'activité d'un routeur dans Brutlag (2000)

Les lignes rouges représentent les limites de contrôle et les barres jaunes indiquent qu'elles ont été franchies. Lorsqu'elles sont franchies, c'est que la valeur observée n'est pas prévisible à partir de la modélisation de la série chronologique. On peut alors conclure à un comportement anormal. Les personnes en charge de la gestion de l'infrastructure technologique peuvent alors être averties. A elles ensuite d'analyser le phénomène, pour en dégager, si nécessaire, les bénéfices attendus présentés en Figure 12. Dans le cadre de nos recherches, ces bénéfices consistent à comprendre en quoi un hors-contrôle statistique peut être exploité pour améliorer la modélisation prospective du système.

Une limitation peut être soulevée. En effet, un certain nombre de contraintes opérationnelles (présence d'observations atypiques, nombre de variables contrôlées simultanément, etc.) peuvent nuire à l'efficacité des méthodes statistiques de contrôle automatique. En conséquence, il est nécessaire d'adapter la méthode statistique de contrôle aux spécificités locales d'application de la détection de comportements anormaux. C'est ce dont traite cette thèse : quelles peuvent être les méthodes statistiques appropriées à la définition d'une stratégie de contrôle automatique chez STMicroelectronics ? Répondre à cette question correspond à la prise en compte de l'enjeu de gestion n°1 (gérer un système complexe) dans le cadre de nos travaux. L'enjeu de gestion n°2 (intégrer l'activité métier) pourrait faire l'objet d'une réflexion approfondie concernant le choix des variables à mettre sous contrôle. Néanmoins, l'activité de contrôle étant considérée comme annexe dans le cadre de nos recherches, ce dernier point ne sera pas détaillé dans cette thèse.

## Section 2. Formulation de la problématique scientifique de thèse : vers un processus socio-statistique d'aide à la décision.

Par cette analyse de la littérature, nous avons montré deux enjeux majeurs des activités de *capacity planning* : gérer la complexité des systèmes technologiques contemporains et tenir compte de la dimension métier de leur activité. Il semble que le recours aux méthodes statistiques soit approprié à la gestion de la complexité (cette proposition sera discutée plus en détail ultérieurement). Pourtant, cette proposition n'est pas suffisante pour implémenter une solution opérationnelle. La spécification de solutions statistiques dans un contexte de *capacity planning* requiert de s'inscrire dans une démarche méthodologique qui reste entièrement à construire. Cette démarche rigoureuse inclut notamment un double cadrage qui semble nécessaire :

- *D'une part, pour décrire et rendre explicite le besoin décisionnel général* : que cherche-t-on à faire ? Dans le cadre de nos recherches, c'est ce questionnement qui permet d'identifier les deux besoins, principal et annexe, de modélisation prospective et de détection de comportements anormaux, identifiés chez STMicroelectronics et pour lesquels nous chercherons à développer des réponses quantitatives appropriées. Cette explicitation du besoin décisionnel général constitue une étape générique, pouvant être appliquée à tout autre contexte d'application ;
- *D'autre part, pour spécifier le besoin détaillé et sélectionner le type de réponse statistique à apporter* : pour chaque besoin, il convient de détailler le choix des méthodes statistiques, des variables, etc. Ce sont autant de questions auxquelles la littérature ne semble pas apporter de réponses explicites. Si la littérature est riche en propositions de méthodes et en exemples d'application, le praticien peut malgré tout rester démuné : quelle est la méthode appropriée à son besoin décisionnel précis ? Comment et pourquoi doit-il l'employer ?

La nécessité de ce cadrage méthodologique est naturel : face à la complexité et la diversité des situations de gestion contemporaines, les modèles à utiliser, les méthodes, les possibilités d'applications, etc. sont pléthoriques. Il est alors délicat de vouloir proposer des solutions statistiques « clés en main » directement applicables. Cette proposition apparaît clairement dans le cas de modélisation prospective, au cœur de nos recherches. Dans les années 1970, ère des *mainframes* monolithiques et centralisés, évoluant dans des environnements stables, l'hégémonie des modèles de performances était compréhensible. Or, aujourd'hui les solutions statistiques opérationnelles sont *ad hoc*, locales, constituées autour de problèmes de gestion et d'objectifs décisionnels ciblés : les solutions doivent être contextualisées. Le choix des méthodes et des variables ne peut se faire que face à un problème industriel clairement exprimé. Bien sûr, le recours à des solutions *ad hoc* n'empêche pas d'employer une approche générique, dans les pratiques de constitution de modèles utiles à la prise de décision. Simplement cette approche générique doit impliquer la prise en compte de la dimension locale et contextuelle du besoin décisionnel.

Nos travaux visent à développer une telle approche. Ils consistent à mettre au point une méthodologie rigoureuse et générique, pour la conception de solutions décisionnelles, ayant recours à des méthodes statistiques et adaptées à un contexte organisationnel

précis. Nous dénommons cette méthodologie « processus socio-statistique d'aide à la décision », pour les raisons suivantes :

- *Recours au terme « socio-statistique »* : selon nous, utiliser des statistiques dans un contexte industriel n'est pas qu'une problématique strictement technico-mathématique. Il faut également être capable de mobiliser les acteurs, afin de :
  - \* *Décrire la situation organisationnelle pour laquelle une aide à la décision quantitative est attendue* : les acteurs participent pleinement à la contextualisation de la future utilisation des méthodes statistiques, en exprimant eux-mêmes les préoccupations industrielles décisionnelles,
  - \* *Définir la solution quantitative attendue* : en impliquant les acteurs dans la constitution d'une aide à la décision statistique, on peut espérer inscrire l'utilisation des méthodes quantitatives et des modèles mathématiques dans leurs pratiques ;
- *Recours au terme « processus d'aide à la décision »* : la méthodologie que nous proposons ne doit pas être intuitive et informelle. Elle doit être scientifique, rigoureuse, et exprimer une volonté de rationalité. On attend pour cela un séquençement d'étapes et une structuration claire des éléments qui conduiront à apporter une réponse formalisée à un besoin décisionnel.

L'objectif de la thèse est de définir, de spécifier puis d'appliquer concrètement ce processus dans les limites de cas décisionnels opérationnels, déterminés pour répondre à des besoins de STMicroelectronics. Les contributions scientifiques de cette thèse concernent ainsi plusieurs dimensions : des contributions en terme de spécification d'une méthodologie ; des contributions en terme de modélisation prospective et de contrôle de l'activité d'un système informatique ; des contributions en terme d'applications décisionnelles.

Au final, au travers d'une application chez STMicroelectronics, nos travaux de recherches visent à montrer comment une approche structurée peut permettre :

- *D'analyser et comprendre des contextes décisionnels particuliers* : comme nous le développerons dans le chapitre 2, c'est cette analyse qui permet d'identifier le besoin « cœur de thèse » de modélisation prospective, et le besoin « annexe » de détection des comportements anormaux, puis de préciser quelles en sont les contraintes et les conditions de mise en œuvre, dans le contexte particulier de l'activité de STMicroelectronics ;
- *De développer des solutions innovantes* : on traite ici du questionnement central de cette thèse, qui s'intéresse à l'usage des statistiques pour améliorer les pratiques de gestion des capacités. Une fois un contexte particulier défini, on peut en effet envisager la réponse mathématique la plus adaptée, en choisissant les variables et les traitements quantitatifs qui peuvent y être appliquées. Les deux besoins décisionnels identifiés ne seront pas également traités :



- \* *Pour le besoin « cœur de thèse » de modélisation prospective* : une solution globale sera proposée. Quelles variables pour répondre au besoin ? Quelles méthodes statistiques ? Remarquons d'ores et déjà que les méthodes statistiques qui seront utilisées sont assez classiques. En revanche, l'originalité de notre proposition consiste à les inscrire dans le cadre d'une démarche de modélisation complète et structurée, permettant de répondre pleinement au problème décisionnel considéré,
  - \* *Pour le besoin « annexe » de détection des comportements anormaux* : considérant ce besoin comme complémentaire à la modélisation prospective (cf. état de l'art), notre proposition sera plus restreinte. On se limitera à développer une solution statistique, adaptée au contexte industriel de STMicroelectronics. Nous monterons comment ce contexte contraignant a imposé de développer une approche quantitative innovante ;
- *D'inscrire ces résultats dans une connaissance et un agir organisationnels* : en effet, selon nous, l'application obtenue par un processus structuré permet de formaliser les mécanismes de définition de solutions locales d'aide à la décision, tout en impliquant les acteurs de l'organisation. Ces mécanismes et leurs résultats sont donc des objets pouvant être discutés et partagés. En ce sens, nous pensons qu'ils s'inscrivent dans un processus de gestion des connaissances organisationnelles, actionnables dans les pratiques. Cette question sera traitée de façon complète pour le besoin « cœur de thèse » (formalisation et usage de la connaissance organisationnelle). Concernant le besoin « annexe », on ne considérera la détection de comportements anormaux que comme un complément, permettant de faire évoluer de la connaissance formalisée dans le cadre d'un modèle prospectif.

Les bénéfices attendus de nos travaux de recherches peuvent être résumés par la Figure 14.

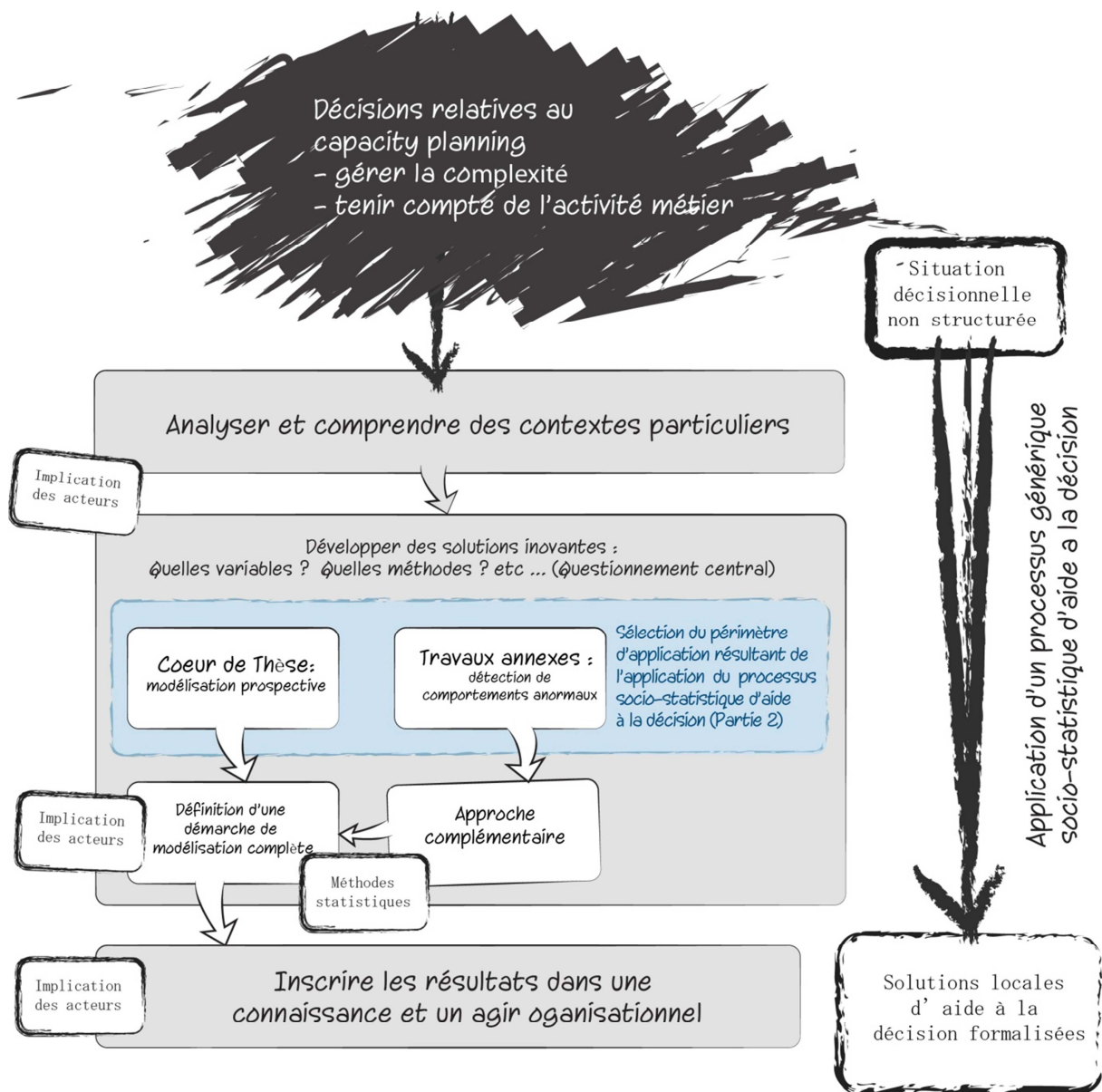


Figure 14. Apports attendus de la thèse.

Pour aboutir à ces solutions locales d'aide à la décision, formalisées et inscrites dans les pratiques organisationnelles, il faut développer un couplage entre méthodes scientifiques quantitatives et qualitatives. Ce couplage est le concept central de la définition du processus socio-statistique d'aide à la décision, pour favoriser les synergies entre expertises locales (les acteurs de l'organisation) et expertises statistique et méthodologique (l'analyste/le chercheur).

Ainsi, si le questionnement central de nos travaux concerne le développement d'aide à la décision sur la base de méthodes statistiques, ce questionnement doit inclure une réflexion plus générale sur la place de ces méthodes dans l'ensemble du cheminement du développement d'une aide à la décision. C'est ce que propose le processus socio-statistique d'aide à la décision. Sa construction et mise en action représente la problématique de cette thèse, telle que formulée ci-après.

## Problématique générale de thèse

*Comment définir, spécifier et appliquer un processus socio-statistique d'aide à la décision :*

- *Permettant de formaliser des connaissances organisationnelles au service de l'action organisationnelle ;*
- *Fondées sur une aide à la décision basée sur des méthodes statistiques (questionnement central), qui doivent être couplées à l'implication des acteurs de l'organisation, dans le cadre d'une démarche de modélisation structurée ;*
- *A partir d'un besoin décisionnel contextualisé et clairement exprimé par les acteurs de l'organisation ;*
- *De façon à contribuer à la gestion des capacités de l'infrastructure des TI, en tenant compte des enjeux de gestion identifiés ?*

Traiter cette problématique devrait permettre de faciliter l'implémentation *in situ* du processus ITIL de gestion des capacités et d'apporter une contribution au bon fonctionnement et à la compétitivité de STMicroelectronics. Cette implémentation sera réalisée sur les besoins identifiés comme étant les plus importants chez STMicroelectronics : la modélisation prospective (cœur de thèse) et la détection de comportements anormaux (travaux annexes).

### Conclusion du chapitre 1

Ce chapitre a permis de situer la place de l'architecture des technologies de l'information dans le cadre l'activité d'une organisation. Nous avons montré qu'elle est un contributeur à part entière à son bon fonctionnement et à sa performance. Deux enjeux de gestion doivent être surmontés : être capable de gérer un système complexe et de tenir compte pleinement de l'activité métier de l'organisation. Pour ce faire, un ensemble de bonnes pratiques peuvent être utilisées, dont le couple ISO 20000/ITIL.

Ce couple offre un certain nombre de recommandations utiles à la gestion des capacités. Il introduit notamment les deux activités traitées dans cette thèse : modélisation et analyse de tendances (cœur des recherches) et détection de comportements anormaux (travaux complémentaires). Nous avons montré qu'elles posent un certain nombre de limitations d'applications. Nous avons expliqué que, de façon générale, la mise en œuvre de solutions quantitatives de *capacity planning* ne peut être envisagée par la seule considération de méthodes statistiques décontextualisées, ou par l'étude de cas issus d'autres organisations. Au contraire, une solution locale et contextuelle, couplant expertises statistiques et organisationnelles, rendue possible par la combinaison de méthodes scientifiques quantitatives et qualitatives, doit être favorisée.

C'est pourquoi nous avons suggéré la nécessité de définir et spécifier un processus générique qualifié de « processus socio-statistique d'aide à la décision ». Son application

doit permettre de définir les besoins décisionnels à traiter et de produire des résultats pouvant être inscrits dans la connaissance et les bonnes pratiques organisationnelles.

Le chapitre suivant va décrire les fondements théoriques qui sous-tendent la définition de ce processus socio-statistique d'aide à la décision. Ces fondements permettront alors de le spécifier plus en détail.

**Chapitre 2. Définition et spécification du processus socio-statistique  
d'aide à la décision.**

## Introduction

Ce chapitre donne l'ensemble des éléments de définition et de spécification du processus socio-statistique d'aide à la décision. Son plan est résumé dans la Figure 15.

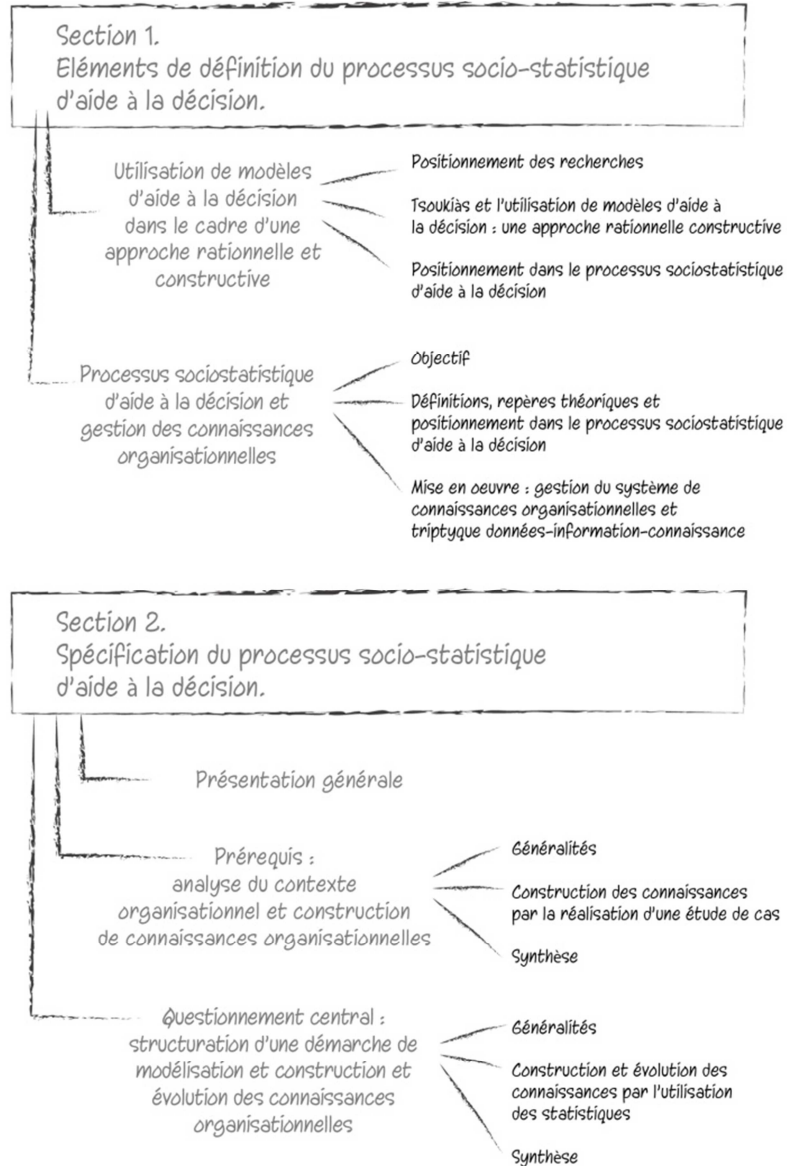


Figure 15. Plan du chapitre 2.

Tout d'abord, un ensemble d'éléments de définition théoriques sont proposés. Pour cela, la section 1 précise le positionnement des recherches menées : on parle d'utilisation de modèles dans le cadre d'une aide à la décision rationnelle et constructive. Il nous permet de mobiliser un cadre conceptuel général, pour lequel nous nous référerons notamment aux travaux d'Alexis Tsoukiàs. Nous expliquons quelle est la place de ce cadre général dans le processus socio-statistique d'aide à la décision.

Par ailleurs, nous indiquons que le processus socio-statistique d'aide à la décision nécessite également une réflexion sur ses implications cognitives. En effet, le processus implique un certain nombre d'artefacts (propositions de Tsoukiàs et utilisation de

modèles statistiques) : il convient de s'interroger sur leurs conditions de constitution. Un certain nombre d'indications théoriques concernant la gestion des connaissances organisationnelles sont précisées pour cela, ainsi que leurs relations au processus socio-statistique d'aide à la décision. Enfin, nous présentons deux concepts théoriques complémentaires, utiles à la mise en œuvre de la gestion des connaissances organisationnelles : la gestion des systèmes de connaissances, proposée par Jean-Louis Ermine, ainsi que le triptyque données-informations-connaissances.

Sur ces bases théoriques, la section 2 développe notre proposition d'un processus socio-statistique d'aide à la décision. Après une présentation générale, ses deux principales étapes, visant à agir sur les connaissances organisationnelles, font l'objet d'une discussion plus détaillée : un certain nombre de considérations méthodologiques sont fournies, pour l'analyse du contexte organisationnel et pour la structuration d'une démarche de modélisation. Pour la première étape (travail préliminaire de contextualisation), on présentera le recours à la constitution d'une étude de cas ; pour la seconde (utilisation de modèles statistiques), on s'interrogera sur les conditions d'utilisation des méthodes statistiques pour la constitution de connaissances organisationnelles, la prise de décision, et la modélisation d'un système complexe.

## Section 1. Eléments de définition du processus socio-statistique d'aide à la décision.

### 1.1. Utilisation de modèles d'aide à la décision dans le cadre d'une approche rationnelle et constructive

#### 1.1.1. Positionnement des recherches

##### i. La décision dans les organisations : diversité des approches

L'étude des processus de prise de décision dans les organisations est un champ de recherche vaste, recouvrant de nombreuses dimensions et intérêts académiques. De façon générale, décider, c'est s'engager à agir d'une certaine façon, selon des choix distincts et identifiables (Mintzberg et *al.*, 1976 ; Langley et *al.*, 1995 ; Sven et Björn, 2011). Le processus de prise de décision est défini par un ensemble d'actions et de facteurs, commençant par un élément déclencheur et finissant par un engagement à agir spécifique (Mintzberg et *al.*, 1976 ; Nutt, 1993, 1998 ; Sven et Björn, 2011). Cet ensemble peut être analysé selon des perspectives très différentes. Miller et *al.* (*in* Salaman, 2002) proposent deux composantes permettant de classer la littérature académique :

- *Les modes d'action de la prise de décision* : la prise de décision peut être envisagée selon un *continuum* allant du processus cohérent et structuré au processus chaotique ;
- *Les modes d'intérêt de la prise de décision* : la prise de décision peut être envisagée selon un *continuum* allant du processus orienté « résolution de problème » au processus de négociation politique.

Langley et *al.* (1995) distinguent quant à eux :

- *Les approches rationnelles* : la prise de décision est considérée comme un processus cognitif, pouvant être décomposé en étapes successives ;
- *Les approches anarchiques* : la prise de décision est considérée selon ses dimensions collectives, impliquant négociations et interactions sociales.

Ils suggèrent également une troisième composante, impliquant deux conceptions de la prise de décision :

- *Une conception dite « étroite » de la prise de décision* : la prise de décision est vue comme un acte réifié, déshumanisé et isolé ;
- *Une conception dite « ouverte » de la prise de décision* : la prise de décision est vue comme un acte d'inspiration, de convergence et d'interconnexions d'idées.

Ces classifications permettent de positionner l'abondante littérature relative à la prise de décision. Par exemple, le modèle du « *garbage can* » (voir Cohen et *al.* *in* March, 1991) est une description populaire d'un mode d'action chaotique, où les organisations (ou les situations décisionnelles) sont vues comme des anarchies organisées faites de



préférences incertaines, de procédures tâtonnantes et de participation des acteurs fluctuantes. Crozier et Friedberg (2004) ou March (1991) décrivent quant à eux les mécanismes de négociation politique, où la prise de décision est conditionnée par des jeux politiques et des conflits de pouvoir.

Ces classifications nous permettent également de situer les travaux présentés dans cette thèse car évidemment, toutes les composantes suggérées ne peuvent être traitées à la fois. Dans le cadre d'un soutien aux activités quantitatives que représente le *capacity planning*, par la définition d'un processus socio-statistique d'aide à la décision structuré, ce sont clairement les aspects rationnels qui nous intéressent. De même, la conception « ouverte » est hors de nos propos. La Figure 16 résume ce positionnement au sein des différentes approches possibles.

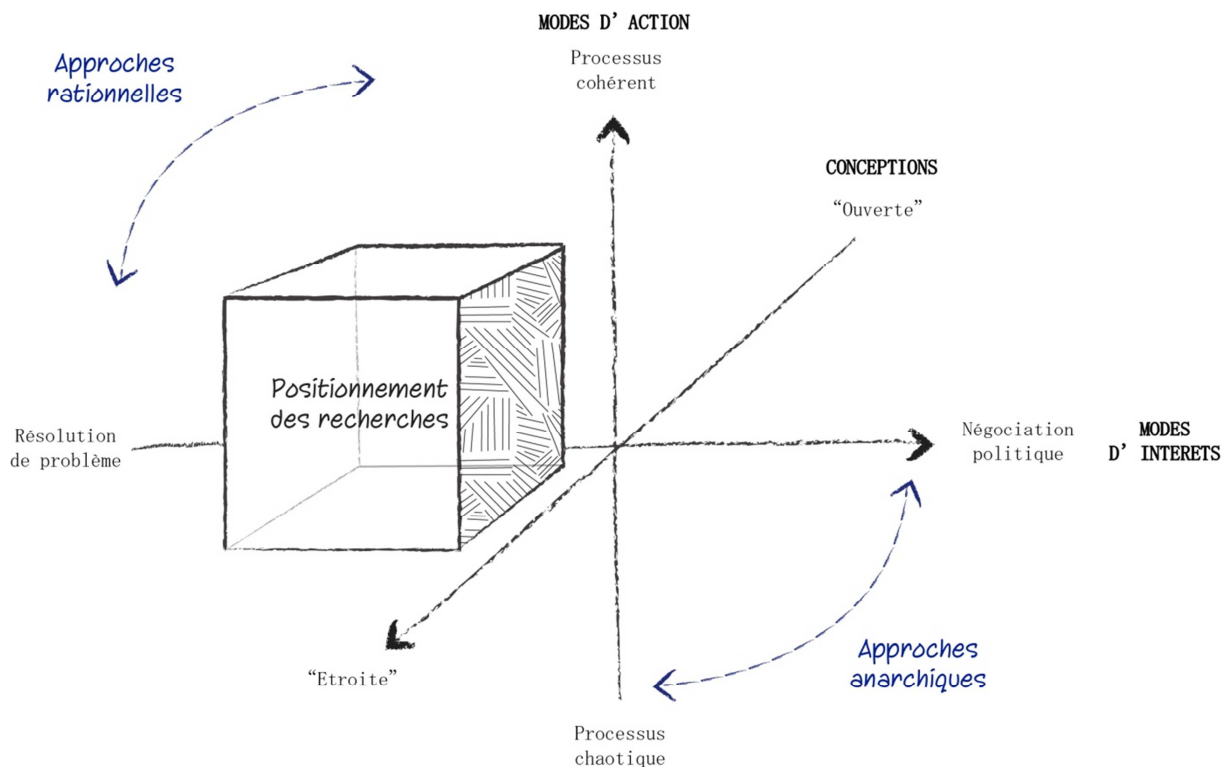


Figure 16. Le processus de prise de décision : approches et positionnement des recherches.

Nous sommes tout à fait conscient que des phénomènes de conflits d'intérêts et de jeux de pouvoir, de pur hasard et d'irrationalité, d'émulation sociale et d'inspiration individuelle, peuvent avoir lieu à chaque prise de décision. Toutefois, ces phénomènes dépassent le cadre de nos préoccupations. Selon notre positionnement, quelle est alors la place des modèles dans la prise de décision ?

ii. Recherche menée : modèles d'aide à la décision et approche rationnelle constructive

Appuyer une décision sur l'utilisation et la construction rationnelles de modèles quantitatifs, basés sur l'exploitation de données, est qualifié de « prise de décision avec utilisation de modèles d'aide à la décision » par Azoulay et Ponthus (1974) (Figure 17).

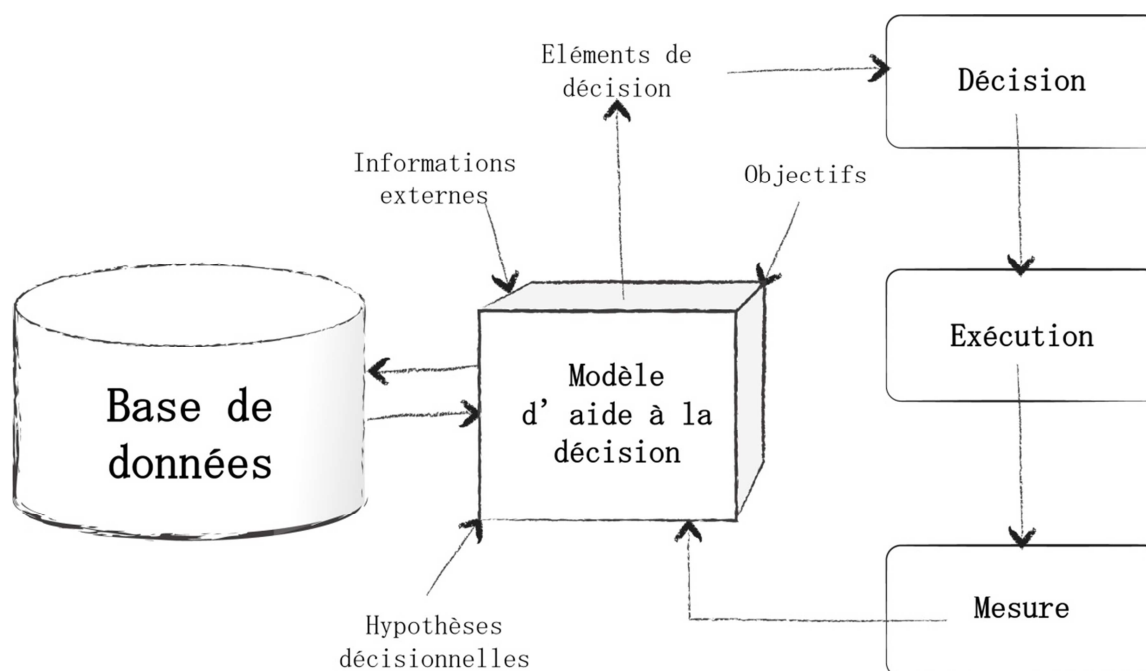


Figure 17. Prise de décision avec utilisation de modèles d'aide à la décision (Azoulay et Ponthus, 1974).

Selon les auteurs, une organisation appuie ses décisions sur des modèles qui devraient représenter d'aussi près que possible la réalité, en s'appuyant sur les données disponibles. Ces modèles fournissent des résultats numériques, dont l'interprétation se fait par comparaisons à des objectifs et des hypothèses décisionnelles. Ces résultats fournissent des éléments de décision qui permettent de décider, puis d'exécuter la décision et de mesurer son application. Dans ce cadre général peuvent s'inscrire la plupart des modèles rationnels de la prise de décision, dont le fondamental et très célèbre modèle « intelligence-modélisation-choix » de Simon (Le Moigne, 1974 ; Carluer et Richard, 2002). Ce cadre reste d'actualité, mais doit être nuancé. On cherche désormais moins à produire une :

« « copie » [du réel] ou modèle représentant d'aussi près que possible la réalité [...] » (Azoulay et Ponthus, 1974)

qu'un :

« schéma qui, pour un champ de questions, est pris comme représentation d'une classe de phénomènes plus ou moins habilement dégagés de leur contexte par un observateur pour servir de support à l'investigation et/ou à la communication » (Roy, 1992a<sup>17</sup>)

Ainsi, le modèle reste un élément important, mais, conformément à la proposition faite précédemment, il s'intègre dans un contexte et dans une activité sociale<sup>18</sup> d'investigation et de communication. En conséquence, contextualisation et dimension humaine font

<sup>17</sup> Roy (1993) pour une version en anglais.

<sup>18</sup> Notons que cette prise en compte de la dimension sociale n'est pas en contradiction avec la typologie des courants de recherche présentée. On peut avoir une approche rationnelle, structurée et orientée résolution de problème ainsi qu'une vision réifiée de la décision, tout en s'intéressant aux acteurs de l'organisation. L'anarchie et l'ouverture n'est pas le propre de la dimension humaine.

partie intégrante de l'aide à la décision. Roy (1992a) propose ainsi la définition suivante de l'aide à la décision :

*« Activité de celui qui, par des voies dites scientifiques, aide à obtenir des éléments de réponses à des questions que se posent des acteurs impliqués dans un processus de décision, éléments concourants à éclairer la décision en vue de favoriser un comportement des acteurs de nature à accroître la cohérence entre l'évolution du processus d'une part, les objectifs et/ou les systèmes de valeurs au service desquels ces acteurs se trouvent placés d'autre part »*

A côté du modèle (la voie scientifique), la question des objectifs et des valeurs des acteurs est posée. Ostanello et Tsoukiàs (1993) rapportent la citation suivante de Jacquet-Lagrèze :

*« Une des premières formes d'aide à la décision réside, pensons-nous, dans l'aide à l'explicitation des représentations de l'action collective. Si seulement les acteurs en présence avaient une meilleure intelligence de la situation [...] »*

Nos travaux se placent dans cette perspective de la prise de décision avec utilisation de modèles d'aide à la décision, que certains auteurs qualifient de méthodologies « *soft* » de recherche opérationnelle/aide à la décision (David, 1999). Telle que nous venons de la décrire, l'aide à la décision sur la base de modèles implique donc (Roy, 1992a) :

- « *des voies dites scientifiques* » : l'aide à la décision a recours à une approche scientifique impliquant des méthodes suffisamment rigoureuses pour guider la prise de décision, par l'élaboration de concepts, de modèles, de procédures et de résultats. Cette composante constitue le cœur de nos recherches : quels modèles quantitatifs pour décider ? ;
- « *l'explicitation des représentations de l'action collective* » : la vertu première d'un modèle est l'aide à la décision qu'il apporte, c'est-à-dire son adéquation à un besoin décisionnel bien identifié. Ce besoin n'est pas toujours donné de soi : il doit bien souvent être construit dans le cadre d'un processus d'aide à la décision. Cette composante est un prérequis indispensable au cœur de nos recherches et a donc été intégrée à nos travaux.

La nécessité de coupler de ces deux composantes, formelle et organisationnelle, est très souvent soulignée dans la littérature (Roy, 1992b, 2006 ; Le Moigne *in* Alcaras et *al.*, 2004 ; Tsoukiàs, *in* Bouyssou et *al.*, 2006a<sup>19</sup>, 2006b ; Tsoukiàs, 2007). Tsoukiàs (*in* Bouyssou et *al.*, 2006a ; 2006b ; Tsoukiàs, 2007) propose un cadre théorique permettant de les structurer et les formaliser.

---

<sup>19</sup> Tsoukiàs (2008) pour une version en anglais.

### 1.1.2. Tsoukiàs et l'utilisation de modèles d'aide à la décision : une approche rationnelle constructive

#### i. Présentation générale des travaux de Tsoukiàs

Tsoukiàs suggère une prise de décision basée sur un modèle d'évaluation, qui permet d'en guider le processus. Ce modèle est en lien au problème, aux préférences, d'un décideur. Ce lien implique l'utilisation d'un modèle de rationalité, permettant de traduire une information informelle en une représentation formelle du contexte décisionnel. Opérer ce passage peut se fonder sur différentes approches. D'une part, le problème et les préférences du décideur peuvent préexister à la conduite de l'aide à la décision. Dans ce cas, on peut s'appuyer sur les approches suivantes de l'aide à la décision :

- *Normative* : c'est l'approche « classique » de la théorie de la décision. On fait appel à une rationalité exogène (des modèles connus, extérieurs à la situation décisionnelle considérée), qui va fournir une solution à laquelle le décideur doit adapter ses choix ;
- *Descriptive* : on a également recours à une rationalité exogène, mais en se basant sur l'observation *in situ* du décideur. L'aide à la décision est donc adaptée à un cas particulier ;
- *Prescriptive* : dans ce cas, il n'y a plus mobilisation de modèles externes. On cherche à identifier le modèle de rationalité endogène du décideur pour l'aider à décider.

D'autre part, il est possible que le modèle de rationalité du décideur ne puisse être défini aussi directement, parce que le problème n'est pas clairement exprimé. La dimension contextuelle entre alors en jeu. Sa prise en compte correspond à une quatrième approche de l'aide à la décision. L'objectif est alors de construire progressivement le problème et les préférences du décideur, afin de proposer le modèle d'évaluation approprié. Ce modèle est donc le résultat d'un processus d'apprentissage, impliquant le décideur mais aussi l'analyste ou le chercheur qui va l'aider à aboutir à une aide à la décision. Le décideur apprend à comprendre son problème d'un point de vue abstrait et formel. L'analyste/le chercheur apprend à comprendre le problème du décideur et le point de vue de celui-ci. Tsoukiàs illustre cette approche par la célèbre formule de Bachelard :

*« Rien n'est donné. Tout est construit »*

Cette construction implique de concevoir la décision comme un processus cognitif distribué<sup>20</sup>. La rationalité du décideur y est locale et procédurale<sup>21</sup> et la conduite du

---

<sup>20</sup> Tsoukiàs *in* Bouyssou et al. (2006b) suggère des exemples simples, et Ostanello et Tsoukiàs (1993), Paschetta et Tsoukiàs (2000), Stamelos et al. (2000), Stamelos et Tsoukiàs (2003) proposent des applications dans le cadre de travaux scientifiques concernant des problématiques décisionnelles d'évaluation de logiciels.

<sup>21</sup> Cf. la célèbre distinction de Simon (Simon, 1992 ; Simon, 2004), qui définit : 1) la rationalité substantive comme la détermination de la bonne façon d'agir ; 2) la rationalité procédurale comme la recherche d'un mode de détermination d'une bonne façon d'agir, dans un contexte de capacités computationnelles et cognitives limitées. Il nous semble que, par ces travaux, Simon représente réellement un auteur charnière en ce qui concerne l'étude des mécanismes décisionnels, qui permet d'articuler sciences de l'ingénieur et sciences de gestion. Introduisant cette

processus d'aide à la décision fait partie intégrante de l'aide à la décision. Ce processus est un acte social et partagé : plusieurs participants essaient de produire une représentation commune de la réalité. Cette représentation est cristallisée par la production d'artefacts. Tsoukiàs en identifie quatre. Nous suggérons que les deux premiers correspondent à la contextualisation de l'aide à la décision, les deux autres au développement d'une aide à la décision basée sur l'utilisation de modèles. Nous les détaillons ci-après.

ii. *Contextualisation de l'aide à la décision : représentation de la situation problématique et formulation du problème*

Représentation de la situation problématique

Représenter la situation problématique est le point d'ancrage du processus d'aide à la décision et de l'intelligence de la situation. Il permet de décrire, voire d'expliquer la situation, et ainsi de cerner ce que l'on cherche à faire. Il permet de comprendre la position du décideur par rapport au processus de décision et le rôle de l'analyse/du chercheur. Tsoukiàs définit formellement cet artefact selon le triplet :

$$P = \langle A, O, S \rangle$$

avec :

- *A* : acteurs impliqués dans le processus de décision ;
- *O* : objets et enjeux apportés par chacun des acteurs (problèmes, intérêts, opportunités...) ;
- *S* : engagements et ressources apportés par chacun des acteurs pour chacun des objets le concernant (connaissances, ressources financières, pouvoirs décisionnels, etc.)<sup>22</sup>.

Cet artefact n'est pas figé, il est susceptible d'évoluer au fur et à mesure du processus de décision.

Formulation du problème de décision

A partir de la représentation de la situation problématique, l'analyste va pouvoir proposer une formulation du problème (ou des problèmes, si nécessaire). La construction de cet artefact fait intervenir le modèle de rationalité du décideur. Il s'agit de réduire la réalité du processus de décision par l'utilisation d'un langage formel et abstrait. La préoccupation du décideur devient ainsi un problème formel sur lequel une méthode d'aide à la décision peut être appliquée. Cette formulation est importante car c'est une première orientation de la décision finale.

---

dimension que David (1999) qualifie de « *soft* », il permet effectivement de mieux situer l'usage de méthodes formelles et mathématiques dans un contexte organisationnel et humain.

<sup>22</sup> Dans Ostanello et Tsoukias (1993), la typologie suivante est proposée : ressources quantifiables, non-quantifiables et comportementales passive ou active.

Cet artefact est représenté par le triplet :

$$\tau = \langle \mathcal{A}, \mathcal{V}, \Pi \rangle$$

avec :

- $\mathcal{A}$  : alternatives (actions) potentielles à considérer dans le cadre de  $P$  ;
- $\mathcal{V}$  : points de vue sous lesquels il est envisagé d'observer, d'analyser, de comparer les actions potentielles ;
- $\Pi$  : problématique décisionnelle, périmètre d'évaluation (choix, tri, description, etc.) envisagé sur l'ensemble  $\mathcal{A}$ .

Lorsque cet artefact est approuvé par le décideur, la situation décisionnelle est clarifiée. On passe d'une compréhension floue à une formalisation précise d'un problème, sur lequel on va pouvoir appliquer une méthode formelle d'aide à la décision.

*iii. Aide à la décision basée sur l'utilisation de modèles : modèle d'évaluation et recommandation finale*

#### Modèle d'évaluation

La formulation du problème permet de construire le modèle d'évaluation. On retrouve les modèles « classique » d'aide à la décision, définis selon un large éventail de méthodes (recherche opérationnelle<sup>23</sup>, théorie des jeux, cybernétique, statistiques, intelligence artificielle...). Dans les approches normatives d'aide à la décision, les degrés de libertés sont faibles pour la construction d'un modèle en réponse à un problème donné. Les autres approches laissent plus de latitude pour définir les éléments et les paramètres du modèle. Le modèle d'évaluation est défini par le n-uplet :

$$M = \langle \mathcal{A}, D, E, H, U, R \rangle$$

avec :

- $\mathcal{A}$  : alternatives sur lesquelles va s'appliquer le modèle d'évaluation ;
- $D$  : dimensions sous lesquelles les éléments de  $\mathcal{A}$  sont considérés ;
- $E$  : échelles associées à  $D$  (métriques) ;
- $H$  : critères d'évaluation de  $\mathcal{A}$ , pour prendre en compte les préférences du décideur ;
- $U$  : incertitude associée à l'information disponible (relative à  $D$  et/ou  $H$ ) ;
- $R$  : procédures d'agrégation des préférences, des mesures, des incertitudes, qui permettent d'obtenir une information synthétique sur les éléments de  $\mathcal{A}$ .

---

<sup>23</sup> Nous en profitons pour préciser ici qu'en ce sens, le recours aux modèles de performance basés sur la théorie des files d'attente n'est pas incompatible avec une approche constructive de l'aide à la décision. Ces modèles ne sont tout simplement pas adaptés aux contextes décisionnels dont traite cette thèse.

Une grande partie des modèles d'aide à la décision sont représentables via ce n-uplet. On comprend donc que les différences entre les approches d'aide à la décision ne se situent pas au niveau des méthodes employées, mais de la manière dont on va les ancrer dans les activités de l'organisation.

### Recommandation finale

Le modèle d'évaluation va produire un résultat  $\Phi$ . La recommandation finale est la traduction de  $\Phi$  en langage courant. Cela correspond à un retour à la réalité organisationnelle de la prise de décision. Un certain nombre de précautions doivent être prises lors de la construction de cet artefact (validation du modèle, légitimité de la recommandation...).

#### *1.1.3. Positionnement dans le processus socio-statistique d'aide à la décision*

Le cadre théorique offert par Tsoukiàs, visant à impliquer acteurs et méthodes formelles pour développer une aide à la décision contextualisée, est conforme aux objectifs du processus socio-statistique que l'on souhaite définir. En résumé, la conduite d'un processus d'aide à la décision, ayant recours à une telle approche constructive, repose sur la production de quatre artefacts<sup>24</sup> : la situation problématique, la formulation du problème, le modèle d'évaluation et la recommandation finale.

Evidemment, une telle « *check-list* » peut parfois masquer la complexité réelle de certains processus d'aide à la décision. Les qualités didactique et structurante de ce cadre conceptuel en font néanmoins un outil fort utile. C'est pourquoi, nous suggérons qu'un processus d'aide à la décision, permettant la définition successive des artefacts définis par Tsoukiàs, doit être favorisé pour améliorer les pratiques de *capacity planning* évoquées précédemment.

Le cœur de notre travail de recherche intègre la construction d'un artefact additionnel, matérialisé par un modèle statistique. Nous le proposons en complément aux artefacts suggérés par Tsoukiàs. Ce modèle statistique pourra être mis en usage dans le cadre d'un modèle d'évaluation quantitatif, qui permettra de proposer une recommandation finale. On rappelle également que la structuration de la situation problématique et de la formulation du problème sont des étapes de contextualisation préliminaires, indispensables à la définition d'une solution *ad hoc*. Ce positionnement du cadre de Tsoukiàs et au sein de nos travaux est résumé dans le Tableau 1.

---

<sup>24</sup> Selon Tsoukiàs, tous les artefacts ne sont pas nécessairement produits pendant le processus d'aide à la décision : une bonne analyse de la composante contextualisation suffit parfois à prendre une décision.

<b>Etape de l'aide à la décision</b>	<b>Apport attendu</b>	<b>Artefact</b>	<b>Positionnement dans les travaux de recherche</b>
<i>Contextualisation</i>	Analyser et comprendre des contextes particuliers	Situation problématique ( <i>cadre de Tsoukiàs</i> )	Prérequis
		Formulation du problème ( <i>cadre de Tsoukiàs</i> )	Prérequis
<i>Utilisation de modèles</i>	Développer des solutions innovantes et inscrire les résultats dans une connaissance et un agir organisationnel	Modèle statistique ( <i>artefact additionnel</i> )	Questionnement central
		Modèle d'évaluation ( <i>cadre de Tsoukiàs</i> )	Usage
		Recommandation finale ( <i>cadre de Tsoukiàs</i> )	Usage

Tableau 1. Place des artefacts de l'approche constructive de l'aide à la décision dans les travaux de recherche.

Il nous semble pertinent d'intégrer ce cadre de Tsoukiàs dans la définition du processus socio-statistique d'aide à la décision : d'une part, il implique un séquençement d'étapes matérialisé par la constitution d'artefacts (« processus d'aide à la décision ») ; d'autre part, il contribue à impliquer les acteurs de l'organisation, pour aboutir à un modèle formel et son utilisation (« socio-statistique »). Par ailleurs, il vise à contextualiser l'aide à la décision, permettant ainsi d'aboutir à une recommandation locale.

En revanche, Tsoukiàs donne peu d'indications sur la constitution des artefacts. Nous proposons un cadre théorique apportant des éléments méthodologiques pour ce faire, basé sur la gestion des connaissances organisationnelles.

## **1.2. Processus socio-statistique d'aide à la décision et gestion des connaissances organisationnelles**

### *1.2.1. Objectif*

Avant toute chose, citons Ermine (1989) :

*« pour rendre toute sa dimension à l'analyse de la connaissance, il faut recourir à une pluridisciplinarité qui peut donner le vertige. »*

Face à ce constat toujours d'actualité, il convient de rester modeste dans le cadre des travaux présentés ici. Le cœur de nos recherches consiste à utiliser des méthodes statistiques, dans le but de répondre à une problématique industrielle. Toutefois, nous cherchons à nous appuyer sur certains concepts, issus du domaine de la gestion des connaissances organisationnelles. Ils pourront nous aider à mieux structurer la constitution des artefacts de l'aide à la décision selon Tsoukiàs, ainsi que le recours à la modélisation statistique.



En premier lieu, avant d'envisager l'utilisation de modèles quantitatifs, il faut construire une intelligence de la situation (contextualisation). Une approche structurée est nécessaire pour définir avec rigueur cette étape. Les concepts de la gestion des connaissances organisationnelles peuvent nous aider pour cela.

Ensuite, à partir de cette contextualisation, nous souhaitons apporter un éclairage original sur l'usage des bases de données pour la gestion des connaissances, dans le but de constituer *in fine* des modèles pour l'aide à la décision (étape d'utilisation de modèle de l'aide à la décision). Il est en effet connu que ces bases de données sont des sources non triviales de connaissances latentes, devant être mises en action à des fins décisionnelles (Fouet, 1997 ; Lebraty, 1998). Citons pour exemple les approches dites « *knowledge discovery in database* » (Fayyad et al., 1996, 1996b ; Fayyad et Stolorz, 1997 ; Cios et al., 2007), qui proposent un cadre général d'application des méthodes de *data-mining*, afin de les orienter explicitement vers la constitution de connaissances. Néanmoins, dans le cadre de ces approches, la notion de connaissance reste essentiellement définie comme une régularité dans les données (Cios et al., 2007). Notre objectif est de porter un regard moins technique sur le problème, pour pourvoir parler de création de connaissances organisationnelles. Par ailleurs, il nous semble également important d'envisager la possible évolution dans le temps de ces connaissances : la création de connaissances nouvelles est une question qu'il est pertinent de compléter, par des mécanismes visant à en assurer la validité future.

### *1.2.2. Définitions, repères théoriques et positionnement dans le processus socio-statistique d'aide à la décision*

#### *i. Artefacts et connaissances organisationnelles*

Les artefacts sont des objets conçus et construits par l'homme<sup>25</sup>, fonctions de l'environnement dans lequel ils vivent (Perrin *in* Jacot, 1994 ; Simon, 2004). Ils sont adaptés à des objectifs fixés. Ils peuvent être de différentes natures : objets techniques, organisations sociales... Dans le cadre de nos travaux, ils correspondent à des représentations des différentes étapes du processus d'aide à la décision : artefacts du cadre de Tsoukiàs et modèles statistiques. Leur création relève d'un processus cognitif. En effet, construire un artefact nécessite de construire une représentation de la réalité et cette représentation nécessite de disposer de connaissances sur la réalité, qui doivent également être construites si non disponibles. Comme l'indique Perrin, « approche artefact » et cognition sont liés.

Cette notion de connaissance doit être précisée. Il est communément considéré qu'une connaissance est une croyance vraie et justifiée (Le Ny *in* Tiberghien, 2002 ; Nonaka, 1994 ; Nonaka et von Krogh, 2009).

---

<sup>25</sup> Le terme artefact est dérivé de la racine latine *facere*, qui signifie « faire » (Segal, 1990).

Dans le cadre de nos travaux, on s'intéresse plus spécifiquement à la question de la connaissance organisationnelle, que Kim et *al.* (2003) définissent comme :

*« a set of fact, know-hows, patterns and models validated to be useful for solving organizational problems »*

Tsoukas et Vladimirou (2001) proposent quant à eux la définition plus générale suivante :

*« the capability members of an organization have developed to draw distinctions in the process of carrying out their works, in particular concrete contexts, by enacting sets of generalizations whose application depends on historically evolved collective understandings »*

On comprend que la connaissance organisationnelle est le résultat d'une activité humaine collective volontaire et contextualisée, guidée par un objectif précis. Elle peut provenir de diverses origines : expériences individuelles, apprentissages dans le cadre de routines, traitements d'informations, activités de résolution de problèmes, culture organisationnelle (Kim et *al.*, 2003). Elle peut aussi être individuelle et/ou collective (Dieng-Kuntz et Matta, 2002 ; Yang et *al.*, 2010).

Dans le cadre du processus socio-statistique d'aide à la décision, on cherche à développer une méthodologie qui permette de créer et, si besoin, de faire évoluer des connaissances organisationnelles<sup>26</sup> utiles à la constitution des artefacts de l'aide à la décision. Avant d'aller plus avant sur ce point, il convient de préciser les mécanismes de création de connaissances organisationnelles : ils font effectivement l'objet de nuances qui nécessitent d'être éclaircies.

## *ii. Création de connaissances organisationnelles*

Kim et *al.* (2003) proposent de considérer la création de connaissance organisationnelle selon deux dimensions :

- *Le mode de création de la connaissance* : il peut être expérientiel et dans ce cas, il provient de l'accumulation d'expériences individuelles ou organisationnelles. Il peut aussi être analytique. La création se base alors sur des activités d'analyse de données ou d'informations, pour détecter des régularités et/ou construire des modèles. Selon notre volonté de construire une démarche de modélisation permettant de construire un modèle d'évaluation, dans le cadre de la conduite d'un processus d'aide à la décision explicite, on se situe clairement dans ce dernier cas de figure ;
- *Le statut épistémologique de la connaissance* : la question porte ici sur la définition de la connaissance sous-jacente à la création des connaissances organisationnelles. Ce point doit être discuté plus en détail.

---

<sup>26</sup> Notons que cette création est une branche de l'apprentissage organisationnel (Lopez-Nicolas et Soto-Acosta, 2010) : tel que nous souhaitons conduire l'aide à la décision, nous sommes dans le contexte d'individus confrontés à une situation problématique et entamant une investigation au nom de l'organisation (Argyris et Schön, 2001), dans le cadre d'un phénomène collectif d'acquisition et d'élaboration de compétences (Koenig, 2006), destinés à améliorer la prise de décision *via* la constitution de démarches de modélisation et de modèles.

En effet, ce statut épistémologique de la connaissance organisationnelle fait l'objet d'une abondante littérature et de nombreux débats. Sans vouloir en donner un aperçu exhaustif, qui dépasserait largement le cadre de cette thèse, nous évoquons quelques éléments de littérature, utiles pour comprendre les mécanismes sous-jacents au processus socio-statistique d'aide à la décision. Ces débats s'ancrent sur une proposition originelle de Polanyi, qui distingue deux types de connaissances :

- *La connaissance explicite*, qui peut être exprimée et transmise formellement, par des langages formels, des équations, des symboles ;
- *La connaissance tacite*, qui représente une connaissance qui n'est pas externalisée de la sorte : elle est dans l'« esprit » des acteurs de l'organisation.

La distinction entre ces deux types de connaissances est quasiment unanimement acceptée dans la littérature, qui y fait largement référence (Nonaka, 1994 ; Fouet, 1996 ; Tsoukas, 1996 ; Nonaka et Konno, 1998 ; Tsoukas et Vladimirou, 2001 ; Tsoukas *in* Easterby-Smith et Lyles, 2002 ; Dieng-Kuntz et Matta, 2002 ; Kim et *al.*, 2003 ; Nonaka et von Krogh, 2009 ; Yang et *al.*, 2010 ; Lopez-Nicolas et Soto-Acosta, 2010). D'autres distinctions de la connaissance organisationnelle existent (par exemple, procédurale *versus* déclarative, compétence *versus* connaissance, *know-how versus know-what*, etc.) mais, selon Kim et *al.* (2003), l'ensemble de ces typologies peuvent être classifiées selon les perspectives tacites ou explicites

Les travaux de Nonaka (Nonaka, 1994 ; Nonaka et Konno, 1998 ; Nonaka et von Krogh, 2009) ont fortement contribué à la diffusion de ces deux perspectives, en suggérant un *continuum* allant de connaissance tacite à connaissance explicite, ainsi qu'en précisant des modes de conversion (ou de traduction) des connaissances. Ces modes de conversion proposés par Nonaka permettent de préciser les conditions de création de connaissances organisationnelles nouvelles : convertir une connaissance tacite en une nouvelle connaissance tacite correspond au mode dit de « socialisation », convertir une connaissance tacite en une connaissance explicite correspond au mode « externalisation », etc., comme le résume la Figure 18.

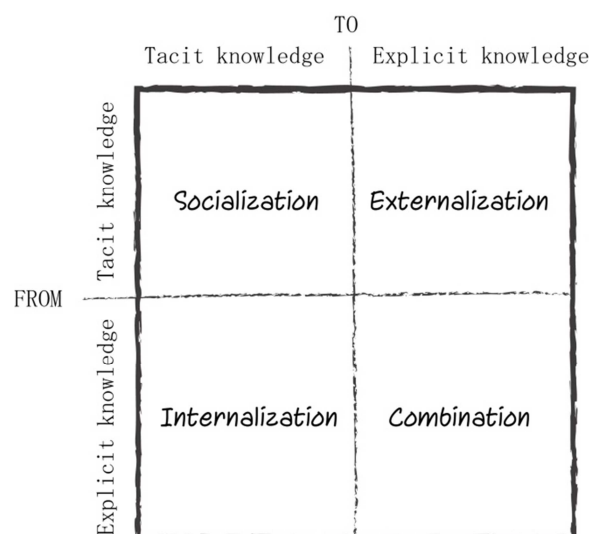


Figure 18. Les modes de conversion de la connaissance organisationnelle selon Nonaka (1994).

Ces mécanismes pourraient être sollicités lors de l'application du processus socio-statistique d'aide à la décision. On cherche en effet à construire une représentation commune d'un problème, à partir des connaissances des acteurs impliqués, pour progressivement aboutir à une recommandation, basée sur un modèle formel donc une connaissance explicite, qui s'ancrera dans les pratiques. Ainsi :

- *Les acteurs partageront entre eux des connaissances tacites* : on peut parler de socialisation ;
- *Des connaissances tacites seront formalisées au travers d'artefacts* : on peut parler d'externalisation ;
- *Si ces acteurs ou d'autres acteurs s'approprient ensuite les artefacts constitués dans leurs pratiques* : on peut parler d'internalisation.

Cette proposition est séduisante. Néanmoins, plusieurs auteurs rejettent les définitions de Nonaka. Se voulant plus fidèles à la pensée de Polanyi, ils arguent l'impossibilité de convertir une connaissance tacite en connaissance explicite : la connaissance tacite n'est pas qu'une connaissance qui n'est pas externalisée, c'est une connaissance qui ne peut pas l'être (Tsoukas et Vladimirou, 2001 ; Kim et al., 2003 ; Tsoukas, 1996, 2003). Dans le cadre de nos travaux, on ne peut alors plus parler d'externalisation lors de la création d'artefacts.

Pour sortir de ce conflit, Kim et al., 2003 proposent un troisième type de connaissance organisationnelle : la connaissance implicite, qui n'est pas externalisée mais qui peut l'être. Tsoukas (Tsoukas et Vladimirou, 2001 ; Tsoukas in Easterby-Smith et Lyles, 2002) développe quant à lui une vision moins segmentée : connaissances tacite et explicite ne sont pas les deux *extremums* d'un *continuum*, mais plutôt les deux faces d'une même pièce. Toute connaissance explicite reste liée à une connaissance implicite sous-jacente : la connaissance tacite ne peut pas être traduite ou convertie, mais uniquement mise en évidence, grâce à la connaissance déjà existante. Ainsi, Tsoukas propose une vision de la création de connaissance, où la connaissance existante est mise en action pour faire évoluer les pratiques, par des interactions sociales. La création de connaissances organisationnelles passe par la mise en évidence de pratiques individuelles, qui fera conséquemment évoluer les pratiques. En définitive, selon cette acceptation de la création de connaissance, si les artefacts du processus d'aide à la décision sont structurés autour de nouvelles connaissances, ils ne constituent pas en eux-mêmes de nouvelles connaissances : c'est leurs acceptations par les acteurs et leurs mises en usage, dans le cadre des activités des organisations, qui en seront. Evidemment, cela n'empêche pas que la connaissance mise en évidence soit capturée hors des schèmes cognitifs individuels, c'est-à-dire documentée, mais la documentation qui en résultera ne sera plus une connaissance en tant que telle (Dieng-Kuntz et Matta, 2002).

En vérité, il n'est pas de notre compétence de nous positionner dans un tel débat académique. Par contre, il semble important de retenir :

- *Que la construction de connaissances requiert, dans tous les cas, un enracinement dans un processus social*, impliquant les acteurs de l'organisation et les connaissances (tacites ou explicites) déjà en leur possession ;
- *Le caractère local de la connaissance créée* (Tsoukas, 1996) : problème traité, acteurs impliqués, et nombre de connaissances déjà existantes sont effectivement locaux. C'est un avantage par rapport au déploiement d'une solution « clés en mains » : il est souvent considéré que la mise en place de processus locaux d'apprentissage (Cohendet et Llerena *in* Ancori, 1992) et de prise de décision (Glodas et Micaelli *in* Jacot, 1994 ; Quaadgras et Edwards, 2012) est une excellente solution pour la viabilité organisationnelle et la qualité décisionnelle.

Ces deux remarques sont en phase avec les objectifs du processus socio-statistique d'aide à la décision. D'une part, il vise à impliquer des acteurs de l'organisation ; d'autre part, il vise à créer une connaissance en réponse à un problème de décision local.

Néanmoins, la question de l'opérationnalisation de ces concepts théoriques se pose désormais : que l'on veuille convertir une connaissance ou mettre en évidence et éventuellement documenter une connaissance, quel procédé utiliser ? Dans quelle mesure peut-elle être mise en usage ? etc. Pour cela, on peut s'appuyer sur un ensemble de concepts théoriques présentés ci-après.

### *1.2.3. Mise en œuvre : gestion du système de connaissances organisationnelles et triptyque données-information-connaissance*

#### *i. Gestion du système de connaissances organisationnelles et processus socio-statistique d'aide à la décision*

Pour la gestion de l'ensemble des artefacts du processus socio-statistique d'aide à la décision, on propose de s'appuyer sur les travaux d'Ermine (Ermine, 1996 ; Barthelmé, 1998). Il considère les organisations comme des systèmes complexes, maillés de sources de connaissances potentielles, humaines ou non, explicites ou implicites. Il développe alors une théorie systémique considérant la gestion des systèmes de connaissances des organisations comme une problématique d'ingénierie, visant à mieux maîtriser les flux de connaissances organisationnelles (construction et formalisation) et de cognition (utilisation des connaissances organisationnelles). Toujours en partant du principe que les systèmes de connaissances doivent permettre de construire des représentations utiles pour répondre à un problème, il propose un cadre méthodologique afin de les conceptualiser et de les gérer.

D'une part, un système de connaissances doit être considéré selon trois points de vue :

- *Ontologique* : ce que le système est, sa structure, sa constitution ;
- *Phénoménologique* : ce que le système fait, sa fonction ;
- *Génétique* : ce que le système devient, son évolution.

D'autre part, pour chacun de ces points de vue, la connaissance doit être abordée selon trois aspects :

- *Syntaxique* : la forme sous laquelle se traduit la connaissance, le code qu'elle utilise pour prendre forme ;
- *Sémantique* : le fond de la connaissance, la structure qu'elle utilise pour prendre sens ;
- *Pragmatique* : l'environnement que la connaissance utilise pour se mettre en contexte.

Dans le cadre du processus socio-statistique d'aide à la décision, on s'intéressera en premier lieu aux interactions entre ontologie du système de connaissances et aspects syntaxique, sémantique et pragmatique : l'objectif est avant tout de formaliser les artefacts de l'aide à la décision, en impliquant la construction d'une nouvelle connaissance quantitative actionnable grâce à un modèle prospectif (cœur de thèse). Les points de vues phénoménologique et génétique seront abordés, mais plus superficiellement :

- *Concernant le point de vue phénoménologique*, on montrera comment la connaissance créée peut être utilisée (usage) ;
- *Concernant le point de vue génétique*, on montrera comment on peut mettre en place des mécanismes favorisant le maintien à jour et l'évolution de la connaissance (notamment par les travaux annexes de détection de comportements anormaux).

En ce qui concerne le point de vue ontologique, il nous faut donner plus détail, afin de clarifier comment nous favoriserons les flux de connaissances organisationnelles. Ce point de vue définit les différents aspects de la connaissance comme suit :

- *L'aspect syntaxique du point de vue ontologique* est une donnée informationnelle, une « information » brute non signifiante ;
- *L'aspect sémantique du point de vue ontologique* est une structuration de la donnée informationnelle en donnée sémantique. Cette structuration se construit et se pérennise dans les structures mentales de l'homme, prenant une signification de façon à satisfaire un objectif ou résoudre un problème ;
- *L'aspect pragmatique du point de vue ontologique* est l'intégration de la donnée informationnelle dans un contexte donné, dans un système général. Ermine parle d'identifier l'enveloppe qui donne toute leur richesse aux données et significations.

On retrouve cette structuration dans l'« approche artefact » (Jacot, 1994) : en présence d'un problème de gestion, une structure conceptuelle est nécessaire pour apporter une réponse. Cette réponse passe par une représentation de la réalité, qui nécessite des conditions structurelles et fonctionnelles de traitement de l'information. Elle aboutit à la constitution d'artefacts, construits en fonction d'objectifs, pour gérer la complexité.

Ces aspects de l'ontologie du système de connaissances suggèrent trois niveaux d'organisation et de compréhension croissants de l'information. Cette idée se retrouve souvent sous une autre forme dans la littérature : le triptyque données-information-connaissance.

*ii. Triptyque données-information-connaissance et processus socio-statistique d'aide à la décision*

On fait usuellement référence dans la littérature à une grille de lecture conceptuelle, qui permet de distinguer les trois notions suivantes (Tuomi 1991 ; Gordon et Paugam-Moisy, 1997 ; Lebraty, 1998 ; Alavi et Leidner, 2001 ; Tsoukas et Vladimirou, 2001 ; Dieng-Kuntz et Matta, 2002 ; Kim et *al.*, 2003) :

- Les « données », qui sont des chiffres et faits bruts ;
- Les « données » deviennent des « informations », lorsqu'elles sont traitées et mise en relation<sup>27</sup> ;
- Les « informations » deviennent des « connaissances », lorsqu'elles sont interprétées par des individus, et mises en contexte et/ou en théorie.

Certains auteurs suggèrent que les frontières entre ces notions peuvent être plus ou moins délimitées, et ainsi s'organiser selon un *continuum* (Tsoukas et Vladimirou, 2001) ; d'autres indiquent une hiérarchie « classique » (des données vers la connaissance) et « inversée » (de la connaissance vers les données), qui peuvent coexister (Tuomi, 1999). Ces notions, ainsi que leurs relations à l'ontologie du système de connaissances selon Ermine, sont résumées le Tableau 2.

<b>Aspect de l'ontologie du système de connaissances</b>	<b>Niveau de la connaissance organisationnelle</b>	<b>Description</b>
<i>Syntaxique</i>	Données	Matériau brut
<i>Sémantique</i>	Informations	Structuration et mise en relation des données
<i>Pragmatique</i>	Connaissances	Interprétation et mise en contexte des informations

*Tableau 2. Ontologie du système de connaissances organisationnelles et triptyque données-information-connaissance.*

<sup>27</sup> Voir aussi Ermine (1989), qui fait référence à l'analyse structurale : les variables ne prennent sens que l'une par rapport aux autres, pour répondre à un problème. Les objets-artefacts construits sont donc issus de l'organisation de ces variables, signaux de l'environnement du décideur. Cette notion d'organisation de signaux perçus est fondamentale dans la littérature relative à la gestion des connaissances, qui semble ainsi fortement nourrie des thèses constructivistes (Watzlawick, 1988 ; Segal, 1990 ; Le Moigne, 2003) et structuralistes (Piaget, 1968).

Le processus socio-statistique d'aide à la décision doit donc permettre d'évoluer au sein de ces différents niveaux, pour construire les artefacts, donc un ensemble de connaissances, utiles à la prise de décision. La transition d'un aspect de l'ontologie du système de connaissances à un autre correspond à une gymnastique cognitive favorisant les conversions de connaissance (Nonaka) ou la mise en évidence/la documentation des connaissances (Tsoukas).

Cette gymnastique implique les acteurs de l'organisation et leurs connaissances déjà existantes, ainsi que le recours à un certain nombre de méthodes permettant de mettre en exergue et formaliser ces connaissances, apportées par l'analyste/le chercheur. Par la mise en œuvre de cette activité cognitive, le processus socio-statistique d'aide à la décision doit permettre d'apporter des contributions nouvelles à l'ontologie du système de connaissances de l'organisation, en cherchant à construire les artefacts de l'aide à la décision. Cette contribution pourra ensuite être utilisée pour décider (phénoménologie du système de connaissances). Elle pourra aussi être maintenue à jour et évoluer (génétique du système de connaissances).

Cette mécanique cognitive représente le cœur du processus socio-statistique d'aide à la décision. Les bases théoriques présentées nous permettent de le spécifier plus en détail, dans la section suivante.



## Section 2. Spécification du processus socio-statistique d'aide à la décision.

### 2.1. Présentation générale

La Figure 19 fournit un aperçu général du processus socio-statistique d'aide à la décision, proposé dans le cadre de nos travaux. Il sera appliqué dans les parties 2 et 3 de la thèse.

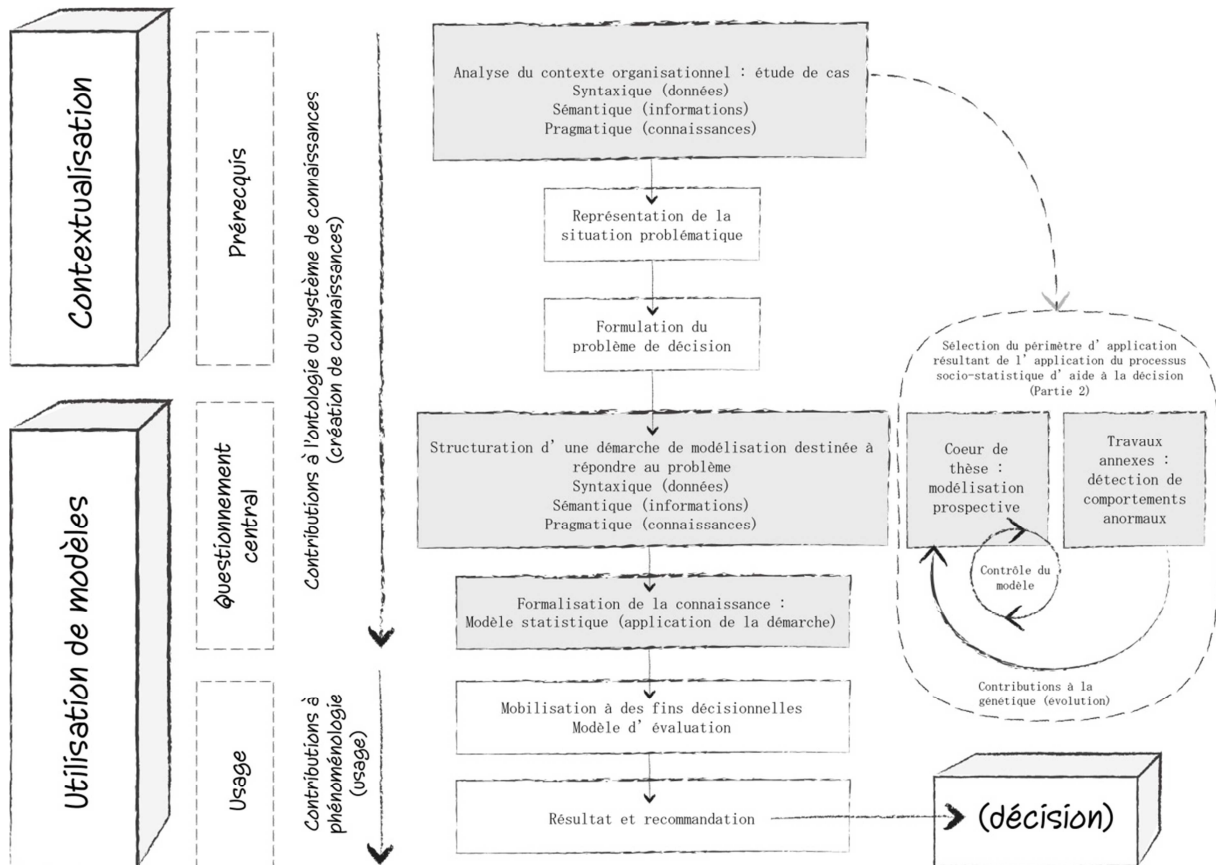


Figure 19. Processus socio-statistique d'aide à la décision : proposition.

On peut y voir les artefacts de Tsoukiàs (Représentation de la situation problématique, Formulation du problème de décision, Modèle d'évaluation, Résultat et recommandation). Les apports principaux de la thèse sont mis en évidence en gris. Ils englobent les phases suivantes, qui visent à agir sur la connaissance organisationnelle :

- En prérequis, une construction de connaissance est nécessaire pour contextualiser la situation décisionnelle. Nous proposons pour cela d'utiliser une méthode de structuration d'étude de cas, qui permet de sélectionner le périmètre d'application de nos travaux, de formaliser les artefacts « représentation de la situation problématique » et « formulation du problème de décision » de Tsoukiàs ;

- *Le questionnement central consiste ensuite à apporter des moyens de réponse à la situation décisionnelle explicitée.* Pour cela, il faut proposer une démarche de modélisation, pour répondre au problème de décision. Cette démarche doit permettre de structurer une connaissance organisationnelle, sous la forme d'un modèle statistique. Elle vise principalement à proposer un modèle prospectif (cœur de thèse). On propose également de lui adjoindre des mécanismes visant à favoriser la génétique (évolution de la connaissance organisationnelle) ; on distingue :
  - \* *Le contrôle du modèle prospectif (cœur de thèse) :* l'objectif est de s'assurer que la connaissance formalisée dans le cadre du modèle de prévision reste valide,
  - \* *La détection de comportements anormaux (travaux annexes) :* conformément aux bénéfices attendus de l'activité de contrôle, on cherche à détecter des comportements inattendus qui permettraient d'améliorer le modèle prospectif.

Le modèle de prévision statistique proposé pourra être utilisé et mis en usage (phénoménologie du système de connaissances) dans le cadre d'un modèle d'évaluation, afin de suggérer une recommandation supportant la prise de décision relative au *capacity planning* à Crolles 2. Comme indiqué précédemment, chacune des étapes de construction de connaissance implique la mobilisation des diverses expertises de l'organisation et de l'analyse/chercheur. Pour cela, nous précisons ci-dessous les étapes d'analyse du contexte organisationnel et de structuration d'une démarche de modélisation.

## **2.2. Prérequis : analyse du contexte organisationnel et construction de connaissances organisationnelles**

### *2.2.1. Généralités*

La méthodologie de contextualisation de l'aide à la décision ne constitue pas le cœur de nos travaux. C'est toutefois une étape préliminaire indispensable dans le cadre du processus socio-statistique d'aide à la décision. C'est pourquoi un certain nombre d'éléments méthodologiques sont proposés ici.

On cherche à constituer une représentation de la situation problématique et une formulation du problème à traiter, à partir des représentations plus ou moins floues que peuvent en avoir des acteurs impliqués dans une situation décisionnelle. On vise donc à comprendre un phénomène social particulier (la prise de décisions relatives au *capacity planning*), dans son environnement naturel (Crolles 2). Une telle analyse relève de ce que Coutelle (2005) et Thietart (2007) qualifient d'enquête ethnographique. Le cadre de Tsoukiàs constitue une grille de lecture permettant de structurer l'analyse de ce phénomène.

Conformément à l'ontologie du système de connaissances, on cherche à guider cette analyse selon les différents aspects/niveaux de la connaissance. La question est alors la suivante : quelles données utiliser et comment opérer les transformations entre données brutes, information et connaissances, pour contribuer à l'ontologie du système de connaissances, afin d'être capable de contextualiser l'aide à la décision souhaitée ?

Pour y répondre, on propose de construire une étude de cas. Cette méthode est recommandée par Coutelle (2005) et Thietart (2007), dans le cadre d'une enquête ethnographique. Les principales étapes de la construction d'une étude de cas peuvent être considérées comme une contribution à l'ontologie du système de connaissances. Sauf mention expresse, ces étapes sont précisées ci-après à partir de notre interprétation des sources suivantes : Wacheux (1996), Coutelle (2005), Thietart (2007), Revillard (2012).

### 2.2.2. Construction des connaissances par la réalisation d'une étude de cas

#### i. Récolte des données (aspect syntaxique)

La première étape de l'étude de cas concerne la récolte de données, d'un matériau brut, permettant de constituer une intelligence de la situation. Une analyse de type ethnographique se base sur un matériau qualitatif. Plusieurs méthodes peuvent être employées, en fonction des sensibilités de l'analyste et des contraintes opérationnelles : entretiens, observations, etc. Dans le cadre de nos recherches, nous avons pris le parti de récolter les données brutes à partir d'entretiens individuels, avec les acteurs impliqués dans le processus décisionnel. Pour favoriser la discussion sans être trop directif, nous avons recours à des entretiens semi-directifs. Ils sont basés sur un guide d'entretien. Ce guide est structuré autour des éléments des artefacts permettant de contextualiser l'aide à la décision, selon le cadre de Tsoukiàs :

- *Situation(s) problématique(s)* : quels sont les acteurs impliqués, leurs objets et enjeux, engagements ressources ?
- *Formulation(s) du (des) problème(s)* : quels sont les alternatives potentielles, les points de vues, la problématique décisionnelle ?

Une fois le guide validé<sup>28</sup>, il peut ensuite être utilisé pour la récolte effective des données, par la conduite des entretiens sur le terrain. Le matériau récolté lors de ces discussions prend la forme d'un enregistrement audio, retranscrit sous forme de *verbatim*. Ces entretiens peuvent faire l'objet d'une première micro-analyse, sous la forme d'une synthèse communiquée à l'interviewé, permettant de valider que l'interviewer a bien compris le discours de son interlocuteur, que l'interviewé valide son discours, et d'identifier d'éventuels ajustements nécessaires pour la suite de l'étude (guide d'entretien, conduite des entretiens, etc.). Lors de la conduite d'entretiens et des phases de validation, d'autres matériaux d'analyse complémentaires peuvent être récoltés (courriels, compte-rendu de réunion...). Cette méthode de collecte de données est résumée dans la Figure 20.

---

<sup>28</sup> Dans le cadre de nos travaux, sa cohérence a été validée par d'autres chercheurs, familiers de la problématique traitée.

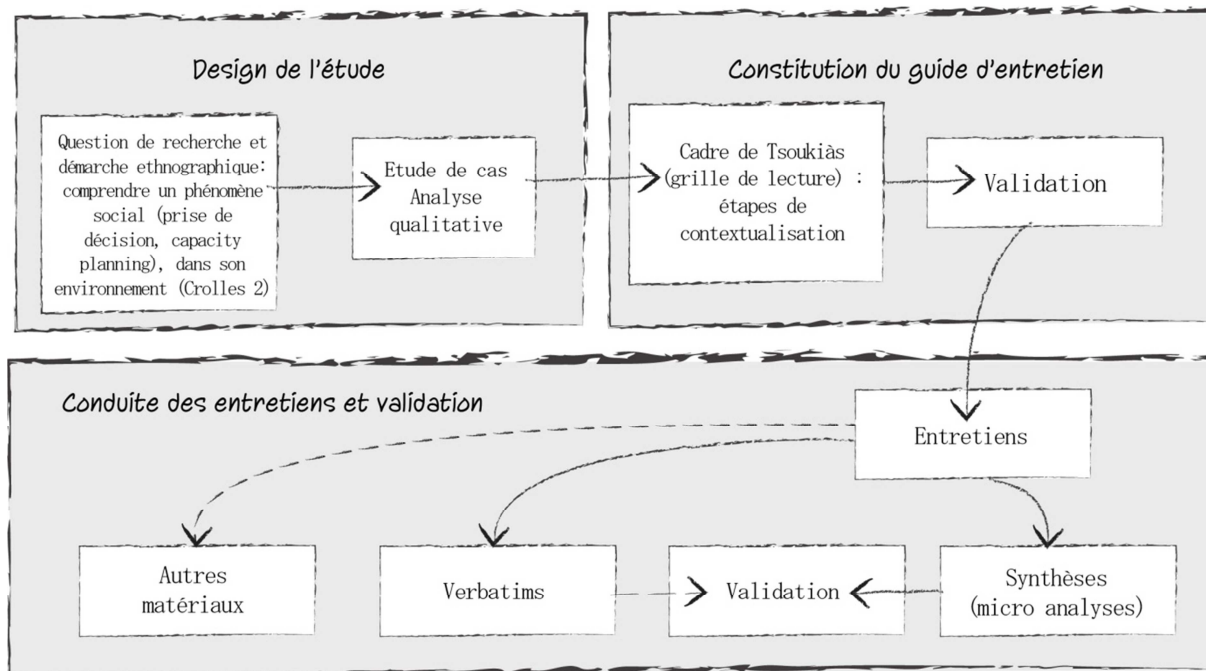


Figure 20. Contextualisation de l'aide à la décision : récolte des données pour la construction d'une étude de cas.

Seules, ces données isolées ne suffisent pas à définir les artefacts du processus d'aide à la décision. Les acteurs donnent des réponses non structurées et personnelles. L'objectif est de les confronter entre elles, pour aboutir à une représentation partagée de la situation.

## ii. Structuration des données - informations (aspect sémantique)

Les données récoltées sont ensuite exploitées qualitativement. Il n'est pas fait usage de procédures quantifiées : dans le cadre de la constitution d'une étude de cas, c'est la recherche de sens qui prime, par l'analyse approfondie et la proximité avec le matériau récolté. Cette recherche de sens passe par l'organisation et la structuration du matériau brut. Trois objectifs principaux sont recherchés : mettre en exergue les principaux thèmes rencontrés lors de la conduite des entretiens, les catégoriser de façon descriptive (ou causale, mais cette question n'entre pas dans le cadre de nos recherches) et établir les liens avec la théorie (le cadre de Tsoukiàs).

Cette approche analytique qualitative correspond à une fonctionnalité de traitement des données que Lejeune (2010) qualifie de réflexive. Le recours à un logiciel spécifique d'aide à l'analyse qualitative est envisageable, mais pas indispensable. Dans un premier temps, l'ensemble du matériau est agencé autour des thèmes identifiés lors des micro-analyses, afin de réduire le volume des données récoltées, en les structurant et les mettant en relation. Dans un second temps, la question de recherche est éclairée par la constitution de catégories d'analyse, permettant d'organiser le matériau selon le cadre théorique mobilisé et l'objectif de l'étude terrain. A ce stade de nos travaux, nous nous sommes cantonnés à la considération manuelle de l'exploitation de ces données, mais

des pistes de réflexions pourraient être envisagées pour situer ce travail dans une dimension plus automatisée et quantitative.

Ce travail permet d'aboutir à la construction de l'étude de cas, rédigée sous la forme d'une synthèse complète, proposant une définition structurée des artefacts de permettant de contextualiser l'aide à la décision, par la formalisation de la représentation de la situation problématique et du problème de décision.

### *iii. Interprétation des informations - connaissances (aspect pragmatique)*

Conformément aux niveaux de la connaissance organisationnelle, il ne suffit pas d'acquérir une information pour parler de connaissance : il est d'abord nécessaire d'interpréter et de mettre en contexte l'information. Pour cela, l'étude de cas rédigée est diffusée aux interviewés pour une discussion et une validation finale. On s'assure de la sorte que le cas construit, la représentation de la situation problématique, la formulation du problème de décision, sont bien caractéristiques de la réalité perçue par les acteurs impliqués dans le processus d'aide à la décision.

Précisons que ce travail ne vise aucune généralisation. C'est un travail idiosyncratique et contextuel : sa validité externe n'a pas à être justifiée. Concernant la validité interne, un certain nombre de critères généralement admis doivent être vérifiés : méthode de recherche identifiée et justifiée, multiplication des sources, analyse produite à partir d'une théorie acceptée, relecture du cas par les interviewés, honnêteté dans le processus de recherche, etc.

### *2.2.3. Synthèse*

Cette première phase de contextualisation du processus socio-statistique d'aide à la décision mobilise donc à la fois les acteurs de l'organisation et l'analyste. Chacun amène un certain nombre de connaissances déjà en leur possession (plutôt la connaissance du contexte décisionnel pour les premiers, la connaissance des méthodes construction d'étude de cas et d'analyse qualitative pour ce dernier), pour en créer une nouvelle. Le Tableau 3 résume cette proposition.

<b>Aspect du système de connaissances</b>	<b>Etapes de la construction de connaissance</b>	<b>Principales connaissances existantes à mobiliser</b>
<i>Syntaxique</i>	<i>Récolte de données</i> : conduite d'entretiens semi-directifs	Contexte décisionnel ; Méthodes de construction d'étude de cas
<i>Sémantique</i>	<i>Structuration des données - informations</i> : analyse thématique et analyse par catégories	Contexte décisionnel ; Méthodes de construction d'étude de cas et d'analyse qualitative
<i>Pragmatique</i>	<i>Interprétation des informations - connaissances</i> : acceptation de l'étude de cas et de la représentation de la (ou des) situation(s) problématique(s) et de la (ou des) formulation(s) du (des) problème(s)	Contexte décisionnel ; Méthodes de construction d'étude de cas

Tableau 3. *Analyse du contexte organisationnel et construction des connaissances organisationnelles.*

Cette étape préliminaire permet de considérer l'utilisation de modèles dans l'aide à la décision : en favorisant la compréhension du contexte organisationnel et des besoins décisionnels, le passage vers une seconde phase de résolution de problème est facilité. Cette deuxième phase est quant à elle basée sur des modèles, mobilisant des techniques quantitatives et non plus uniquement qualitatives.

### **2.3. Questionnement central : structuration d'une démarche de modélisation et construction/évolution des connaissances organisationnelles**

#### *2.3.1. Généralités*

La question posée est celle de l'exploitation des données issues de bases de données (en l'occurrence le CMIS ITIL), pour la prise de décision. Dans une perspective cognitive, ces données sont considérées comme des opportunités pour créer de nouvelles connaissances (Lebraty, 1998 ; Alavi et Leidner, 2001 ; Lee et Kim, 2001 ; Lopez-Nicolas et Soto-Acosta, 2010). Néanmoins, la littérature ne considère pas suffisamment, à notre connaissance, la place des statistiques dans un tel processus. La question est alors : comment s'appuyer sur des méthodes statistiques pour opérer les transformations entre données brutes, information et connaissances, afin de contribuer à l'ontologie du système de connaissances ?

La réponse est nécessairement contextualisée : le choix des variables et des méthodes statistiques dépend nécessairement du problème de décision local. Ces choix dépendront donc de l'application du processus socio-statistique d'aide à la décision à un cas particulier. Néanmoins, certains concepts généraux peuvent être précisés : dans le cadre du processus socio-statistique d'aide à la décision et des deux activités traitées dans le cadre de cette thèse, en quoi les statistiques peuvent-elle contribuer à la création de connaissances organisationnelles, afin de supporter les décisions de *capacity planning* ?

Nous verrons aussi que les statistiques peuvent contribuer à l'évolution de cette connaissance, c'est-à-dire à la génétique du système de connaissances.

### *2.3.2. Construction et évolution des connaissances par l'utilisation des statistiques*

#### *i. Récolte des données (aspect syntaxique)*

##### Objectif

Tout d'abord, la question des données, du matériau brut, doit être abordée. La première étape de tout travail statistique est celle de la quantification, c'est-à-dire du codage du réel, par la construction d'un espace de mesure (Desrosières, 2010). Dans le cadre de nos travaux, il convient de s'interroger sur les variables pertinentes pour la constitution d'un modèle de prévision (variables qui pourraient ensuite, entre autres, être mises sous contrôle de détection des comportements anormaux).

##### Place de la récolte des données dans la construction de connaissances : cas d'un système complexe

Compte-tenu de la complexité des systèmes à gérer, ce questionnement n'est pas anodin. Effectivement, aborder des systèmes complexes implique une façon particulière d'appréhender le réel (Le Moigne, 1999). D'une part, la modélisation est la seule voie qui permette de gérer de tels systèmes : alors qu'un système compliqué peut être simplifié pour être expliqué, un système complexe doit être modélisé pour en construire l'intelligibilité. D'autre part, appréhender un système complexe suggère un positionnement téléologique et constructiviste. En effet, une réflexion concernant le système à modéliser est un prélude au travail de modélisation. Face à la complexité, cette question n'est pas neutre : la définition du système n'est pas une hypothèse de travail évidente et objective. Le modélisateur doit conceptualiser la réalité qu'il souhaite comprendre, en fonction de ses objectifs. Un modèle d'un système complexe est finalisé et la réalité qu'il représente construite (Le Moigne, 1974, 1999, 2003 ; Morin, 2005).

Cette approche constructiviste de la modélisation des systèmes complexes est largement développée dans la littérature systémique. Krob (*in* Appriou, 2009) montre comment les couplages entre systèmes techniques et systèmes organisationnels peuvent être étudiés comme une intégration d'abstractions de sous-systèmes. Ces abstractions dépendent du niveau de granularité requis. Un système peut être décomposé en un ensemble de convenance de sous-fonctions (pouvant être elles-mêmes décomposées, de façon arborescente), dont on ne cherche pas nécessairement à décrire les mécanismes (Simon, 2004). Le Moigne (1999) parle de réseau de processeurs pour qualifier ces boîtes noires en interrelation. Guy (2012) illustre comment tout appareil complexe est un assemblage de composants, dont on admet les lois de fonctionnement. Accepter ces lois sans chercher à les démontrer permet d'observer un système à un niveau plus macroscopique. De façon générale, trois grandes idées caractéristiques définissent un système complexe (Walliser, 1977) : c'est un ensemble en rapport avec son environnement, il est composé de sous-systèmes en interactions et il peut se modifier dans le temps.

Appréhender un système complexe, c'est donc stabiliser la compréhension de sous-systèmes, d'interactions entre sous-systèmes, et de relations à l'environnement. Le Moigne (1999) parle de :

« *projet de complexifications fonctionnel croissant par étapes stabilisatrices* »

#### Conséquences pour le processus socio-statistique d'aide à la décision

Dans le cadre de l'aspect syntaxique de l'ontologie du système de connaissances, l'objectif est donc d'arriver à identifier les données pertinentes, existantes ou non, permettant de caractériser les sous-systèmes, l'environnement, etc., en fonction de l'aide à décision à construire. On se base pour cela sur les connaissances préexistantes des acteurs de l'organisation, experts et référents dans le domaine du projet de gestion considéré. Cette approche permet de répondre aux enjeux de gestion de l'infrastructure des TI : le système informatique complexe peut être considéré grâce à une approche statistique et systémique (enjeu n°1), impliquant de prendre en compte son environnement, c'est-à-dire l'activité métier (enjeu n°2).

Il est important d'étendre le périmètre fonctionnel des acteurs impliqués dans cette discussion au-delà du seul service en charge de l'infrastructure des technologies de l'information. En effet, du fait de la décentralisation des architectures des TI (voir plus haut), les managers en charge de l'infrastructure des TI n'ont plus nécessairement le contrôle de l'ensemble des données produites par le SI. Ces données peuvent être dispersées au sein de multiples sous-systèmes hétérogènes (Lloyd et Rudd, 2007). Elles peuvent aussi être éparses, incomplètes ou erronées (Martin et al., 2010 ; Shah et al., 2012). En conséquence, les managers en charge de l'infrastructure des TI peuvent avoir du mal à localiser, évaluer l'utilité, interpréter et valider les données susceptibles d'être utiles et intégrées au CMIS ITIL.

Remarquons que si les données sont inexistantes, il faudra être capable de les créer : nous sommes dans le cadre de la hiérarchie inversée du *continuum* données-informations-connaissance (Tuomi, 1999). Basée sur une connaissance préalable, cette nouvelle donnée correspond à une formalisation et/ou documentation de connaissance, qui peut très bien perdre son statut de connaissance si l'on en perd la maîtrise ou l'usage.

Pour ces raisons, le processus socio-statistique d'aide à la décision devra donc intégrer une réflexion qualitative approfondie afin de conceptualiser préliminairement l'objet à modéliser, puis d'identifier ou de constituer les données en conséquence.

#### *ii. Structuration des données - informations (aspect sémantique)*

##### Objectif

Ensuite, la question de la structuration des données doit être abordée. Nous avons pris explicitement le parti de recourir à des méthodes statistiques, en accord avec les recommandations ITIL et l'évolution des pratiques relevées dans la littérature<sup>29</sup>. Les

---

<sup>29</sup> Et conformément à la volonté des managers à l'origine de ces travaux de thèse.



statistiques sont en effet un outil tout à fait adapté à la structuration des données, sous un angle quantitatif (Desrosières, 2010). Elles permettent :

« D'une part, [de] fixer des objets, en convenant de leurs équivalences par des définitions standardisées, ce qui permet de les mesurer, en précisant par le langage probabiliste le degré de confiance que l'on peut attacher à ces mesures. D'autre part, [de] décrire les relations entre les objets ainsi construits, et éprouver la consistance de ces liens. » (Desrosières, 2010)

Ce travail de « mathématisation », d'exploitation des données identifiées, est la deuxième étape d'une démarche statistique<sup>30</sup>. Il consiste à traiter mathématiquement les mesures, afin d'en analyser les relations et régularités significatives. Nos travaux sur le *capacity planning* impliquent trois types de traitements à envisager<sup>31</sup> :

- Dans le cadre de l'activité (Q) Modélisation et analyse des tendances (cœur des recherches) : pour mettre en relation les variables identifiées afin de constituer le modèle prospectif souhaité ;
- Lors de la détection de comportements anormaux, dans le cadre de l'activité (K) Contrôle (travaux annexes) : pour modéliser les séries chronologiques et les résidus de la modélisation, conformément à la méthode usuelle de contrôles de séries de données dépendantes du temps, afin de détecter des comportements anormaux.
- Par ailleurs, un troisième type de traitement statistique nous paraît important : compte-tenu de la possible instabilité d'un système complexe (cf. la définition de Walliser ci-dessus), un traitement intermédiaire entre les activités (Q) et (K) semble nécessaire. (Q) fournit un support de représentation d'un système complexe, mais on sait que le système peut changer : cette représentation peut alors devenir caduque. Or, la prise de décision sur la base d'un modèle statistique de prévision implique que le système se comportera dans le futur en conformité avec les phénomènes observés et modélisés (Stock et Watson, 1996). Il faut donc être capable d'identifier un éventuel changement, en contrôlant la validité dans le temps du modèle prospectif.

Ainsi, la combinaison de ces trois types de traitement semblent répondre à la recommandation de Morin (2005), qui indique que la combinaison de séquences maîtrisées et la capacité à réagir à l'inattendu est une des voies pour s'orienter dans la complexité :

- Construire un modèle prospectif vise à fournir des représentations de caractéristiques maîtrisées du système complexe, en fournissant une connaissance stabilisée et formalisée ;
- La détection de comportement anormaux et le contrôle du modèle prospectif permettent de détecter des comportements inattendus. La conjonction de ces contrôles permettra notamment de faire évoluer la connaissance stabilisée et formalisée, contribuant ainsi à la génétique du système de connaissances.

---

<sup>30</sup> Armatte (2010) indique que cette distinction entre quantification et mathématisation peut être arbitraire, et que le clivage entre ces deux notions n'est pas forcément catégorique. Néanmoins, nous ne tiendrons pas compte de ce nuancement épistémologique, qui ne semble pas nécessaire à ce stade de nos travaux.

<sup>31</sup> On rappelle que les activités (Q) et (K) ont été situées dans le cadre du référentiel ITIL, dans la section 1 du premier chapitre.

### Place de la structuration des données dans la construction de connaissances : cas d'un système complexe

Toutefois, il convient de s'interroger : la statistique est-elle vraiment un outil approprié à la modélisation des systèmes complexes et à la constitution de modèles d'aide à la décision ? Une justification est nécessaire, car certains auteurs ont pu critiquer le recours à la statistique et aux probabilités pour appréhender les systèmes complexes (De Rosnay, 1977 ; Bruter, 1982). Néanmoins, cette critique semble plus marginale de nos jours. Tout d'abord, Barbin et Lamarche (2004), comme Desrosières (2010), indiquent bien que la modélisation statistique est pleinement adaptée aux problématiques décisionnelles. Elle permet en effet d'appuyer la rationalité d'une décision sur des observations stables et censées, permettant d'établir des comparaisons et des équivalences.

Par ailleurs, sans y faire directement allusion, Desrosières (2010) propose un positionnement épistémologique de la statistique qui semble faire écho à la modélisation téléologique et constructiviste des systèmes complexes, évoquée précédemment. Selon lui, lors de la construction d'un modèle, l'objectif est d'apporter des éléments de réponse à un besoin de décision (hypothèse téléologique). Peu importe que le modèle ne vise pas à fournir une description complète et objective de la réalité. L'essentiel est de constituer un espace cognitif partageable, pour décrire et décider, correspondant à une prise de position sur la contingence (hypothèse constructiviste).

D'autres auteurs confirment cette proposition en affirmant explicitement la validité des statistiques en tant qu'outils propres à appréhender les systèmes complexes (Shalizi, *in* Deisboeck et Kresh, 2006 ; Laflamme, 2008). Pour cela, l'essentiel est d'être lucide sur ce que l'on cherche à modéliser. Si les données proviennent d'une conceptualisation finalisée d'un système, composé de boîtes noires (cf. i. *Récolte des données*), il est normal qu'on ne puisse considérer le modèle statistique sous un angle réaliste (Armatte *in* Barbin et Lamarche, 2004). Il faut alors être conscient que l'on modélise des relations de covariations, plus que de causalité (Laflamme, 2008). Pour autant, la valeur du modèle n'en est pas moindre : c'est une description résumée du réel, utilisable pour l'action, dont l'essentiel est qu'elle soit mathématiquement suffisamment solide pour répondre aux objectifs fixés (Desrosières, 2010). Comme l'indique Le Moigne (1999), à l'explication causale de la modélisation analytique, on peut substituer la compréhension téléologique de la modélisation systémique. Voilà pourquoi le recours à la statistique semble être approprié au contexte de nos travaux et à l'enjeu de gestion « gérer un système complexe ».

### Conséquences pour le processus socio-statistique d'aide à la décision

Nous suggérons ici la place importante de l'analyste/du chercheur, qui doit apporter ses connaissances en matière de méthodologie d'analyses statistiques, pour une bonne exploitation des données. Effectivement, s'il est recommandé que les experts en charge de l'infrastructure des TI aient une bonne formation en statistiques, pour assurer eux-mêmes cette tâche (Topi et *al.*, 2010), il semblerait que les pratiques professionnelles restent souvent trop basiques : Banerjee et Igbaria (1993) montrent que les praticiens se contentent souvent de visualiser les données disponibles. Nous déplorons de ne pas avoir d'étude plus récente sur ce sujet, ni plus spécifique aux pratiques industrielles (l'article est essentiellement focalisé sur les universités nord-américaines, mais précise que les résultats seraient assez similaires dans l'industrie). Ces conclusions semblent

néanmoins toujours d'actualité et sont par ailleurs constatées au sein de notre terrain de recherche, comme nous le verrons dans la partie 2.

C'est pourquoi il est important d'avoir recours à des méthodes statistiques fidèles à l'état de l'art, mais qui restent exploitables et appropriées pour un usage industriel. Comme nous l'avons expliqué, ces méthodes doivent permettre de construire une connaissance formalisée au travers de relations mathématiques, ainsi que d'en favoriser l'évolution.

### *iii. Interprétation des informations - connaissances (aspect pragmatique)*

#### Objectif

Comme le souligne le *continuum* données-informations-connaissances, l'information obtenue par structuration des données nécessite une phase d'interprétation qualitative. C'est alors que l'on peut parler de connaissance. Cette phase est peu développée dans la littérature, tant et si bien que les termes « information » et « connaissance » sont souvent indifféremment utilisés (Kim et al., 2003). L'interprétation, la contextualisation de l'information est cependant indispensable si l'on veut disposer d'une connaissance, utile à l'action intelligente (Tsoukas, 1997).

#### Place de l'interprétation des informations dans la construction de connaissances

La phase d'interprétation des construits mathématiques est une étape cruciale. Nous avons vu précédemment que l'objectif de la statistique est de fournir une description résumée du réel, utilisable pour l'action. Desrosières (2010) va plus loin en invoquant le concept d'objectivation : le travail statistique permet d'engendrer des réalités, que l'on peut considérer comme telles à partir du moment où les acteurs y font référence pour coordonner leurs actions. Il propose ainsi de résoudre la tension entre deux conceptions de l'usage de la statistique : celle stipulant que la statistique correspond à la modélisation d'une réalité indépendante de l'activité de mesure ; l'autre avançant qu'elle n'est que pures conventions sociales. Ce positionnement permet de sortir de ce débat, qui dépasse le cadre de cette thèse (on pourra par exemple se référer à Barbin et Lamarche (2004), Quinio Benamo (2009), Armatte (2010), Desrosières (2010), pour d'autres éléments de réflexions sur ce sujet).

En vérité, notre préoccupation n'est pas de discourir de la réalité d'un modèle, mais de définir les conventions qui permettront de créer un espace de débat pour alimenter le processus décisionnel, matérialisé par une connaissance formalisée dans un modèle mathématique. Ainsi, la pratique de la statistique dépasse le traitement quantitatif d'un ensemble de mesures : la solidité mathématique, la validité quantitative, ne suffit pas à garantir la qualité d'un modèle. Il est essentiel d'analyser, de comprendre et d'explicitier le modèle mathématique. Un sens commun, un accord concernant l'interprétation et la crédibilité des résultats, doivent être dégagés. Face à un modèle construit et finalisé, il faut pouvoir évaluer qualitativement dans quelle mesure l'on accepte :

*« de considérer que le modèle adopté simule convenablement la réalité et que la réalité va se comporter comme le modèle » (Meusnier, in Barbin et Lamarche, 2004)*

### Conséquences pour le processus socio-statistique d'aide à la décision

En ce qui concerne la modélisation prospective, cette phase correspond à l'acceptation du modèle de prévision, qui pourra alors être inscrit dans les schèmes cognitifs et utilisé en toute confiance. Ensuite, la génétique de cette connaissance doit être assurée, par l'interprétation des résultats de la détection de comportements anormaux et du contrôle du modèle prospectif : en quoi ces résultats statistiques peuvent-ils amener à reconsidérer le modèle de prévision ?

Dans le cadre de l'application du processus socio-statistique d'aide à la décision, il faut pour cela se baser sur les connaissances existantes des acteurs organisationnels adéquats. Ce n'est qu'après leurs analyses des résultats des traitements statistiques qu'on pourra dire que l'exploitation des données brutes aura permis la constitution de croyances vraies et justifiées, et que le terme de connaissance pourra être employé. Cette connaissance constitue alors un espace cognitif partagé utilisable pour décrire et décider, basé sur une formalisation de la connaissance au travers d'un modèle mathématique. Ce modèle peut tout à fait retourner à l'état de simple information si l'on en perd la compréhension ou l'usage. Par ailleurs, il est possible que cette nouvelle connaissance puisse aboutir à la reconsidération des données initialement disponibles et à la nécessité d'en créer des nouvelles, donc de ré-envisager l'aspect syntaxique du système de connaissances, comme le suggère la hiérarchie inversée du *continuum* données-informations-connaissances (Tuomi, 1999).

Tout comme pour la phase de quantification, il est recommandé d'étendre le périmètre fonctionnel des acteurs impliqués dans cette réflexion. Effectivement, la décentralisation des architectures des TI peut induire une perte de maîtrise pour les acteurs directement en charge de l'infrastructure des TI, des relations entre le système informatique technique et les usages applicatifs et métiers qui en sont faits. Dans ce cas, la capacité de ces acteurs à interpréter et évaluer les résultats de travaux de modélisation intégrant la dimension métier est réduite. Il faut donc considérer cette phase avec beaucoup de précautions, en impliquant les acteurs organisationnels appropriés.

#### 2.3.3. Synthèse

Nous avons montré que la construction d'une démarche statistique, visant à construire et maintenir à jour des connaissances organisationnelles, implique un ensemble de réflexions : conceptualisation du système et choix des variables, choix des méthodes statistiques, interprétation et acceptation de traitements quantitatifs. Ces étapes ne sont pas neutres. Elles s'inscrivent dans un processus social intégrant les acteurs de l'organisation, le modélisateur, leurs connaissances. Le processus socio-statistique d'aide à la décision doit intégrer ces questionnements. Pour cela, analyses quantitatives et qualitatives doivent être couplées, pour aboutir à la création d'une connaissance nouvelle et utile pour l'aide à la décision, basée sur la connaissance existante, tout en en assurant les conditions de mises à jour. Le Tableau 4 synthétise cette proposition.

	<i>Aspect du système de connaissances</i>		
	<b>Syntaxique</b>	<b>Sémantique</b>	<b>Pragmatique</b>
<i>Etape de la construction et de l'évolution des connaissances</i>	Récolte de données :  Modélisation systémique et identification des données	Structuration des données – informations :  Identification de méthodes quantitatives appropriées au contexte industriel et décisionnel, pour structurer les données. On distingue :  <i>Cœur de thèse</i> : Modélisation prospective (construction) et contrôle du modèle (évolution)  <i>Travaux annexes</i> : Détection de comportements anormaux (évolution)	Interprétation des informations – connaissances :  Inscrire les phases de quantification et de mathématisation dans une pleine dimension cognitive, afin de constituer une croyance vraie et justifiée partagée par les acteurs impliqués dans le processus de construction de connaissance et de décision
<i>Type d'approche</i>	Qualitative et quantitative si construction de données à envisager	Quantitative	Qualitative
<i>Principales connaissances existantes à mobiliser</i>	* Artefacts de contextualisation de l'aide à la décision * Système et son environnement * Données disponibles	* Artefacts de contextualisation de l'aide à la décision * Méthodes statistiques	* Artefacts de contextualisation de l'aide à la décision * Système et son environnement

*Tableau 4. Structuration d'une démarche de modélisation et construction et évolution des connaissances organisationnelles.*

Pour constituer une connaissance nouvelle, les acteurs de l'organisation apporteront principalement la connaissance du système et de son environnement, tandis que l'analyste/le chercheur apportera sa connaissance des méthodes statistiques. Pour construire le modèle, tous pourront s'appuyer sur les artefacts de contextualisation de l'aide à la décision préalablement formalisée. Ces artefacts permettront de concrétiser les phases de quantification, mathématisation et validation, à partir d'une application bien définie de l'aide à la décision qui guidera le choix des données, des méthodes statistiques et les interprétations.

## **Conclusion du chapitre 2**

Ce chapitre a fourni une trame générale visant à répondre à la problématique générale de cette thèse. A partir d'un ensemble de considérations théoriques relatives aux processus d'aide à la décision et à la gestion des connaissances industrielles, un processus socio-statistique d'aide à la décision a été défini et spécifié.

Ce processus vise à formaliser des connaissances organisationnelles utiles à la prise de décision. Pour cela, il suggère de tenir compte des différents aspects de l'ontologie du système de connaissances, aussi bien pour la constitution d'un modèle statistique qui pourra être mis en usage, que pour la contextualisation de la définition de ce modèle. Ce processus générique est autant social que mathématique, et implique des analyses quantitatives et qualitatives. La dynamique du système considéré a également été prise en compte : des mécanismes de contrôle et de maintien à niveau de la connaissance ont été suggérés, pour intégrer l'évolution du système dans le temps. Cette proposition est conforme à la complémentarité entre les besoins décisionnels considérés dans cette thèse, tels qu'évoqués dans le premier chapitre.

Un certain nombre d'éléments de ce processus socio-statistique d'aide à la décision restent non détaillées : acteurs à mobiliser, choix des variables, des méthodes statistiques, etc. : c'est parce qu'ils dépendent d'une application locale du processus. Les deux parties suivantes vont permettre d'apporter ces indications complémentaires. Elles montrent en effet comme le processus socio-statistique d'aide à la décision a pu être appliqué à un cas concret d'application chez STMicroelectronics.

**Deuxième partie.** Application du processus socio-statistique d'aide à la décision : analyse du contexte organisationnel et structuration d'une démarche de modélisation.

## Plan détaillé de la deuxième partie.

### Chapitre 3. Introduction : présentation générale du contexte d'application STMicroelectronics.

#### Introduction.

#### Section 1. Gestion de l'infrastructure des technologies de l'information à Crolles 2 et enjeu de gestion

1.1. Enjeu n°1 : gérer un système complexe.

1.2. Enjeu n°2 : soutenir pleinement l'activité de l'organisation.

#### Section 2. Limites de mise en œuvre de la gestion des capacités des technologies de l'information à Crolles 2.

2.1. Une approche statistique.

2.2. Limites de l'exploitation de données.

#### Conclusion du chapitre 3.

### Chapitre 4. Analyse du contexte organisationnel : étude de cas.

#### Introduction.

#### Section 1. Réalisation d'une étude de cas.

1.1. Une étude *a posteriori* : un cas particulier d'application du processus socio-statistique d'aide à la décision.

1.2. Création de connaissances et identification du périmètre industriel d'application des recherches.

##### 1.2.1. Création de connaissances.

*i. Récolte de données (aspect syntaxique).*

*ii. Structuration des données – informations (aspect sémantique).*

*iii. Interprétation des informations (aspect pragmatique).*

##### 1.2.2. Aperçu général du cas et définition du périmètre industriel d'application des recherches.

*i. Aperçu général du cas.*

*ii. Définition du périmètre industriel d'application des recherches.*

#### Section 2. Situation problématique et problèmes de décision : présentation du périmètre industriel d'application des recherches.

2.1. Représentation de la situation problématique.

2.2. Formulation des problèmes de décision et limites d'application.

##### 2.2.1. Cœur des recherches : choix du dimensionnement des composants d'infrastructure (activité ITHL (Q) Modélisation et analyse des tendances).

*i. Formulation du problème de décision.*

*ii. Limites des pratiques en usage lors de la migration MES.*

##### 2.2.2. Travaux complémentaires : identification du risque de sous-capacité (détection de comportements anormaux dans le cadre de l'activité (K) Contrôle).

*i. Formulation du problème de décision.*

*ii. Limites des pratiques en usage lors de la migration MES.*

#### Conclusion du chapitre 4.



## Chapitre 5. Structuration d'une démarche de modélisation destinée à répondre aux problèmes.

### Introduction.

#### Section 1. Structuration d'une démarche de modélisation pour la constitution d'un modèle de prévision ( $II_1$ : dimensionnement des composants d'infrastructure).

1.1. Principes généraux : mobilisation des expertises locales et structuration progressive de connaissances organisationnelles.

*1.1.1. Mobilisation des expertises locales.*

*1.1.2. Structuration progressive des connaissances concernant les interactions entre système des technologies de l'information et environnement industriel.*

1.2. Structuration progressive des connaissances : méthodes quantitatives employées.

*1.2.1. Introduction.*

*1.2.2. Méthodes descriptives, utilisées pour l'étape (A) intermédiaire d'observation n°1.*

*1.2.3. Méthodes explicatives et prédictives, utilisées pour l'étape (B) intermédiaire d'observation n°2 et l'étape (C) d'aide à la décision.*

*1.2.4. Méthodes d'évaluation et de contrôle, utilisées pour l'étape (B) intermédiaire d'observation n°2 et l'étape (C) d'aide à la décision.*

1.3. Synthèse : démarche de modélisation pour la constitution du modèle de prévision attendu.

#### Section 2. Contributions à la détection de comportements anormaux ( $II_2$ : identification du risque de sous-capacité).

2.1. Apport attendu des travaux de recherche menés.

*2.1.1. Contexte général.*

*2.1.2. Axes d'améliorations.*

*i. Amélioration de la robustesse du contrôle.*

*ii. Prise en compte des changements structurels.*

*iii. Mise en œuvre d'un contrôle multivarié.*

2.2. Solutions quantitatives mises en œuvre.

*2.2.1. Amélioration de la robustesse du contrôle : présentation de l'approche implémentée et évaluation théorique de la performance.*

*i. Approche implémentée.*

*ii. Évaluation théorique de la performance.*

*2.2.2. Prise en compte des changements structurels : présentation de l'approche implémentée.*

*2.2.3. Mise en œuvre d'un contrôle multivarié : présentation de l'approche implémentée.*

### Conclusion du chapitre 5.

\*  
\* \*

Cette partie concerne l'application de l'ensemble des étapes préparatoires à la formalisation d'une connaissance, selon un modèle statistique, impliquées par le processus socio-statistique d'aide à la décision. La Figure 21 indique le plan général de cette deuxième partie de la thèse, qui inclut les chapitres 3, 4 et 5.

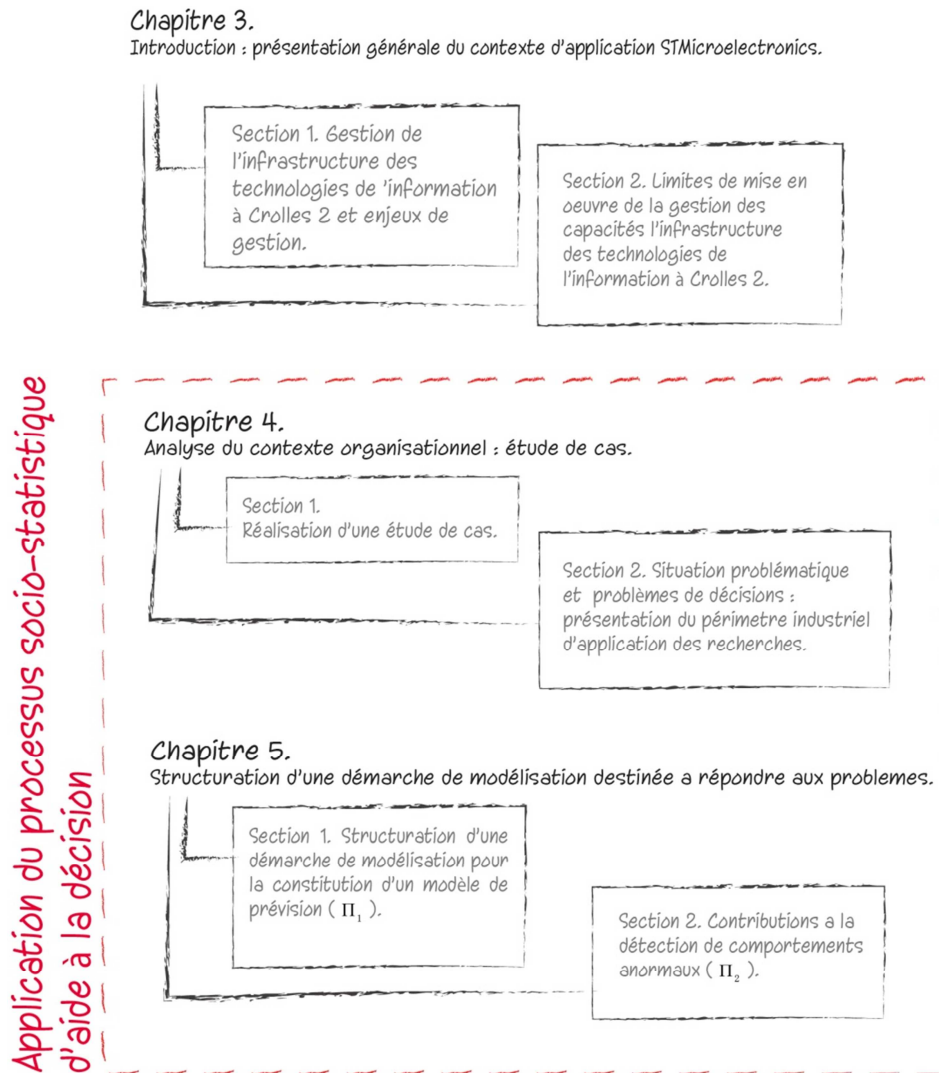


Figure 21. Plan général de la partie 2.

Avant l'application directe du processus, une introduction au contexte STMicroelectronics est proposée (chapitre 3) : on montre que d'une part, les managers de l'infrastructure TI-C2 sont confrontés aux enjeux de gestion mis en exergue précédemment (section 1) et que d'autre part, ils rencontrent un certain nombre de difficultés pour construire et déployer un processus de gestion des capacités (section 2). Pour ces raisons, l'application du processus socio-statistique d'aide à la décision devrait permettre d'améliorer cette situation.

En conséquence, les chapitres 4 et 5 proposent une application du processus à un cas STMicroelectronics. La Figure 22 indique comment ces deux chapitres, ainsi que leurs sections, se situent au sein du processus socio-statistique d'aide à la décision.

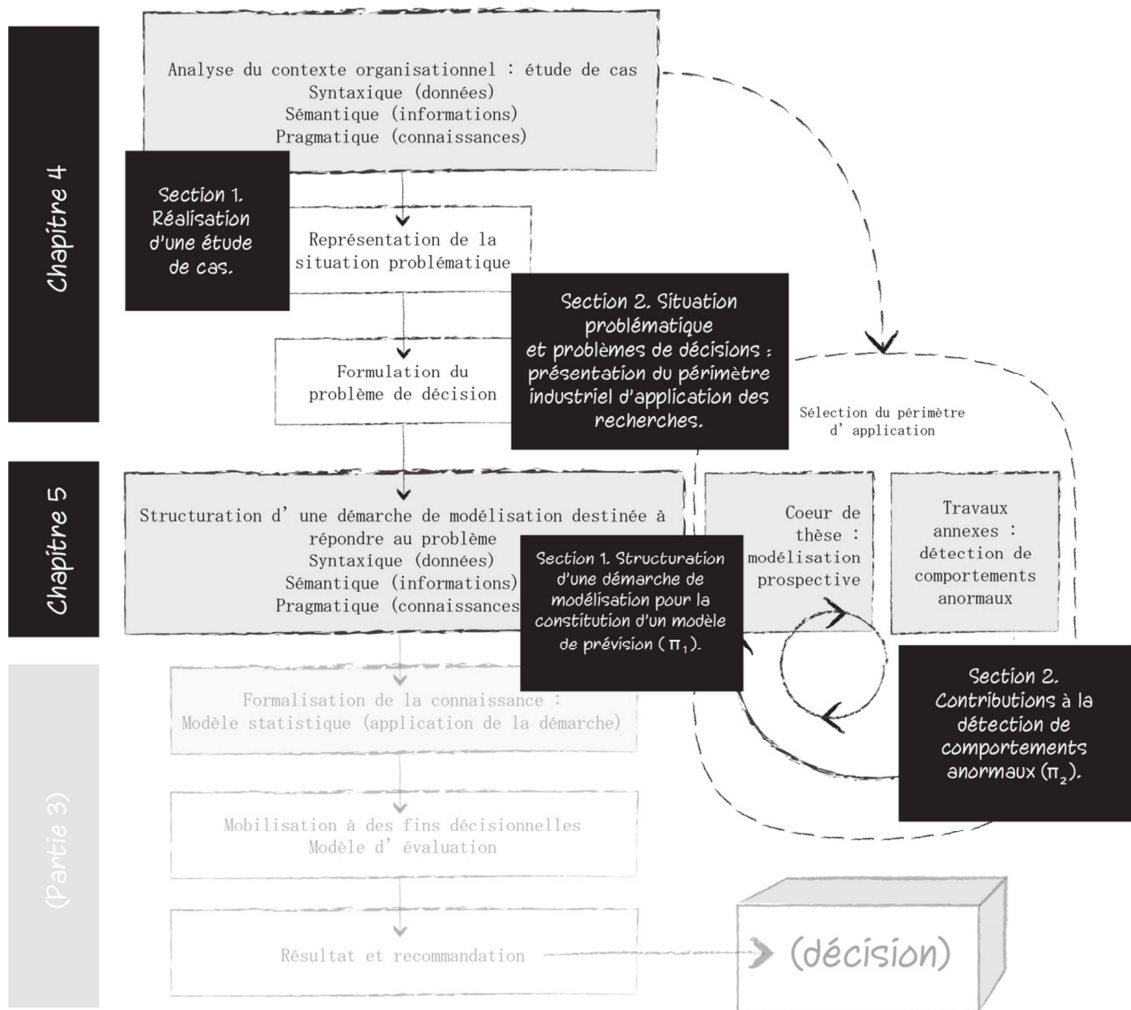


Figure 22. InSCRIPTION des chapitres 4 et 5 au sein du processus socio-statistique d'aide à la décision.

Ainsi, le chapitre 4 présente l'application de l'ensemble de l'étape de contextualisation du processus, à un cas STMicroelectronics. La section 1 détaille la manière dont une connaissance a été construite pour aboutir à la constitution d'une étude de cas. Cette étude de cas permet de définir le périmètre d'application industrielle des recherches. La situation problématique et les questions décisionnelles afférentes, peuvent alors être formulées.

Conformément au schéma du processus socio-statistique d'aide à la décision, une démarche de modélisation destinée à répondre aux problèmes de décision est ensuite structurée. C'est l'objet du chapitre 5, qui se décompose en deux temps : d'abord, il traite de la structuration de la démarche de modélisation, qui permettra la constitution d'un modèle de prévision (cœur des recherches) ; puis, en complément, il présente un ensemble de contributions additionnelles, proposées pour améliorer la détection de comportements anormaux (travaux annexes).

Ces propositions pourront ensuite être appliquées, pour améliorer concrètement la gestion des capacités chez STMicroelectronics. Cette application sera développée dans la troisième partie de la thèse.

\*  
\* \*

**Chapitre 3. Introduction : présentation générale du contexte  
d'application STMicroelectronics.**

## Introduction

Ce chapitre introductif fournit des précisions quant au contexte général d'application du processus socio-statistique d'aide à la décision chez STMicroelectronics. La Figure 23 en présente le plan.

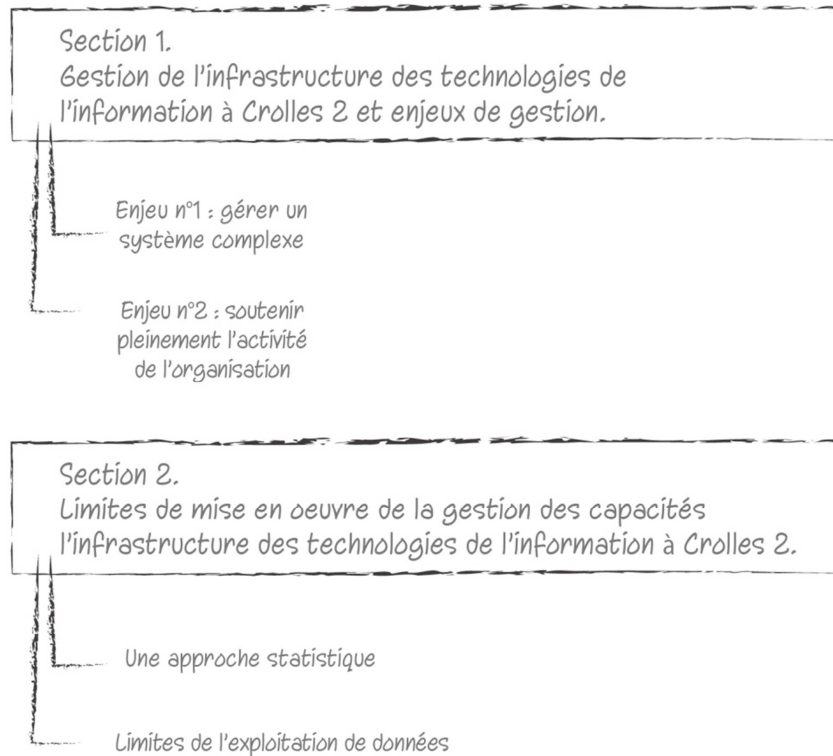


Figure 23. Plan du chapitre 3.

Dans un premier temps (section 1), les enjeux de gestion, suggérés en première partie de thèse, sont comparés aux conditions d'exercice des missions des managers de l'infrastructure TI-C2. Nous constaterons que ces derniers sont en effet confrontés à la nécessité de gérer un système complexe, tout en devant prendre pleinement en considération l'activité métier de l'organisation.

La section 2 montre ensuite que les pratiques managériales, mises en œuvre pour gérer les capacités de l'infrastructure des technologies de l'information de Crolles 2, pourraient être améliorées. Même si l'on observe une volonté de gérer le système par la mise en œuvre d'une approche statistique, on constate que les pratiques en usage restent aujourd'hui insuffisantes. C'est pourquoi les travaux présentés dans cette thèse devraient permettre d'apporter un progrès significatif.

## Section 1. Gestion de l'infrastructure des technologies de l'information à Crolles 2 et enjeux de gestion.

### 1.1. Enjeu n°1 : gérer un système complexe

Pour soutenir l'activité du système d'information, les managers de l'architecture de l'infrastructure TI-C2 gèrent un système technologique complexe distribué de type client/serveur. Pour l'ensemble des domaines fonctionnels couverts, il englobe plus de 60 applications, environ 430 serveurs physiques et plus de 12 téraoctets de données, contenues dans des bases de données applicatives. Les serveurs sont généralement des *clusters* composés de nœuds multiprocesseurs, à plusieurs cœurs. Les données sont majoritairement gérées par de nombreuses bases Oracle. Chacun de ces composants peut être tantôt client ou serveur d'autres composants, qui interagissent en temps réel grâce à une solution *middleware* adaptée.

A l'heure où cette thèse est rédigée, il est difficile de donner un aperçu synthétique de la complexité de cette architecture : un projet de cartographie du système est encore en cours de développement. A titre d'exemple, la Figure 24 représente une illustration possible d'une partie de l'architecture de l'infrastructure des TI, supportant une partie de l'architecture applicative gérée par le service *Manufacturing Execution System* (cf. Introduction générale de la thèse).

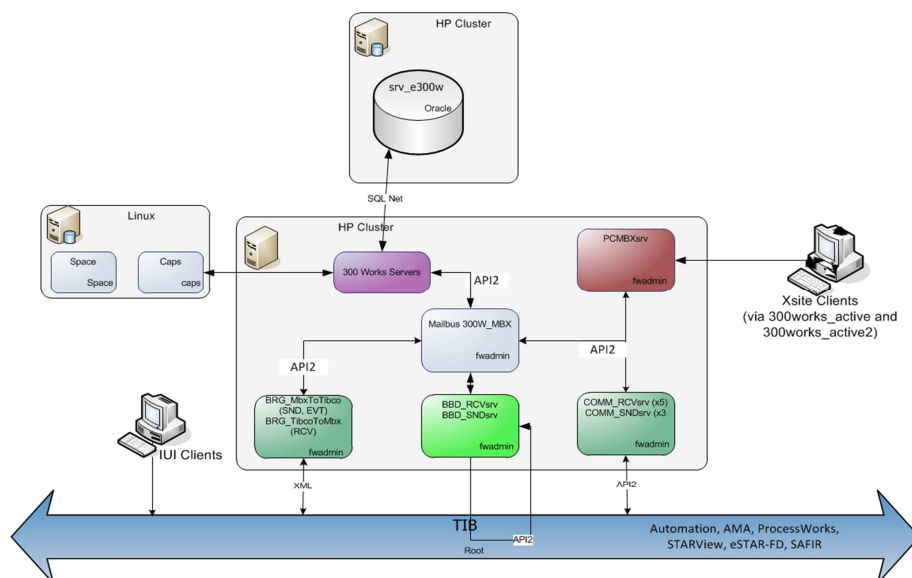


Figure 24. Exemple d'une partie de l'architecture de l'infrastructure des TI de Crolles 2.

La Figure 24 montre qu'un *cluster* HP héberge des serveurs applicatifs *300Works* (une des applications du service *Manufacturing Execution System*). Ces serveurs communiquent directement avec un serveur Linux supportant les applications *Space* et *Caps* (service *Process Control and Automation*). Les serveurs *300Works* communiquent également avec un *cluster* HP de base de données, pour accéder à une base nommée *srv\_e300w*. D'autres serveurs assurent la communication vers des PC<sup>32</sup> utilisateurs (clients *Xsite*, gérés par le

<sup>32</sup> Personal Computers

service *Manufacturing Execution System*) et vers un bus TIBCO, qui assure la communication entre *300Works* et plusieurs autres applications.

Cette illustration, qui ne montre qu'une petite partie de l'infrastructure des technologies de l'information du SI de Crolles 2, semble suffisamment convaincante pour faire comprendre que l'on est bien face à un système technique complexe, répondant à la description de Mitleton-Kelly et Land présentée dans le premier chapitre. Face à ce système complexe, les managers de l'infrastructure TI-C2 doivent être en mesure de se doter des bons moyens de gestion et de prise de décision, notamment pour la gestion des capacités des capacités informatiques, comme nous le préciserons plus loin.

## **1.2. Enjeu n°2 : soutenir pleinement l'activité de l'organisation**

Incontestablement, l'usage du système d'information est profondément ancré dans toutes les étapes de l'activité métier de l'usine Crolles 2. Les activités métiers supportées par le système d'information de Crolles 2 sont résumées dans le Tableau 5, selon les principaux groupes d'applications. Ces groupes sont gérés par les services en charge de l'architecture applicative, présentés lors de l'introduction générale de la thèse.

<b>Groupe d'applications</b>	<b>Service en charge de l'architecture applicative</b>	<b>Activités métiers supportées</b>
SAFIR ( <i>Single Application Fab InteRface</i> )	<i>Decisional Solutions</i>	Gestion de l'activité de production, de la maintenance des équipements et gestion du modèle de production (environnement <i>web</i> )
SPACE	<i>Process Control and Automation</i>	Gestion des alarmes des équipements
R2R ( <i>Run to Run</i> )	<i>Process Control and Automation</i>	Métriologie et ajustement dynamique du processus de production
Reporting/TGV ( <i>Tool for Global Visibility</i> )	<i>Decisional Solutions</i>	<i>Datawarehouse</i> : collecte de données en provenance de la fabrication, traitement et historisation sous la forme d'indicateurs servant au pilotage de l'activité
MES ( <i>Manufacturing Execution System</i> )	<i>Manufacturing Execution System</i>	Gestion de production : suivi de l'état des lots en production, des équipements de production, archivage de l'historique des données de production, stockage de l'ensemble des processus de production...
<i>Automation</i>	<i>Process Control and Automation</i>	Gestion de la communication entre les équipements de production et le reste du système d'information
FDC ( <i>Fault Detection and Classification</i> )	<i>Process Control and Automation</i>	Collecte de données de production et contrôle des dérives et de la conformité de la production
MFA ( <i>Move to Full Auto</i> )	<i>Process Control and Automation</i>	Automatisation de l'activité de production : <i>dispatching</i> (assignation d'un lot à un équipement), <i>scheduling</i> (planification de la production) et <i>monitoring</i> (détection d'anomalies). Implique la gestion des équipements, des lots de production et du système de transport
MACS (MURATEC AMHS - <i>Automated Material Handling System - Control System</i> )	<i>Process Control and Automation</i>	Gestion des transports physiques des lots de production
RMS ( <i>Recipe Management System</i> )	<i>Process Control and Automation</i>	Gestion des versions de recettes équipement
<i>Alarm Management</i>	<i>Process Control and Automation</i>	Gestion des alarmes émises (équipement, FDC ou automation) et pilotage des actions en conséquence, sur les équipements
<i>Klarity</i>	<i>Process Control and Automation</i>	Stockage et analyse des données concernant les défauts de production (recherches de causes, analyse de tendances...)

Tableau 5. Activités métier supportées par le système d'information de Crolles 2.

Ces groupes d'applications produisent et utilisent les 12 téraoctets de données applicatives évoquées précédemment. Ces données correspondent à la représentation informatisée de l'activité de l'usine.



Par ailleurs, tous ces groupes d'applications interagissent en temps réel, comme le représente la Figure 25.

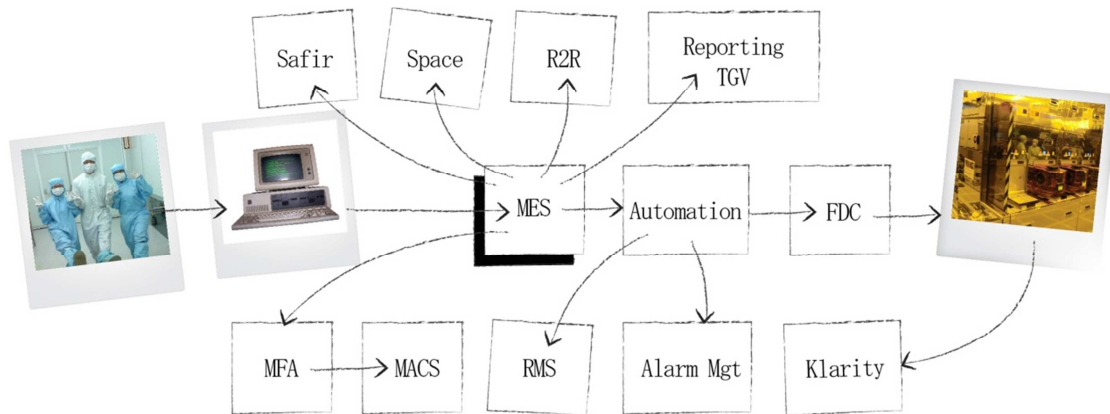


Figure 25. Interactions entre les groupes d'applications sous-jacentes à l'activité métier de Crolles 2.

Par exemple, le traitement d'un lot de production recourra à des applications du groupe MES. Les groupes *Automation*, FDC, SPACE leur fourniront un certain nombre de données, nécessaires au traitement. De plus, si le traitement est réalisé de façon automatique, la gestion du lot et de l'équipement de production sera réalisée *via* les applications MFA, et le groupe MACS sera sollicité pour charger physiquement le lot dans l'équipement. Pour cela, il faudra s'appuyer sur des données fournies par le groupe *Reporting/TGV*. En conséquence, il est indispensable d'assurer sans interruption le service fourni par les technologies de l'information, conformément aux attentes des clients et sur tous les domaines applicatifs, afin de garantir le bon fonctionnement du système productif. Toute rupture d'activité de l'un des groupes applicatifs peut mettre en péril l'ensemble du fonctionnement de l'usine de Crolles 2.

Si les services en charge de l'architecture applicative sont responsables du maintien et du développement fonctionnel des solutions applicatives, c'est le service infrastructure TI-C2 qui assure une infrastructure technologique globale cohérente et apte à soutenir l'ensemble du système d'information. Aussi est-il fondamental que, lors de toute prise de décision quant à l'infrastructure des TI, ce service puisse établir des liens clairs entre le système technique dont il a la charge et les différents pans de l'activité métier de l'usine et de son SI.

Nous pouvons donc affirmer que les managers de l'infrastructure TI-C2 sont bien confrontés aux enjeux de gestion qui ont été identifiés. Les questions suivantes se posent alors : quels moyens ces managers ont-ils mis en œuvre pour intégrer ces enjeux dans leurs procédures de gestion des capacités ? Ces moyens sont-ils suffisants ?

## Section 2. Limites de mise en œuvre de la gestion des capacités des technologies de l'information à Crolles 2.

### 2.1. Une approche statistique

Le service infrastructure TI-C2 a pour volonté de déployer les principes du référentiel ITIL dans ses pratiques. Une politique soutenue de formation et de certification ITIL de son personnel en atteste.

Concernant la mise en œuvre du processus de gestion des capacités, les managers de l'infrastructure TI-C2 ont perçu l'intérêt d'établir une approche statistique. Cela s'est concrétisé en 2008 par la mise en place de ce qu'ils nomment la « base de performance », élément central de leurs activités de *capacity planning*. Cette base de données s'appuie sur un mécanisme de collecte de données d'activité de l'infrastructure des TI. Ce mécanisme analyse automatiquement les fichiers de *log* générés par les applications (*parsing*)<sup>33</sup>, ainsi qu'un certain nombre de rapports concernant l'activité des composants de l'infrastructure (sar Linux<sup>34</sup>, statistiques Oracle<sup>35</sup>, etc.). Deux niveaux de collecte de données, représentatifs de l'activité du système des technologies de l'information, sont considérés. Ils permettent de caractériser :

- *L'activité de l'infrastructure des TI*, correspondant à l'activité des divers composants informatiques (utilisation des serveurs, tailles des bases de données, etc.) ;
- *L'activité de l'architecture applicative*, correspondant à l'activité transactionnelle du système d'information (volume de transactions utilisées, temps de réponse, etc.).

Un serveur dédié traite ces mesures pour générer des statistiques concernant l'activité du système : totaux, valeurs moyennes, maximums, quantiles 90, selon différentes agrégations temporelles (journée, semaine, mois...). Ces statistiques sont stockées dans la base de performance. Ce mécanisme est représenté en Figure 26.

---

<sup>33</sup> Un fichier de *log* est un fichier enregistrant tous les événements recensés par un serveur ou une application. Par exemple, pour une application, un fichier de *log* pourrait contenir : les noms des transactions exécutées, associés aux utilisateurs les ayant appelées, les temps de début et de fin de traitement, les serveurs utilisés, les éventuelles erreurs, etc. Le *parsing* de ces fichiers permet d'en extraire des données utiles (dans le cas de notre exemple, ce pourrait être le calcul du temps d'exécution de chacune des transactions exécutées).

<sup>34</sup> Cette commande Linux fournit un certain nombre de statistiques sur l'activité du système d'exploitation, notamment l'utilisation des processeurs (LINUX, 2012)

<sup>35</sup> Quelques exemples : *CPU used by this session* (temps processeur utilisé par Oracle lors des appels utilisateurs), *RedoSize* (activité en écriture sur la base), *SessionLogicalReads* (activité en lecture sur la base), etc. Voir ORACLE (2012) pour un aperçu exhaustif.

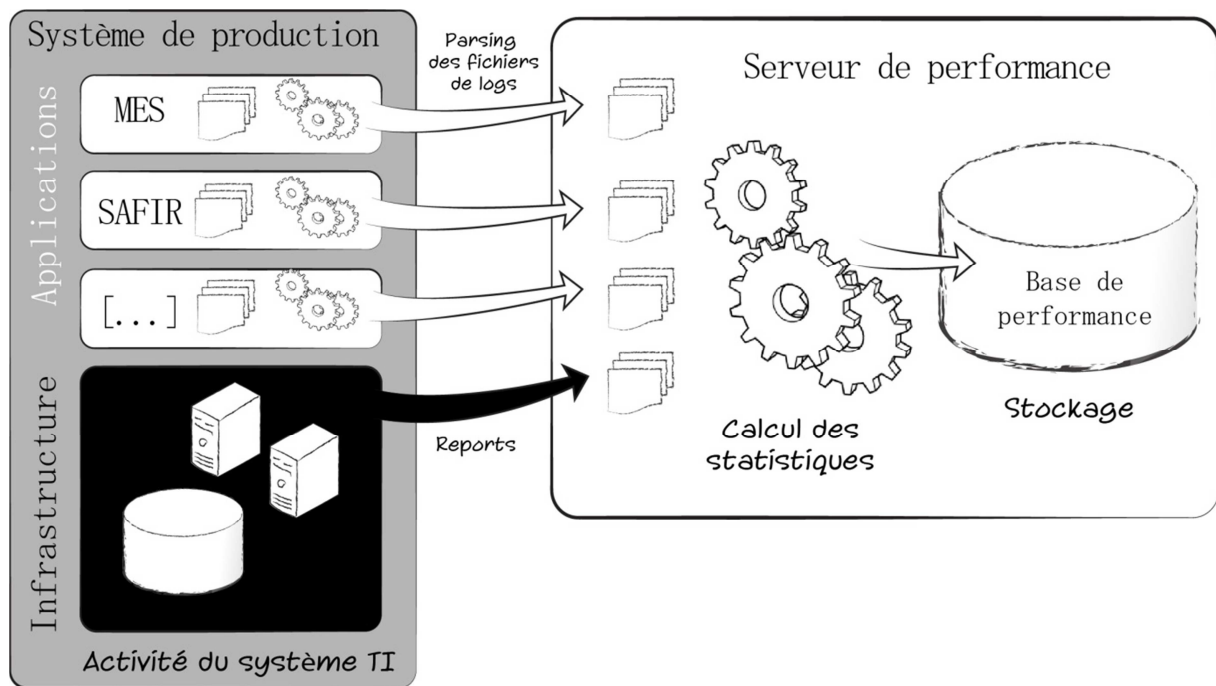


Figure 26. La base de performance de Crolles 2.

Plusieurs centaines de variables sont ainsi stockées, pour un volume de données d'environ 370 giga-octets, régulé par un mécanisme de purge automatique.

Nous analysons ci-après certaines limites de cette approche statistique, constatées au début de nos travaux de recherches.

## 2.2. Limites de l'exploitation de données

La Figure 27 est un exemple typique de la façon dont la base de performance était exploitée par le service infrastructure TI-C2 pour prendre des décisions, avant le début de nos travaux.

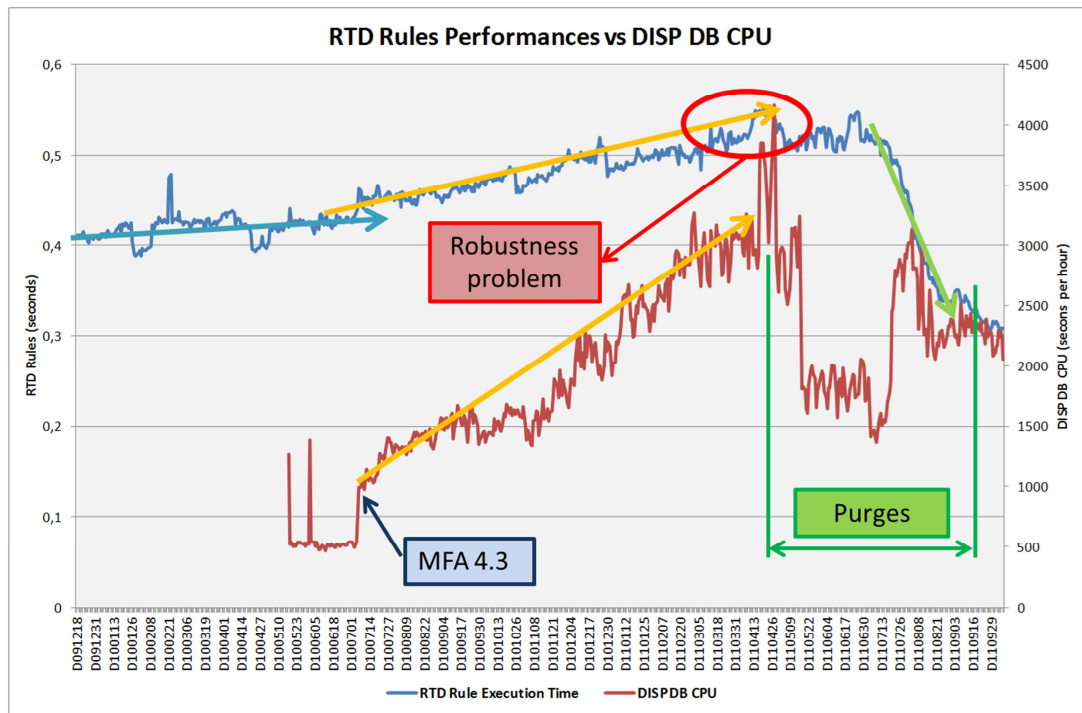


Figure 27. Exemple d'exploitation des données à Crolles 2.

Qualité de service (temps de réponse applicatif) et utilisation d'un serveur sont tracés dans un même graphique, discuté entre experts techniques. Cet exemple illustre bien les possibilités d'amélioration des pratiques de gestion des capacités. En effet, le recours aux méthodes statistiques paraît insuffisant. La Figure 27 est une simple représentation graphique des données disponibles, effectuée avec Microsoft Excel. Ce mode d'utilisation des données est l'approche prédominante chez les experts du service infrastructure TI-C2, lorsque nos recherches débutent. Nous n'observons pas de travaux structurés visant à quantifier formellement des relations ou des régularités dans les données recueillies. Or, comme indiqué précédemment, des méthodes statistiques correctement employées et bien contextualisées pourraient certainement apporter une aide à la gestion du système informatique complexe de Crolles 2, en tenant compte de sa situation industrielle particulière. Les pratiques de *capacity planning* en usage semblent donc tout à fait améliorables.

Précisons que cette critique des pratiques décisionnelles du service infrastructure TI-C2 ne doit pas être perçue comme une volonté de dénigrer son travail. Avant nos travaux, les décisions sont prises principalement en s'appuyant sur l'avis des experts du service. Ayant des compétences techniques excellentes et une connaissance profonde du système en place, ils sont capables de donner intuitivement des recommandations pertinentes, sans forcément avoir recours à des analyses quantitatives approfondies.

Toutefois, la volonté des managers d'améliorer l'exploitation des données, pour leurs activités de *capacity planning* n'en est pas moins forte : un meilleur usage des données permettrait de compléter les avis des experts et d'augmenter le poids de leurs arguments, lorsqu'ils communiquent avec d'autres entités de STMicroelectronics. Il permettrait également de mieux partager et communiquer les connaissances des experts, au travers de représentations formelles du système.

### **Conclusion du chapitre 3**

Nous avons vu que l'usine de Crolles 2 est concernée par les enjeux de gestion des technologies de l'information : gérer un système technologique complexe, tout en garantissant le bon fonctionnement des processus métier, est un défi quotidien pour les managers de l'infrastructure TI-C2. Pour cela, ils tentent de renforcer leurs pratiques de gestion des capacités, par le déploiement d'une démarche basée sur l'exploitation de données. Toutefois, il semble que cette volonté mérite d'être soutenue plus franchement, grâce à un perfectionnement des pratiques en usage.

Le processus socio-statistique d'aide à la décision vise à apporter une telle amélioration. C'est pourquoi il va désormais être appliqué à un cas réel de STMicroelectronics. Cette application débute dans le chapitre suivant.

## **Chapitre 4. Analyse du contexte organisationnel : étude de cas.**

## Introduction

Ce chapitre présente l'application de la phase préliminaire du processus socio-statistique d'aide à la décision, dont l'objectif est de contextualiser l'aide à la décision à développer. Le plan du chapitre 4 est indiqué dans la Figure 28.

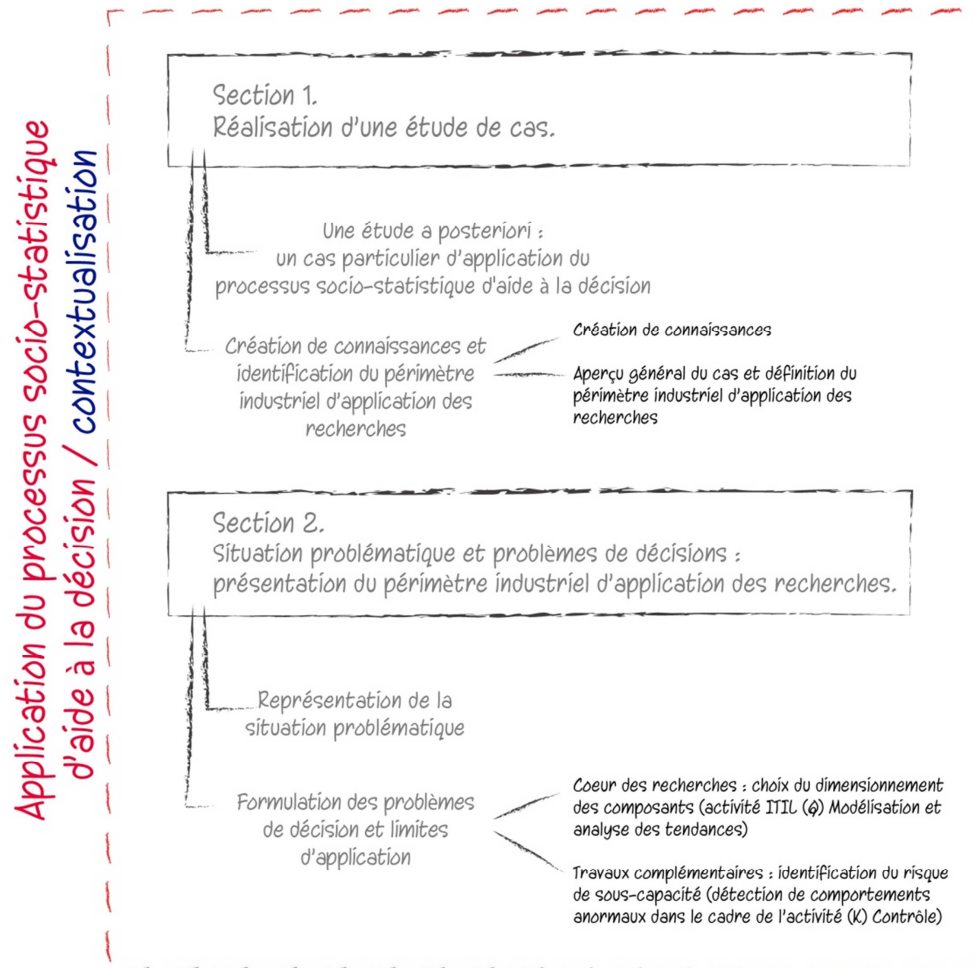


Figure 28. Plan du chapitre 4.

Tout d'abord, la section 1 présente comment une étude de cas est constituée. C'est une analyse *a posteriori* d'un cas passé qui sera présentée ici. Cette spécificité permet de démontrer en quoi le processus socio-statistique d'aide à la décision aurait pu être utile lorsque s'est déroulé le cas considéré. Après avoir précisé le contexte et la particularité de cette analyse, nous décrivons la façon dont une connaissance est créée grâce à la constitution d'une étude de cas, selon les trois aspects de l'ontologie du système de connaissances. Ce cas présente la propriété d'être constitué d'un enchaînement de situations décisionnelles successives. Les besoins d'aide à la décision qu'elles impliquent sont alors hiérarchisés. Ensuite, un contexte décisionnel circonscrit, auquel sera appliquée la suite du processus d'aide à la décision, est défini. C'est à ce stade qu'est véritablement identifié le périmètre industriel d'application de nos recherches, déjà évoqué tout au long de cette thèse : la modélisation prospective (besoin principal) et la détection de comportements anormaux (besoin secondaire).

Une fois ce périmètre borné, la situation problématique et les problèmes de décision des deux besoins traités sont décrits, conformément aux propositions de Tsoukiàs (section 2). Pour chacun des problèmes de décisions, les modèles de rationalité des décideurs sont exprimés formellement. Nous expliquons également pourquoi ils n'ont pu être satisfaits lors de la réalisation effective du cas. Cette insuffisance justifie la nécessité de développer des réponses statistiques appropriée. C'est pour cela que la phase d'utilisation de modèles du processus socio-statistique d'aide à la décision sera développée dans le chapitre 5.



## Section 1. Réalisation d'une étude de cas.

### 1.1. Une étude *a posteriori* : un cas particulier d'application du processus socio-statistique d'aide à la décision

Le processus socio-statistique d'aide à la décision est appliqué à un cas bien connu à Crolles 2 : la gestion de l'infrastructure technologique du groupe d'application MES. Ce groupe est la colonne vertébrale de l'activité industrielle de Crolles 2. Il intervient à toutes les étapes du processus de production (cf. Tableau 5). Sa gestion pose un certain nombre de problèmes de décision, qui se sont particulièrement révélés lors d'une migration, finalisée en 2010 et qui a conduit à l'initiation des travaux présentés dans cette thèse. Cette migration visait à ajuster les capacités de l'infrastructure informatique de Crolles 2, en prévision d'une hausse de l'activité de production de l'usine. L'ensemble du projet s'est déroulé sur la période 2008-2010. La migration a entraîné de lourds changements concernant :

- *L'architecture applicative* : changements de versions du logiciel de gestion de production *300Works* et de plusieurs applications satellites ;
- *L'architecture de l'infrastructure des technologies de l'information* :
  - \* *Au niveau des bases de données* : changement de version Oracle (passage de 9i à 10g),
  - \* *Au niveau des serveurs physiques* : remplacement des serveurs physiques HP *Superdome* par une technologie plus récente HP *Itanium*.

Nous souhaitons analyser ce cas *a posteriori*. Il présente un intérêt décisionnel indéniable et non trivial. Y appliquer le processus socio-statistique d'aide à la décision permet en quelque sorte de « rejouer » la prise de décision, et de voir en quoi il aurait pu y apporter une contribution significative. Ce choix permet donc de comparer la connaissance générée grâce à la l'application du processus socio-statistique d'aide à la décision et les décisions prises à l'époque de la migration. Nous pouvons en conséquence discuter la pertinence et la valeur-ajoutée de l'aide à la décision fournie. Par ailleurs, remarquons que la connaissance générée a une portée prédictive, réutilisable pour traiter un ensemble d'occurrences de la même situation décisionnelle dans l'avenir.

La compréhension du contexte décisionnel de ce projet passé de migration est rendue possible par l'application de la première étape du processus socio-statistique d'aide à la décision, avec une légère modification du guide d'entretien général présentée dans le premier chapitre. En effet, le processus est censé s'appliquer en « temps réel », dans le vif de l'activité organisationnelle. Or nous l'utilisons ici pour effectuer une analyse *a posteriori* d'un cas. Cette migration a nécessité un processus décisionnel complexe, au sein duquel se sont intégrées les deux activités de modélisation prospective et de détection des comportements anormaux, considérées dans cette thèse. Pour mieux comprendre le déroulement de la migration dans son ensemble, nous avons cherché à en analyser le processus décisionnel dans sa globalité. Les deux activités étudiées dans cette thèse pourront ainsi être situées dans l'ensemble du déroulement du projet MES.

Conformément au processus socio-statistique présenté dans premier chapitre de la thèse, la première étape de son application consiste à créer une connaissance permettant de contextualiser l'aide à la décision.

## **1.2. Création de connaissances et identification du périmètre industriel d'application des recherches**

### *1.2.1. Création de connaissances*

#### *i. Récolte des données (aspect syntaxique)*

Afin de récolter les données nécessaires à la création de connaissances, il a fallu tenir compte de la dimension *a posteriori* du cas. Pour cela, le guide d'entretien, structuré autour de la représentation de la situation problématique et de la formulation du problème de décision, a été complété par les recommandations de Nutt (1993, 1998). Elles permettent de reconstituer un processus décisionnel aussi exhaustivement que possible, par l'emploi systématique de questions du type « que s'est-il passé avant/après ? ».

Cela a permis d'étudier l'enchaînement des choix réalisés lors de la migration MES, et d'aboutir à plusieurs représentations de situations problématiques et formulations de problèmes successives. Les activités de modélisation/analyse des tendances et de contrôle peuvent être situées dans cet enchaînement chronologique.

Six personnes ont été interrogées<sup>36</sup>. Pour des raisons de garantie d'anonymat, nous ne pouvons donner plus de détail à leur sujet. Les entretiens ont été menés en face-à-face, dans des bureaux isolés et calmes, pendant une durée approximative d'une heure.

#### *ii. Structuration des données - informations (aspect sémantique)*

Le matériau récolté a ensuite été analysé manuellement. Les thèmes se rapportant aux différents choix du processus décisionnel, fonctions du temps, ont d'abord été identifiés. Une arborescence de thèmes a ainsi été constituée. Elle ne peut être fournie dans son intégralité pour des raisons de confidentialité, mais la Figure 29 en donne un aperçu.

---

<sup>36</sup> Nous aurions aimé rencontrer plus de personnes, mais cela n'a pas été possible. Néanmoins, nous arrivons à saturation du cas sur plusieurs idées évoquées lors des entretiens.

CHOIX N°3 : DEFINITION DE LA PLATEFORME HARDWARE		Interviewés					
		#1	#2	#3	#4	#5	#6
<b>I Construction de scénarios de solutions techniques (à partir d'octobre 2008)</b>							
<b>A Participants</b>							
A1	#1	X	X		X		X
A2	Acteurs ICT site concernés	X	X		X		X
A3	Fournisseurs hardware (renseignements et cotations haut-niveau)	X	X				X
A4	#3				X		X
A5	IT corp. Central Solution (indirectement : fourni standards)		X		X		
<b>B Mode de travail</b>							
B1	Réunions régulières #1 / #3 et #4 / équipe MES)		X				X
B2	Durée du travail : 1 an		X				
B3	Plusieurs scénarios sont construits		X		X		
<b>C Objectif 1 : construction de scénarios opérationnels d'architecture</b>							
C1	Décision multi-critères	X	X		X		
C2	Critère : coûts d'achats et de maintenance	X	X		X		
C3	Critère : sécurité de l'architecture (fiabilité, stabilité, redondance en cas de panne : disponibilité et capacité à assurer le service attendu)	X			X		
C4	Critère : sécurité de la technologie (ne pas être cobaye d'une technologie pas suffisamment mature)	X			X		
C5	Critère : disponibilité de la technologie au jour de la migration	X			X		
C6	Critère : assurer une capacité suffisante (cf. D)	X			X		
C7	Critère : respect des standards corporate		X		X		
C8	Critère : solution connue par les équipes (limite les coûts d'administration et de formation des équipes ; facilite la maintenance et la réactivité en cas de problème)		X		X		
C9	Critère : évolutivité de la solution		X				
C10	Critère : criticité du domaine applicatif				X		
C11	Critère : capacité à répondre à la demande du client				X		
C12	Critère : capacité des équipes à mettre en œuvre la solution (compétence et temps de travail)				X		
C13	Critère : avenir des technologies				X		
C14	Critère : performance de la solution				X		
C15	Critère : limiter les risques de contamination en cas de problème (silos)				X		
C16	Pondération des critères : disponibilité, stabilité, réactivité de la maintenance et non-contamination sont les critères les plus importants. Le principal objectif des équipes IT est de faire tourner les fab et d'empêcher toute perturbation due à l'IT					X	
<b>D Objectif 2 : dimensionnement de la capacité des serveurs pour chacun des scénarios</b>							
D1	Travail commun #1 / #4		X		X		X
D2	Etape 1 : évaluation de la charge actuelle		X		X		X
D3	Etape 2 : dimensionnement des plate-formes futures		X		X		
D4	Hypothèse : les nouveaux serveurs devront être utilisés à moins de 30-40% de leur capacité maximal lorsque le FBO sera atteint		X				
D5	Ce seuil garantit un taux de service satisfaisant (dégradation à partir de 60-70%), tout en garantissant une marge de sécurité		X				

Figure 29. Analyse thématique du matériau récolté : extrait de l'arborescence constituée.

Ces thèmes ont ensuite pu être catégorisés, de manière à décrire les représentations des diverses situations problématiques et les formulations des problèmes conséquents, en fonction des décisions successives prises lors de la migration (Figure 30).

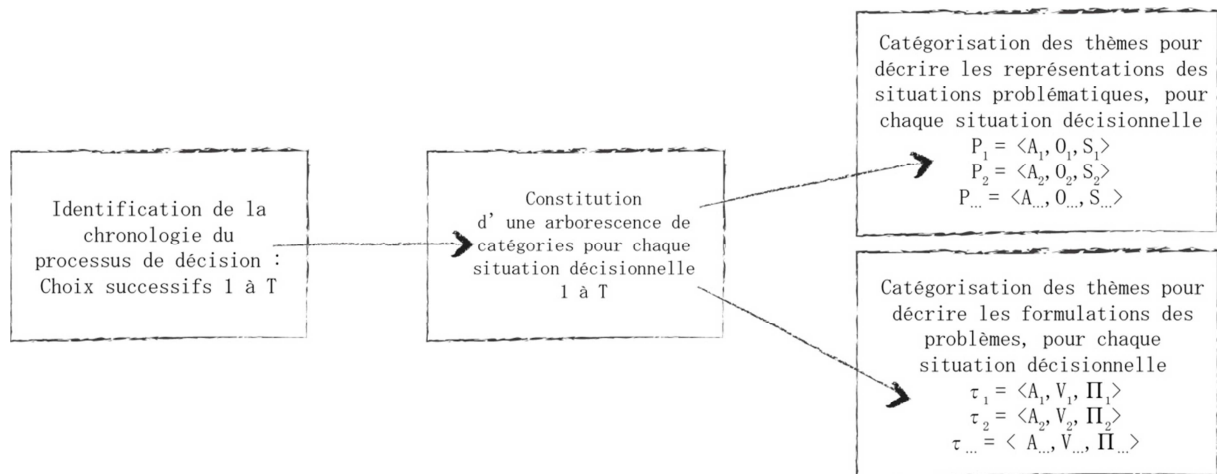


Figure 30. Méthode de structuration du matériau récolté.

iii. *Interprétation des informations - connaissances (aspect pragmatique)*

Une synthèse complète de cette analyse est finalement rédigée, sous la forme d'une étude de cas (non diffusable pour des raisons de confidentialité). Elle est partagée et discutée avec les interviewés, pour une validation finale.

Un certain nombre d'artefacts cognitifs sont ainsi constitués. Ils permettent de définir les différentes situations du processus décisionnel de la migration MES. Nous en donnons un aperçu général ci-après, sachant que toutes ne seront pas traitées.

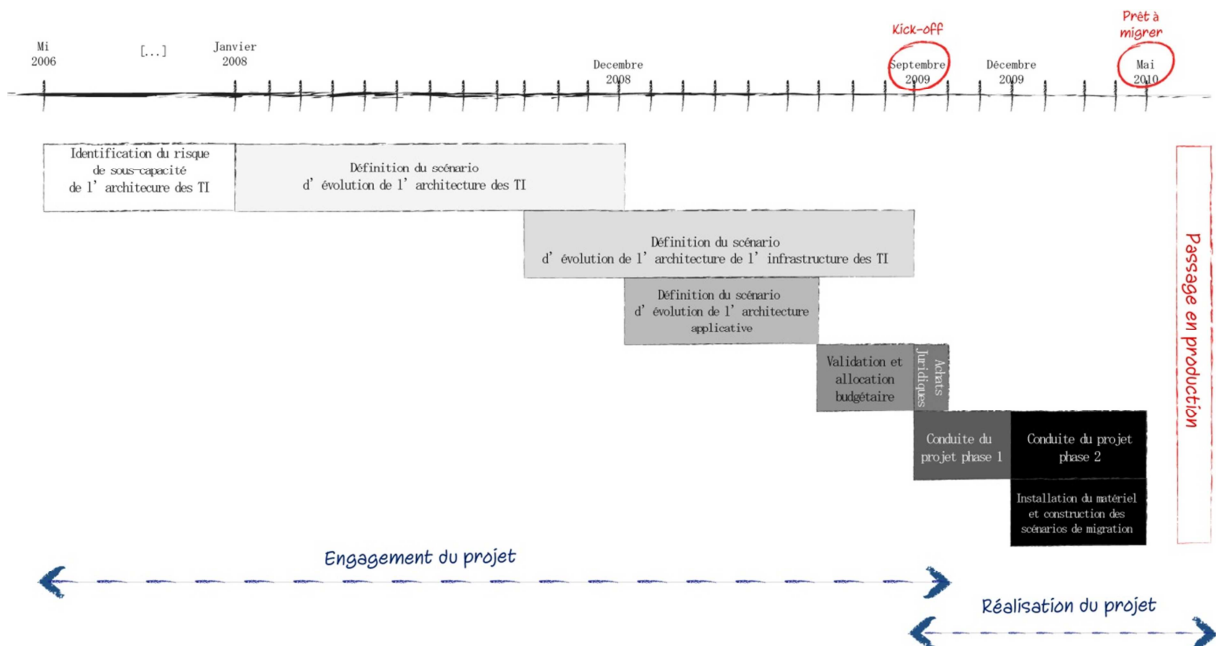
1.2.2. *Aperçu général du cas et définition du périmètre industriel d'application des recherches*

i. *Aperçu général du cas*

Neuf situations décisionnelles, constitutives du processus de migration, ont été identifiées. Elles peuvent être divisées en deux grandes phases :

- *L'engagement du projet*, qui consiste à identifier la nécessité de procéder à un ajustement et à définir la solution technique appropriée ;
- *La réalisation du projet*, qui consiste à mettre en œuvre concrètement la solution identifiée, la tester et l'ajuster, avant la migration (achat des composants, installations, tests, etc.) et après la migration (passage en production, c'est-à-dire mise à la disposition des utilisateurs).

La Figure 31 représente la chronologie de ces neuf situations décisionnelles.



Toutes ces situations ne nous intéressent pas. Tout d'abord seule la phase d'engagement est considérée : c'est à ce stade qu'est abordée la question de la gestion des capacités. Ensuite, la situation décisionnelle « *Définition du scénario d'évolution de l'architecture applicative* » n'est pas étudié en détail : dépendante surtout des services applicatifs, elle dépasse le périmètre organisationnel de nos travaux. Restent les trois situations décisionnelles suivantes, en rapport avec la gestion des capacités de l'infrastructure des TI :

- « *Identification du risque de sous-capacité de l'architecture des TI* » : la question est d'observer dynamiquement l'activité du système des technologies de l'information, afin de savoir si elle pourrait présager d'un risque de sous-capacité. Si un tel risque est identifié, il doit être maîtrisé et des décisions destinées à le prévenir doivent être prises ;
- « *Définition du scénario d'évolution de l'architecture des TI* » : cette situation consiste à décider des grandes lignes de l'ensemble de la solution technologique (applications et infrastructure), destinée à soutenir l'activité métier de l'usine. Deux étapes sont nécessaires :
  - \* *D'une part, la définition d'une architecture générale cohérente* : il faut assurer la compatibilité entre les différents composants (logiciels, bases de données, *middleware*, architecture des serveurs, etc.),
  - \* *D'autre part, la détermination d'une stratégie de migration* : c'est un compromis à trouver, entre charge de travail et maîtrise des risques liés à la migration ;
- « *Définition du scénario d'évolution de l'architecture de l'infrastructure des TI* » : cette situation implique le choix précis des composants de l'infrastructure des TI. C'est une décision en deux temps :
  - \* *Tout d'abord, le choix des technologies*, sur la base de critères tels que le coût, la performance, la sécurité, la pérennité, le respect des standards STMicroelectronics, etc.,
  - \* *Ensuite, le choix du dimensionnement des composants technologiques*, qui consiste à sélectionner, pour chacune des technologies retenues, des composants d'infrastructure ayant suffisamment de capacités pour soutenir une activité industrielle anticipée.

Cet aperçu général montre qu'un cas de migration, et une analyse de *capacity planning* en particulier, entraînent de multiples problèmes de décision. Pour en améliorer les pratiques, il est nécessaire de bien comprendre l'aide à apporter, donc de délimiter clairement des périmètres d'aide à la décision. C'est là toute la valeur-ajoutée de l'étape de contextualisation du processus socio-statistique d'aide à la décision.

ii. *Définition du périmètre industriel d'application des recherches*

Les trois situations retenues et leurs étapes ne posent pas toutes les mêmes difficultés en termes de prise de décision. Une hiérarchisation des besoins décisionnels est opérée sur la base de cette étude de cas, avec les managers de l'infrastructure TI-C2 (Tableau 6).

<b>Situation décisionnelle</b>	<b>Etape</b>	<b>Criticité du besoin d'aide à la décision</b>
<i>1. Identification du risque de sous-capacité de l'architecture des TI</i>	-	<b>Modérée</b> : maîtrisée mais amélioration possible
<i>2. Définition du scénario d'évolution de l'architecture des TI</i>	<i>a. Définition de l'architecture générale</i>	Faible : consensus facile à obtenir
	<i>b. Détermination de la stratégie de migration</i>	Faible : consensus facile à obtenir
<i>3. Définition du scénario d'évolution de l'architecture de l'infrastructure des TI</i>	<i>a. Choix des technologies</i>	Faible : consensus facile à obtenir
	<i>b. Choix du dimensionnement des composants d'infrastructure</i>	<b>Forte</b> : manque de maîtrise

*Tableau 6. Hiérarchisation de la difficulté des besoins décisionnels de l'étude de cas pour l'identification du périmètre d'application industriel des recherches.*

Cette hiérarchisation explique le choix des activités de *capacity planning* discutées dans le premier chapitre de thèse :

- *Un besoin décisionnel modéré est relevé pour la situation 1. Identification du risque de sous-capacité de l'architecture des TI* : ce besoin correspond tout à fait aux activités ITIL (K) Contrôle et (R) Gestion et contrôle des seuils. ;
- *Un besoin décisionnel fort est relevé pour l'étape b. Dimensionnement des composants d'infrastructure de la situation 3. Définition du scénario d'évolution de l'architecture de l'infrastructure des TI* : ce besoin consiste en l'activité ITIL (Q) Modélisation et analyse des tendances.

Ces deux besoins seront précisés plus loin. Dans les deux cas, et compte-tenu des limitations évoquées dans le chapitre 1, aucune solution prête à l'emploi ne peut être appliquée pour aider les managers de l'infrastructure TI-C2. Les deux activités peuvent en conséquence être améliorées, grâce l'application de la suite du processus socio-statistique d'aide à la décision (utilisation de modèles). Compte-tenu du caractère *a posteriori* du travail proposé, il sera possible d'en comparer les bénéfices par rapport aux pratiques effectivement en usage lors de la migration MES.

Nos recherches ont principalement porté sur le besoin décisionnel ayant une forte criticité, correspondant à l'activité ITIL (Q). C'est pourquoi il sera traité prioritairement dans l'ensemble de cette thèse et conditionne l'articulation de la présentation de nos travaux. Le besoin décisionnel de criticité modérée, relatif aux activités de contrôle, est quant à lui abordé à la marge, en tant qu'activité complémentaire consistant à maintenir à niveau le modèle de prévision.

## Section 2. Situation problématique et problèmes de décision : présentation du périmètre industriel d'application des recherches.

### 2.1. Représentation de la situation problématique

Grâce à l'étude de cas constituée, nous pouvons préciser la représentation de la situation problématique du périmètre industriel d'application. Elle est commune aux deux activités traitées dans cette thèse.

A Crolles 2, que cela soit pour dimensionner les composants de l'architecture des TI ou pour identifier un risque de sous-capacité, on observe un processus informel d'échanges entre le service applicatif MES et le service architecture TI-C2. Les managers du service infrastructure TI-C2 sont responsables de la décision finale. Ils s'appuient pour cela sur les experts techniques de leur service (experts Oracle, UNIX, architecture matérielle, etc.), et sur des éléments chiffrés issus de la base de performance. Les managers des services applicatifs fournissent quant à eux des éléments de réflexion suggérés par leurs propres experts applicatifs, qui peuvent eux-mêmes se baser sur l'analyse des bases de données applicatives dont ils ont la maîtrise. La Figure 32 résume ce processus.

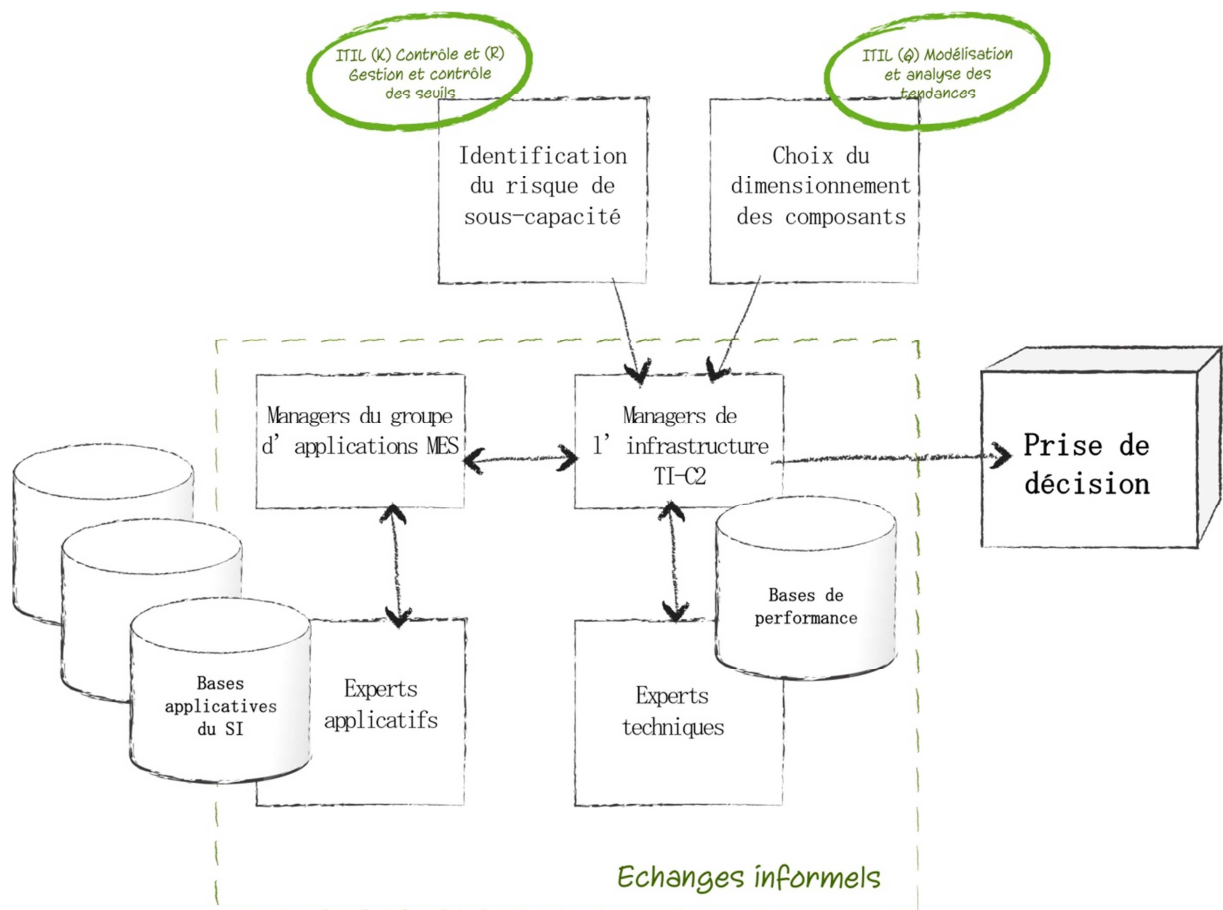


Figure 32. Représentation de la situation problématique : processus informel d'échanges et acteurs impliqués.

Chacun des services amène donc un ensemble de ressources propres, mais également des enjeux spécifiques : les services applicatifs auront plus tendance à favoriser la

performance applicative, tandis que le service infrastructure sera plus focalisé sur la cohérence globale de l'ensemble du système informatique. En conséquence, les éléments de la situation problématique, telle que définie par le cadre de Tsoukiàs, peuvent être résumés par le Tableau 7.

<b>A (Acteurs)</b>	<i>Managers du service infrastructure TI-C2</i>	<i>Experts techniques du service infrastructure TI-C2</i>	<i>Managers du service applicatif MES</i>	<i>Experts applicatifs du service applicatif MES</i>
<b>O (Objets / enjeux)</b>	Garantir une <u>architecture de l'infrastructure</u> des TI, fournissant un service satisfaisant, apte à soutenir l'activité présente et future de l'usine par un juste dimensionnement des composants de l'infrastructure		Garantir une <u>architecture applicative</u> fournissant un service satisfaisant, apte à soutenir l'activité présente et future de l'usine par un juste dimensionnement des composants de l'infrastructure	
<b>S (Engagements / ressources)</b>	Décisionnaire final ; Appui des experts techniques ; Relation avec les managers des services applicatifs ; Base de performance	Expertise technique (Oracle, Unix, matériel...) ; Base de performance	Appui des experts applicatifs	Expertise applicative et métier ; Base de données applicatives

Tableau 7. La gestion des capacités des TI à Crolles 2 : représentation de la situation problématique (P).

Après cette représentation de la situation problématique, nous pouvons fournir une formulation des problèmes de décision des décideurs (les managers du service infrastructure TI-C2), pour les deux besoins décisionnels considérés dans cette thèse, ainsi que leurs limites d'application lors du projet MES.

## 2.2. Formulation des problèmes de décision et limites d'application

### 2.2.1. Cœur des recherches : choix du dimensionnement des composants d'infrastructure (activité ITIL (Q) Modélisation et analyse des tendances)

#### i. Formulation du problème de décision

L'étude de cas permet de formaliser le problème de décision du besoin de dimensionnement des composants de l'infrastructure des TI. Les alternatives à considérer (élément *A* du problème de décision, selon le cadre de Tsoukiàs) sont les suivantes : à un horizon *b* donné, faudra-t-il oui ou non modifier l'infrastructure technologique et, si oui, dans quelle mesure ? Répondre avec justesse à cette question permet de bien dimensionner les ressources technologiques, afin d'éviter tout risque de sous-capacité (risque de rupture de service), mais également de surcapacité, qui n'est pas



plus souhaitable (surcoût, risques technologiques liés à l'acquisition de matériel trop récent, etc.).

Pour observer, analyser et comparer les alternatives potentielles ( $V$ ), les managers de l'infrastructure TI-C2 cherchent à évaluer un accroissement du besoin en ressource technologique, que nous appellerons  $F_b$ . Dans l'absolu, il peut varier de 1 à l'infini (besoin stable à besoin illimité).  $F_b$  correspond à un rapport entre une estimation du besoin en ressource technologique au moment où l'évaluation est réalisée ( $\widehat{Y}_0$ ) et une estimation à l'horizon  $b$  ( $\widehat{Y}_b$ ). Ces estimations doivent être fournies par un modèle de prévision permettant de prédire l'activité  $\widehat{Y}_t$  ( $t$  représente le temps) en fonction d'un certain nombre de variables  $X_{i,t}$  ( $i$  allant de 1 à  $n$ ) de pilotage de l'activité industrielle. Les variables  $X_{i,t}$  doivent être non triviales et connues de la Direction Générale de STMicroelectronics. Elles doivent favoriser le dialogue avec le management de haut niveau et la prise de décisions sur la base de scénarios industriels stratégiques. Ce modèle de rationalité est formalisé dans le Modèle 1.

$F_b = \frac{\widehat{Y}_b}{\widehat{Y}_0}$ <p style="text-align: center;">et</p> $\widehat{Y}_t = f(X_{i,t})$ <p style="text-align: center;">avec :</p> <p style="text-align: center;"><math>F</math> : facteur d'accroissement du besoin en ressource à estimer,</p> <p style="text-align: center;"><math>\widehat{Y}</math> : estimation du besoin en ressource technologique,</p> <p style="text-align: center;"><math>X_i</math> : variables industrielles de pilotage (<math>i : 1 \text{ à } n</math>)</p> <p style="text-align: center;"><math>t</math> : temps</p> <p style="text-align: center;"><math>0</math> : moment de réalisation de l'estimation,</p> <p style="text-align: center;"><math>b</math> : horizon temporel du besoin à estimer</p>
---

*Modèle 1. Modèle de rationalité du besoin décisionnel « dimensionnement des composants de l'infrastructure ».*

En conséquence, la problématique décisionnelle (II) peut être exprimée comme suit :

*$\Pi_1$  : comment constituer un modèle de prévision qui permettra de quantifier  $F_b$ , c'est-à-dire de prévoir le facteur d'accroissement d'un besoin en ressource technologique à l'horizon  $b$ , en fonction de variables industrielles de pilotage ?*

Cette formulation du problème de décision permet de constater que l'on est bien en présence d'une préoccupation très locale, jamais rencontrée sous cette forme dans la littérature dédiée au *capacity planning*. Une réponse *ad hoc* est donc souhaitable.

ii. *Limites des pratiques en usage lors de la migration MES*

Lors de la migration MES, l'évaluation de  $F_b$  s'est faite à partir d'une communication de la Direction Générale de STMicroelectronics : l'activité de l'usine devait passer d'un volume d'activité actuel estimé à 2000 plaques produites par semaine en moyenne (unité : wo/w, *wafers-out/week*), à 4500 à l'horizon  $b$ .

Pour intégrer cette information, le modèle de rationalité exprimé dans le Modèle 1 n'a pu être satisfait. Les managers de l'infrastructure TI-C2 ne disposaient pas de modèle de prévision, permettant d'estimer le besoin en ressources technologiques à partir d'une activité métier. Ils alors simplement appliqué une règle de trois, complétée par une marge d'incertitude arbitraire, comme l'indique l'Application numérique 1.

$$F_b = \frac{4500}{2000} + 0,25 = 2,5$$

2000 est une estimation de l'activité de l'usine au moment de l'estimation de  $F_b$ ,

4500 la prévision d'activité en  $b$  de l'usine fournie la Direction Générale de STMicroelectronics,

0,25 est une marge d'incertitude arbitraire

*Application numérique 1. Evaluation de  $F_b$  lors de la migration MES : application d'une règle de trois pour pallier à l'impossibilité d'appliquer le modèle de rationalité des décideurs.*

Cette méthode a semblé trop naïve, trop peu formelle, et a été passablement mise en question lors du projet de migration. On peut citer quelques extraits des *verbatim* de l'étude de cas pour en attester :

« [Interviewé n°3] *avait dit* « 2,5 ». *Mais ça semblait sorti du chapeau.* »

« *Le facteur 2,5, certes, on y est allé un peu au doigt mouillé, on n'avait pas d'outils très élaborés, de modèles existants [...]* »

Clairement, cette solution ne semblait pas être à même de tenir compte des enjeux de gestion identifiés dans le chapitre premier de cette thèse :

- *Enjeu de gestion n°1 « gérer un système complexe »* : à aucun moment, l'activité de l'infrastructure des TI n'a été réellement prise en compte. L'Application numérique 1 se basait simplement sur l'évolution de l'activité industrielle, sans essayer de mesurer ses impacts sur une ressource technologique. Compte-tenu de la complexité de l'infrastructure des TI de Crolles 2, il aurait été préférable d'inclure une modélisation explicite du comportement du composant technologique lors de la prise de décision ;

- *Enjeu de gestion n°2 « soutenir pleinement l'activité de l'organisation »* : cet enjeu a été insuffisamment considéré. L'activité industrielle a été intégrée par le nombre de plaques produites par semaine. Mais était-il suffisant de considérer l'activité métier sous cette seule dimension ? On peut en douter : il est probable que d'autres dimensions puissent impacter l'activité du serveur. Les identifier aurait permis de mieux modéliser les interactions entre activité métier et ressource technologique et de permettre un support à la décision plus fin, permettant de tester des scénarios métier plus pertinents.

Ainsi, pour mieux tenir compte de ces enjeux de gestion et conformément au modèle de rationalité des décideurs, il aurait été plus rigoureux de disposer d'un modèle d'évaluation pour l'estimation  $F_b$ , fondé sur un modèle de prévision permettant de prédire l'activité de composants d'infrastructure, en fonction de scénarios métier multidimensionnels.

### *2.2.2. Travaux complémentaires : identification du risque de sous-capacité (détection de comportements anormaux dans le cadre de l'activité (K) Contrôle)*

#### *i. Formulation du problème de décision*

Les alternatives à considérer ( $\mathcal{A}$ ) sont les suivantes : existe-t-il oui ou non un risque de sous-capacité du système ? Y répondre permet d'enclencher ou de ne pas enclencher l'ensemble d'un processus décisionnel plus global, qui pourrait aboutir à faire évoluer l'infrastructure des technologies de l'information. Ces alternatives sont envisagées dans le cadre d'une activité générale de contrôle de l'activité de l'infrastructure des TI, telle que présentée dans la première partie de cette thèse. Comme nous l'avons vu, deux points de vue ( $\mathcal{V}$ ) sont envisageables pour automatiser une telle vérification :

- *La détection automatique de franchissement de seuils*, qui correspond à l'activité ITIL (R) Gestion et contrôle des seuils, qui ne requiert pas d'intelligence statistique et qui n'est pas considérée dans cette thèse ;
- *La détection automatique de comportements anormaux*, dans le cadre de l'activité ITIL (K) Contrôle, et pour laquelle nous souhaitons apporter une contribution. Son principe général a déjà été évoqué (Partie 1) : l'objectif est d'être capable d'observer dynamiquement une variable  $x$  et de tester l'hypothèse nulle  $H_0$  voulant que, pour chaque unité de temps  $t$ , les résidus  $e$  de la modélisation de l'évolution de cette variable dans le temps sont approximativement indépendants et identiquement distribués (i.i.d.), de loi normale  $N(0, \sigma^2)$  constante. Le Modèle 2 formalise ce modèle de rationalité.

Contrôle automatique de  $H_0 : e_t$  sont i.i.d. de loi  $N(0, \sigma^2)$

$H_0$  acceptée  $\rightarrow$  comportement normal de la série observée

$H_0$  rejetée  $\rightarrow$  comportement anormal

avec :

$$e_t = x_t - \hat{x}_t,$$

$x_t$  : série chronologique observée,

$\hat{x}_t$  : modélisation de  $x_t$

*Modèle 2. Modèle de rationalité du besoin décisionnel « identification du risque de sous-capacité ».*

Pour analyser correctement l'activité du système et de son environnement, une bonne stratégie de contrôle doit permettre de saisir toutes les dimensions de l'activité du système, c'est-à-dire l'ensemble des trois niveaux ITIL (I) « service », (J) « ressource » et (H) « métier ». Toutefois, compte-tenu du caractère annexe de ce problème de décision dans le cadre de nos recherches, la question du choix des variables ne sera pas discutée. Nous considérerons une sélection arbitraire de variables, considérée comme complémentaires au maintien à niveau de la connaissance formalisée dans le cadre du modèle de prévision. Notre questionnement scientifique portera uniquement sur la définition d'un traitement statistique approprié aux variables sélectionnées. Ainsi, la problématique décisionnelle (II) peut être définie comme suit :

*$\Pi_2$  : comment modéliser l'ensemble des variables  $x_{it}$ , caractérisant l'activité du système, et détecter d'éventuels comportements anormaux, identifiés comme des écarts significatifs entre la modélisation des séries  $x_{it}$  et leurs valeurs observées ?*

Pour y répondre, des problèmes locaux d'application se posent et nécessitent le développement de réponses *ad hoc*.

ii. *Limites des pratiques en usage lors de la migration MES*

Lors du projet MES, le modèle de rationalité exprimé par le Modèle 2 n'était tout simplement pas envisageable. Aucun mécanisme de contrôle automatisé n'était à l'œuvre : un certain nombre de variables étaient vérifiées manuellement, selon l'intuition des experts. La citation suivante, extraite d'un *verbatim*, indique que l'approche en œuvre était peu formelle :

*« Ce sont des calculs très approximatifs. Une règle très simple que j'utilise, c'est que les gros serveurs ne doivent jamais être au-delà de 60% de leur capacité. »*

Cette affirmation peut être mise en regard de l'enjeu de gestion n°1 (« gérer un système complexe ») : dans le premier chapitre, nous avons vu que la présence d'un système complexe implique des procédures automatisées, devant permettre d'effectuer des contrôles, même lorsqu'aucune hypothèse *a priori* sur les limites de contrôles n'est

disponible. L'usage de méthodes statistiques appropriées permettrait donc d'améliorer cette situation.

En ce qui concerne l'enjeu de gestion n°2 (« soutenir pleinement l'activité de l'organisation »), on rappelle que la question du choix des variables n'est pas approfondie dans cette thèse. C'est pourquoi les éventuelles limites liées à cet enjeu de gestion ne seront pas abordées ici.

Ainsi, pour opérationnaliser le modèle de rationalité des décideurs et pour surmonter l'enjeu de gestion n°1, il aurait fallu être capable de mettre en œuvre *in situ* le Modèle 2 de détection de comportements anormaux, par un traitement statistique approprié.

#### **Conclusion du chapitre 4**

L'application du processus socio-statistique d'aide à la décision a permis de constituer une étude de cas. Cette étude de cas s'est avérée utile pour formaliser une connaissance, permettant notamment de décrire avec précision les problèmes de décisions de deux activités particulières du projet de migration MES : la modélisation prospective et la détection de comportements anormaux. Cette connaissance est une base stable sur laquelle construire une aide à la décision, utilisant des méthodes quantitatives. Les modèles de rationalité des décideurs sont clairement exprimés et leurs limites connues : on sait vers quelle forme doit tendre l'aide à la décision attendue.

La définition de cette aide à la décision va désormais être précisée dans le chapitre suivant. Une démarche de modélisation, intégrant les deux besoins décisionnels considérés dans le cadre du périmètre industriel d'application de cette thèse, est structurée et spécifiée. La structure générale de cette démarche a été définie lors de la présentation du processus socio-statistique d'aide à la décision, dans la première partie de cette thèse. Le chapitre suivant va quant à lui préciser certaines spécificités méthodologiques, nécessaires à la mise en œuvre *in situ* d'une démarche de modélisation, adaptée à une aide à la décision contextualisée.

**Chapitre 5. Structuration d'une démarche de modélisation destinée à répondre aux problèmes.**

## Introduction

Le chapitre 5 concerne la première étape du questionnement central de cette thèse : la définition des conditions de développement de modèles, qui permettront d'apporter une aide à la décision sur la base d'éléments quantitatifs. Pour cela, ce chapitre présente l'ensemble des réflexions, qui permettent de construire, d'expliquer et de justifier une démarche de modélisation complète et structurée, destinée à répondre aux problèmes de décisions présentés dans le chapitre précédent. Le plan proposé dans la Figure 33 sera suivi.

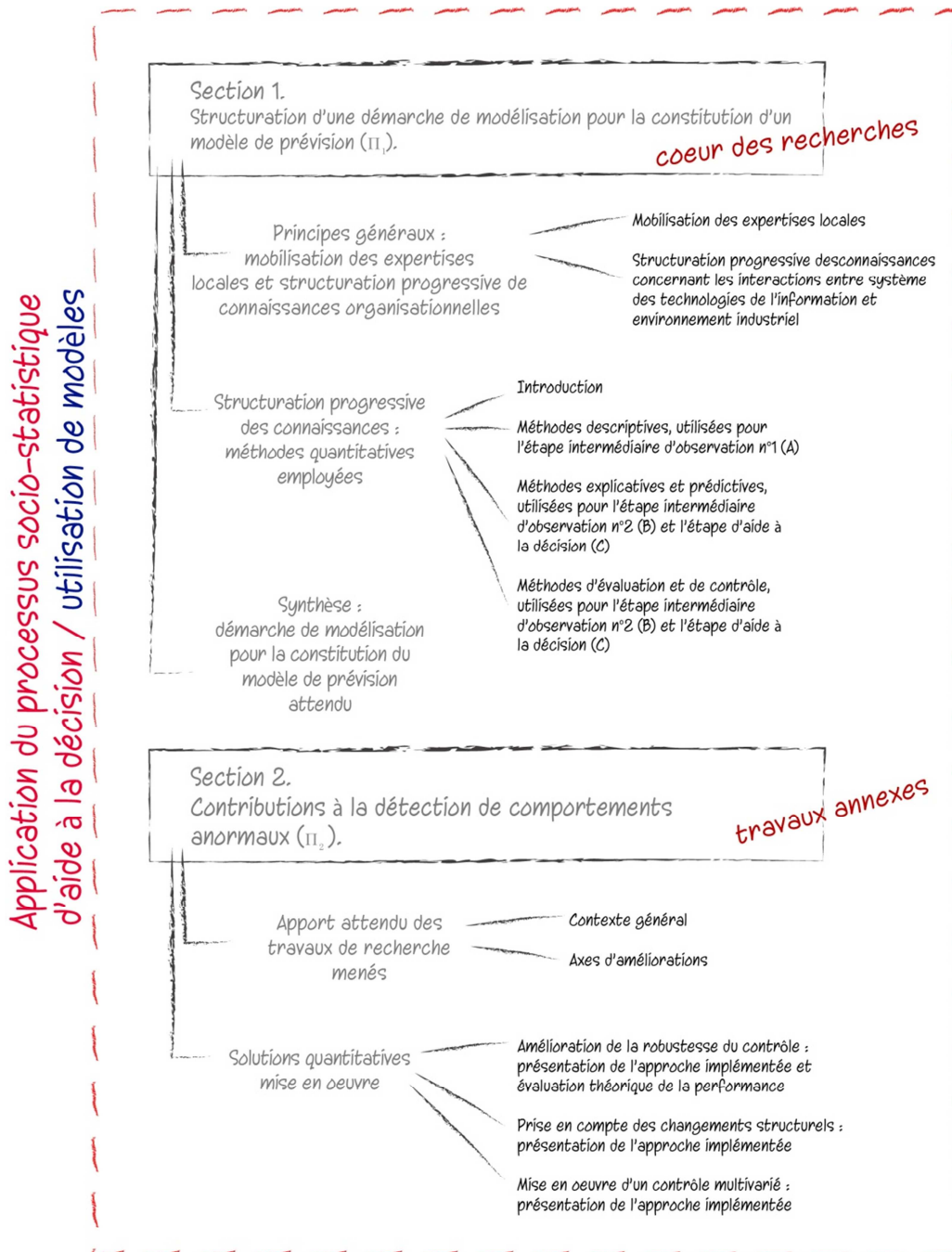


Figure 33. Plan du chapitre 5.

Le chapitre 5 est constitué de deux sections bien distinctes. La première section est dédiée au travail de thèse principal : la construction d'un modèle de prévision. Tout d'abord, les principes sous-jacents à la constitution d'une démarche de modélisation, liés au contexte d'application, sont précisés : quels sont les acteurs de l'organisation impliqués dans la modélisation ? Comment parvient-on à construire le modèle souhaité, sachant que les interactions que l'on cherche à modéliser sont méconnues des experts de STMicroelectronics ? Pour répondre à cette deuxième question, nous aurons recours à une démarche progressive de structuration de connaissances nouvelles, en plusieurs étapes. Deux étapes intermédiaires d'observation et une étape d'aide à la décision seront effectivement mobilisées. Chacune de ces étapes répond à des objectifs spécifiques, qui seront précisés. Chacune d'elles fait appel à un certain nombre de méthodes statistiques particulières (méthodes descriptives, explicatives/prédictives et d'évaluation/de contrôle), qui seront présentées. Ces considérations méthodologiques locales permettront de compléter la démarche de modélisation générique, présentée dans la première partie de la thèse. Ainsi, en guise de synthèse de la section 1, une démarche de modélisation complète, permettant la constitution du modèle de prévision attendu et sa mise sous contrôle (génétique du système de connaissances, cf. présentation générale du processus socio-statistique), pourra être proposée.

La section 2 concerne les travaux de thèse annexes, relatifs à la détection de comportements anormaux. Nous rappelons que nous les considérons exclusivement comme une contribution à la génétique du système de connaissances, en complément de la connaissance formalisée par le modèle de prévision construit par l'application de la démarche de modélisation présentée en section 1. Forts de ce principe, nous indiquons les apports attendus pour l'amélioration de l'activité de détection des comportements anormaux, dans le cadre de nos travaux de recherches. Ces améliorations concernent exclusivement les méthodes statistiques employées. Elles sont destinées à tenir compte du contexte local d'application de la détection automatique de comportements anormaux. Trois axes de progrès sont identifiés et traités : l'amélioration de la robustesse des contrôles, la prise en compte des changements structurels, et la mise en œuvre d'une stratégie de contrôle multivariée.

L'application de l'ensemble des propositions des sections 1 et 2 permettra, conformément au séquençage du processus socio-statistique d'aide à la décision, de formaliser une connaissance pouvant être mobilisée à des fins décisionnelles, ainsi que de garantir certaines conditions d'évolution dans le temps. Cette application fera l'objet de la troisième partie de la thèse.



**Section 1. Structuration d'une démarche de modélisation pour la constitution d'un modèle de prévision (II<sub>1</sub>: dimensionnement des composants d'infrastructure).**

**1.1. Principes généraux : mobilisation des expertises locales et structuration progressive de connaissances organisationnelles**

*1.1.1. Mobilisation des expertises locales*

Selon le processus socio-statistique d'aide à la décision, la phase d'utilisation de modèles repose sur les acteurs de l'organisation. Ils apportent leurs connaissances concernant les différents niveaux du système des TI, l'environnement, les données, etc. En conséquence, lors de l'application du processus, on s'appuiera autant que possible sur les expertises identifiées lors de la représentation de la situation problématique (P, Tableau 7), comme l'indique le Tableau 8.

Experts mobilisés	Connaissances apportées	
	Niveaux infrastructure TI/environnement	Source de données
<i>Experts techniques infrastructure TI-C2</i>	Architecture de l'infrastructure de TI	Base de performance
<i>Experts applicatifs MES</i>	Architecture applicative	Base de performance
	Activité industrielle	Bases de données applicatives (représentation informatisée de l'activité industrielle)

*Tableau 8. Experts locaux à mobiliser et connaissances apportées.*

La mobilisation de ces expertises est au cœur du processus socio-statistique d'aide à la décision et des mécanismes de conversions/documentations de connaissances qu'il implique. Pour ce faire, ces experts sont mobilisés autant que possible, par la conduite de groupes de travail, principalement lorsque les aspects syntaxique et sémantique de l'ontologie du système de connaissances sont considérés.

*1.1.2. Structuration progressive des connaissances concernant les interactions entre système des technologies de l'information et environnement industriel*

Cette section 1 est destinée à structurer et justifier la démarche de modélisation de connaissances, construite pour traiter la problématique II<sub>1</sub>. Rappelons que cette démarche aura pour objet de produire un modèle de prévision, répondant au besoin de gestion des capacités des technologies de l'information.

Conformément aux difficultés d'application de l'activité ITIL (Q) Modélisation et analyse des tendances, discutées dans le premier chapitre de la thèse, les interactions entre l'activité des composants d'infrastructure des TI et l'activité métier sont méconnues des experts locaux. Il faut malgré tout parvenir à structurer progressivement une connaissance organisationnelle, pour constituer le modèle de prévision orienté

métier, attendu dans le cadre de la problématique décisionnelle. Deux approches pourraient être envisagées pour construire cette connaissance :

- *Approche « descendante »* : elle partirait d'une modélisation de l'activité industrielle, que l'on tenterait de rattacher à l'activité du système d'information, puis aux composants d'infrastructure. Connaissant la complexité des processus industriels considérés (Cooper et al., 1992 ; Montoya-Torres, 2007) et le manque de cartographie du SI disponible au sein de notre terrain de recherche, cette approche n'est pas réaliste dans le contexte de notre recherche ;
- *Approche « ascendante »* : elle consisterait à centrer l'analyse sur le problème de décision, c'est-à-dire sur l'observation du composant d'infrastructure, pour lequel on souhaite constituer un modèle de prévision. On en restituerait alors progressivement l'activité, dans le cadre de son environnement applicatif, puis industriel. C'est cette méthode, opérationnellement plus réalisable, qui sera retenue.

Pour mettre en œuvre l'approche ascendante, trois étapes de structuration progressive de connaissances ont été définies :

- *(A) Etape intermédiaire d'observation n°1* : cette étape à une visée descriptive. Son objectif est de restituer l'utilisation du composant d'infrastructure, pour lequel on souhaite constituer un modèle de prévision (Y), dans son contexte industriel. L'objectif est d'établir les liens entre l'activité du composant et l'activité applicative du SI, puis de comprendre en quoi ces activités sont le reflet d'une activité industrielle sous-jacente. Cette activité industrielle, déductible de l'activité transactionnelle, est donc plutôt une activité opérationnelle, représentative de l'activité de production courante. Il en résulte un modèle descriptif, qui exprime de manière structurée une connaissance sur le système étudié, et qui permet de guider la phase ultérieure ;
- *(B) Etape intermédiaire d'observation n°2* : cette étape à une visée explicative et prédictive. Son objectif est de quantifier l'utilisation du composant informatique, dans le cadre de l'activité industrielle opérationnelle. A partir du modèle d'observation n°1, on va chercher à mettre en relation l'activité industrielle opérationnelle et l'activité de la ressource informatique. Il en résulte un modèle explicatif, qui permet une première caractérisation des liens entre le niveau architecture de l'infrastructure des TI et le niveau métier. Sur la base de ce modèle, on peut alors chercher à identifier l'activité industrielle de pilotage, sous-jacente à l'activité opérationnelle modélisée. Cette activité de pilotage peut être plus difficile à déduire, car son impact sur l'activité courante du système de production n'est pas forcément direct. Ce modèle explicatif et prédictif exprime lui aussi une connaissance structurée sur système étudié, qui sera utilisée lors de l'étape suivante ;
- *(C) Etape d'aide à la décision* : c'est l'étape finale, qui mènera au modèle de prévision attendu. Elle a une visée explicative et prédictive. En s'appuyant sur la connaissance générée par le modèle d'observation n°2, on va mettre en relation variables industrielles de pilotage ( $X_i$ ) et utilisation du composant d'infrastructure (Y), en définissant la forme finale du modèle de prévision.

Le périmètre de compréhension de l'activité de la ressource informatique est ainsi progressivement étendu à l'activité métier opérationnelle de l'usine, puis à des indicateurs industriels de pilotage généraux. Cette construction progressive permet une conceptualisation de plus en plus macroscopique de l'activité du système. Les liens entre le composant d'infrastructure et l'activité applicative sont directs et conceptuellement évident. Les liens entre le composant d'infrastructure et l'environnement industriel sont moins intuitifs, mais sont néanmoins tout à fait intelligibles lorsqu'on a compris les interactions décrites ou expliquées par les modèles préalables. Cette formalisation progressive de connaissances est résumée par la Figure 34.

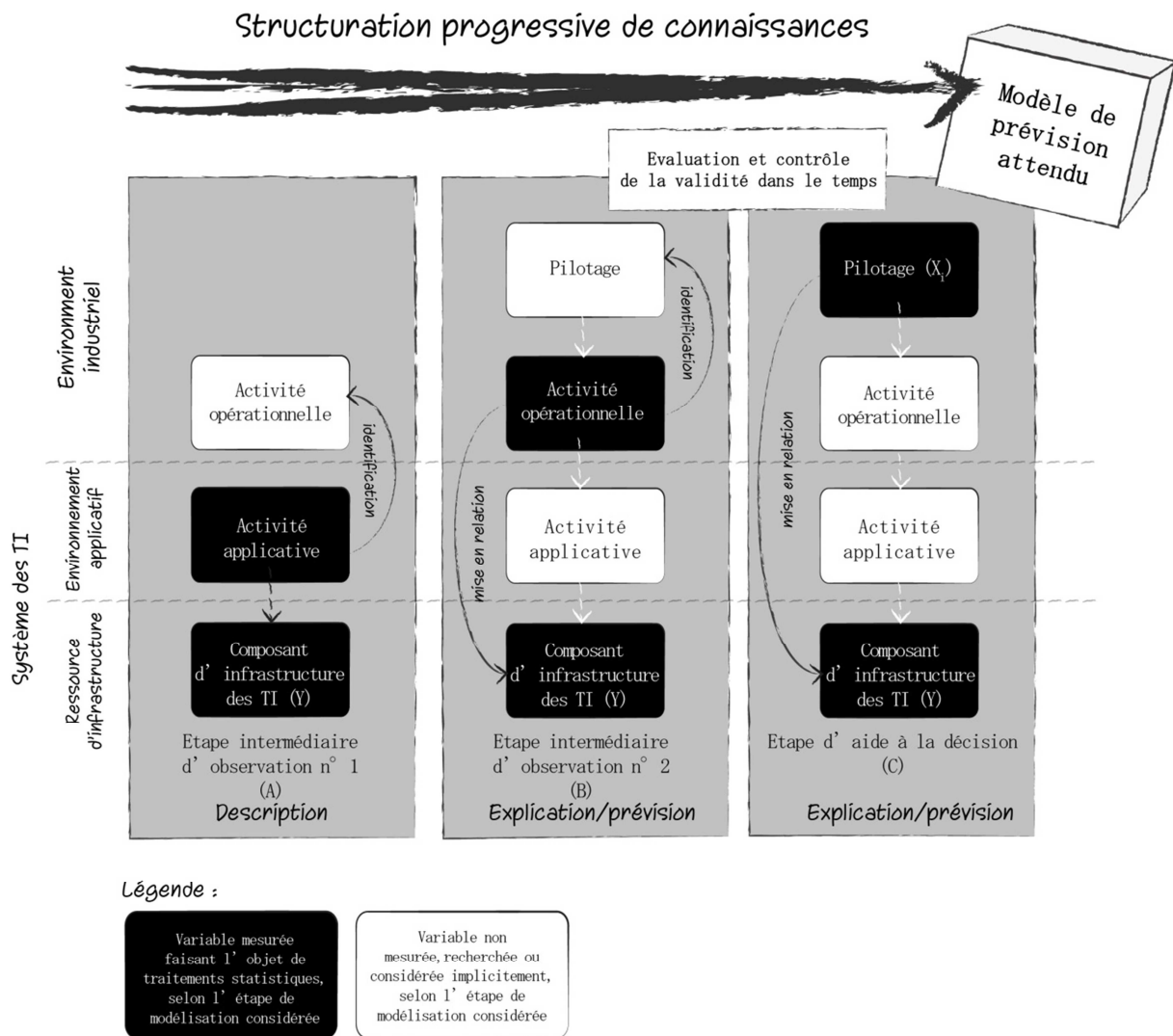


Figure 34. Réalisation du modèle de prévision : structuration progressive de connaissances organisationnelles.

Comme indiqué dans la Figure 34, les modèles explicatifs et prédictifs constitués impliquent deux précautions additionnelles :

- *D'une part, il convient de les évaluer, afin de s'assurer qu'ils soient effectivement en mesure de fournir des prévisions fiables ;*
- *D'autre part, il convient d'en contrôler la validité dans le temps, tel que nous l'avons indiqué dans le premier chapitre de thèse, afin d'assurer la génétique du système de connaissances organisationnelles.*

La constitution, l'évaluation et le contrôle des modèles impliqués par ces différentes étapes nécessitent un certain nombre de traitements quantitatifs des données disponibles. Nous les précisons ci-après.

## **1.2. Structuration progressive des connaissances : méthodes quantitatives employées**

### *1.2.1. Introduction*

Cette section indique quelles sont les méthodes statistiques que nous proposons de mobiliser, pour mettre en œuvre la structuration progressive de connaissances suggérée en Figure 34. Trois familles de traitements quantitatifs sont identifiées :

- *Des méthodes descriptives, utilisées pour l'étape (A) intermédiaire d'observation n°1. Leur objectif est de décrire l'activité de l'environnement applicatif du composant d'infrastructure, pour lequel on souhaite constituer le modèle de prévision. Les données exploitées sont issues de la base de performance, afin de mesurer l'activité de l'infrastructure ainsi que l'activité applicative ;*
- *Des méthodes explicatives et prédictives, utilisées pour l'étape (B) intermédiaire d'observation n°2, et l'étape (C) d'aide à la décision. Leur objectif est d'expliquer et de prédire les interactions entre le composant d'infrastructure pour lequel on souhaite constituer le modèle de prévision et :*
  - \* *L'environnement industriel opérationnel, pour l'étape (B),*
  - \* *L'environnement industriel caractérisé par des variables de pilotage, pour l'étape (C),*

La base de performance est utilisée pour mesurer l'activité de l'infrastructure. Les bases de données applicatives sont quant à elles exploitées, pour quantifier l'activité industrielle opérationnelle et les variables de pilotage.

- *Des méthodes d'évaluation et de contrôle, utilisées pour l'étape (B) intermédiaire d'observation n°2, et l'étape (C) d'aide à la décision. Ces traitements, complémentaires à la réalisation des modèles explicatifs et prédictifs, visent à :*

- \* *Evaluer la qualité des modèles par :*
  - ❖ *L'évaluation de leurs vertus prédictives*, en appréciant leur pertinence pour un usage en prévision (a),
  - ❖ *L'évaluation de leurs vertus explicatives*, en validant la constance de leurs paramètres sur la période observée (b),
- \* *Contrôler dynamiquement dans le temps la constance des modèles*, pour toutes nouvelles observations du système modélisé (c). Cela correspond à l'idée de contrôle du modèle prospectif, intermédiaire entre les activités ITIL (Q) et (K), évoquée dans le premier chapitre de la thèse. Il vise à détecter un changement de comportement du système considéré, après la constitution et l'évaluation d'un modèle de prévision permettant de le caractériser,

La base de performance est utilisée (activité de l'infrastructure), ainsi que les bases de données applicatives (activité industrielle opérationnelle et variables de pilotage).

Ci-dessous, nous présentons les principales caractéristiques des trois familles de traitements quantitatifs mentionnées.

### *1.2.2. Méthodes descriptives, utilisées pour l'étape (A) intermédiaire d'observation n°1*

Pour comprendre l'activité industrielle opérationnelle sous-jacente à l'activité applicative, en analysant les centaines de variables transactionnelles caractérisant cette dernière, la statistique descriptive et exploratoire peut être mobilisée (Lebart et *al.*, 2006). Les outils usuels d'analyse multivariée sont utilisés, pour constituer un modèle en dimension réduite des variables originales, afin qu'elles puissent être appréhendées et discutées avec les experts. On s'intéressera aux méthodes factorielles qui permettent, comme l'indique Revelle (2011), d'identifier des phénomènes sous-jacents non observés. C'est ce type de phénomènes latents que nous souhaitons identifier, supposant qu'ils correspondent à l'activité industrielle opérationnelle de l'usine.

Comme l'indique Rodriguez et *al.* (2009), lorsque le modèle n'est pas spécifié et que certaines hypothèses sur les données (indépendance, distribution normale, homoscedasticité – c'est-à-dire constance de la variance –) ne peuvent être vérifiées *a priori*, l'Analyse en Composante Principale (ACP) et/ou la régression *Partial Least Squares* (PLS) sont recommandées. L'ACP classique (Lebart et *al.*, 2006 ; Abdi et Williams, 2010 ; Revelle, 2011) vise à construire des axes factoriels (ou composantes principales), permettant de réduire la complexité des données originales, en les projetant dans un espace de dimension réduite.

Ces axes sont :

- Définis par une équation linéaire, caractérisée par les variables d'origine qui ont contribué à leur construction ;
- Caractérisés par le niveau d'information restitué des variables d'origine ;
- Ordonnés par rapport à ce niveau d'information ;
- Non-corrélés entre eux.

L'ACP classique peut être complétée par une rotation *Varimax* (Abdi *in* Lewis-Beck et *al.*, 2003 ; Lebart et *al.*, 2006 ; Abdi et Williams, 2010). En effet, l'ACP classique vise à restituer le maximum de variance dans le premier axe factoriel. En conséquence, toutes les variables ont tendance à être orientées dans la direction de cet axe. La rotation des axes factoriels atténue ce phénomène et facilite leur interprétation. Le critère de rotation orthogonal, qui préserve la non corrélation des composantes, le plus utilisé est le critère *Varimax*, proportionnel à la variance empirique des carrés des coordonnées.

La régression PLS est également une méthode factorielle, qui permet quant à elle de superviser la constitution des axes factoriels, de façon à maximiser l'explication de la variabilité d'une ou plusieurs autres variables (Tenenhaus et *al.*, 1995 ; Tenenhaus, 1998 ; Wold et *al.*, 2001). Cette méthode pourrait par exemple être employée de façon à constituer des axes factoriels à partir des variables applicatives, visant explicitement à favoriser l'explication de la variabilité de l'activité de l'infrastructure.

L'application de ces méthodes au sein de notre terrain de recherche a révélé que les résultats les plus intéressants sont obtenus avec l'ACP classique avec rotation *Varimax*<sup>3738</sup>. En effet, on observe que, quelle que soit la variante de l'algorithme d'ACP utilisée, les principaux axes factoriels sont généralement d'excellents prédicteurs de l'activité de la ressource informatique : il n'est donc pas nécessaire d'aller plus en avant par l'emploi d'une méthode supervisée de type régression PLS. Il faut par contre privilégier la méthode qui facilitera l'interprétation des axes factoriels, pour identifier au mieux les phénomènes sous-jacents non observés à l'origine de ces axes. Or, nous constatons que les experts, avec qui sont discutés les résultats des analyses statistiques, ont beaucoup plus de facilités à interpréter les composantes issues des ACP avec rotations *Varimax*.

### *1.2.3. Méthodes explicatives et prédictives, utilisées pour l'étape (B) intermédiaire d'observation n°2 et l'étape (C) d'aide à la décision*

La statistique inférentielle et confirmatoire peut être employée (Lebart et *al.*, 2006), et notamment la méthode de la régression linéaire multiple (Makridakis et *al.*, 1998 ;

---

<sup>37</sup> Nous avons également eu recours à une méthode d'ACP robuste, limitant les biais en cas de valeurs aberrantes dans les données (Hubert et *al.*, 2005). Nous n'en avons pas tiré de bénéfice particulier dans le cadre de nos travaux, c'est pourquoi nous nous limitons à la mentionner dans cette note de bas de page.

<sup>38</sup> Remarque : un recours exclusif à l'analyse de relations linéaires est une limite des méthodes employées. Cependant, c'est une première approximation suffisante (Husson et *al.*, 2009), dans le cadre d'une étude qui n'a qu'un caractère purement exploratoire.

Johnston et Dinardo, 1999 ; Gujarati, 2004 ; Montgomery et *al.*, 2006 ; Kennedy, 2010). Cette méthode bien connue permet de quantifier les relations linéaires entre une variable expliquée<sup>39</sup> (Y) et une ou plusieurs<sup>40</sup> variables explicatives<sup>41</sup> (caractérisation de l'environnement industriel). Elle est particulièrement adaptée à la problématique *II*<sub>1</sub>. En effet, elle a un double objectif d'explication, en permettant de comprendre la nature des liaisons entre les variables, et de prédiction. Dans ce dernier cas, les liaisons entre variables peuvent être utilisées pour l'anticipation et la prise de décision, à partir de valeurs connues des variables explicatives, qui permettront de prédire la valeur inconnue de la variable expliquée.

La régression linéaire multiple consiste à définir un plan multidimensionnel, quantifiant une relation linéaire entre la variable expliquée et les variables explicatives. Cet hyperplan est un modèle, caractérisé dans le cas général par les paramètres suivants :

- Une ordonnée à l'origine ;
- Les coefficients de régression partiels de la variable expliquée sur les variables explicatives, qui mesurent le taux de variation de la variable expliquée, pour une variation d'une variable explicative, lorsque toutes les autres variables explicatives sont maintenues constantes ;
- Un terme d'erreur (ou résiduel), qui mesure l'écart entre les observations réelles et les relations linéaires du modèle. Il résume tout ce que le modèle n'explique pas.

Dans ce cas général, on utilise habituellement la méthode des moindres carrés ordinaires pour calculer ces paramètres.

Pour être valide, le modèle linéaire implique un certain nombre d'hypothèses. Nous en rappelons ici deux principales, qui peuvent poser problème au sein de notre terrain de recherche :

- *Non corrélation des variables explicatives* : dans le cas contraire, on parle de risque de colinéarité ou de multicollinéarité, qui peut nuire à l'estimation des paramètres. Plusieurs outils peuvent être employés pour évaluer ce risque (Gujarati, 2004 ; O'Brien, 2007 ; Chatelain, 2010) : indicateurs synthétique tels que le *Variance Inflation Factor* (VIF), analyse des régressions auxiliaires (part de la variabilité de chacune des variables explicatives expliquée par les autres variables explicatives), etc. ;
- *Indépendance et homoscélasticité des erreurs* : les résidus sont supposés être indépendamment distribués selon une loi normale de paramètres  $N(0, \sigma^2)$ , de variance constante. Dans ce cas favorable, on parle de résidus indépendants et identiquement distribués (i.i.d.).

---

<sup>39</sup> Aussi qualifiée d'endogène, dépendante ou à prédire.

<sup>40</sup> Le cas à une variable explicative correspond à la régression linéaire simple, cas particulier de la régression linéaire multiple.

<sup>41</sup> Aussi qualifiées d'exogènes, indépendantes, prédictives, ou encore de prédicteurs.

Pour contourner certaines violations de ces deux hypothèses, rencontrées lors de l'exploitation des données issues de la base de performance et des bases de données applicatives de Crolles 2, les outils suivants peuvent être utilisés :

- *Variables explicatives plus ou moins corrélées entre elles* : on peut traiter le risque de colinéarité en procédant à des traitements manuels, par l'analyse des variables explicatives (O'Brien, 2007 ; Chatelain, 2010). Par exemple, des méthodes telles que le regroupement de variables ou l'orthogonalisation (calcul d'un modèle intermédiaire entre variables explicatives et intégration des résidus dans le modèle final) peuvent être employées. La régression PLS peut également être utilisée dans cette situation (Gauchi, 1995 ; Tenenhaus, 1998) ;
- *Résidus dépendants du temps* : en analysant les autocorrélations<sup>42</sup> du terme d'erreur, on peut le modéliser comme un processus temporel, défini par un modèle ARIMA (Makridakis et al., 1998 ; Johnston et Dinardo, 1999 ; Gujarati, 2004 ; Kennedy, 2010). Trois paramètres sont pris en compte :
  - \* *Auto-régression (AR)* : à un instant  $t$ , la série chronologique est définie par une fonction linéaire des valeurs précédentes, modélisée par régression linéaire,
  - \* *Différenciation (I)* : le modèle ARIMA nécessite de travailler sur une série stationnaire, c'est-à-dire de moyenne et variance constantes dans le temps. Si ce n'est pas le cas, on transforme la série en la différenciant, ce qui consiste à remplacer la série originale par une série de différences adjacentes appropriée (plusieurs degrés de différenciation sont possibles),
  - \* *Moyenne mobile (MA)* : à un instant  $t$ , la variable est définie par des fluctuations autour d'une valeur moyenne, modélisées par moyenne mobile,

Lorsqu'on modélise ainsi les résidus d'un modèle de régression, on parle de régression multiple avec erreurs ARIMA. Dans ce cas de figure, les moindres carrés ordinaires ne sont plus adaptés pour l'estimation des paramètres : on a recours aux méthodes des moindres carrés généralisés ou du maximum de vraisemblance ;

- *Présence de phénomènes saisonniers (forme particulière de dépendance au temps parfois observée dans les autocorrélations résiduelles)* : on peut les modéliser par l'emploi de variables indicatrices (Hardy, 1993 ; Makridakis et al., 1998 ; Gujarati, 2004 ; Kennedy, 2010), qui permettent d'intégrer un codage qualitatif dans le modèle de régression, afin d'identifier certains phénomènes connus, sans pour autant avoir à les mesurer.

Cette méthode est simple et souple d'utilisation, comme en atteste par exemple ses capacités d'adaptation à la temporalité des phénomènes observés. Elle est donc bien adaptée aux contraintes d'un usage industriel. Par ailleurs, comme nous le verrons dans le chapitre 6, elle a fourni de très bons résultats dans le cadre des variables considérées lors de nos recherches. En conséquence, il n'a pas été nécessaire d'employer d'autres

---

<sup>42</sup> Corrélation d'une série temporelle avec elle-même, selon un décalage défini.



méthodes plus sophistiquées : modèle linéaire généralisé, réseaux de neurones, arbres de régression, approches non-paramétriques, etc. (voir par exemple Hand et *al.* (2001) pour un aperçu des diverses formes fonctionnelles de ces autres modèles explicatifs et prédictifs).

#### *1.2.4. Méthodes d'évaluation et de contrôle, utilisées pour l'étape (B) intermédiaire d'observation n°2 et l'étape (C) d'aide à la décision*

En ce qui concerne l'évaluation des modèles pour un usage en prévision (a), on peut avoir recours aux méthodes de validation croisée. Ces méthodes permettent de mesurer la qualité prédictive d'un modèle. Elles consistent à partitionner les données utilisées en plusieurs groupes. L'un des groupes est retiré. Les paramètres du modèle sont estimés sur l'ensemble des groupes restants, puis on utilise ce résultat pour effectuer une prévision sur le groupe qui a été retiré. On peut alors comparer les valeurs prédites aux données réelles. La procédure est répétée plusieurs fois, avec des groupes différents (Hastie et *al.*, 2001). Lorsque les données présentent une composante temporelle, Hydman (2011) propose deux algorithmes de validation croisée adaptés :

- *Estimation du modèle sur un groupe dont la taille augmente séquentiellement*, en tenant compte de l'ordre temporel des données, la prévision étant faite sur les valeurs postérieures à cet échantillon ;
- *Estimation du modèle sur un groupe défini par une fenêtre glissante de taille fixe*, la prévision étant réalisée sur les valeurs postérieures à cet échantillon.

Au sujet de l'évaluation de la constance des paramètres des modèles sur la période observée (b), on peut employer des méthodes de test de la stabilité structurelle. Ils font l'objet d'une vaste littérature technique (cf. Maddala et Kim, 1996 ; Kennedy, 2010 pour un aperçu de la littérature). L'objectif général de ces tests est de tester une hypothèse nulle de stabilité structurelle, *versus* l'hypothèse alternative d'instabilité des coefficients du modèle. On peut les regrouper en trois grandes familles (Zeileis, 2005) :

- *Les tests d'égalité des paramètres*, entre un ou plusieurs échantillons issus du jeu de données initial, basés sur la statistique de Fisher (exemple : tests de Chow et de Quandt, voir Johnston et Dinardo, 1999 ; Hansen, 2001) ;
- *Les tests fondés sur le maximum de vraisemblance* (exemple : test de Nyblom-Hansen, voir Zeileis, 2005) ;
- *Les tests de fluctuations*. Ces derniers sont commodes d'utilisations car ils sont faciles à mettre en œuvre et impliquent peu d'hypothèses *a priori* sur les changements structurels à tester. Le test peut se baser sur les fluctuations des coefficients ou des résidus d'un modèle. Trois modes d'estimation de ces fluctuations sont possibles (Zeileis, 2005) :

- \* *Estimation récursive* : sur la base d'un échantillon des données initiales, dont la taille augmente séquentiellement,
- \* *Estimation glissante* : à partir d'un échantillon mobile, défini par une fenêtre glissante de taille fixe,
- \* *Estimation globale* : en utilisant l'ensemble de l'échantillon en une seule fois,

A chacun de ces couples type de fluctuation-mode d'estimation, correspondent une ou plusieurs statistiques de test de la stabilité structurelle. Par exemple, si l'on se base sur la fluctuation des résidus : le test du CUSUM (Brown et *al.*, 1975 ; voir aussi Johnston et Dinardo, 1999) correspond à une estimation récursive, le test du MOSUM (Chu et *al.*, 1995) à une estimation glissante, et le test de l'OLS-based CUSUM (Ploberger et *al.*, 1989) à une estimation globale.

L'immense éventail de tests possibles et leur hétérogénéité ne facilite pas leur usage en contexte industriel. C'est pourquoi les travaux d'Achim Zeileis (Zeileis, 2005 ; Zeileis et *al.*, 2010) sont très utiles. Ils montrent que les trois familles de tests peuvent être ramenées à l'estimation d'un processus de fluctuation empirique (à l'exception des tests de fluctuations récursifs et glissants). Une fois ce processus calculé, il peut être soumis à un test de significativité statistique, basé sur une statistique en rapport avec l'un des tests suscités. Le choix de la statistique dépend de la situation du test, chaque statistique étant plus ou moins adaptée à certains cas de figure : autocorrélation, changements structurels multiples... Le praticien dispose ainsi d'un cadre très général et souple, qui ne présuppose pas d'hypothèses quant aux changements structurels potentiels, et qui peut être adapté en fonction du modèle à tester. Il est résumé dans la Figure 35.

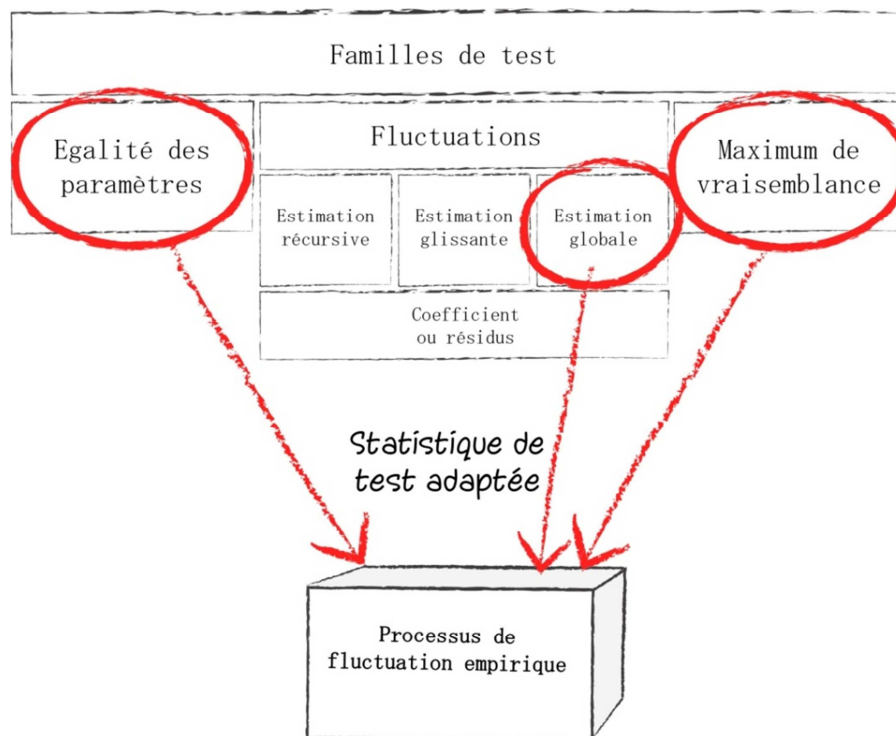


Figure 35. Un cadre général de tests de stabilité structurelle basé sur les processus de fluctuation empirique, proposé par Zeileis (inspiré de Zeileis, 2005 ; Zeileis et *al.*, 2010).

Ce cadre est mobilisé lors de la démarche de modélisation afin de vérifier que, une fois un modèle constitué, il est bien structurellement stable, c'est-à-dire que les coefficients du modèle ne changent pas pendant la période modélisée. En cas d'instabilité, il est possible de dater le point de rupture qui marque un changement dans les paramètres du modèle. Pour ce faire, on peut avoir recours aux travaux de Bai (Bai, 1997 ; Bai et Perron, 2003), basés sur un algorithme itératif permettant d'identifier séquentiellement les points de ruptures potentiels, par division successive de l'échantillon de départ (voir aussi Hansen, 2001, Zeileis et *al.*, 2010). Il faut alors comprendre pourquoi le modèle a changé, et estimer les paramètres du modèle sur des périodes identifiées comme étant stables.

Enfin, ces tests de stabilité structurels peuvent également être utilisés de façon dynamique (c), afin de contrôler dans le temps que la modèle calculé reste valide, après que la modélisation ait été effectuée (Zeileis et *al.*, 2005, 2010). L'utilisation industrielle de cette approche est assez novatrice si l'on se réfère à Zeileis et *al.* (2005), qui indiquent que l'on trouve peu d'exemples d'applications de ces méthodes dans la littérature. Elle nous semble pourtant essentielle : nous évoluons dans des environnements industriels et technologiques très évolutifs. Il faut donc pouvoir détecter le plus rapidement possible tout changement du système considéré, afin de faire évoluer son modèle en conséquence le cas échéant.

Concrètement, un tel contrôle est initialisé sur une période historique où la stabilité structurelle est supposée (période utilisée pour la modélisation). Ensuite, pour tout nouveau point ou groupe de points, on teste l'hypothèse nulle de stabilité à partir d'un test de fluctuation<sup>43</sup>.

De façon générale, toutes ces méthodes (validation croisée et cadre général de test de stabilité structurelle de Zeileis) offrent un ensemble basé sur des pratiques statistiques conformes à l'état de l'art, pour évaluer et contrôler les modèles réalisés par régression linéaire multiple. Leur usage *in situ* chez STMicroelectronics a permis de démontrer la qualité des modèles proposés aux managers de l'infrastructure TI-C2, mais aussi d'en assurer la pérennité.

### **1.3. Synthèse : démarche de modélisation pour la constitution du modèle de prévision attendu**

La démarche de modélisation générique du processus socio-statistique d'aide à la décision, présentée dans la première partie de thèse, est ici complétée par les spécificités contextuelles de l'aide à la décision à apporter (expertises, structuration progressive de connaissances, méthodes statistiques), qui viennent d'être développées.

La structuration progressive de connaissances implique, pour chacune des trois étapes, de naviguer entre les différents aspects de l'ontologie du système de connaissances. Chacun des aspects requiert un certain nombre d'actions, dont la réalisation dépend de la mobilisation des connaissances de l'expertise locale (connaissance du système et son

---

<sup>43</sup> Pour l'instant, Zeileis propose uniquement des outils de contrôle dynamique basés de tests de fluctuations classiques.

environnement et des données disponibles) et de l'analyste/du chercheur (méthodes statistiques). Cette démarche de modélisation est synthétisée dans le Tableau 9.

<i>Gestion des connaissances</i>			
	<b>Récolte des données</b> (aspect syntaxique)	<b>Structuration des données - informations</b> (aspect sémantique)  <i>Cœur de thèse : modélisation prospective (construction) et contrôle du modèle (évolution)</i>	<b>Interprétation des informations - connaissances</b> (aspect pragmatique)
<i>Principales connaissances existantes mobilisés</i>	Expertise locale : experts techniques infrastructure TI-C2 et experts applicatifs MES	Analyste/chercheur	Expertise locale : experts techniques infrastructure TI-C2 et experts applicatifs MES
<i>(A) Etape intermédiaire d'observation n°1</i>	Mesurer (base de performance) l'environnement applicatif de la variable à prédire (Y)	Réaliser un modèle descriptif par ACP classique avec rotation <i>Varimax</i>	Interpréter et valider le modèle et identifier l'activité industrielle opérationnelle sous-jacente
<i>(B) Etape intermédiaire d'observation n°2</i>	Mesurer (bases de données applicatives) l'activité industrielle opérationnelle identifiée en (A)	Réaliser un modèle explicatif/prédictif des interactions entre Y et l'activité industrielle opérationnelle par régression linéaire multiple, puis évaluer et contrôler le modèle (validation croisée, test de stabilité structurelle, mise sous contrôle)	Interpréter et valider le modèle, puis identifier l'activité industrielle de pilotage sous-jacente
<i>(C) Etape d'aide à la décision</i>	Mesurer (bases de données applicatives) l'activité industrielle de pilotage ( $X_i$ ) identifiée en (B)	Réaliser un modèle explicatif / prédictif des interactions entre Y et $X_i$ par régression linéaire multiple, puis évaluer et contrôler le modèle (validation croisée, test de stabilité structurelle, mise sous contrôle)	Interpréter et valider le modèle, ainsi que son intérêt pour la prise de décision

Tableau 9. Démarche de modélisation pour la constitution du modèle de prévision attendu : application des spécificités contextuelles au processus socio-statistique d'aide à la décision.

On constate des liens forts entre les trois étapes de la démarche et les différents aspects de l'ontologie du système de connaissances de l'organisation. Par exemple, l'étape (A) va s'appuyer sur une connaissance existante qui permettra, grâce à l'analyse quantitative et l'interprétation des experts, à générer une nouvelle connaissance. Cette nouvelle connaissance amènera ensuite à considérer l'aspect syntaxique de l'étape (B), en s'appuyant sur les connaissances existantes, pour progressivement constituer une nouvelle connaissance qui pourra être mobilisée lors de l'étape (C), etc. On remarque que le passage de l'aspect syntaxique à l'aspect pragmatique, au sein d'une étape, correspond à la hiérarchie classique du *continuum* données-informations-connaissances. Lorsque le passage d'une étape à une autre nécessitera la création *ad hoc* de données

nouvelles, non disponibles dans les bases de données, on parlera de la hiérarchie inversée du *continuum*.

## **Section 2. Contributions à la détection de comportements anormaux ( $\Pi_2$ : identification du risque de sous-capacité).**

### **2.1. Apport attendu des travaux de recherche menés**

#### *2.1.1. Contexte général*

L'objectif de cette contribution est de favoriser la détection de comportements anormaux dans l'activité du système d'information de STMicroelectronics. Une fois les variables devant mises sous contrôle identifiées, cette détection se fait sur des critères purement statistiques, c'est-à-dire qu'elle ne nécessite pas un avis d'expert *a priori*, qui préciserait ce que doit être un comportement normal ou anormal. Dans le cadre de  $\Pi_2$ , cette détection peut permettre d'identifier automatiquement un risque de sous-capacité du système. Plus généralement, les méthodes présentées ci-dessous sont génériques et applicables dès qu'il s'agit de mettre une série chronologique sous contrôle. Cette contribution est complémentaire à l'activité de modélisation prospective. En effet, nous avons vu qu'un modèle de prévision doit être contrôlé dans le temps et mis à jour en cas de besoin. En cas de modification structurelle du système, la détection de comportements anormaux fournira un faisceau d'indices, qui permettront de comprendre en quoi le modèle de prévision doit être modifié. C'est pourquoi nous considérons l'activité ITIL (K) Contrôle comme complémentaire à l'activité (Q) Modélisation et analyse des tendances.

Comme indiqué plus haut, ce problème est considéré à la marge dans cette thèse. Il n'est intégré dans l'application du processus socio-statistique d'aide à la décision et dans la démarche de modélisation proposée qu'en tant que complément au maintien à jour du modèle de prévision. Pour cela, un certain nombre de contributions apportées dans le cadre de nos travaux sont présentées ici. Elles portent sur les méthodes quantitatives pouvant être mises en œuvre : les aspects qualitatifs (sémantique et pragmatique de l'ontologie du système de connaissances) liés à cette activité ne feront pas l'objet d'une réflexion spécifique. Le choix des variables et l'interprétation des alertes sont donc simplement considérés dans le cadre de la complémentarité entre les activités de contrôle et de modélisation/analyse des tendances, telle qu'évoquée dans le premier chapitre de la thèse.

Au sujet des méthodes quantitatives, un certain nombre de problèmes d'application locaux, propres au terrain de recherche STMicroelectronics ont été identifiés et ont dû être traités pour mettre en œuvre le modèle de rationalité des décideurs (voir Modèle 2).

En effet, afin de satisfaire ce modèle de rationalité, une première procédure de détection automatique de comportements anormaux a été mise en place au sein de notre terrain de recherche, en utilisant des méthodes statistiques classiques destinées à mettre en œuvre les deux étapes du contrôle, à savoir :

- *Modélisation de l'évolution de chaque variable en tant que série chronologique* : cette modélisation a été réalisée par l'utilisation d'un algorithme de lissage exponentiel de type Holt-Winters (Makridakis et al., 1998 ; ONS, 2008). C'est une version élaborée de lissage exponentiel simple, qui permet de tenir compte d'une tendance dans la série (méthode de Holt), ainsi que de saisonnalités additives ou multiplicatives (Holt-Winters). Comparée aux méthodes stochastiques de type ARIMA, celle-ci est particulièrement adaptée dans des contextes organisationnels où l'on a besoin de modéliser beaucoup de séries, sans avoir le temps de procéder à des analyses préliminaires systématiques. Elle est de plus une bonne approximation pour bon nombre de séries temporelles, souvent suffisante dans le cadre d'usages industriels (Makridakis et al., 1998 ; De Gooijer et Hyndman, 2006 ; ONS, 2008 ; Croux et al., 2011). Par ailleurs, nous avons constaté que ce modèle est plutôt bien validé au sein de notre terrain de recherche : une grande majorité de séries correspondent à des modèles de type lissage exponentiel simple ou Holt. Or, dans ces cas de figure, les méthodes Holt-Winters et ARIMA sont directement équivalentes (Makridakis et al., 1998 ; Gardner, 2006) : le lissage simple correspond à un modèle ARIMA (0,1,1) et le modèle de Holt à un modèle ARIMA (0,2,2) ;
- *Modélisation des résidus du modèle obtenu lors de l'étape 1, par carte de contrôle sous hypothèse de normalité* : dans le cas usuel de contrôle des résidus d'une série temporelle, on a recours à une carte de contrôle de Shewart pour les individus classique, étalonnée sur la base d'une variance définie à partir d'un estimateur usuel (Montgomery, 2005), comme par exemple la méthode du *moving range*, basée sur les variabilités observées entre deux observations successives.

Néanmoins, cette approche générale a rapidement révélé plusieurs limitations d'applications, compte-tenu d'un certain nombre de caractéristiques propres au terrain de recherche considéré.

### 2.1.2. Axes d'améliorations

#### i. Amélioration de la robustesse du contrôle

Tout d'abord, nous avons constaté la présence fréquente de valeurs aberrantes dans les données (une valeur aberrante est une observation qui se trouve numériquement loin des autres données). Ces valeurs peuvent poser des problèmes de robustesse lors d'une estimation ou d'une modélisation statistique, c'est-à-dire que les paramètres estimés sont excessivement altérés par ces valeurs aberrantes. Effectivement, le manque de robustesse de l'approche générale a pu être observé : chacune des étapes du contrôle s'est montrée sensible aux valeurs aberrantes, qui pouvaient être présentes dans les séries considérées. En résultait une baisse significative de la qualité des contrôles statistiques.

Une première amélioration a été implémentée, en modifiant la carte de contrôle par l'utilisation d'un estimateur robuste de la variance. L'estimateur  $Q_n$ , proposé par Rousseeuw et Croux (1993), a été utilisé pour ce faire. Il est défini proportionnellement au quartile des distances  $\{|e_i - e_j|; i < j\}$ . Néanmoins, le reste de l'approche usuelle était conservé et il a semblé nécessaire de proposer une approche plus globale, améliorant également la robustesse de la modélisation de la série temporelle, tel que nous l'expliquons dans le paragraphe 2.2.

*ii. Prise en compte des changements structurels*

Ensuite, on peut observer régulièrement la présence de changements structurels dans les séries considérées. Ils se manifestent essentiellement par la modification brusque de la moyenne de la série, ou de sa pente, ou des deux.

Dans ce cas de figure, l'application d'une modélisation globale à l'ensemble de la série peut s'avérer problématique : on cherche en effet à déterminer des paramètres de lissage pour caractériser une unique série, qui s'avère en fait être composée de plusieurs séries successives et différentes.

*iii. Mise en œuvre d'un contrôle multivarié*

Enfin, il a semblé nécessaire de tenir compte du fait que, lors d'une implémentation industrielle, l'objectif n'est pas de mettre sous contrôle une seule série, mais un ensemble de séries réalisées simultanément. Nous sommes donc en présence d'une problématique de tests multiples, ou contrôle multivarié, qui n'est pas sans incidence sur la stratégie de contrôle. En effet, en testant  $H_0$  avec un intervalle de confiance de 95%, on admet un risque de fausses alarmes de 5% par variable. En conséquence, si l'on effectue  $p$  contrôles simultanés, en supposant les variables indépendantes, la probabilité d'émettre une fausse alarme devient  $\{1 - (1 - 0,05)^p\}$ . La Figure 36 illustre ce problème pour  $n \leq 100$ .

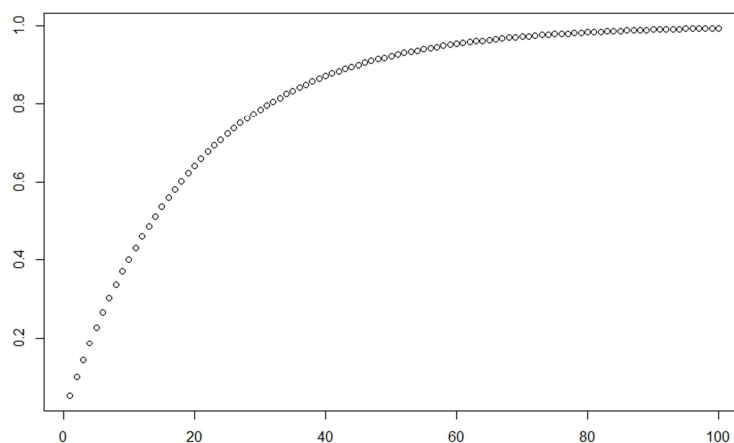


Figure 36. Evolution du risque de fausses alarmes en fonction du nombre de séries contrôlées (cas de variables indépendantes)



Dans le cas du terrain de recherche considéré,  $n > 100$  : on observera donc, de façon quasi-systématique, une fausse alarme par jour, si l'on ne tient pas compte de la dimension du problème.

Notre objectif a donc été d'implémenter des solutions à ces problèmes locaux d'application de l'approche générale de détection de comportements anormaux.

## 2.2. Solutions quantitatives mises en œuvre

*Remarques : les contributions présentées ci-après sont le fruit d'une collaboration scientifique réalisée dans le cadre du Master d'Espéran Padonou. Nous le remercions chaleureusement pour son aide précieuse, ainsi que pour la qualité de son travail.*

*Par ailleurs, compte-tenu des contraintes de production de cette thèse, la présentation et les tests de performance complets des méthodes évoquées ci-dessous ne peuvent être exhaustivement développés ici. Pour plus de détail, on se référera à Padonou et al. (2013)<sup>44</sup>.*

### 2.2.1. Amélioration de la robustesse du contrôle : présentation de l'approche implémentée et évaluation théorique de la performance

#### i. Approche implémentée

On cherche à rendre la détection automatique de comportements anormaux moins sensible aux valeurs aberrantes. Pour cela, l'amélioration de l'approche générale s'est faite par l'implémentation d'une méthode robuste de détection automatique de comportements anormaux, basée sur un lissage Holt-Winters robuste (Gelper et al., 2009 ; Croux et al., 2011). L'objectif de ce lissage robuste est de corriger les estimations  $\hat{x}_t$ , données par l'algorithme de Holt-Winters classique, en fonction de l'erreur de modélisation  $e_t$ . Pour cela, une formule récursive est utilisée afin de prédire l'échelle d'erreur  $\sigma_t$  attendue en date  $t$ . En cas d'erreur observée trop importante par rapport à  $\sigma_t$ , l'estimation est corrigée avant la poursuite du lissage. Pour cette limitation de l'influence d'une grandeur, l'algorithme robuste se base sur une fonction filtre de Huber de correction des erreurs (H), représentée dans la Figure 37.

---

<sup>44</sup> Version de prépublication en annexe 6.

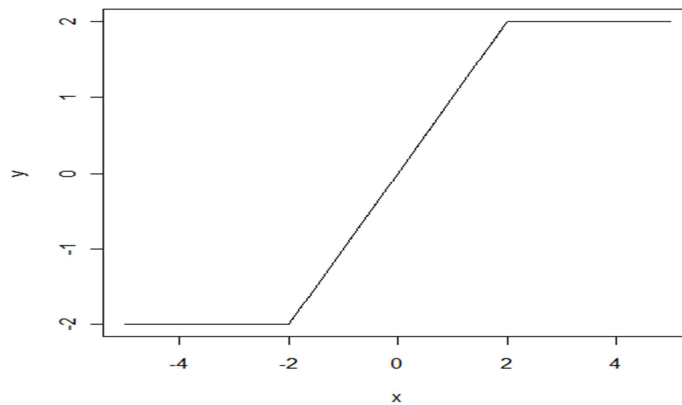


Figure 37. Fonction H de correction des erreurs de Huber ( $k = 2$ ).

Lorsque  $e_t$  dépasse  $k$  fois l'erreur prévue  $\sigma_t$  ( $k = 2$  dans la Figure 37), la partie excédante n'est pas prise en compte pour l'estimation du modèle. Le modèle est donc estimé, non pas à partir des vraies valeurs de  $x_t$ , mais de valeurs corrigées destinées à réduire l'influence des valeurs aberrantes. En conséquence, cet algorithme dépend fortement de l'historique de la série et nécessite d'être initialisé de manière robuste sur une période d'apprentissage, qui permet de déterminer :

- Les paramètres de lissage robustes, par l'emploi d'une méthode de régression robuste ;
- Un ordre de grandeur de résidus, grâce à un estimateur robuste. Les auteurs proposent d'employer le  $\tau$ -scale estimator, basé sur la médiane des erreurs.

Cet estimateur robuste est également utilisé pour calibrer les limites de contrôle de la carte de contrôle des résidus. L'implémentation de cette approche a notablement amélioré les résultats du contrôle de comportements anormaux, en maîtrisant mieux la détection de valeurs hors-contrôle et les risques de fausses alarmes. La Figure 38 montre un cas concret d'application de l'approche classique *versus* robuste (les points rouges représentent les observations hors-contrôles du point de vue de la méthode statistique appliquée).

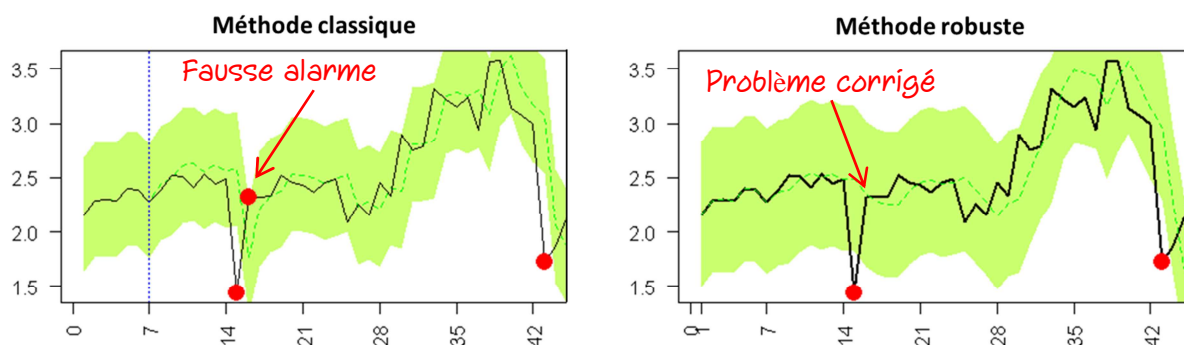


Figure 38. Illustration des bénéfices apportés par l'application de la méthode robuste : limitations des fausses alertes consécutives à la présence de points atypiques.

On peut voir que la valeur hors-contrôle est bien détectée pour l'observation 15. Le point suivant (16) correspond à un retour à la normale de la valeur. Dans le cas de la

méthode classique, cette observation 16 est détectée elle-aussi comme hors-contrôle, car le contrôle est biaisé par la valeur atypique précédente : c'est une fausse alarme. On constate que la méthode robuste permet de corriger ce problème.

La performance théorique de cette approche robuste a été évaluée en comparaison de celle de la version classique.

ii. *Evaluation théorique de la performance*

Protocole

Pour chacune des méthodes, classique et robuste, nous cherchons à évaluer et comparer deux caractéristiques :

- *La puissance du contrôle* : c'est la probabilité qu'une observation hors-contrôle soit effectivement détectée ;
- *Le taux de fausse détection* : c'est la probabilité qu'une observation sous contrôle soit identifiée étant hors-contrôle. Ce risque doit être comparé au niveau de confiance de la carte de contrôle. Par exemple, pour un niveau de confiance égale à 95%, tel que nous le fixerons ci-après, on s'attend à un taux de fausse détection de 5%.

Les données utilisées pour quantifier ces deux caractéristiques sont obtenues selon la procédure suivante :

- *Simulation d'une série temporelle*, de la forme ARIMA(0,2,2) et de longueur 100 ;
- *Contamination de cette série temporelle*, à partir d'un taux de contamination  $R$ . On tire au hasard  $100R$  dates qui vont être transformées en valeurs anormales. Pour cela, on leur rajoute ou soustrait (choix au hasard) une valeur  $e$ , définie selon quatre cas de figure :
  - \*  *$e$  est constant de valeur 10,*
  - \*  *$e$  est constant de valeur 20,*
  - \*  *$e$  est constant de valeur 30,*
  - \*  *$e$  est une valeur tirée au hasard parmi les valeurs  $\{10, 20, 30\}$  ;*
- *Contrôle statistique des séries contaminées*, puis mesure :
  - \* *De la puissance du contrôle*, en calculant le pourcentage de dates contaminées détectées comme valeurs hors-contrôle,
  - \* *Du taux de fausse détection*, en calculant le pourcentage de dates non contaminées détectées comme valeurs hors-contrôle,

Ce contrôle est réalisé de façon dynamique, tel qu'appliqué en réalité chez STMicroelectronics, c'est-à-dire que les paramètres de contrôle (paramètres de lissage, limites de contrôle) sont estimés pour chacun des jours contrôlés.

Pour chacun des cas de figure de définition de  $e$  (10, 20, 30 ou tirage au hasard), la procédure est répétée cent fois. La moyenne de la puissance du contrôle et celle du taux de fausse détection de chacune des cent répétitions donnent ainsi une estimation des caractéristiques des deux types de méthodes de contrôle, selon les différents cas de figure de  $e$ .

### Résultats

Les résultats du protocole d'évaluation des méthodes classique et robuste sont résumés dans le Tableau 10.

	Méthode de contrôle classique			Méthode de contrôle robuste		
	Puissance du contrôle					
	R = 2%	R = 5%	R = 10%	R = 2%	R = 5%	R = 10%
$e = 10$	99,6%	95,4%	80,0%	100,0%	100,0%	99,7%
$e = 20$	100,0%	93,6%	80,0%	100,0%	100,0%	100,0%
$e = 30$	100,0%	97,0%	78,6%	100,0%	100,0%	100,0%
$e \in \{10, 20, 30\}$	95,0%	70,0%	52,0%	100,0%	100,0%	100,0%
	Taux de fausse détection					
	R = 2%	R = 5%	R = 10%	R = 2%	R = 5%	R = 10%
$e = 10$	4,6%	4,3%	4,2%	6,3%	6,8%	8,6%
$e = 20$	3,3%	2,2%	2,0%	6,0%	7,1%	9,1%
$e = 30$	2,5%	2,2%	2,4%	6,2%	6,9%	9,3%
$e \in \{10, 20, 30\}$	3,0%	3,2%	2,0%	6,1%	7,0%	7,2%

Tableau 10. Méthodes de contrôle classique et robuste : évaluation théorique de la performance par simulation.

Les simulations ont montré que le contrôle classique est soumis à deux effets :

- D'une part, les limites de contrôle s'élargissent en fonction du nombre de points atypiques présents dans la série. En conséquence, le taux de fausse détection s'en trouve réduit, mais la puissance du contrôle est fortement dégradée ;
- D'autre part, les prédictions consécutives à un point hors-contrôle sont généralement biaisées. Ce phénomène peut amener à augmenter le nombre de fausses alarmes (cf. Figure 38).

En l'occurrence, lors des simulations, c'est le premier des deux effets constatés qui domine. Cela explique un taux de fausse alarme un peu plus bas que les 5% attendus ; ainsi qu'une puissance de contrôle qui diminue fortement lorsque les séries sont contaminées. C'est particulièrement vrai dans le cas  $e \in \{10, 20, 30\}$ , qui est le plus représentatif des séries réellement rencontrées chez STMicroelectronics.

Au sujet du contrôle robuste, on peut observer que la puissance du contrôle n'est pas altérée par la contamination des séries. Le taux de détection reste égal à 100%. En contrepartie, on constate que le taux de fausse alarme est, dans le cas robuste, légèrement supérieur aux 5% attendus. Néanmoins, de façon générale, on peut conclure que la procédure de contrôle robuste est bien meilleure que la procédure classique, et ce particulièrement dans le cas réaliste où  $e \in \{10, 20, 30\}$  : la légère hausse du risque de fausses alarmes (qui reste toutefois proche du niveau théorique attendu) est très largement compensée par une hausse très significative de la puissance du test.

### 2.2.2. Prise en compte des changements structurels : présentation de l'approche implémentée

L'objectif est de limiter les fausses alertes liées à la présence de changements structurels dans les séries observées : en cas de changement structurel, le lissage robuste présenté ci-dessus pose problème. Effectivement, lors d'un tel changement, le principe du lissage étant de corriger les valeurs observées en tenant compte de l'historique de la série, il entraînera une succession d'erreurs importantes qui se manifesteront par une « rafale » de fausses alarmes (Figure 39).

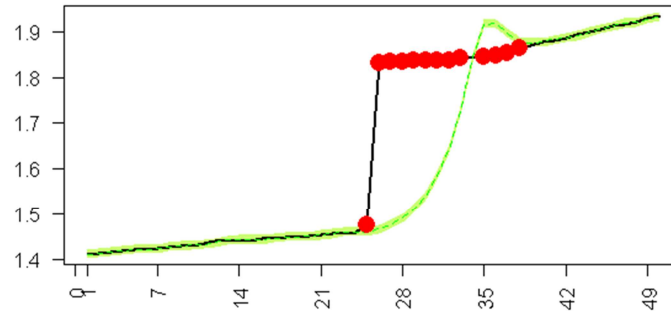


Figure 39. Problème lié au lissage robuste en cas de changement structurel : succession de fausses alarmes.

Il est donc nécessaire de pouvoir caractériser et traiter de tels changements structurels. Padonou et *al.* (2012, 2013) ont montré que, pendant un changement structurel, les valeurs estimées par l'algorithme Holt-Winters robuste ne tiennent plus compte de la série observée, mais sont données par une fonction du temps, déterministe et monotone, tracée en rouge dans la Figure 40.

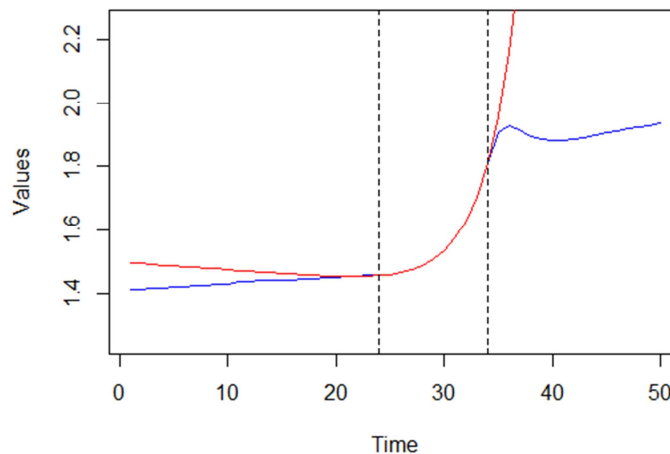


Figure 40. Problème lié au lissage robuste en cas de changement structurel : les prévisions ne tiennent plus compte des valeurs observées.

En réponse à ce problème, une procédure basée sur le test de Chow (voir Johnston et Dinardo, 1999 ; Hansen, 2001) a été employée pour détecter ce genre de changement. Ce test peut poser des difficultés d'utilisation car il suppose de connaître *a priori* le point de changement structurel. Dans le cas ici considéré, on peut contourner la difficulté en

identifiant ces points à l'avance. En effet, comme le suggèrent Padonou et *al.* (2012, 2013), on peut calculer un coefficient  $Q_t$  donnée par :

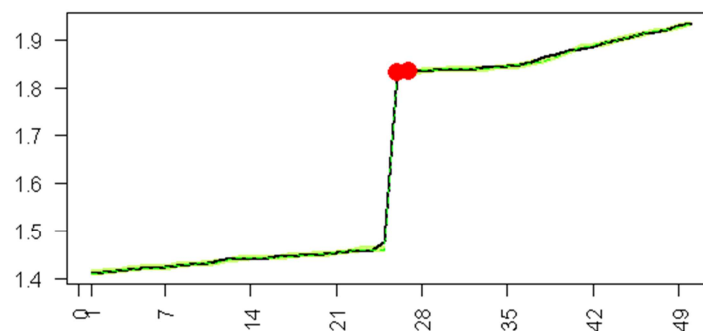
$$Q_t = \frac{\hat{x}_t^* - \hat{x}_t}{\sigma_t}$$

où  $\hat{x}_t^*$  est l'estimation fournie par le lissage robuste,  $\hat{x}_t$  l'estimation fournie par le lissage classique, et  $\sigma_t$  l'erreur prévue en date  $t$ .

Or, lors d'un changement structurel, on peut émettre l'hypothèse que l'erreur commise reste assez grande pour être toujours tronquée, c'est-à-dire que la fonction  $H$  de correction des erreurs reste égale à  $k$  ou  $-k$  (Figure 37). Dans ce cas,  $Q_t$  restera constant et vaudra  $k$  ou  $-k$ . Fondée sur cette hypothèse, la procédure de réapprentissage suivante, permettant de tenir compte des changements structurels lors de la modélisation automatique, peut être proposée :

- *Etape 1* : lissage robuste, calcul de  $Q_t$ , détection des sous vecteurs de  $Q_t$  constants égaux à  $k$  ou  $-k$  ;
- *Etape 2* : tests de monotonie sur les sous vecteurs trouvés. La plus grande erreur de chaque sous vecteur marque le début d'un changement structurel potentiel ;
- *Etape 3* : au(x) point(s) de changement(s) structurel(s) potentiel(s) identifié(s), test de Chow (emploi d'une méthode de régression linéaire robuste) ;
- *Etape 4* : si le test de Chow est positif, réapprentissage des paramètres de lissage avant et après le point de changement structurel identifié ;
- *Etape 6* : actualisation de  $Q_t$ .

L'implémentation sur le terrain de cette procédure a significativement amélioré la détection automatique de comportements anormaux, en limitant les fausses alarmes consécutives à un changement structurel. La Figure 41 en fournit l'illustration. Elle doit être mise en regard de la Figure 39, qui n'intègre pas la procédure de réapprentissage présentée.



*Figure 41. Traitement du problème lié au lissage robuste en cas de changement structurel : illustration des effets de la procédure de réapprentissage.*

On constate que, grâce à la procédure de réapprentissage, les fausses alarmes successives ont bien disparues. Les points détectés comme hors-contrôles sont effectivement des valeurs anormales, à la vue des observations qui leurs sont antérieures.

### 2.2.3. Mise en œuvre d'un contrôle multivarié : présentation de l'approche implémentée

Le but est de limiter les fausses alertes liées au nombre de séries contrôlées simultanément. Face à une telle problématique de tests multiples, trois solutions ont été envisagées :

- *Elargissement des intervalles de confiance*, par corrections de type Šidàk-Bonferroni (Abdi *in* Salkind, 2007) ;
- *Contrôle statistique sur variables latentes*, calculées par ACP ou régression PLS (Mastrangelo *et al.*, 1996 ; Montgomery, 2005) ;
- *Carte de contrôle multivariée de Hotelling* (Montgomery, 2005).

Les corrections de type Šidàk-Bonferroni n'ont pas été retenues. Comme l'indique Abdi (*in* Salkind, 2007), elles peuvent devenir très restrictives lorsque la dimension du problème augmente et que les variables ne sont pas indépendantes. Effectivement, les intervalles de confiance peuvent s'élargir au point de nuire significativement au taux de détection. Nous avons donc préféré envisager des méthodes plus courantes en contrôle statistique industriel. D'une part, le contrôle statistique sur variables latentes : cette méthode n'a pas été explorée en détail, car les axes factoriels utilisés ne sont pas des variables physiques et posent des problèmes d'interprétations des comportements anormaux détectés. D'autre part, la carte de contrôle Hotelling, basée sur la statistique du khi-deux ( $\chi^2$ ) : c'est sur cette méthode qu'ont porté nos travaux.

Ainsi, nous utilisons la carte de contrôle Hotelling pour contrôler simultanément les  $p$  résidus de  $p$  modélisations, faites grâce à la méthode robuste présentée précédemment.  $E$  est alors un vecteur aléatoire, qui correspond à une réalisation simultanée de l'ensemble de ces résidus.  $E$  suit une loi multidimensionnelle de dimension  $p$ , supposée normale, de moyenne nulle et de matrice de covariance  $S$ . Cette hypothèse est cohérente avec la prise en compte de résidus d'une modélisation Holt-Winters. Par ailleurs, elle a été vérifiée par l'analyse des séries observées sur le terrain. Sous cette hypothèse, on peut réduire les  $p$  dimensions de  $E$  à une seule statistique  $T^2$ , qui suit une loi de type  $\chi^2$  lorsque les paramètres sont supposés connus, ou de type Fisher dans le cas contraire. On peut ensuite définir des intervalles de confiance pour cette statistique, en connaissant les quantiles de sa loi.

Tout comme dans le cas univarié, les limites de la carte de contrôle doivent être estimées de façon robuste, pour limiter la sensibilité à d'éventuelles valeurs aberrantes. Pour cela, les recommandations de Rousseeuw et Leroy (1987) et Pison *et al.* (2002) ont été suivies, pour une estimation robuste de l'ellipsoïde de confiance. La Figure 42 montre les apports de cette approche robuste.

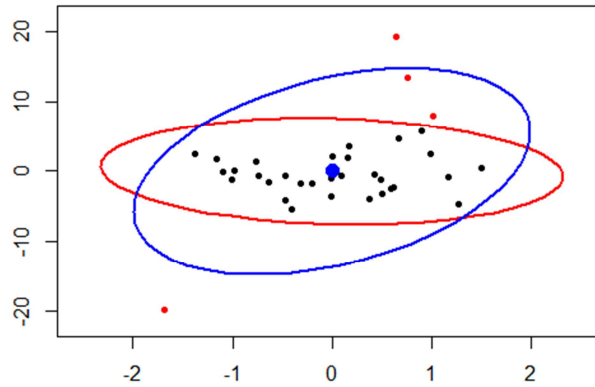


Figure 42. Ellipsoïde de confiance classique et robuste : meilleure orientation dans le 2ème cas.

A cause des valeurs aberrantes (points rouge), on constate que l'ellipse de confiance classique, en bleue, a une aire trop grande. Il y a trop d'espace entre ses bords et les points ; de plus, cet ellipsoïde est incliné comme si les variables étaient corrélées. Or, on voit que ces deux problèmes sont traités par l'estimation robuste de l'ellipse de confiance (en rouge).

Finally, a last element has to be treated to allow an industrial implementation of this multivariate control method. In fact, the alarm given by a group of variables, starting from the  $T^2$  statistic, is a first indicator of abnormal behavior. However, in addition, it is important that users be able to know which of the  $p$  variables of origin can be at the origin of the alarm.

To facilitate this analysis, Montgomery (2005) introduces the notion of partial contributions, which can complete the Hotelling control chart. The partial contribution of a variable of origin given in the multivariate signal, corresponds to the difference between the global  $T^2$  statistic and the  $T^2$  statistic that would have been obtained if this variable did not exist. The higher the partial contribution of a variable, the more one can suspect it of presenting abnormal behavior. For a confidence level  $\alpha$ , one can therefore decree, in view of the group of variables, that a variable of origin presents abnormal behavior if its partial contribution exceeds the value given by the  $\alpha$  quantile of the  $\chi^2$  law with a degree of freedom.

The two robust approaches, univariate and multivariate, have been implemented jointly at STMicroelectronics, in order to offer a general analysis framework to users. These approaches can emit contradictory signals. In fact, it is possible that the multivariate control chart emit an alarm not corresponding to any out-of-control value of the univariate case (for example, if two correlated variables become independent, while remaining in a stable order of magnitude). In the framework of the results presented in this thesis, one will limit oneself to assuming that it is for the users to analyze by themselves these results and to make their own interpretation of the quantitative results.



## Conclusion du chapitre 5

Dans ce chapitre, l'application du processus socio-statistique, défini précédemment, a permis de construire des solutions répondant aux exigences des deux activités ITIL considérées dans cette thèse. Ces solutions sont élaborées afin d'être pleinement adaptées aux conditions locales de STMicroelectronics.

Concernant l'activité de modélisation prospective, une démarche complète de modélisation a été définie. Elle tient compte des recommandations génériques développées dans la première partie de la thèse, afin d'inscrire la démarche de modélisation dans une réelle perspective de création de connaissance. Pour cela, les acteurs de l'organisation sont impliqués autant que nécessaire. De plus, en étendant les acteurs mobilisés au-delà du strict périmètre des experts de l'infrastructure TI-C2, l'intégration de la dimension métier dans la construction de connaissances est favorisée. Cela permet d'assurer la prise en compte de l'enjeu de gestion n°2 (nous avons vu que l'emploi de méthodes statistiques est une réponse en tant que telle à l'enjeu n°1). Par ailleurs, en proposant une structuration progressive de connaissances, la prise en compte de cet enjeu de gestion est garantie grâce à une étape de modélisation, explicitement dédiée à l'intégration de variables métier utiles à l'aide à la décision (intégration des variables de pilotage industrielles). Le large panel de méthodes statistiques existantes nous a permis de spécifier un certain nombre de méthodes adaptées à chacune des étapes de la démarche de modélisation. Chacune de ces méthodes permet de structurer les données brutes, mises à disposition dans la base de performance et les bases de données applicatives, pour constituer une information, qui permettra de tendre vers la construction de connaissances nouvelles. De plus, certaines méthodes offrent aussi la possibilité de contrôler dans le temps la validité de ces connaissances, contribuant ainsi à leur génétique.

L'activité de détection de comportements anormaux est un moyen additionnel de contribuer à cette génétique du système de connaissances, en alertant les experts de STMicroelectronics de tout comportement inhabituel du système. Ces alertes peuvent être révélatrices d'évolutions ou de modifications, devant être apportées à la connaissance formalisée dans le cadre du modèle de prévision. Pour cela, nous avons proposé un certain nombre de méthodes statistiques permettant de garantir une bonne détection de comportements anormaux, dans un contexte perturbé (présence de points atypiques et de changements structurels) et de grande dimension.

Nous pensons que l'ensemble de ces propositions socio-statistiques auraient pu apporter une contribution significative au processus décisionnel en action lors de la migration MES. Pour le démontrer, une aide à la décision *a posteriori* va être apportée dans la partie suivante, en s'appuyant sur les travaux développés dans cette deuxième partie. La troisième partie s'appliquera ainsi à formaliser la connaissance attendue dans le cadre du projet de migration MES, par l'application de la démarche de modélisation proposée. Cette connaissance sera ensuite utilisée à des fins décisionnelles, tout en la mettant sous contrôle et en assurant la génétique.

**Troisième partie.** Application du processus socio-statistique d'aide à la décision : application de la démarche de modélisation, usage de la connaissance créée et évolution.

## Plan détaillé de la troisième partie.

### Chapitre 6. Application de la démarche de modélisation : création et formalisation de connaissances.

#### Introduction.

#### Section 1. Application à la migration MES : constitution d'un modèle de prévision.

1.1. Etape (A) intermédiaire d'observation n°1.

1.1.1. Récolte des données (aspect syntaxique).

1.1.2. Structuration des données – informations (aspect sémantique).

1.1.3. Interprétation des informations – connaissances (aspect pragmatique).

1.2. Etape (B) intermédiaire d'observation n°2.

1.2.1. Récolte des données (aspect syntaxique).

1.2.2. Structuration des données – informations (aspect sémantique).

1.2.3. Interprétation des informations – connaissances (aspect pragmatique).

1.3. Etape (C) d'aide à la décision.

1.3.1. Récolte des données (aspect syntaxique).

1.3.2. Structuration des données – informations (aspect sémantique).

1.3.3. Interprétation des informations – connaissances (aspect pragmatique).

#### Section 2. Quelle généralité de la démarche de modélisation ?

2.1. Justification du questionnement et présentation d'un nouveau cas d'application.

2.1.1. Généralité de la méthode : proposition.

2.1.2. Un nouveau cas d'application.

2.2. Structuration progressive de connaissances pour la modélisation du serveur RTD.

2.2.1. Etape (A) intermédiaire d'observation n°1.

2.2.2. Etape (B) intermédiaire d'observation n°2.

2.2.3. Etape (C) d'aide à la décision.

#### Conclusion du chapitre 6.

### Chapitre 7. Usage du modèle de prévision pour l'aide à la décision.

#### Introduction.

#### Section 1. Construction du modèle d'évaluation : mobilisation de la connaissance pour l'aide à la décision.

1.1. Formalisation du modèle d'évaluation.

1.2. Considération méthodologique pour l'utilisation du modèle de prévision dans le cadre de la procédure d'agrégation du modèle d'évaluation.

#### Section 2. Utilisation du modèle d'évaluation : résultat et recommandation pour l'aide à la décision.

2.1. Simulation et recommandation.

2.2. Discussion.

**Conclusion du chapitre 7.**

## **Chapitre 8. Evolution de la connaissance organisationnelle.**

**Introduction.**

**Section 1. Stratégie et architecture technique de maintien à jour de la connaissance du système.**

1.1. Stratégie.

1.2. Architecture technique.

**Section 2. Illustration de l'application du maintien à niveau de la connaissance.**

2.1. Présentation du cas : passage en production de la nouvelle architecture MES.

2.2. Identification de la rupture de la stabilité structurelle de modèle de prévision.

2.3. Détection de comportements anormaux.

2.4. Discussion.

**Conclusion du chapitre 8.**

\*  
\* \*

Cette partie illustre les étapes de formalisation, d'usage et de maintien à jour de connaissances statistiques. Le plan général de cette partie est fourni par la Figure 43.

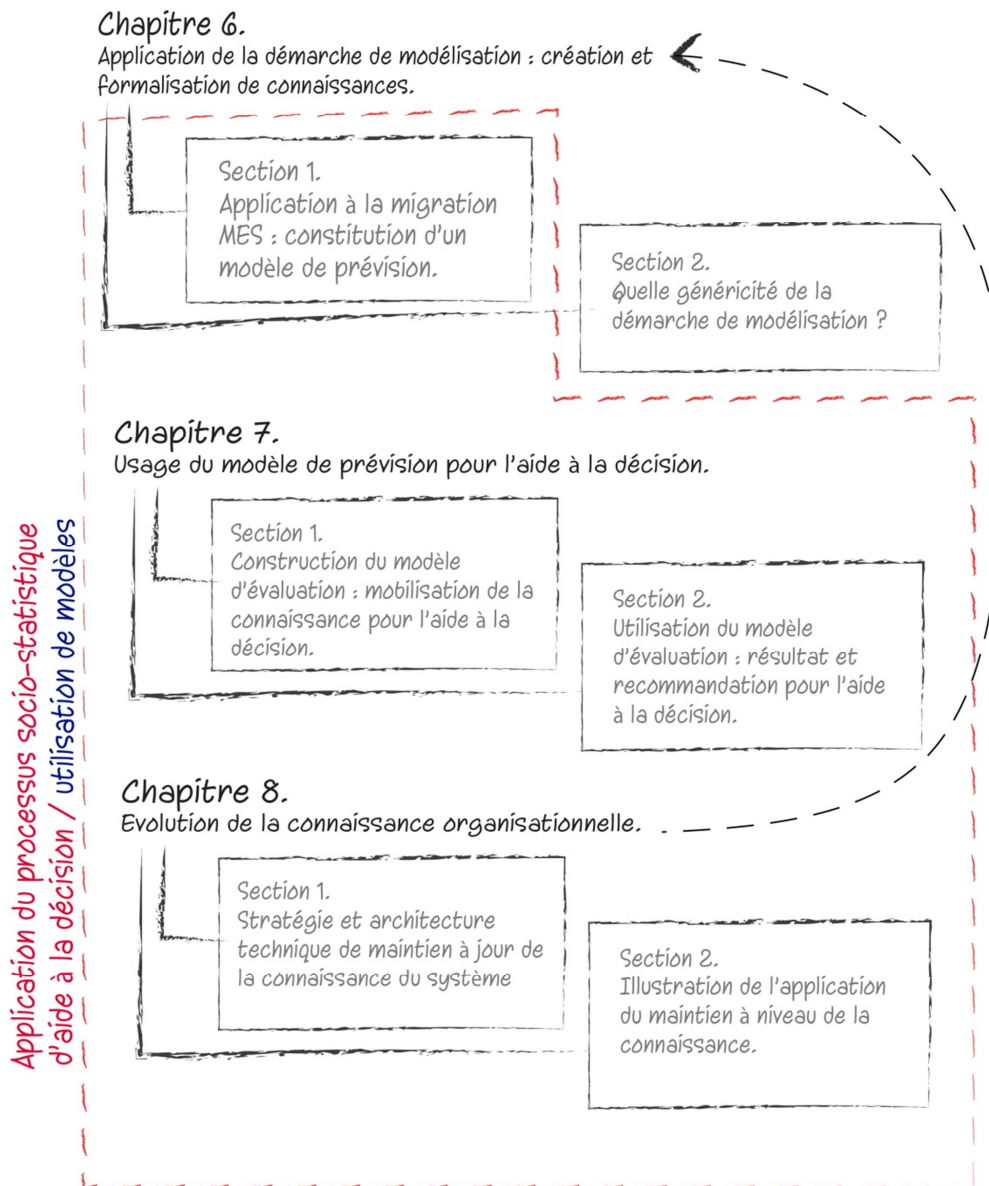


Figure 43. Plan général de la partie 3.

La section 1 du chapitre 6 propose l'application directe de la démarche de modélisation appliquée au cas MES, initié dans la partie précédente, afin de formaliser le modèle de prévision attendu (questionnement central des recherches). La section 2 suggère un questionnement complémentaire, concernant la potentielle généralité de la démarche de modélisation proposée : en effet, il semblerait qu'elle puisse s'appliquer à d'autres cas où le besoin décisionnel est le même. Cela permettrait alors d'apporter des connaissances complémentaires, améliorant ainsi la compréhension des interactions entre les sous-systèmes du système technologique considéré.

Le chapitre 7 concerne la suite de l'application de la démarche de modélisation chez STMicroelectronics : il s'agit de mettre en usage le modèle mathématique de prévision obtenu, dans un contexte opérationnel d'aide à la décision. Dans la section 1, le modèle de prévision MES est mobilisé pour formaliser un

modèle d'évaluation. Ce modèle d'évaluation est destiné à utiliser la connaissance statistique, dans un but d'aide à la décision. Ensuite, la section 2 montre comment le modèle d'évaluation peut être mobilisé pour produire un résultat quantitatif, afin de fournir une recommandation supportant la prise de décision. On cherche pour cela à anticiper la charge des serveurs informatiques MES, face à une évolution des activités industrielles.

Suite à cette application, le chapitre 8 indique comment la dimension génétique du système de connaissances organisationnelles peut être prise en compte au travers du processus socio-statistique d'aide à la décision. On montre comment cet aspect de nos travaux a été concrètement mis en œuvre chez STMicroelectronics (section 1 : stratégie et architecture technique de maintien à niveau de la connaissance) et une illustration est proposée, à partir du cas MES (section 2). En cas d'obsolescence de la connaissance formalisée, la démarche de modélisation pourra être réappliquée, tel que suggéré dans la Figure 43 par une liaison entre les chapitres 8 et 6.

Dans la continuité du cas développé dans le chapitre 2, la Figure 44 montre comment les chapitres et sections de cette troisième partie s'intègrent dans la conduite du processus socio-statistique d'aide à la décision, appliqué au cas MES.

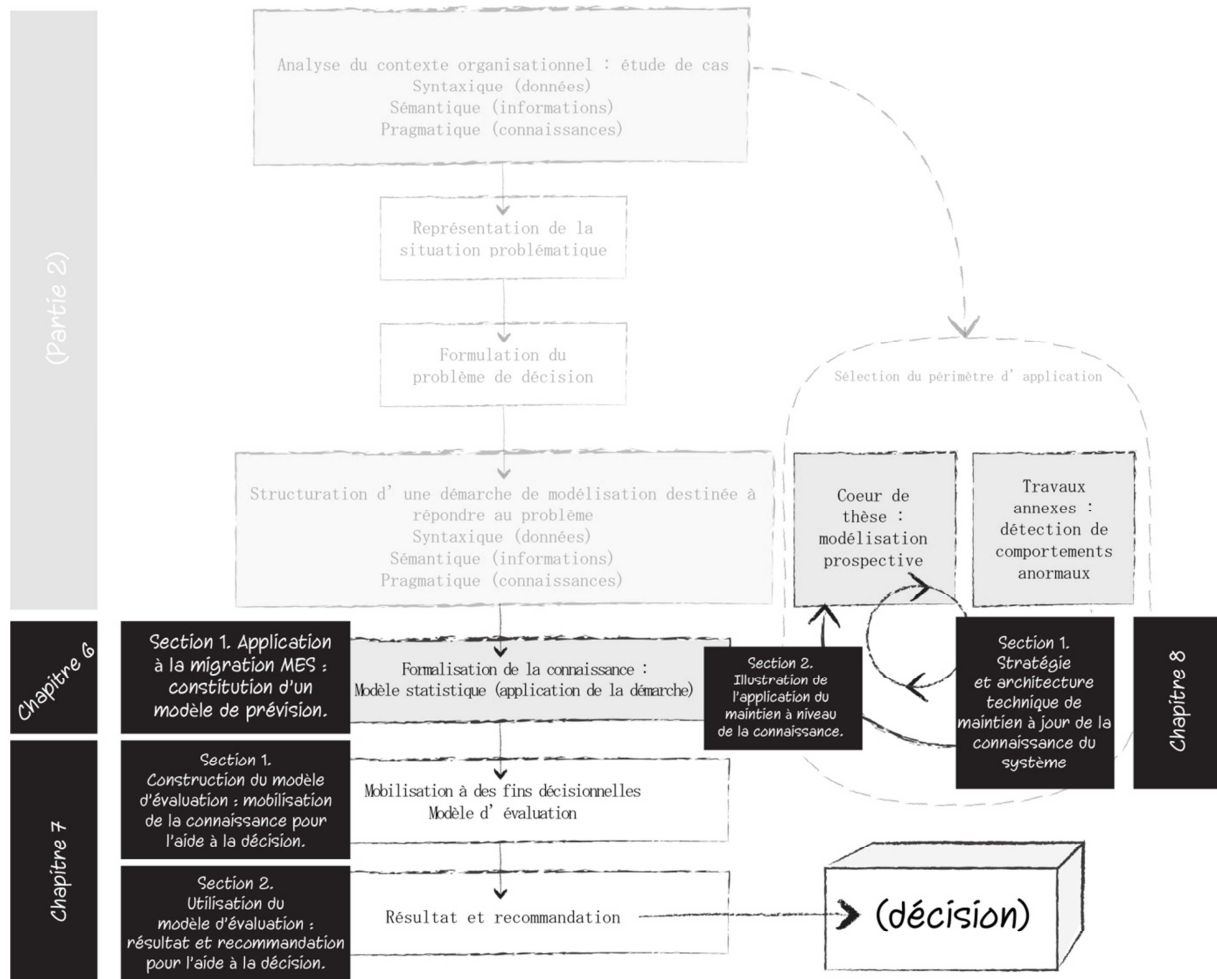


Figure 44. Inscription des chapitres 6, 7 et 8 au sein du processus socio-statistique d'aide à la décision.

\*  
\* \*

## Commentaire préliminaire concernant les traitements quantitatifs

Cette partie s'appuie un ensemble de traitements quantitatifs. Ils ont été réalisés avec le logiciel R (R Development Core Team, 2011), version 2.13.1. Des bibliothèques ou des développements spécifiques ont pu être utilisés, en plus de la distribution standard du logiciel. Dans certains cas, des vérifications additionnelles ont été réalisées avec d'autres logiciels. Tous ces moyens informatiques utilisés sont mentionnés dans le Tableau 11.

Traitement quantitatif	Librairie R ou logiciel spécifique
ACP classique	distribution standard ; <i>rrcov</i> (Todorov et Filzmoser, 2009)
Rotation <i>Varimax</i>	<i>psych</i> (Revelle, 2011)
Régression PLS	<i>mixOmics</i> (Déjean et al., 2012) ; <i>plsrm</i> (Sanchez, 2012) ; vérification avec les logiciels TANAGRA 1.4.40 (Rakotomalala, 2012) et développement spécifique sous Excel 2010
Régression linéaire multiple	distribution standard
Traitement de la colinéarité	distribution standard ; <i>car</i> (Fox et Weisberg, 2011)
Régression linéaire multiple avec résidus ARIMA	distribution standard; <i>forecast</i> (Hyndman et Khandakar, 2008)
Emploi de variable indicatrices	distribution standard
Validation croisée	<i>DiceEval</i> (Dupuy et Helbert, 2012)
Validation croisée pour séries temporelles	développement spécifique
Tests de stabilité structurelle statique	<i>strucchange</i> (Zeileis et al., 2001, 2003 ; Zeileis, 2004 ; Zeileis et Kleiber, 2005)
Datation de points de rupture de stabilité	<i>strucchange</i> (Zeileis et al., 2003 ; Zeileis et Kleiber, 2005)
Tests de stabilité structurelle dynamique	<i>strucchange</i> (Zeileis et al., 2001)
Lissage exponentiel Holt-Winters	distribution standard
Lissage exponentiel Holt-Winters robuste	développement spécifique (E. Padonou)
Carte de contrôle univariée robuste	<i>qcc</i> (Scrucca, 2004) ; développement spécifique (avec E. Padonou)
Carte de contrôle multivariée de Hotelling	<i>qcc</i> (Scrucca, 2004) ; développement spécifique (E. Padonou)
Estimation robuste de l'ellipsoïde de confiance	<i>MASS</i> (Venables et Ripley, 2002), <i>rrcov</i> (Todorov et Filzmoser, 2009)
Analyse des contributions partielles	développement spécifique (E. Padonou)

Tableau 11. Moyens informatiques mobilisés.

**Chapitre 6. Application de la démarche de modélisation : création et formalisation de connaissances.**



## Introduction

Ce chapitre traite de la deuxième étape du questionnement central de cette thèse : l'application de la démarche de modélisation développée dans le cadre du processus socio-statistique d'aide à la décision. Par cette application, nous parviendrons à formaliser une connaissance, sous la forme d'un modèle statistique. Plus tard, cette connaissance pourra être mobilisée à des fins décisionnelles. Ce chapitre fait la démonstration complète de l'application de la démarche de modélisation au cas MES, fil rouge de cette thèse. En marge de ce cas d'application, il pose également, en section 2, la question de la généralité de cette démarche pour des cas décisionnels similaires. Le plan du chapitre est indiqué dans la Figure 45.

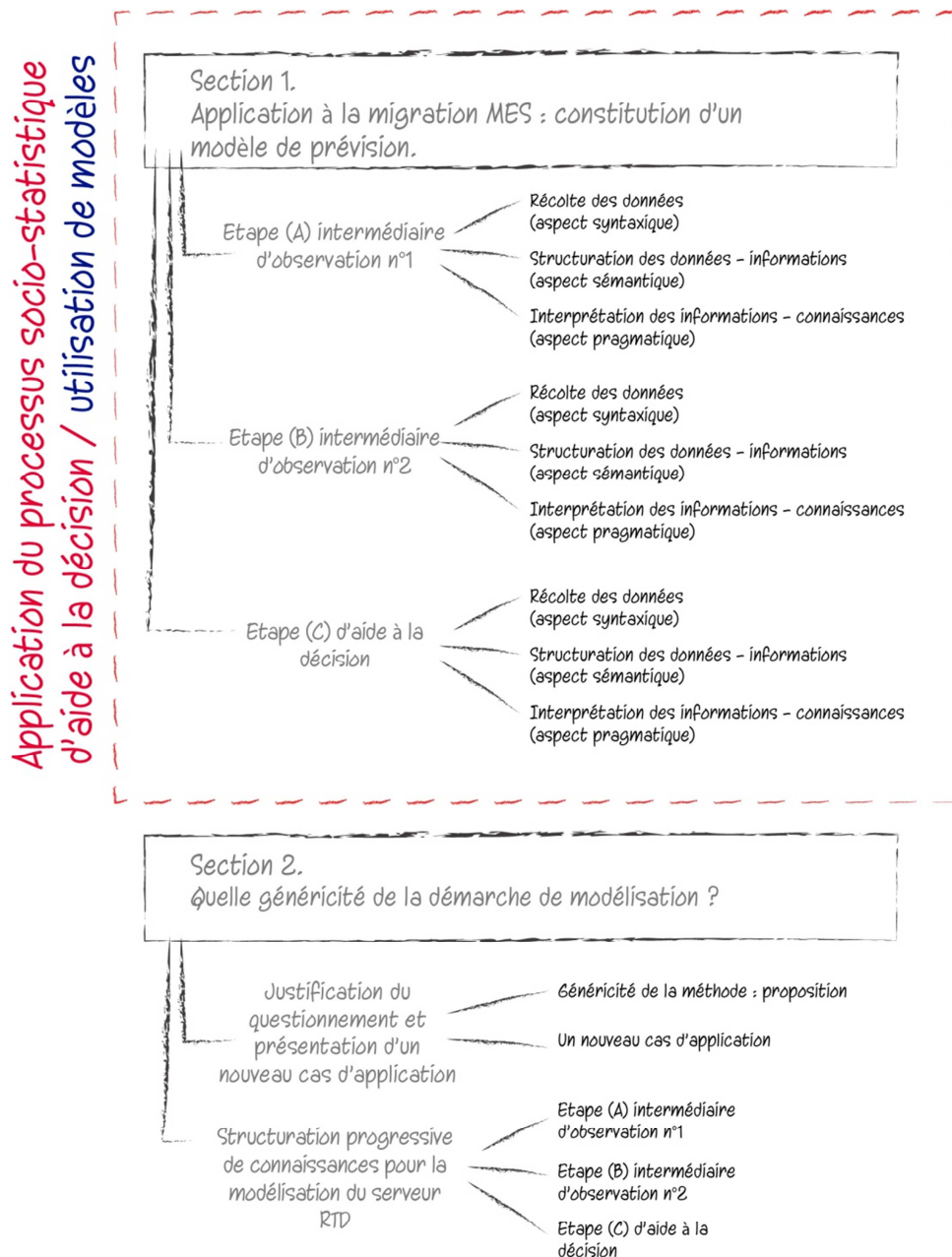


Figure 45. Plan du chapitre 6.

L'articulation de la première section est sans équivoque : elle correspond à l'application au cas MES des étapes de la gestion des connaissances impliquées par la démarche de modélisation, telle que présentée dans le Tableau 9 du Chapitre 5. Il s'agit de constituer deux modèles intermédiaires d'observation et un modèle d'aide à la décision, selon des séquences de récolte des données, de structuration des données et d'interprétation des informations.

La section 2 discute du caractère générique de la démarche. Après avoir indiqué les conditions de cette généricité, la démarche de modélisation est appliquée à un autre cas STMicroelectronics : la modélisation des serveurs RTD (*Real Time Dispatching*), complémentaires aux serveurs MES pour certaines tâches industrielles. Les différentes étapes de la démarche de modélisation seront à nouveau mobilisées pour mener à bien cette analyse.

## Section 1. Application à la migration MES : constitution d'un modèle de prévision.

### 1.1. Etape (A) intermédiaire d'observation n°1

#### 1.1.1. Récolte des données (aspect syntaxique)

On cherche ici à mesurer, grâce aux données de la base de performance, l'environnement applicatif de  $Y_t$  (variable à prédire du modèle final à constituer).

Soit  $Y_t$  l'activité du processeur du serveur MES, mesurée en secondes CPU consommées par jour. L'ensemble des données disponibles avant la migration MES est utilisé, soit 170 observations. L'activité du système d'information supportée par ce serveur mobilise principalement une application informatique dénommée *300Works*, ainsi que quelques applications satellites (maintenance des équipements, gestion des masques<sup>45</sup>, interfaces utilisateurs).

Seule l'activité applicative de *300Works* est enregistrée dans la base de performance, par le volume d'appels journalier de chacune de ses transactions. Les experts confirment que cette activité applicative suffit à appréhender l'activité générale du système d'information pour ce domaine fonctionnel. En effet, plusieurs de ses transactions sont appelées à chaque étape du processus de production des plaques de silicium, pour toutes les tâches manufacturières nécessitant un recours au SI (transport, contrôle des équipements de production, contrôle qualité...).

#### 1.1.2. Structuration des données - informations (aspect sémantique)

L'objectif est désormais de réaliser un modèle descriptif, par ACP classique avec rotation *Varimax*. Le script R utilisé est consultable en annexe 1.

Quatre-cent quarante-cinq transactions sont utilisées durant la période considérée. Chaque jour de la période (les individus) est donc caractérisé par un certain volume d'activité de chacune des transactions *300Works* (les variables). On cherche à observer ces volumes d'appels journaliers, grâce à une ACP complétée par une rotation *Varimax*. Une première analyse (ACP1) est réalisée sur la totalité du jeu de données. La première composante révèle une forte saisonnalité. La Figure 46 trace les valeurs de cette composante dans le temps. On y voit clairement ce phénomène, confirmé par le corrélogramme de la série, qui indique un pic de corrélation tous les sept jours.

---

<sup>45</sup> Un masque est une plaque de verre sur laquelle sont reproduits les motifs correspondant à différents stades d'élaboration des puces. Ils sont utilisés comme des négatifs lors des étapes de photogravure.

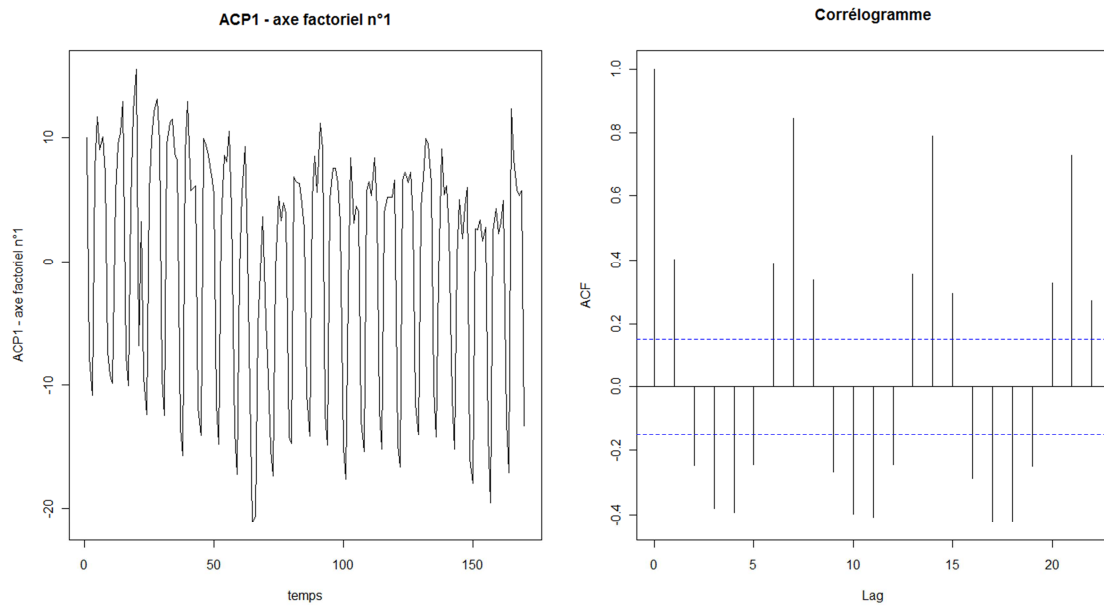


Figure 46. Tracé du premier axe factoriel de l'ACP1 et de son corrélogramme.

En s'intéressant de plus près aux données, on constate une baisse hebdomadaire du volume d'appel d'un ensemble de transactions les weekends. L'échantillon comprend donc deux populations d'individus, non-homogènes en termes d'activité transactionnelle : les jours de la semaine et les weekends. Cela peut nuire à la qualité et à l'interprétation de l'ACP1. La Figure 47, qui trace le premier axe factoriel versus le deuxième axe factoriel de l'ACP1, illustre cette proposition : on voit que les composantes ne sont pas indépendantes si l'on distingue les deux sous-populations identifiées.

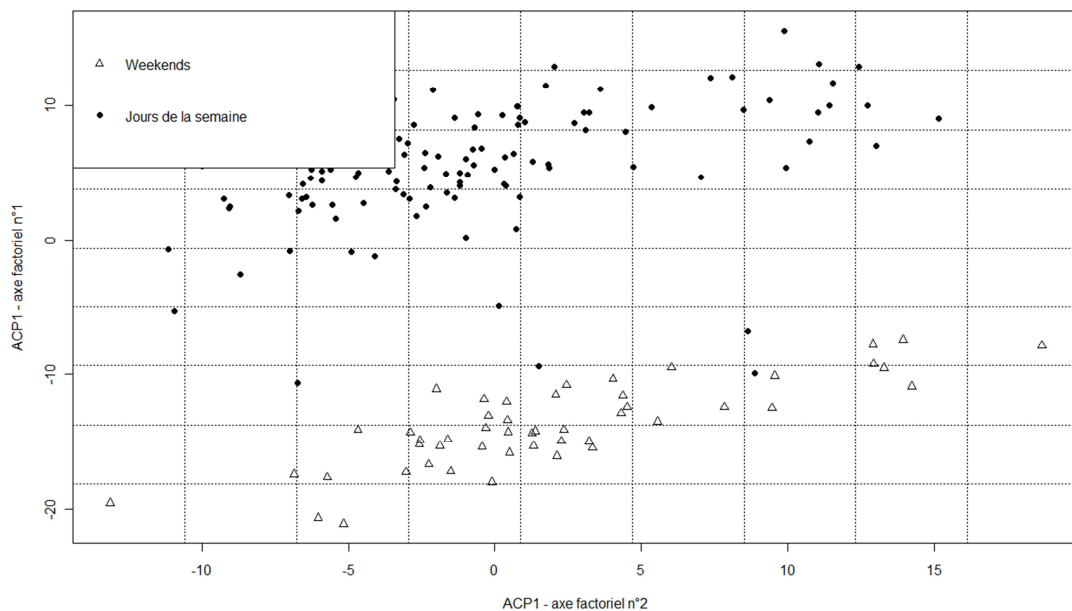


Figure 47. Corrélations « cachées » entre axes factoriels.

Afin de travailler sur une population plus homogène, une deuxième ACP est réalisée, en retirant les samedis et dimanches du jeu de données (ACP2)<sup>46</sup>. En résulte des composantes représentatives de la variabilité du volume d'appel des transactions *300Works*. Les premières composantes sont de très bons prédicteurs de la variable  $Y_t$  (les quatre premières composantes expliquent à elles seules environ 74% de sa variabilité), bien qu'exprimant seulement 28% de la variabilité transactionnelle totale. Ceci peut s'expliquer par le fait qu'un grand nombre de transactions ne sont utilisées que très occasionnellement, et ne sont donc pas bien représentées par les premiers axes factoriels.

### 1.1.3. Interprétation des informations - connaissances (aspect pragmatique)

Les résultats de l'analyse factorielle précédente sont maintenant interprétés et validés, puis l'activité industrielle opérationnelle sous-jacente est identifiée.

Le modèle quantitatif constitué est discuté avec les experts. Des graphiques caractérisant l'espace des variables sont utilisés. Ils permettent de visualiser les coefficients de corrélation entre les variables et les composantes principales. Les experts connaissent suffisamment bien ces variables et leurs usages métier pour identifier les activités industrielles sous-jacentes aux axes factoriels. La Figure 48 donne un exemple de ce travail.

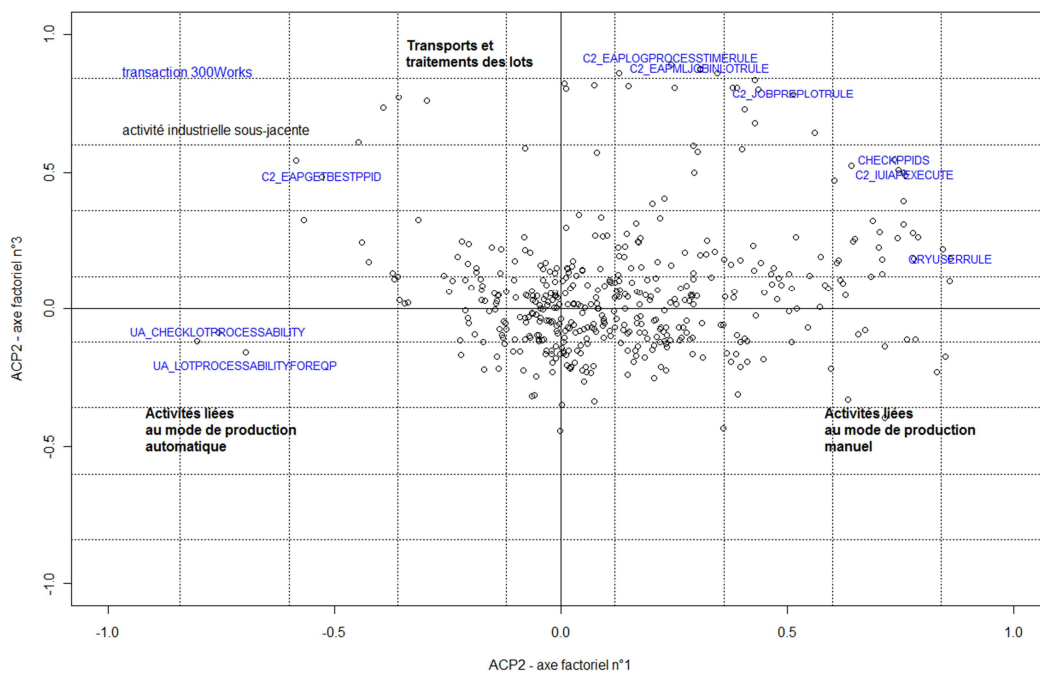


Figure 48. Exemple d'interprétation avec les experts de l'ACP2, après rotation Varimax (espace des variables : projection des variables d'origine sur les axes factoriels 1 et 3).

Le Tableau 12 résume les principaux phénomènes métier ainsi identifiés.

<sup>46</sup> Une analyse complémentaire est également réalisée en ne conservant que les week-ends, mais aucune information complémentaire n'est obtenue.

<b>Activité transactionnelle</b>	<b>Activité industrielle sous-jacente</b>
<i>Saisonnalité ACP1</i>	Activités non effectuées le weekend (création de nouveaux produits, <i>reporting</i> ...)
<i>ACP2, composante 1 (corrélation positive)</i>	Vérification des lots de production à traiter, assignation des lots aux files d'attente des équipements et recherche d'information, pour préparer <u>manuellement</u> le traitement des lots (deux modes de productions manuel/automatique sont identifiés)
<i>ACP2, composante 1 (corrélation négative)</i>	Vérification des lots de production à traiter, assignation des lots aux files d'attente des équipements et recherche d'information, pour préparer <u>automatiquement</u> le traitement des lots (compte tenu des modes de productions manuel/automatique)
<i>ACP2, composante 2</i>	Création de lots non-productifs (lots tests)
<i>ACP2, composante 3</i>	Transports et traitements des lots <sup>47</sup>
<i>ACP2, composante 4</i>	Activités de Recherche et Développement (R&D)

Tableau 12. Composantes principales de l'activité transactionnelle 300Works et principaux phénomènes industriels sous-jacents.

Un modèle descriptif de l'activité du SI, synthétisant la variabilité applicative et inscrit dans le périmètre de l'activité industrielle, est ainsi constitué. Ceci permet de passer à l'étape de modélisation n°2.

## 1.2. Etape (B) intermédiaire d'observation n°2

### 1.2.1. Récolte des données (aspect syntaxique)

Lors de l'étape (B), l'activité industrielle opérationnelle identifiée lors de l'étape (A) est mesurée, par le biais des bases de données applicatives.

A partir des principaux axes factoriels identifiés précédemment, des moyens de quantification de l'activité industrielle opérationnelle identifiée sont recherchés avec les experts. Soient  $x_{it}$  ces variables. Ce travail est résumé dans le Tableau 13.

<sup>47</sup> Par traitement d'un lot, nous entendons l'une des actions élémentaires opérée pour transformer une plaque de silicium : traitement thermique, photolithographie, gravure sèche et humide, implantation ionique, test paramétrique, etc. Plusieurs centaines de ces traitements sont nécessaires pour aboutir au produit fini.

Activité industrielle opérationnelle	Variable de mesure		
	Description	Nom	Disponibilité
Activités non effectuées le weekend (création de nouveaux produits, <i>reporting</i> ...)	Variable indicatrice : '0' pour les jours ouvrés, '1' pour le weekend	$x_{WE,t}$	Création <i>ad hoc</i> d'une variable indicatrice
Vérification des lots de production à traiter, assignation des lots aux files d'attentes des équipements et recherche d'information, pour préparer <u>manuellement</u> le traitement des lots	Nombre total journalier de préparations de lots de production <sup>48</sup> manuelles	$x_{1,t}$	Création <i>ad hoc</i> d'une nouvelle variable de mesure par traitement des données des bases applicatives
Vérification des lots de production à traiter, assignation des lots aux files d'attentes des équipements et recherche d'information, pour préparer <u>automatiquement</u> le traitement des lots	Nombre total journalier de préparations de lots de production automatiques	$x_{2,t}$	Création <i>ad hoc</i> d'une nouvelle variable de mesure par traitement des données des bases applicatives
Transports et traitements des lots	* Volume journalier moyen de plaques en cours de production (production à faire) * Nombre total journalier de lots traités (production effective)	* $x_{3,t}$ * $x_{4,t}$	Directement disponibles dans les bases de données applicatives
Création de lots non-productifs (lots tests)	Non identifiées pour l'instant. Ne seront pas considérées.	-	-
Activités R&D			

Tableau 13. Choix des mesures des phénomènes industriels sous-jacents à la variabilité applicative

On remarque que les variables  $x_{1,t}$  et  $x_{2,t}$  ont fait l'objet de la création d'une nouvelle mesure, non existante en l'état, avant l'application de la démarche de modélisation. Conformément aux bénéfices attendus de l'application du processus socio-statistique d'aide à la décision, on peut parler ici de la hiérarchie inversée du continuum données-informations-connaissances.

Concernant l'activité industrielle « Transports et traitements des lots », il convient de justifier le recours à deux variables de mesures. La mesure de la production effective, c'est-à-dire le nombre de plaques effectivement traité par jour, doit être complétée par la mesure de la production en attente de traitement. En effet, à production effective constante, une hausse de la production en attente de traitement aura tendance à « engorger » l'usine. Cette diminution du débit de production aurait un impact significatif sur l'activité du SI, car elle entraînerait la complexification d'un certain nombre de tâches traitées informatiquement.

<sup>48</sup> Un lot de production contient approximativement entre 20 et 25 plaques de silicium à traiter.

### 1.2.2. Structuration des données - informations (aspect sémantique)

A partir des données identifiées, on peut à ce stade réaliser un modèle explicatif/prédictif des interactions entre  $Y_t$  et l'activité industrielle opérationnelle, par régression linéaire multiple. Ce modèle est évalué et mis sous contrôle. Le script R utilisé est consultable en annexe 2.

Les variables  $x_{i,t}$  sont utilisées pour la constitution du modèle de prévision de  $Y_t$ . Tout d'abord, on vérifie que les variables retenues semblent adaptées à la réalisation d'un modèle linéaire. C'est ce que suggère la Figure 49 : on observe des relations linéaires plus ou moins marquées entre  $Y_t$  et les  $x_{i,t}$  quantitatives retenues.

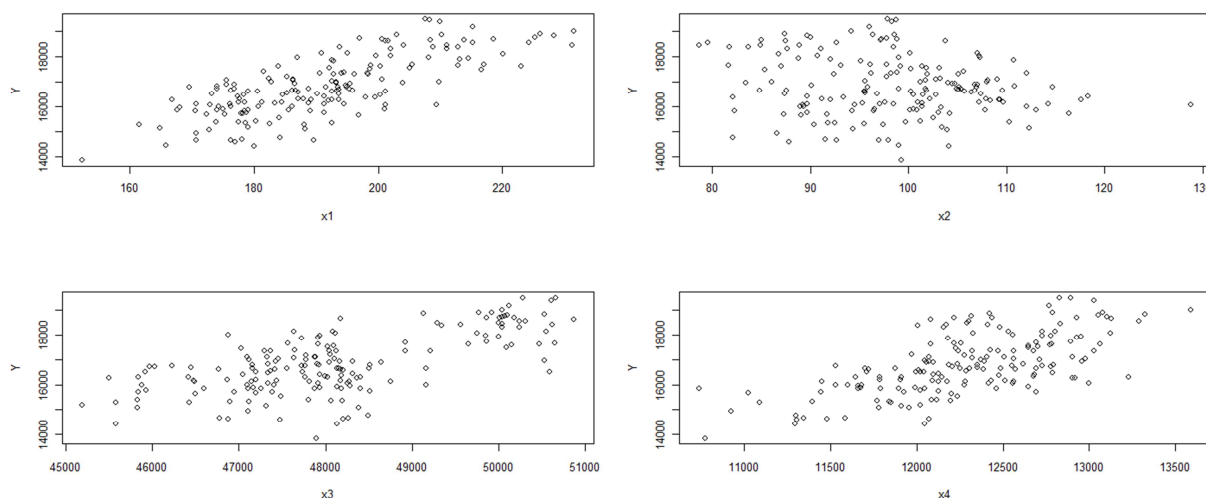


Figure 49. Relations linéaires observées entre  $Y_t$  et les  $x_{i,t}$  quantitatives retenues.

Des corrélations significatives entre les variables explicatives quantitatives sont observées (Tableau 14).

	$x_{1,t}$	$x_{2,t}$	$x_{3,t}$	$x_{4,t}$
$x_{1,t}$	1	-	-	-
$x_{2,t}$	-0,49	1	-	-
$x_{3,t}$	0,70	-0,45	1	-
$x_{4,t}$	0,71	0,24	0,37	1

Tableau 14. Corrélations entre variables explicatives quantitatives identifiées (coefficient de Pearson), arrondies à 0,01 près.

Le risque de colinéarité est limité, mais nous illustrons comment un tel problème peut être traité. Par ailleurs, ce traitement permet de simplifier le modèle final en éliminant une variable.



Les corrélations entre variables exogènes sont traitées de la façon suivante :

- *La variable  $(x_{1,t} + x_{2,t})$  est calculée* : elle correspond au nombre total de préparations de lots, quel que soit le mode de production. Cette nouvelle variable est utilisée dans le modèle final. On remarque qu'elle exprime quasiment la même information que  $x_{4,t}$  (coefficient de corrélation proche de 1)<sup>49</sup>. La variable  $x_{4,t}$  ne sera donc pas utilisée dans le modèle.  $x_{2,t}$  est aussi conservée isolément, pour tenir compte d'effets supplémentaires liés au mode automatique ;
- *Une méthode d'orthogonalisation est utilisée* : le modèle de régression linéaire multiple suivant est posé :

$$x_{3,t} = \gamma_0 + \gamma_1(x_{1,t} + x_{2,t}) + \gamma_2 x_{2,t} + R_t$$

Son coefficient de détermination ( $R^2$ ) ajusté<sup>50</sup> est d'environ 0,50. Le terme résiduel  $R_t$  représente la part de variabilité de  $x_{3,t}$  qui n'est pas expliquée par les fluctuations de  $(x_{1,t} + x_{2,t})$  et  $x_{2,t}$ . En l'occurrence, c'est la fluctuation du niveau d'en-cours de production, qui n'est pas en lien direct avec l'activité de préparation et de traitement effectif des lots (phénomène de baisse de débit évoqué précédemment). C'est cette nouvelle variable que nous utiliserons pour le calcul du modèle final<sup>51</sup>.

Les variables finales du modèle sont donc :  $(x_{1,t} + x_{2,t})$ ,  $x_{2,t}$ ,  $R_t$  et  $x_{WE,t}$ . Le Tableau 15 montre que les corrélations croisées sont désormais très faibles.

	$(x_{1,t} + x_{2,t})$	$x_{2,t}$	$R_t$	$x_{WE,t}$
$(x_{1,t} + x_{2,t})$	1	-	-	-
$x_{2,t}$	0,10	1	-	-
$R_t$	0,00	0,00	1	-
$x_{WE,t}$	0,048	0,01	-0,1	1

Tableau 15. Corrélations entre variables explicatives finales après traitement du risque colinéarité (coefficient de Pearson), arrondies à 0,01 près.

Un modèle provisoire est calculé à partir de ces variables, tel que :

$$Y_t = \alpha_0 + \alpha_1(x_{1,t} + x_{2,t}) + \alpha_2 x_{2,t} + \alpha_3 R_t + \alpha_4 x_{WE,t} + \mu_t$$

L'ordonnée à l'origine n'est pas statistiquement significative. Le coefficient de détermination ajusté est d'environ 0,79. Cependant,  $\mu_t$  n'est pas i.i.d., mais suit un processus temporel de type AR(1). C'est une violation des hypothèses standards du

<sup>49</sup> C'est-à-dire que l'on a autant de préparations de lots manuelles et automatiques que de transports et traitements de lots. C'est normal, car le rapport entre ces deux groupes est généralement de un pour un : pour un transport et traitement de lot, il faut toujours une étape de préparation, qui peut être réalisée soit manuellement, soit automatiquement.

<sup>50</sup> A titre informatif, les estimations des paramètres du modèle sont les suivantes :  $\gamma_0 \approx 398092,37$  ;  $\gamma_1 \approx 54,70$  ;  $\gamma_2 \approx -76,71$ .

<sup>51</sup> Soulignons que cette méthode introduit un biais, car les paramètres liant  $(x_{1,t} + x_{2,t})$ ,  $x_{2,t}$  et  $x_{3,t}$  sont inconnus et estimés à partir du même jeu de données que celui utilisé pour le modèle final utilisé pour  $Y_t$ . Cependant, comme la taille de l'échantillon est assez grande, ce biais est considéré comme négligeable. Par ailleurs, il peut être corrigé en divisant le jeu de données en deux sous-échantillons. Sur les premières  $n1$  données,  $(x_{1,t} + x_{2,t})$  et  $x_{2,t}$  sont régressés sur  $x_{3,t}$ . Le modèle ainsi obtenu est alors utilisé pour prédire  $x_{3,t}$  et calculer les résidus  $R'_t$  sur les  $(n-n1)$  données restantes. Finalement, le modèle pour  $Y_t$  peut être calculé avec ces données. Cette méthode a été appliquée, et des résultats en phase avec ce qui est présenté dans cette thèse sont obtenus.

modèle linéaire. En conséquence, un modèle final avec erreurs ARIMA est estimé par maximum de vraisemblance (Modèle 3)<sup>52</sup>. Il est de la forme :

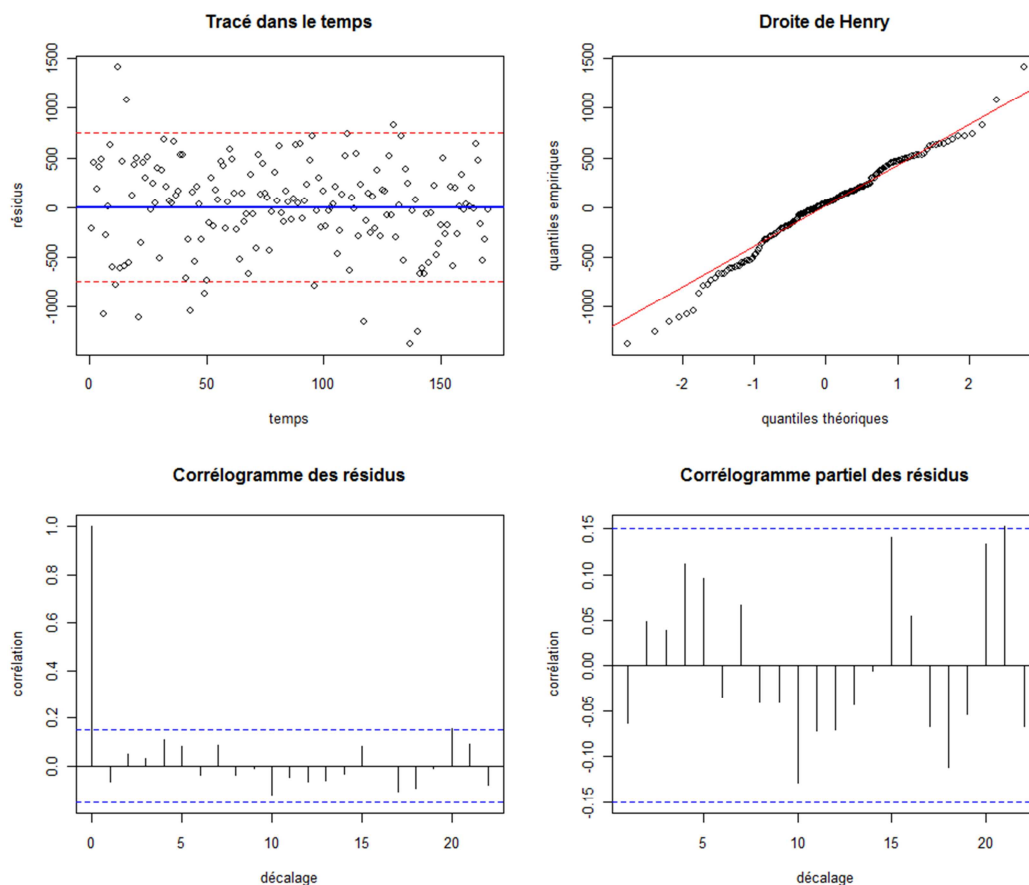
$$Y_t = \beta_1(x_{1,t} + x_{2,t}) + \beta_2x_{2,t} + \beta_3R_t + \beta_4x_{WE,t} + \mu'_t$$

avec :

$$\mu'_t = \varphi\mu'_{t-1} + \varepsilon_t, \varepsilon_t \text{ i.i.d. de loi } N(0, \sigma^2)$$

*Modèle 3. Modèle de prévision de  $Y_t$  (serveur MES) retenu, à partir de l'activité industrielle opérationnelle.*

Le Modèle 3 respecte les hypothèses du modèle linéaire. La Figure 50 montre que les résidus du modèle peuvent être considérés comme homoscédastiques, sans points aberrants particulièrement préoccupants (cf. tracé dans le temps). La droite de Henry, qui permet de comparer les quantiles empiriques des résidus aux quantiles théoriques d'une loi normale de paramètre  $N(0, \sigma^2)$ , montre qu'on peut considérer ce terme d'erreur comme étant i.i.d. de loi normale. Enfin, on constate que les résidus sont bien indépendants du temps (cf. corrélogramme et corrélogramme partiel).



*Figure 50. Vérification des résidus du Modèle 3 : hypothèses du modèle linéaire respectées.*

Par ailleurs, la Figure 51 montre que les résidus de ce modèle semblent bien indépendants vis-à-vis des prédicteurs retenus. La forme du modèle semble donc valide.

<sup>52</sup> Rappel : les modèles 1 et 2 représentent les modèles de rationalité des décideurs. Ils ne correspondent pas à des modèles statistiques tels que formalisés dans ce chapitre.

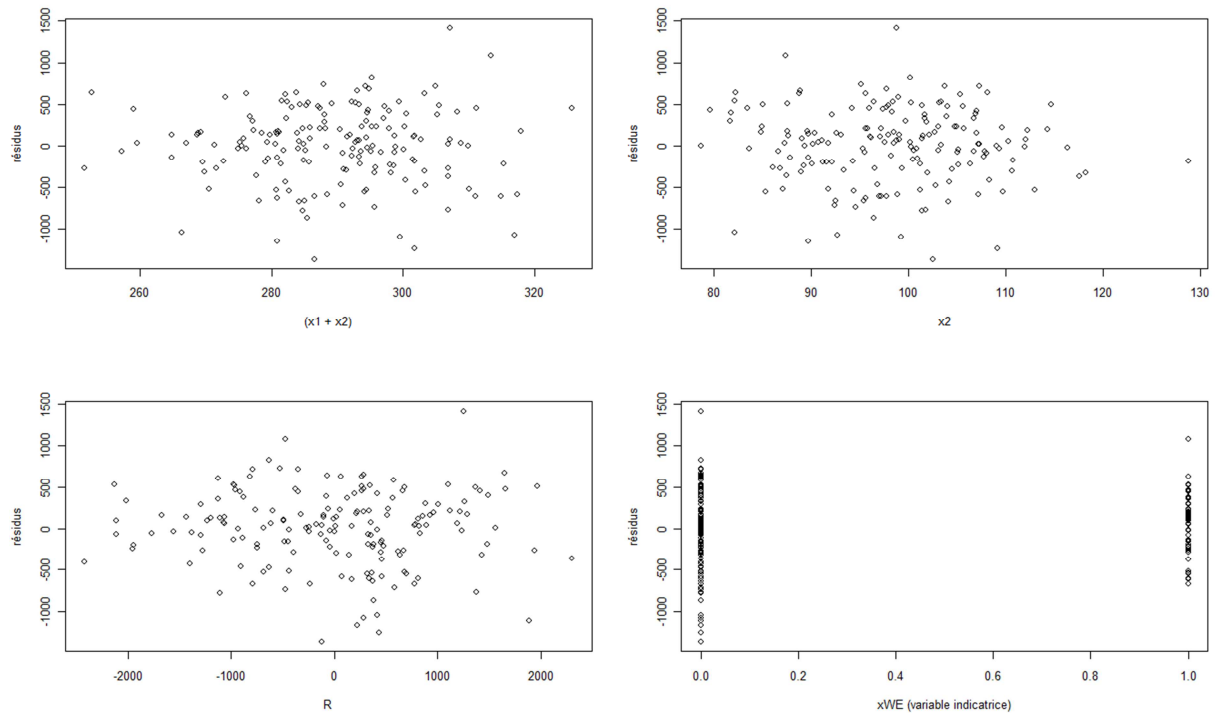


Figure 51. Vérification des résidus du Modèle 3 : résidus contre prédicteurs.

Le Modèle 3 est donc retenu. Il montre que la consommation en secondes CPU du processeur MES ( $Y_t$ ) dépend de trois variables industrielles opérationnelles : le volume de production (préparation et traitement de lots) réalisé en modes manuel ( $x_{1,t}$ ) et automatique ( $x_{2,t}$ ), ainsi que l'en-cours de production non traité ( $R_t$ ). Par ailleurs, un effet saisonnier lié à une modification de l'activité durant les weekends est modélisé ( $x_{WE,t}$ ), de même que certains phénomènes sous-jacents inconnus liés à la temporalité des données observées ( $\mu'_t = \varphi\mu'_{t-1} + \varepsilon_t$ ).

L'estimation ponctuelle de la forme développée du modèle est donnée par l'Application numérique 2.

$$Y_t \approx 68,64x_{1,t} + 39,81x_{2,t} + 0,31R_t - 823,91x_{WE,t} + \mu'_t$$

avec :

$$\mu'_t \approx 0,57\mu'_{t-1} + \varepsilon_t, \varepsilon_t \text{ i.i.d. de loi } N(0; 454,16^2)$$

*Application numérique 2. Estimation du Modèle 3, à 0,01 près.*

La Figure 52 trace les valeurs prédites  $\hat{Y}_t$  du Modèle 3, contre les données réelles connues de la variable  $Y_t$ : le modèle fournit une bonne approximation de la consommation réelle de ressources CPU, peu de points sont hors des intervalles de prévision à 5%.

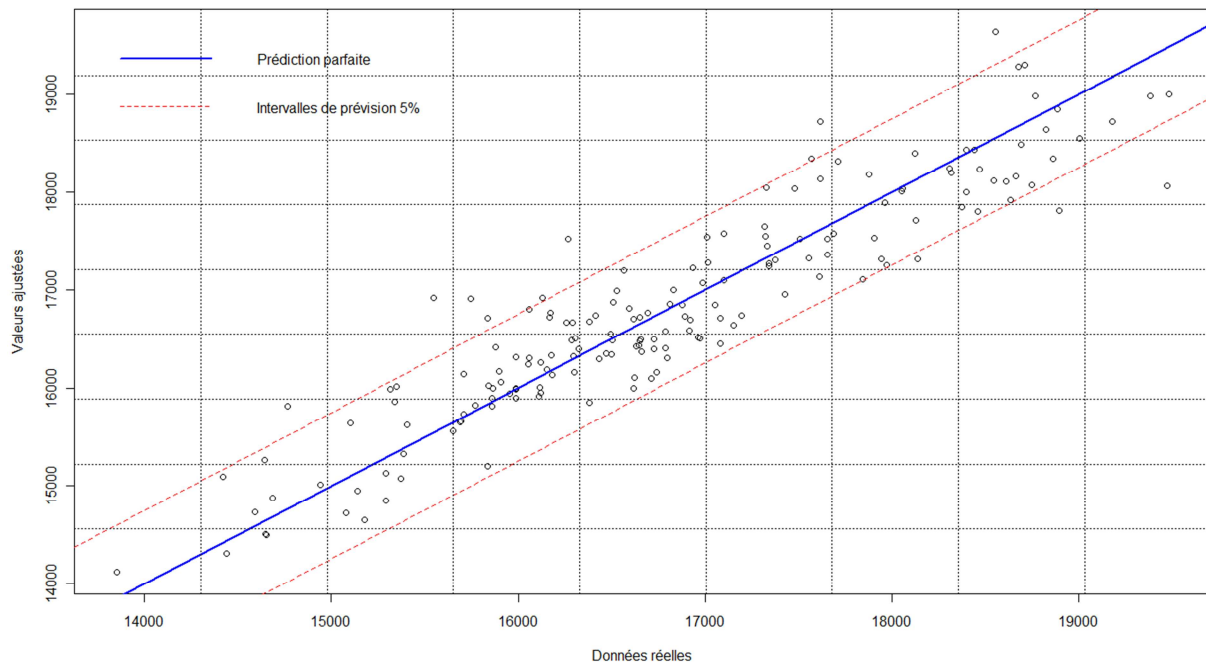


Figure 52. Modèle 3 : valeurs prédites versus données réelles de la consommation CPU du serveur MES.

Ce modèle est validé quantitativement. Tout d’abord, les algorithmes de validation croisée de Hyndman sont appliqués (le temps ayant une importance, on ne peut appliquer une méthode validation croisée classique). Un échantillon de modélisation d’une taille initiale de 50 jours est utilisé pour l’estimation séquentielle, et un échantillon d’une taille fixe de 50 jours est utilisé pour l’estimation glissante. La Figure 53 montre que la moyenne absolue de l’erreur de prévision est toujours faible, inférieure à 900 secondes, soit 5% de la consommation CPU moyenne sur la période.

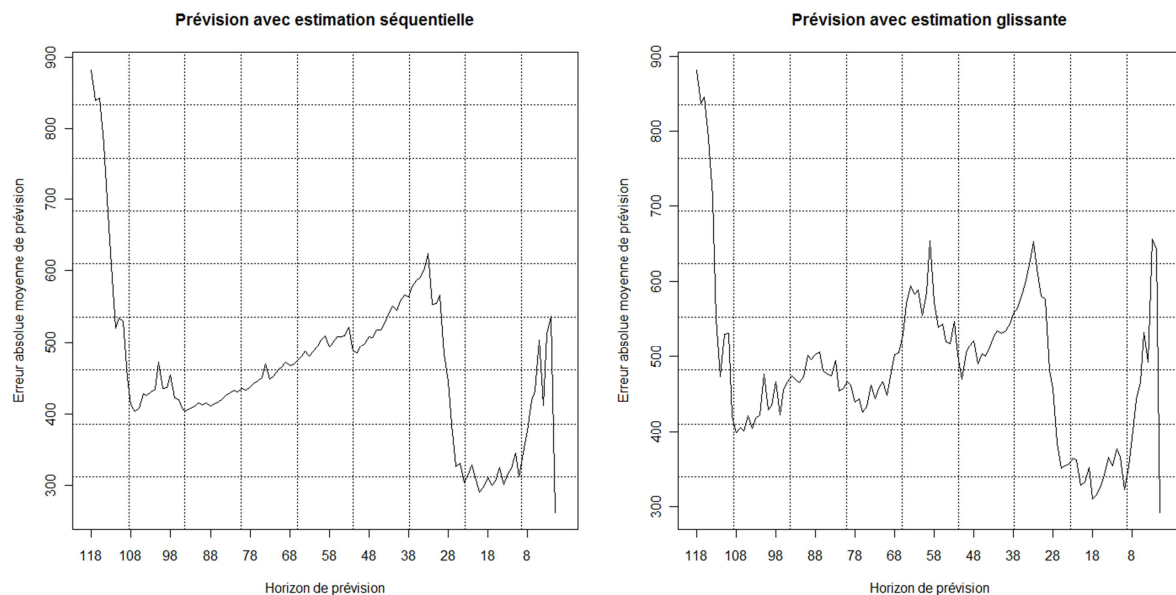


Figure 53. Validation croisée du Modèle 3.

L’hypothèse nulle de stabilité des paramètres du Modèle 3 est testée.

Elle est soumise à deux tests statistiques, sur la base du processus de fluctuation empirique de Zeileis :

- *Statistique de Nyblom-Hansen*, particulièrement adaptée aux modèles comprenant des phénomènes d'autocorrélation et/ou présentant de multiples ruptures ;
- *Statistique SupLM*, particulièrement adaptée pour détecter un changement structurel unique impliquant plusieurs changements simultanés dans les coefficients.

Aucun de ces tests ne suffit à réfuter l'hypothèse nulle : les seuils de rejet ne sont jamais franchement dépassés, et les valeurs critiques sont toujours supérieures au seuil d'erreur de 5%, comme le montre la Figure 54.

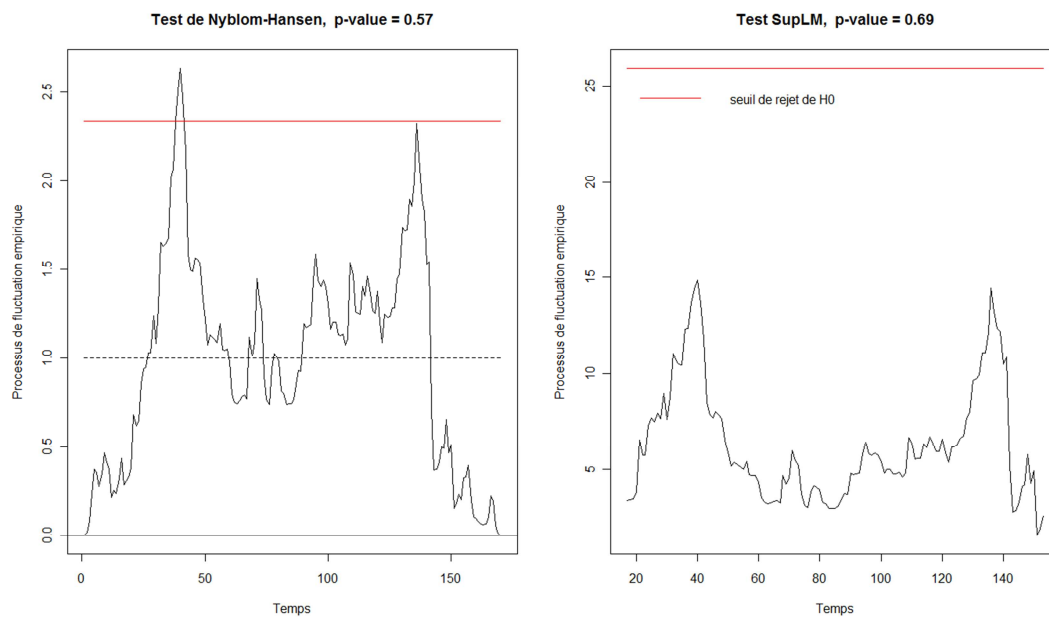


Figure 54. Test de la stabilité structurelle du Modèle 3.

Ces validations quantitatives amènent à conclure à un modèle fiable pour une utilisation en prévision.

Finalement, nous utilisons la procédure de contrôle dynamique de Zeileis pour mettre sous contrôle le modèle de prévision basé sur les variables industrielles opérationnelles<sup>53</sup>. Nous illustrons ci-après le comportement d'un tel contrôle en période de stabilité structurelle. Le contrôle est initialisé sur les 163 premiers jours du jeu de données, puis les sept observations suivantes des variables du modèle sont confrontées à un contrôle dynamique de stabilité structurelle. Nous utilisons deux statistiques de tests, l'une basée sur les résidus globaux du modèle (Figure 55, OLS-based CUSUM), l'autre sur l'estimation récursive des coefficients du modèle (RE).

<sup>53</sup> Par souci de simplification, le modèle mis ici sous contrôle est un peu différent du Modèle 3 : les variables n'ont pas été retravaillées face au risque de colinéarité (nous avons vu que le modèle était valide dans ce cas de figure également). Dans les deux cas, les résultats du contrôle dynamique sont similaires.

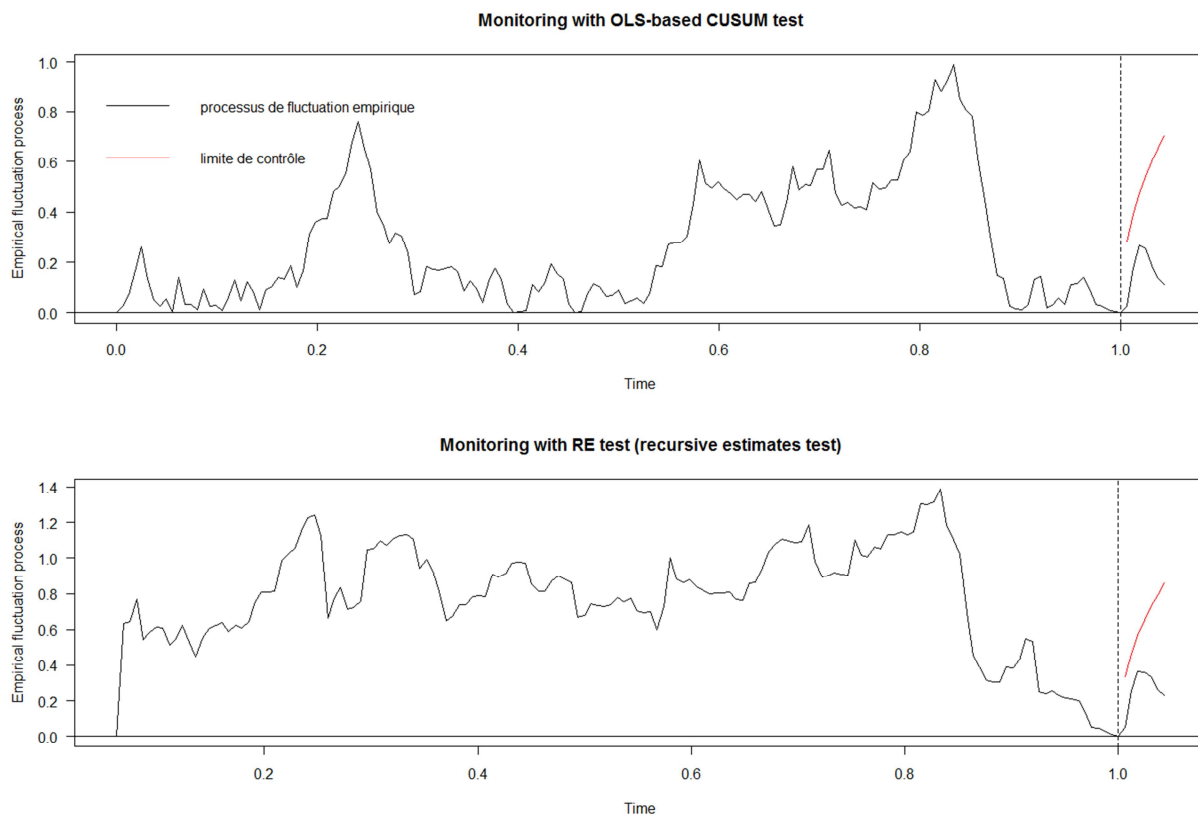


Figure 55. Contrôle dynamique de la stabilité structurelle du modèle de prévision : modèle sous contrôle.

La mise en œuvre de ce contrôle sous R est présentée dans l'annexe 5. On constate qu'après la fin de la période d'initialisation (à gauche de la ligne pointillée verticale), l'hypothèse d'instabilité structurelle est rejetée pour les sept nouvelles observations (à droite de la ligne) : pour les deux contrôles, le processus de fluctuation empirique est sous la limite de contrôle.

### 1.2.3. Interprétation des informations - connaissances (aspect pragmatique)

Le modèle qui vient d'être constitué est désormais interprété et validé. On cherche ensuite à identifier l'activité industrielle de pilotage, qui lui est sous-jacente.

L'Application numérique 2 est discutée avec les experts :

- *Différence entre les coefficients associés à  $x_{1,t}$  et  $x_{2,t}$*  : la préparation des lots en mode de production automatique peut nécessiter moins de ressources CPU MES. En effet, certaines tâches ne sont plus effectuées par les opérateurs, mais par un domaine du SI qui est soutenu un autre serveur physique ;
- *Variable  $R_t$*  : lorsque l'en-cours de production augmente, sans que la fabrication puisse augmenter son débit de production, la complexité de l'activité transactionnelle de *300Works* augmente et nécessite plus de ressources serveur ;

- *Variable indicatrice*  $x_{WE,t}$  : il est normal que l'arrêt d'un certain nombre d'activités le weekend entraîne une baisse de la consommation CPU.

Cette étape indispensable valide la pertinence du Modèle 3. Ses variables explicatives et ses coefficients font sens pour les experts et mettent en avant des éléments d'explication du comportement de  $Y_t$ , à partir de prédictors caractéristiques de l'environnement industriel opérationnel du serveur MES. C'est une base de compréhension pertinente et mathématiquement solide pour aller plus avant vers la constitution du modèle final d'aide à la décision.

Fort de cette connaissance nouvelle, le Modèle 3 est reconsidéré avec les experts. Comment passer d'un modèle caractérisant l'activité quotidienne opérationnelle de l'usine, à un modèle de pilotage basé sur des variables non triviales, connues de la Direction Générale de STMicroelectronics et pouvant être anticipées à un horizon temporel pertinent pour prendre des décisions quant à l'évolution du système informatique ? Cette question est discutée dans le cadre de groupes de travail.

Tout d'abord, l'échelle temporelle est mise en question. Si l'activité journalière est pertinente pour observer le comportement du système d'un point de vue opérationnel, elle ne l'est pas pour l'aide à la décision. Les prévisions managériales sont usuellement exprimées en activité industrielle à la semaine (cf. le modèle de rationalité en usage lors de la migration MES). Il est donc décidé de travailler sur données agrégées à un niveau hebdomadaire. Ainsi, dans la suite de ce travail de modélisation, nous considérons désormais que  $t$  représente des données à la semaine, et non plus à la journée. A ce niveau d'agrégation, les variations saisonnières liées aux weekends n'ont plus à être prises en compte. On peut soit retirer les valeurs des weekends du jeu de données, soit les intégrer à l'agrégation. Nous avons opté pour cette dernière solution. La variable  $x_{WE,t}$  n'a donc plus lieu d'être.

Ensuite, il convient de définir les bonnes variables d'aide à la décision, à ce niveau d'agrégation. Rappelons l'indicateur d'activité décisionnel clé en usage lors de la migration MES, mentionné dans le deuxième chapitre : le nombre de *wafers-outs* par semaine (wo/w), soit le nombre de plaques sortantes de la chaîne de fabrication chaque semaine. Cet indicateur ne peut être utilisé en l'état. Effectivement, ce n'est pas les plaques sortantes de la production qui ont directement un impact sur le SI, mais plutôt ce qui se passe avant qu'elles sortent.

Pour cela, trois indicateurs sont retenus :

- *D'une part, le volume de plaques à produire.* Toute hausse de l'objectif de production fera croître ce volume, et le nombre de *wafers-outs* le fera diminuer. Nous avons déjà rencontré cet indicateur précédemment :  $x_{3,t}$  (Tableau 13). Il caractérise la charge globale de l'usine. Il est bien connu des managers de STMicroelectronics sous le nom *Work In Process* (WIP). C'est une donnée régulièrement anticipée dans le cadre des plans de charge de l'usine ;

- *D'autre part, l'activité effective de l'usine doit être caractérisée.* On parle ici du processus de production en tant que tel, qui permettra de faire passer les plaques de l'état de WIP à celui de produits finis (*wafers-outs*). Les variables relatives à la production effective (préparation et traitement des lots) identifiées plus haut ( $x_{1,t}$ ,  $x_{2,t}$ ,  $x_{4,t}$ ) ne conviennent pas. Elles correspondent à des étapes trop fines du processus de production. Les managers ne peuvent pas les anticiper à suffisamment long terme pour les utiliser dans le cadre d'un modèle d'aide à la décision. Néanmoins, l'analyse de la structure des gammes de production de STMicroelectronics permet de trouver un indicateur plus général, englobant cette notion de production effective. La production d'une plaque de semi-conducteur implique la réalisation de plusieurs « briques ». Chaque brique contient plusieurs « opérations ». Enfin, chaque opération contient plusieurs traitements à réaliser sur les lots, qui impliquent les variables  $x_{1,t}$ ,  $x_{2,t}$  et  $x_{4,t}$ , évoquées précédemment. Plus un produit est complexe, plus il contiendra de briques de production et plus ses briques pourront contenir d'opérations : le nombre d'opérations de production à effectuer sur une plaque de silicium est un indicateur majeur de la complexité d'un produit semi-conducteur. Lorsque toutes les briques, donc toutes les opérations, d'un produit sont complétées, on aboutit à un wafer-out et le WIP diminue. Dans ce contexte, une mesure typique de l'activité industrielle effective est le *move*, qui correspond au passage d'une opération de production à une autre. Le nombre de *moves* représente donc la réalisation d'un ensemble traitements et préparations de lots. C'est un indicateur de pilotage de l'activité industrielle important chez STMicroelectronics. Sa place dans la structure des gammes de production est illustrée par la Figure 56.

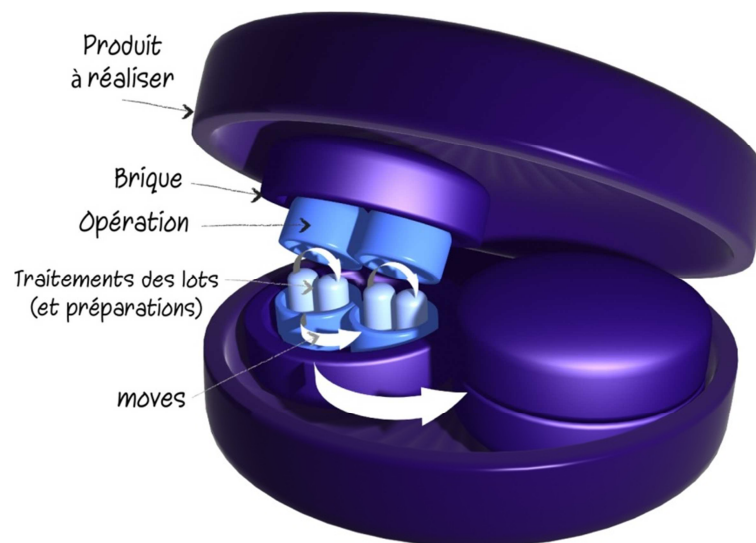


Figure 56. Structure des gammes de production des produits semi-conducteurs.

Par ailleurs, rappelons que le système est sensible au débit de l'usine, caractérisé par  $R_t$  dans le Modèle 3. En mettant en relation le volume de plaques à produire et l'activité effective de l'usine, on exprime une information sensiblement similaire. Par exemple, lorsque le volume de plaques à produire augmente tandis que l'activité effective reste constante, c'est que le débit de la fabrication diminue ;



- Enfin, nous avons identifié dans l'étape précédente deux modes de production, manuel et automatique, qui ont des impacts différents sur les serveurs informatique. Ils doivent donc être intégrés au modèle d'aide à la décision. Ces deux modes de production sont suivis par le management de STMicroelectronics, dans le cadre d'un projet de déploiement des applications *Move to Full Auto* (voir introduction générale). L'objectif est d'automatiser de plus en plus l'usine pour un certain nombre de tâches, relatives à la préparation des lots de production.

La Figure 57 résume cette discussion concernant les variables de pilotage sous-jacentes au modèle intermédiaire d'observation n°2.

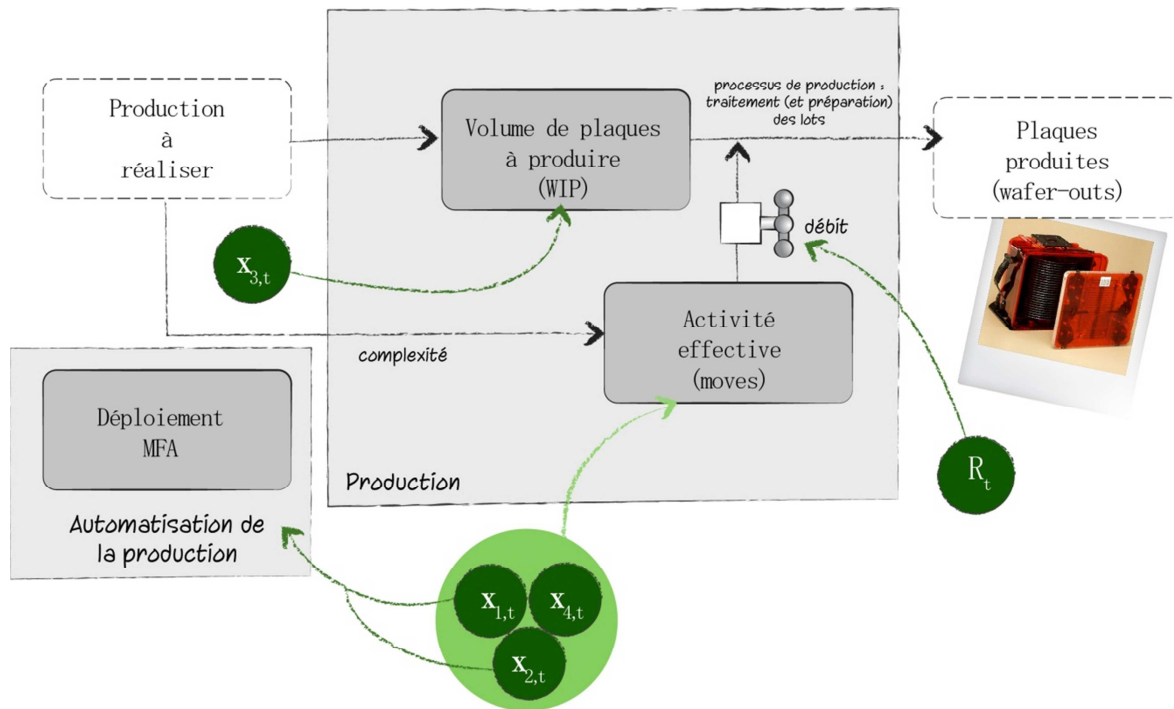


Figure 57. Discussion avec les experts : variables de pilotage sous-jacentes au modèle intermédiaire d'observation n°2.

Ces variables peuvent donc servir de base à la constitution du modèle de prévision d'aide à la décision attendu.

### 1.3. Etape (C) d'aide à la décision

#### 1.3.1. Récolte des données (aspect syntaxique)

On peut à présent mesurer, *via* les bases de données applicatives, l'activité industrielle de pilotage ( $X_i$ ) identifiée lors de l'étape (B).

$Y_t$  caractérise toujours l'activité du processeur du serveur MES, mais l'unité temporelle change : on mesure désormais l'activité moyenne à la semaine. Avec les experts, nous

cherchons à quantifier les variables de pilotage discutées ci-dessus. Le Tableau 16 synthétise ces échanges.

Activité industrielle de pilotage	Variable de mesure		
	Description	Nom	Disponibilité
<i>Volume de plaques à produire (en-cours de production)</i>	Charge globale de l'usine, exprimée en nombre moyen de plaques en cours de traitement (WIP) à la semaine (WIP/w)	$X_{1,t}$	Directement disponibles dans les bases de données applicatives
<i>Activité effective</i>	Nombre moyen d'opérations de production ( <i>moves</i> ) réalisées par semaine ( <i>moves/w</i> )	$X_{2,t}$	Directement disponibles dans les bases de données applicatives
<i>Déploiement MFA</i>	Pourcentage d'automatisation de l'usine moyen sur la semaine	$X_{3,t}$	Création <i>ad hoc</i> d'une nouvelle variable de mesure, par traitement des données des bases applicatives

Tableau 16. Choix des mesures des variables de pilotage.

Tout comme lors de la quantification du modèle d'observation n°2, cette étape a conduit à la création d'une nouvelle mesure ( $X_{3,t}$ ), conformément à la hiérarchie inversée du *continuum* données-informations-connaissances déjà mentionnée.

Ces données sont mesurées à partir de l'ensemble des données disponibles avant la migration MES. Vingt-six semaines d'observations du système et de son environnement sont ainsi disponibles. Ce chiffre peut sembler réduit pour une approche statistique. Néanmoins, le fait de s'appuyer sur le modèle d'observation n°2, qui rend compte plus finement du comportement opérationnel du système à partir d'un jeu de données plus conséquent, permet de soutenir les conclusions obtenues avec le jeu de données hebdomadaire restreint.

### 1.3.2. Structuration des données - informations (aspect sémantique)

Finalement, nous pouvons réaliser le modèle explicatif/prédictif attendu, formalisant les interactions identifiées entre  $Y_t$  et  $X_{i,t}$ , par régression linéaire multiple. Le modèle est évalué et peut être mis sous contrôle (cette dernière application ne sera pas développée, compte-tenu du nombre restreint d'observations qui seront considérées ici ; de toute manière, le principe est strictement le même que pour le Modèle 3). Le script R utilisé est consultable en annexe 3.

Les variables de pilotage  $X_{1,t}$ ,  $X_{2,t}$  et  $X_{3,t}$  sont utilisées pour constituer un modèle de prévision pour  $Y_t$ . La Figure 58 trace les valeurs de  $Y_t$  contre celles des  $X_{i,t}$  : le recours à un modèle linéaire est justifié.

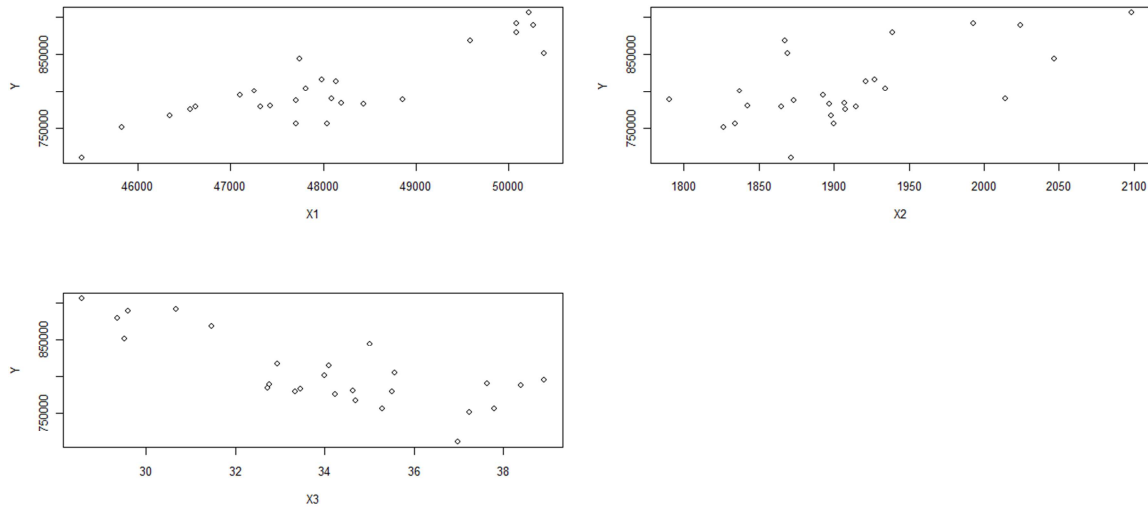


Figure 58. Relations linéaires observées entre  $Y_t$  et les  $X_{i,t}$  quantitatives retenues.

Le modèle obtenu est caractérisé par un coefficient de détermination ajusté d'environ 0,83. L'ordonnée à l'origine n'est pas significative. La forme finale du modèle est donc la suivante :

$$Y_t = \alpha_1 X_{1,t} + \alpha_2 X_{2,t} + \alpha_3 X_{3,t} + \varepsilon_t$$

avec :

$$\varepsilon_t \text{ i.i.d. de loi } N(0, \sigma^2)$$

Modèle 4. Modèle de prévision de  $Y_t$  (MES), à partir des variables  $X_{i,t}$  de pilotage.

Le processus résiduel  $\varepsilon_t$  ne présente plus de composante temporelle et semble satisfaire aux hypothèses du modèle linéaire (Figure 59).

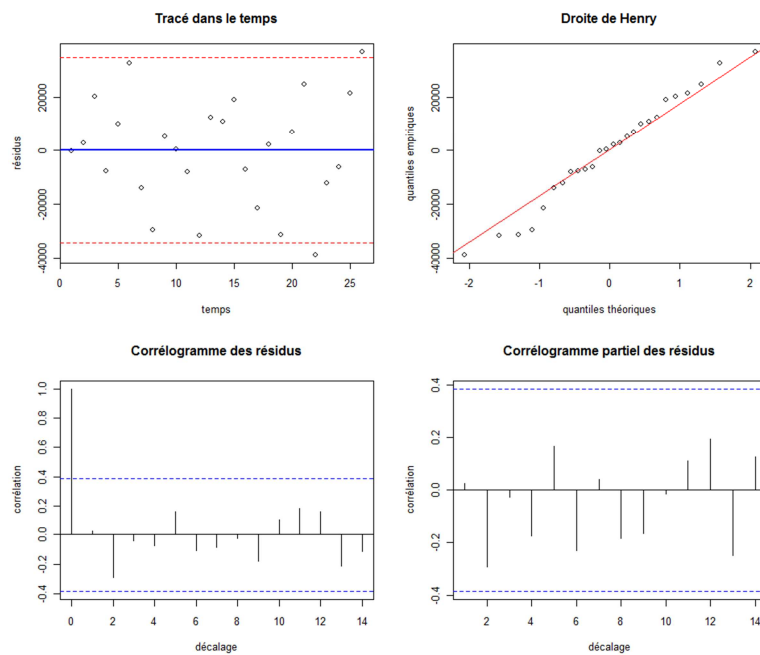


Figure 59. Vérification des résidus du Modèle 4 : hypothèses du modèle linéaire respectées.

De même, l'indépendance des résidus par rapport aux prédicteurs du modèle (Figure 60) montre que la forme de ce modèle est appropriée.

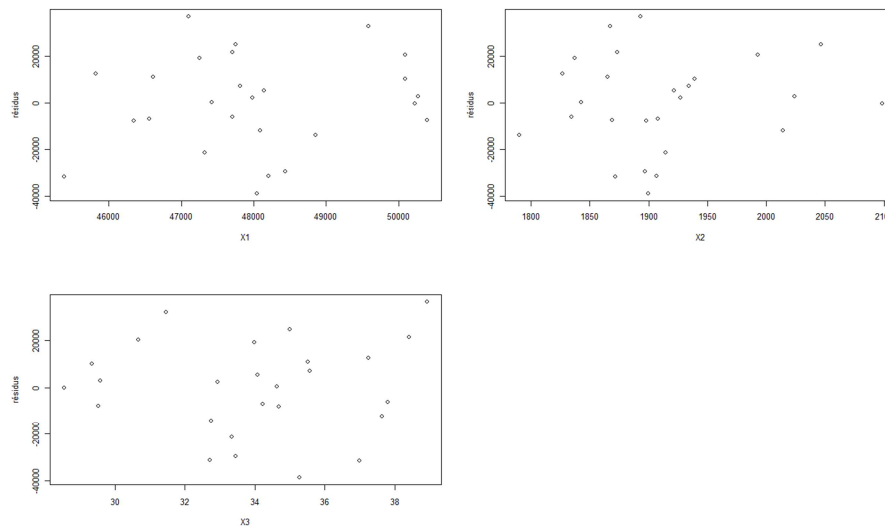


Figure 60. Vérification des résidus du Modèle 4 : résidus contre prédicteurs.

L'estimation ponctuelle du Modèle 4 est donnée par l'Application numérique 3.

$$Y_t \approx 14,02X_{1,t} + 183,54X_{2,t} - 6414,53X_{3,t} + \varepsilon_t$$

avec :

$$\varepsilon_t \text{ i.i.d. de loi } N(0 ; 20958,52^2)$$

*Application numérique 3. Estimation du Modèle 4.*

Il faut souligner qu'un risque de colinéarité subsiste, la corrélation entre  $X_{1,t}$  et  $X_{3,t}$  étant par exemple égale à -0,79. Néanmoins, le recours à deux règles de vérification usuelles indique que ce risque est limité : le VIF (*Variance Inflation Factor*) de chaque variable est inférieur à 3 et le coefficient de détermination de chacune des régressions partielles est inférieur au coefficient de détermination global du modèle. En conséquence, nous choisissons de conserver telles quelles les variables, afin d'éviter des manipulations qui pourraient engendrer un modèle difficile à interpréter. Par ailleurs, on vérifie également la validité du Modèle 4 en effectuant une régression PLS<sup>54</sup>, adaptée au traitement des modèles colinéaires, et des résultats proches de l'Application numérique 3 sont obtenus.

La Figure 61 trace les valeurs réelles de  $Y_t$ , les valeurs prédites  $\hat{Y}_t$  selon le Modèle 4, ainsi que les intervalles de prévision.

<sup>54</sup> Estimée avec le logiciel TANAGRA.

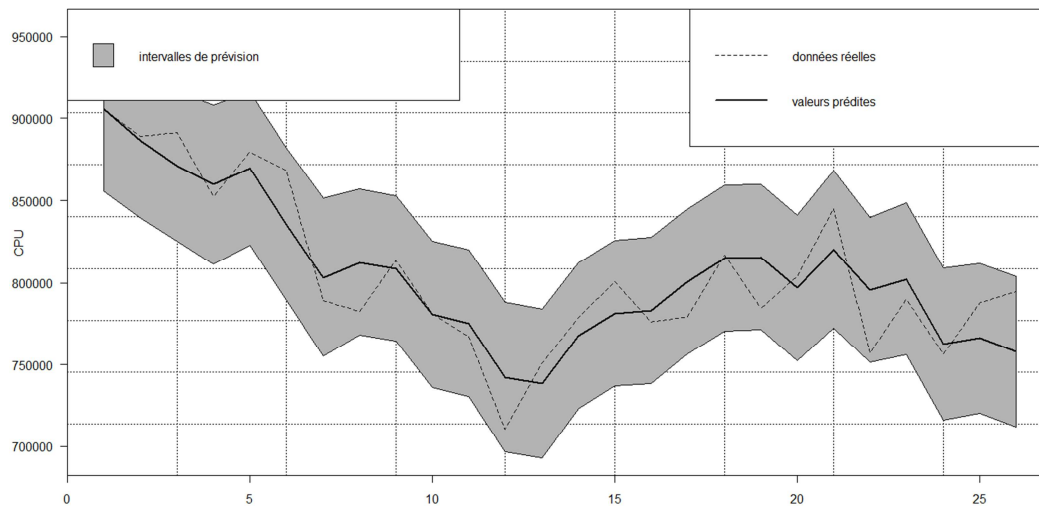


Figure 61. Modèle 4 : données réelles et valeurs prédites.

Aucun point n'est en dehors des intervalles de prévision. Cela semble cohérent car, au seuil d'erreur de 5% et pour 26 observations, on peut s'attendre au plus à une valeur de  $Y_t$  en-dehors des intervalles de prévision. Ce modèle fournit donc une bonne approximation du comportement du système étudié.

Nous procédons ensuite à une validation quantitative. Pour tester la qualité du modèle en prévision, une validation croisée à trois groupes est réalisée (Figure 62, graphique de gauche). Les groupes utilisés pour l'estimation du modèle et la prévision sont constitués aléatoirement, grâce à un algorithme adapté. On constate que les résultats sont proches des données réelles et prédites présentées en Figure 61. On utilise également le même algorithme pour réaliser une validation croisée à  $k$  groupes (*folds*) : l'expérience est cette fois-ci répétée plusieurs fois, pour différents nombres de groupes. Les résultats sont bons, on constate que le  $Q^2$ <sup>55</sup> se stabilise autour de 0,78 (Figure 62, graphique de droite).

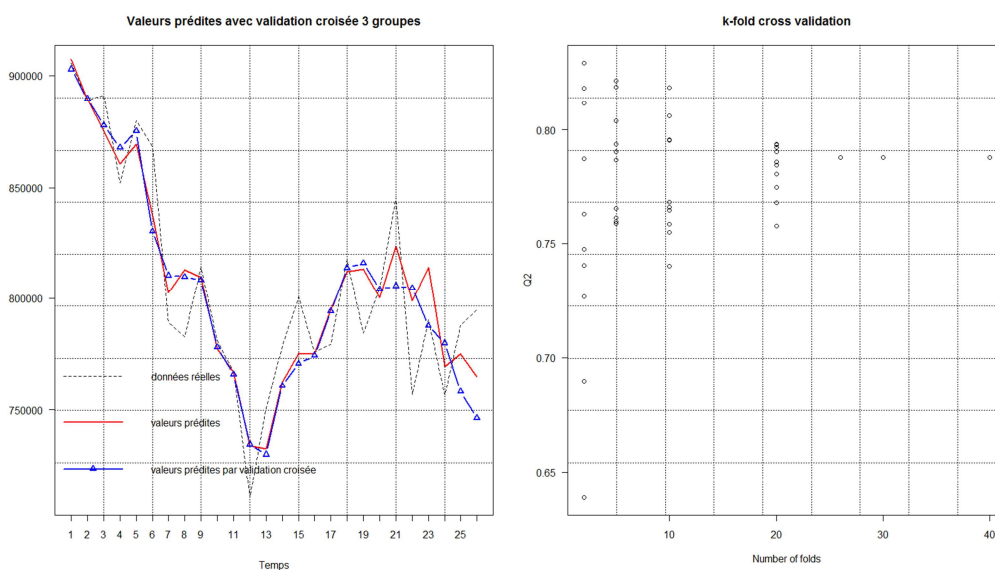


Figure 62. Modèle 4 : validation croisée.

<sup>55</sup> Equivalent au coefficient de détermination  $R^2$  d'une régression linéaire multiple, sur les données de validation croisée.

On teste ensuite l'hypothèse nulle de stabilité des paramètres du Modèle 4. Nous réappliquons la stratégie de test déjà employée pour le Modèle 3. L'hypothèse nulle ne peut être rejetée. Ce modèle semble structurellement stable (Figure 63).

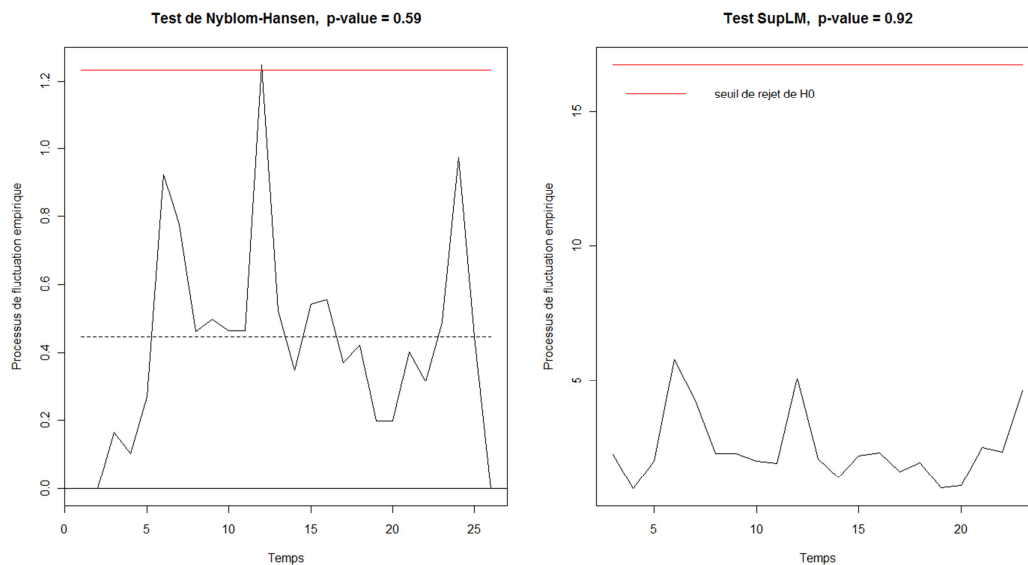


Figure 63. Test de la stabilité structurelle du Modèle 4.

Pour ces raisons, ce modèle semble être une base solide pour une utilisation en prévision.

### 1.3.3. Interprétation des informations - connaissances (aspect pragmatique)

Pour conclure la démarche de modélisation, le modèle de prévision basé sur les variables de pilotage est interprété et validé qualitativement. Son intérêt pour la prise de décision est confirmé.

L'Application numérique 3 est discutée avec les experts. Elle est facilement acceptée, puisqu'elle reflète les conclusions établies lors de l'interprétation qualitative du modèle d'observation n°2 :

- La présence des variables  $X_{1,t}$  et  $X_{2,t}$  permet de bien caractériser l'activité de l'usine.  $X_{2,t}$  reflète la production effective. En la mettant en rapport avec l'en-cours de production  $X_{1,t}$ , on est capable d'évaluer également l'impact de l'engorgement de la fabrication, dont on sait qu'il n'est pas sans conséquences sur le besoin en ressources informatiques MES ;
- $X_{3,t}$  permet d'intégrer au modèle l'effet du déploiement des applications MFA. Son coefficient négatif montre qu'il a tendance à décharger le serveur MES, certaines tâches étant transférées à un autre domaine du SI, lorsque réalisées selon le mode de production automatique.

Le modèle est bien validé. Il est compris et admis par les experts, d'autant qu'il est directement rattachable au modèle basé sur l'activité de production opérationnelle (Modèle 3). Ces variables, bien connues dans l'environnement STMicroelectronics,

prévisibles et usuellement rencontrées dans les discours managériaux, rendent ce Modèle 4 utile pour l'aide à la décision souhaitée.

La connaissance formalisée par ce modèle de prévision peut désormais être mise en usage, en l'intégrant dans un modèle d'évaluation, tel que formulé par Tsoukiàs, et servir ainsi de support à l'aide à la décision. Avant cela, nous nous interrogeons sur le caractère générique de la démarche de modélisation proposée.

## Section 2. Quelle généralité de la démarche de modélisation ?

### 2.1. Justification du questionnement et présentation d'un nouveau cas d'application

#### 2.1.1. Généralité de la méthode : proposition

Par l'application du processus socio-statistique d'aide à la décision au cas MES, nous avons défini une démarche de modélisation. Cette méthode vise à organiser la compréhension de l'activité d'un composant d'infrastructure, dans le cadre d'un environnement industriel défini par des variables de pilotage. Cette méthode a été bâtie dans le cadre d'un besoin idiosyncratique, à savoir la modélisation du comportement du serveur MES. Néanmoins, il semblerait qu'elle puisse offrir une trame générale de modélisation, dès lors que l'on cherche à établir un modèle cherchant à caractériser des interactions entre un composant d'infrastructure et une activité organisationnelle. Cela signifie que, si une nouvelle application du processus socio-statistique d'aide à la décision révèle un problème de décision  $\tau$  similaire à celui exprimé pour le projet MES, la méthode de structuration progressive de connaissance proposée pourrait être appliquée à nouveau. En effet, ses étapes de modélisation et les méthodes quantitatives suggérées semblent fournir un cadre suffisamment général pour définir des modèles de prévision en fonction de variables industrielles de pilotage, pour d'autres sous-systèmes fonctionnels de l'architecture des TI.

Même si cette proposition mériterait d'être confrontée à une étude plus générale, nous apportons ci-après un argument en ce sens : la méthode proposée est appliquée à un autre périmètre applicatif du système d'information de Crolles 2, avec des résultats probants.

#### 2.1.2. Un nouveau cas d'application

Nous appliquons la méthode de structuration de connaissances au serveur de Crolles 2 dédié à l'application RTD (*Real Time Dispatching*), du groupe d'application MFA. Ce serveur supporte plusieurs activités d'automatisation des traitements de lots de production, évoquées lors de l'étude du cas MES. Ces deux périmètres fonctionnels sont complémentaires : lorsque certaines actions sont réalisées en mode automatique, certains traitements réalisés par le SI sont gérés par les serveurs RTD et non plus MES (d'où le coefficient négatif associé à la variable  $X_{3,t}$  dans le modèle d'aide à la décision MES).

Nous cherchons à formaliser une connaissance relative à l'activité de ce serveur, grâce à la démarche de modélisation suggérée. Tout comme pour le cas MES, aucune connaissance *a priori* n'est directement exploitable pour ce modèle. Il convient de stimuler les mécanismes de conversions de connaissances, tel que proposé dans le cadre du processus socio-statistique d'aide à la décision. Dans le cas présent, une difficulté particulière est liée au caractère instable de l'environnement applicatif. Nous montrerons



comment il peut être surmonté et comment l'observation statistique permet de favoriser la compréhension de ce phénomène.

Cette application est présentée ci-après, de façon plus succincte que pour le cas MES : chacun des aspects de l'ontologie du système de connaissances n'est pas présenté, on résume simplement les principales conclusions de l'application des trois étapes de la démarche de modélisation. Les traitements quantitatifs étant les mêmes que ceux utilisés pour le cas MES, le script R n'est pas proposé en annexe de la thèse.

## **2.2. Structuration progressive de connaissances pour la modélisation du serveur RTD**

### *2.2.1. Etape (A) intermédiaire d'observation n°1*

L'activité de 102 transactions de l'application RTD est considérée, sur une période de 176 jours. Une ACP avec rotation *Varimax* est réalisée. Les deux premières composantes expriment à elles seules près de 72% de la variabilité applicative, et sont de très bon prédicteurs de l'activité du serveur. La Figure 64 caractérise l'espace des variables pour les trois premières composantes. Nous présentons la troisième composante à titre indicatif, car elle permet de découvrir un mode de production nouveau, donc une connaissance nouvelle, non considérée lors du cas MES. En effet, les transactions liées à la troisième composante correspondent à des transactions RTD que les experts ont identifié comme étant exclusivement dédiées à certains équipements de production : ceux utilisés pour les tâches de lithographie. Ces équipements utilisent un mode production semi-automatique spécifique, auquel sont rattachées des transactions RTD particulières. Néanmoins, ce mode de production est trivial en volume (moins de 3% de l'activité de production) et son impact sur l'activité des serveurs peut être considéré comme négligeable. Nous n'aurons donc pas à en tenir compte dans le modèle de prévision.

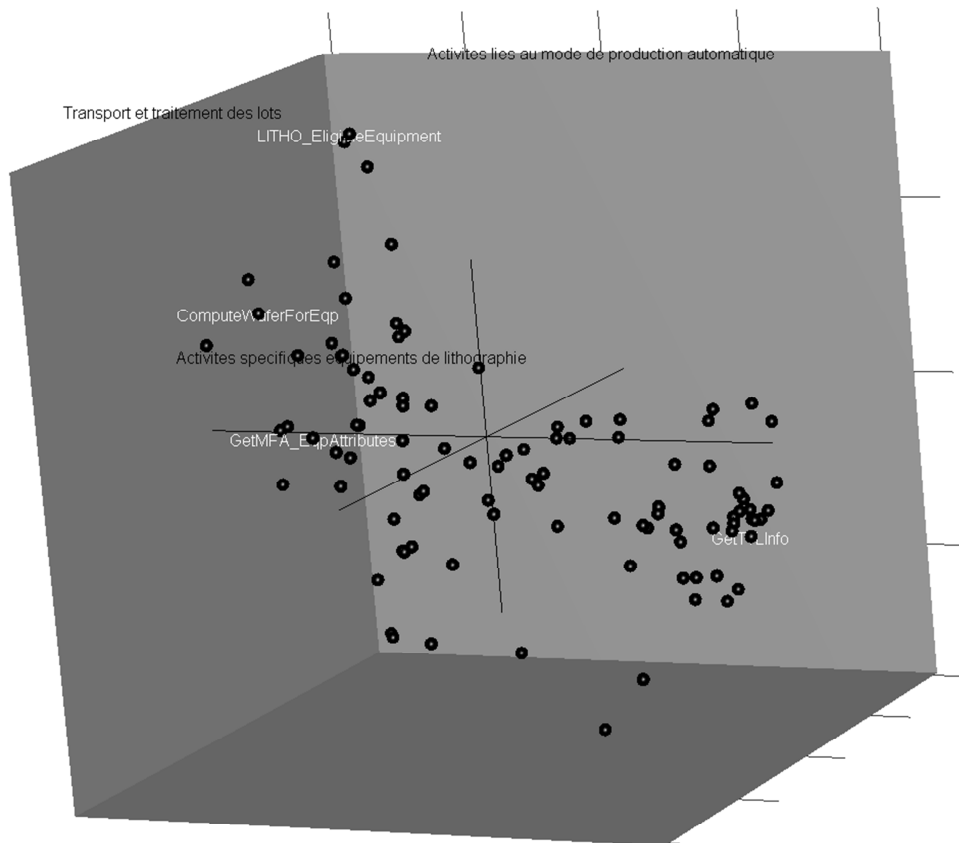


Figure 64. Analyse de l'espace des variables de l'ACP des transactions RTD : découverte d'un nouveau mode de production semi-automatique pour les équipements de lithographie.

En conséquence, deux activités industrielles sous-jacentes, déjà considérées dans le cas MES, sont retenues :

- *Vérification des lots de production à traiter, assignation des lots aux files d'attente des équipements et recherche d'information, pour préparer automatiquement le traitement des lots : RTD seconde le système MES pour un ensemble d'activités de préparation des lots de production, lorsqu'elles sont réalisées automatiquement ;*
- *Transports et traitements des lots : le système RTD est également sollicité lors des étapes de traitement effectif des lots de production, non pas pour effectuer le traitement, mais pour réaliser un ensemble de vérifications et mises à jours utilisées dans le cadre des algorithmes de préparations automatique.*

Cette observation de l'activité applicative permet de constituer le modèle intermédiaire d'observation n°2, basé sur l'activité opérationnelle de production.

### 2.2.2. Etape (B) intermédiaire d'observation n°2

Les variables retenues pour la constitution du modèle opérationnel seront donc les mêmes que les  $x_{i,t}$  du modèle MES, à l'exception de  $x_{WE,t}$ , l'activité RTD ne suggérant pas de saisonnalité hebdomadaire. Seront donc considérées les variables suivantes :

- $(x_{1,t} + x_{2,t})$ , pour caractériser l'activité générale de transports et traitements des lots ;
- $x_{2,t}$ , pour caractériser les effets spécifiques liés au mode de production automatique ;
- $R_t$ , pour mesurer le débit de l'usine tout en évitant les risques de colinéarité du modèle.

Ces variables sont observées sur une période de 237 jours. Les experts indiquent que les effets liés au déploiement de l'activité automatique ( $x_{2,t}$ ) sont difficiles à évaluer directement, car pollués par des instabilités sur la période observée. En effet, deux phénomènes contradictoires accompagnent la hausse constante de  $x_{2,t}$  pendant le laps de temps considérée :

- *D'une part, l'application RTD fait l'objet d'un déploiement soutenu de nouvelles fonctionnalités de traitement automatique, ayant tendance à augmenter la charge sur le serveur ;*
- *D'autre part, l'application RTD fait l'objet d'actions d'optimisation de la performance (tuning), ayant tendance à diminuer la charge sur le serveur.*

Nous tentons d'observer ces phénomènes en intégrant une tendance linéaire dans le modèle. Pour cela, on utilise une variable  $T_t$  dans le modèle, qui représente le temps. Cette astuce augmente nettement la qualité du modèle, mais a pour conséquence de ne plus permettre de capter isolément des effets additionnels liés à  $x_{2,t}$  : cette variable n'est plus significative dans la forme finale du modèle. On note également que le terme d'erreur suit un processus temporel AR(1). En conséquence, on considère le modèle de prévision donné par le Modèle 5, où  $Y_t$  représente l'activité journalière en secondes CPU consommées du serveur RTD.

$$Y_t = \beta_0 + \beta_1(x_{1,t} + x_{2,t}) + \beta_2 R_t + \beta_3 T_t + \mu_t$$

avec :

$$\mu_t = \varphi \mu_{t-1} + \varepsilon_t, \varepsilon_t \text{ i.i.d. de loi } N(0, \sigma^2)$$

*Modèle 5. Modèle retenu de prévision de  $Y_t$  (serveur RTD) à partir de l'activité industrielle opérationnelle.*

Si ce modèle semble bon de prime abord, expliquant près de 85% de la variabilité de  $Y_t$ , il est intéressant de remarquer qu'il présente une instabilité structurelle significative. La Figure 65, qui représente les résultats de l'application des tests de détection de changements structurels basés sur les statistiques de Nyblom-Hansen et SupLM, en atteste.

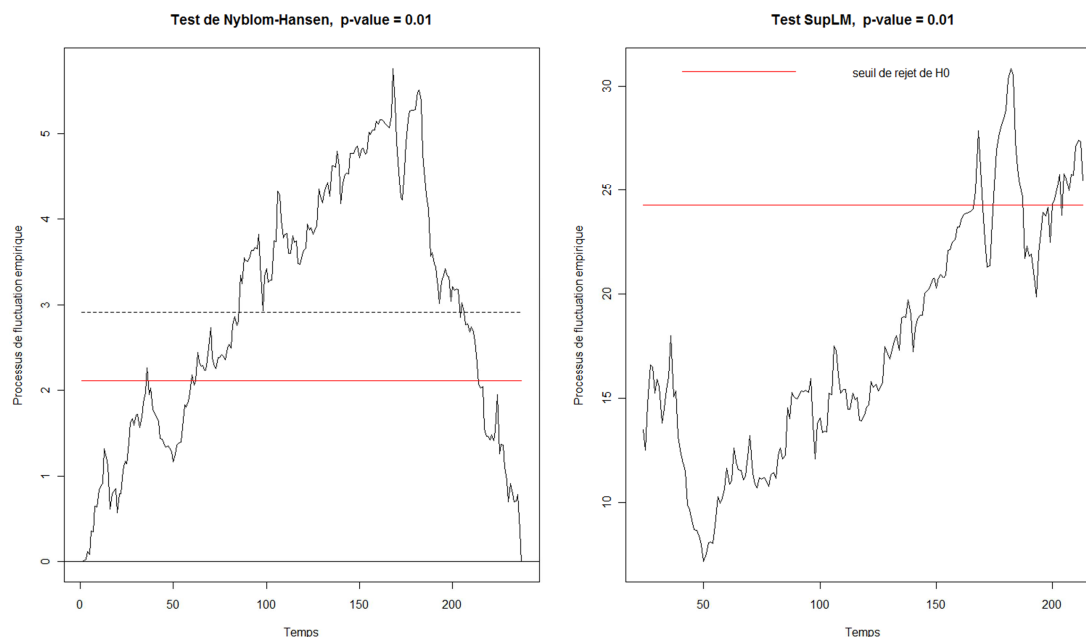


Figure 65. Test de la stabilité structurelle du Modèle 5.

Le seuil de rejet est très nettement franchi. On peut chercher à dater la ou les moments de ruptures de stabilité grâce aux méthodes de Bai<sup>56</sup>. On identifie un, voire deux points de rupture. On considère ici le cas d'un seul changement structurel. L'algorithme de Bai suggère un point de rupture à la 165<sup>ème</sup> observation. L'Application numérique 4 présente l'estimation du modèle sur la première période de stabilité.

$$Y_t \approx 7473,38 + 97,09(x_{1,t} + x_{2,t}) + 0,30R_t + 46,58T_t + \mu_t$$

avec :

$$\mu_t \approx 0,26\mu_{t-1} + \varepsilon_t, \varepsilon_t \text{ i.i.d. de loi } N(0 ; 806,26^2)$$

*Application numérique 4. Estimation du Modèle 5, de la 1<sup>ère</sup> à la 165<sup>ème</sup> observation.*

Ce modèle respecte les hypothèses du modèle linéaire. Il montre que la consommation en secondes CPU du processeur RTD ( $Y_t$ ) dépend de l'activité générale de la fabrication ( $x_{1,t}$  et  $x_{2,t}$ ), de l'en-cours de production non traité ( $R_t$ ), mais aussi d'une tendance linéaire croissante ( $T_t$ ). Cette tendance exprime le fort déploiement de nouvelles fonctionnalités du SI, liées au projet MFA à Crolles 2. Elle intègre à la fois la hausse de la variable  $x_{2,t}$  prise isolément, mais aussi le déploiement de nouvelles fonctionnalités RTD qui tendent à augmenter la charge sur le système. Enfin, certains phénomènes sous-jacents inconnus liés à la temporalité des données observées sont modélisés ( $\mu_t = \varphi\mu_{t-1} + \varepsilon_t$ ).

La Figure 52 trace les valeurs prédites  $\hat{Y}_t$  du Modèle 3, contre les données réelles connues de la variable  $Y_t$ : le modèle fournit une bonne approximation de la consommation réelle de ressources CPU, peu de points sont hors des intervalles de prévision à 5%.

<sup>56</sup> C'est la seule méthode qui n'a pas été employée dans les scripts R présents en annexe. Pour réaliser cette datation, la fonction 'breakpoints' de la librairie R *strucchange* a été utilisée.

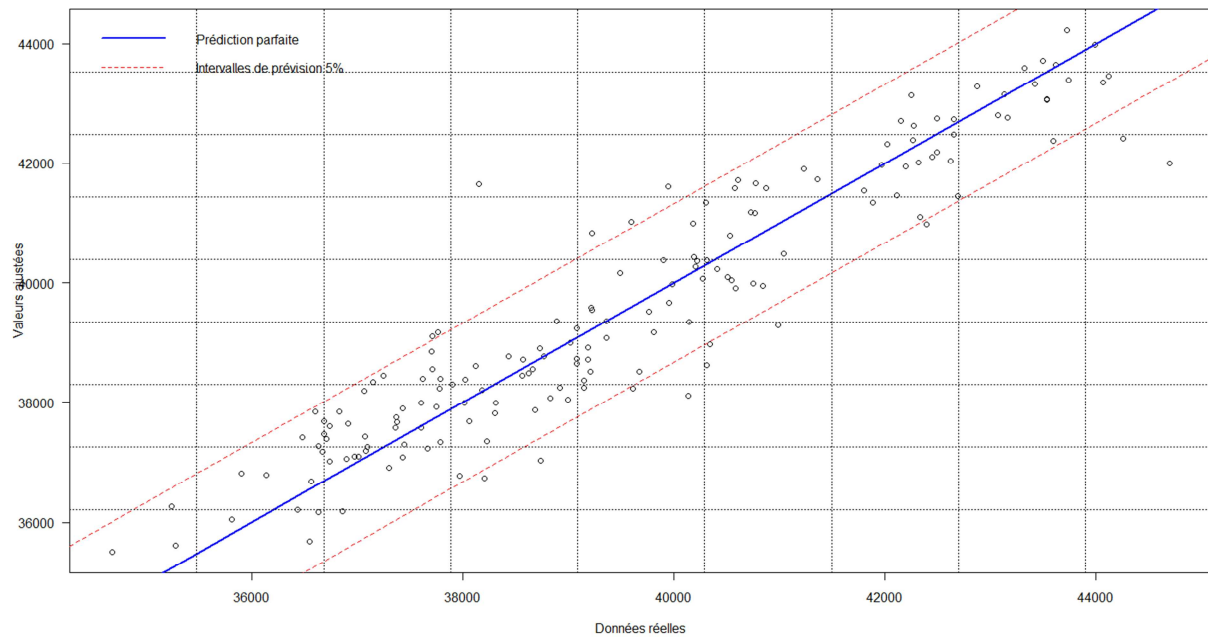


Figure 66. Modèle 5 : valeurs prédites versus données réelles de la consommation CPU du serveur RTD (observations 1 à 165).

Ce modèle est positivement validé par les algorithmes de validation croisée de Hyndman.

Si l'on utilise les observations suivantes (166 à 237) pour estimer le Modèle 5, la qualité de l'ajustement se dégrade et l'utilisation pour la prévision devient moins bonne. Les experts expliquent qu'en effet, c'est une période de fort bouleversement de l'activité du système :

- *Le déploiement de nouvelles fonctionnalités MFA s'arrête*, comme en atteste la variable  $T_i$ , qui cesse d'être statistiquement significative sur cette seconde période ;
- *Le niveau applicatif devient très instable*, fortement perturbé par de nombreuses activités d'optimisation des transactions. Il devient ainsi beaucoup plus difficile d'observer des régularités dans les interactions entre  $Y_i$  et l'activité organisationnelle.

Une étude plus approfondie, destinée à trouver d'autres indicateurs statistiques pertinents, peut être réalisée ; mais cette question ne sera plus traitée dans le cadre de cette thèse. Nous nous limiterons ici à présenter le modèle d'aide à la décision, développé à partir de la première période de stabilité.

### 2.2.3. Etape (C) d'aide à la décision

Les choix de variables est très similaire à celui du modèle MES. On conserve  $X_{1,t}$  (charge globale de l'usine) et  $X_{2,t}$  (nombre moyen d'opérations de production). Seul  $X_{3,t}$  est remplacé par le temps ( $T_i$ ) dans le modèle RTD, afin d'intégrer le déploiement conjoint du mode automatique et des nouvelles fonctionnalités MFA. Sur 26 observations agrégées à la semaine, qui correspondent à la même période de modélisation que celle du

modèle MES, on obtient un excellent modèle. Il explique presque 94% de la variabilité de  $Y_t$ , lorsqu'estimé avec l'ordonnée à l'origine (qui n'est pas statistiquement significative). La forme générale du modèle est donnée par le Modèle 6.

$$Y_t = \alpha_1 X_{1,t} + \alpha_2 X_{2,t} + \alpha_3 T_t + \varepsilon_t$$

avec :

$$\varepsilon_t \text{ i.i.d. de loi } N(0, \sigma^2)$$

*Modèle 6. Modèle de prévision de  $Y_t$  (RTD), à partir des variables  $X_{i,t}$  de pilotage.*

L'estimation de ce modèle est donnée par l'Application numérique 5.

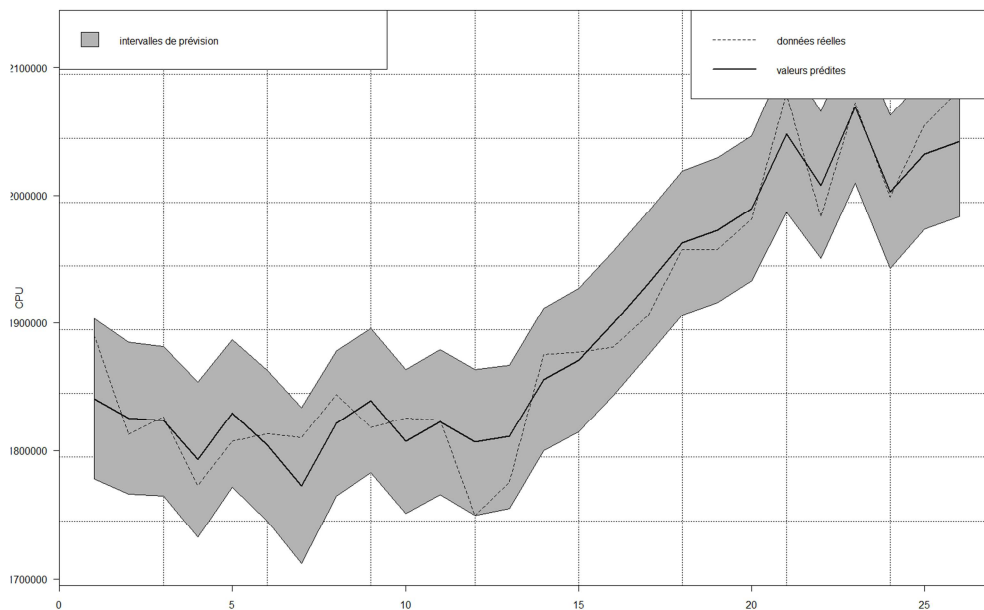
$$Y_t \approx 19,56X_{1,t} + 402,41X_{2,t} + 13808,19T_t + \varepsilon_t$$

avec :

$$\varepsilon_t \text{ i.i.d. de loi } N(0 ; 26506,85^2)$$

*Application numérique 5. Estimation du Modèle 6.*

Les valeurs réelles ( $Y_t$ ) versus les valeurs prédites par le Modèle 4 ( $\hat{Y}_t$ ), ainsi que les intervalles de prévision, sont tracés dans la Figure 67.



*Figure 67. Modèle 6 : données réelles et valeurs prédites.*

Les critères de qualité du modèle sont vérifiés (résidus, utilisation en prévision, etc.). Il représente donc une connaissance solide concernant l'activité du système étudié. Remarquons qu'il se pourrait que l'on ne veuille pas incorporer le temps ( $T_t$ ) dans les variables du modèle, par exemple parce que l'on n'est pas capable d'expliquer la tendance observée. Dans ce cas, la modélisation serait tout de même possible, mais il faudrait stationnariser la série en la différenciant. Il faut savoir que cette manipulation fera augmenter l'incertitude du modèle : en intégrant les erreurs, on va modéliser une

variance de prédiction qui augmentera avec l'horizon de prévision<sup>57</sup> (dans le cas d'une différenciation d'ordre 1, le terme d'erreur correspond à une somme de loi normale de paramètres  $N(0, \sigma^2)$ , c'est-à-dire une marche au hasard).

Finalement, le modèle est validé par les experts. Ils ont notamment apprécié sa capacité à formaliser leur ressenti concernant certaines interactions entre les systèmes MES et RTD. En effet, nous avons vu que le déploiement du mode automatique a tendance à décharger le serveur MES. *A contrario*, il a tendance à charger le serveur RTD, tendance encore accrue du fait du déploiement de nouvelles fonctionnalités de traitements automatiques sur la période modélisée. Les modèles de pilotage permettent bien de formaliser ce phénomène :  $X_{3,t}$  est associé à un coefficient négatif pour le modèle MES et on observe une forte tendance à la hausse ( $T_i$ ) pour le modèle RTD. Cela explique bien les observations qui peuvent être faites quant à l'activité des serveurs. Par exemple, de la semaine 1 à la semaine 12, on observe que l'activité du serveur MES diminue et que celle du serveur RTD reste assez stable, alors que les variables explicatives des deux modèles sont similaires : c'est parce que les deux modèles réagissent différemment à une même activité métier.

## Conclusion du chapitre 6

Ce chapitre a permis de montrer comment la démarche de modélisation proposée peut être appliquée, pour structurer une connaissance visant à satisfaire un besoin décisionnel donné. La démarche a été validée à partir du cas fil rouge de cette thèse : la modélisation prospective des serveurs MES. Qui plus est, nous avons aussi validé sa pertinence sur un autre sous-système fonctionnel, pour un besoin décisionnel identique.

Nous pensons que les limites d'applications des approches usuelles de *capacity planning* évoquées dans la première partie de la thèse ont été surmontées, grâce à l'application de cette démarche de modélisation. En développant les synergies entre les acteurs de l'organisation et l'analyste/le chercheur, un modèle intégrant pleinement l'activité organisationnelle locale et basé sur des méthodes statistiques appropriées, a été proposé. Cette construction progressive a permis de structurer pas-à-pas des connaissances nouvelles, sur les interactions entre le système considéré et son environnement industriel, formalisées dans le cadre de relations quantitatives comprises et acceptées par les acteurs de l'organisation.

En conséquence, ce travail aura permis de tenir compte des enjeux de gestion des technologies de l'information, à savoir : la capacité à gérer un système complexe, par l'emploi d'une approche systémique et statistique, permettant de modéliser certains macro-comportements du système ; la capacité à soutenir pleinement l'activité de l'organisation, en inscrivant ce macro-modèle dans une perspective industrielle, grâce à des variables de pilotage significatives pour la Direction Générale et pour la gestion stratégique de l'appareil productif.

Concernant les apports systémiques de l'approche proposée, la modélisation complémentaire du sous-système RTD apporte un complément de réflexion intéressant.

---

<sup>57</sup> Voir Charpentier (2009).

En effet, par la modélisation de comportements identifiés comme stables dans l'activité des serveurs MES, puis RTD, puis en identifiant des variables communes ainsi que des points d'interactions entre ces deux sous-systèmes, nous suggérons avoir fait un pas additionnel vers l'appréhension d'un système complexe. En effet, observer différents éléments d'un système, puis les connecter progressivement entre eux, par des variables explicatives communes, est une manière d'appréhender un système complexe. C'est une démarche typique de la modélisation systémique, telle mentionnée dans la première partie de la thèse. On peut faire allusion à Le Moigne (1999), qui évoque la modélisation d'un système complexe comme un « *projet de complexifications fonctionnel croissant par étapes stabilisatrices* ». La modélisation MES est une première étape stabilisatrice. La modélisation RTD en est une deuxième. Les deux modèles sont alors connectés par des variables communes, augmentant ainsi le périmètre fonctionnel de compréhension du système complexe. La démarche de modélisation proposée semble alors suffisamment générique pour permettre cette extension progressive du périmètre de compréhension, à condition bien sûr de rester dans un contexte décisionnel identique.

Nous pouvons désormais montrer comment la connaissance formalisée dans ce modèle statistique peut être actionnée à des fins décisionnelles.



## **Chapitre 7. Usage du modèle de prévision pour l'aide à la décision.**

## Introduction

Ce chapitre s'intéresse à la deuxième partie de l'utilisation de modèles dans le cadre de l'aide à la décision. Il montre comment le modèle statistique MES défini dans le chapitre 6 peut désormais fournir un support à la décision, une fois inscrit dans un modèle d'évaluation. Pour cela, le plan ci-dessous sera suivi (Figure 68).

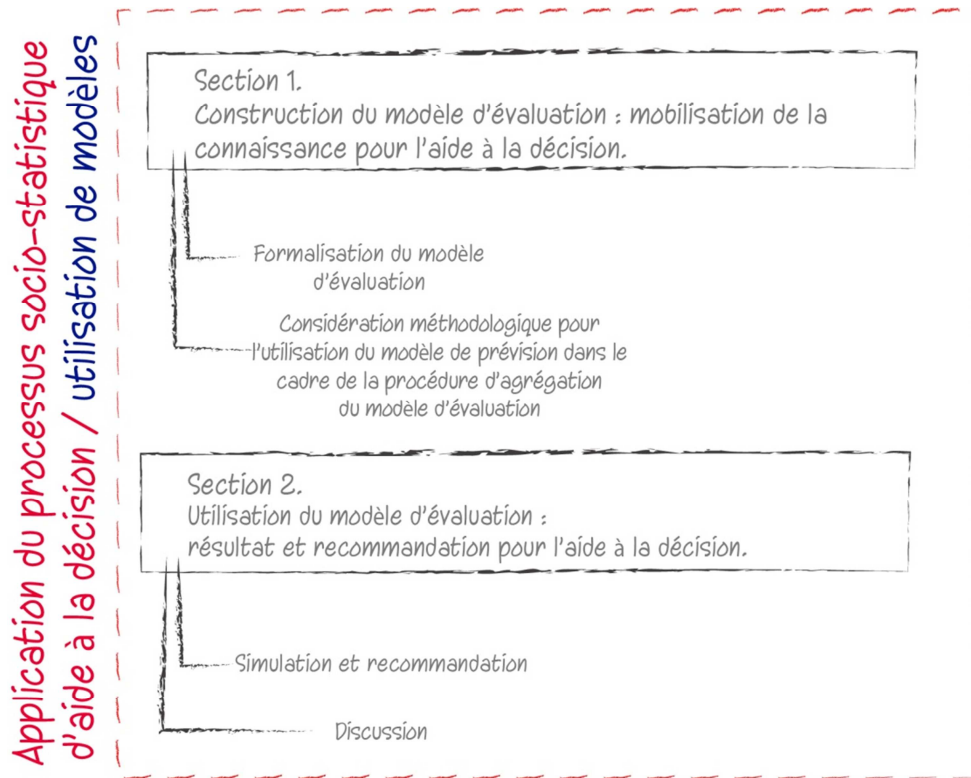


Figure 68. Plan du chapitre 7.

Conformément au cadre de Tsoukiàs, la section 1 montre comment formaliser un modèle d'évaluation, basé sur le modèle prospectif MES. Pour cela, certaines considérations méthodologiques doivent être prises en compte, pour pouvoir exploiter le modèle statistique en réponse à l'aide à la décision attendue.

La section 2 présente concrètement l'utilisation du modèle d'évaluation. Par simulation, nous produirons des résultats quantitatifs pouvant guider la prise de décision. Ils permettront de montrer quelles recommandations auraient pu être suggérées, si le processus socio-statistique d'aide à la décision avait été appliqué au cas de la migration MES. En effet, en 2009, les managers de l'infrastructure TI-C2 appliquèrent une règle de trois pour anticiper l'accroissement de la charge des serveurs en fonction d'une hausse de l'activité industrielle. Or, les résultats quantitatifs obtenus par l'application du processus socio-statistique donnent des résultats ancrés dans une observation plus rigoureuse de l'activité du système et permettent d'affiner les scénarios prévisionnels de l'activité industrielle de Crolles 2. Ces résultats sont discutés plus en détail, afin de mettre en exergue les bénéfices apportés par cette approche formelle et quantitative de l'aide à la décision.

## Section 1. Construction du modèle d'évaluation : mobilisation de la connaissance pour l'aide à la décision.

### 1.1. Formalisation du modèle d'évaluation

Nous montrons désormais comment le modèle de prévision construit pour le serveur MES aurait pu être utilisé lors du projet de migration, s'il avait été inscrit dans le cadre de l'usage d'un modèle d'évaluation  $M$ . On rappelle que  $M$  est défini par le n-uplet :

$$M = \langle \mathcal{A}, D, E, H, U, R \rangle$$

Comme nous l'avons suggéré dans la présentation du problème de décision, les alternatives ( $\mathcal{A}$ ) sur lesquelles va s'appliquer le modèle d'évaluation sont considérées au travers de l'accroissement du besoin en ressource technologique  $F_b$  (voir Modèle 1, modèle de rationalité des décideurs). En effet, selon la valeur de ce facteur, différents dimensionnements des nouveaux serveurs HP *Itanium* peuvent être envisagés (gamme RX660 ou RX8640, nombre de CPU/de cœurs, etc.). Les considérations techniques, menant de l'estimation de  $F_b$  au choix du serveur final, sont hors du périmètre du modèle d'évaluation ici présenté : c'est la définition de  $F_b$  qui nous intéresse. Potentiellement, les alternatives du choix de  $F_b$  s'étendent de un à l'infini.

Les dimensions ( $D$ ) sous lesquelles les alternatives sont considérées et leurs métriques ( $E$ ) correspondent aux variables du Modèle 4 de prévision, rappelées dans le Tableau 17.

Dimensions ( $D$ )	Métriques ( $E$ )
$Y_t$ : activité du processeur du serveur MES	Secondes CPU moyennes consommées à la semaine
$X_{1,t}$ : volume de plaques à produire	WIP/w
$X_{2,t}$ : activité de production effective	<i>moves</i> /w
$X_{3,t}$ : déploiement MFA	%

Tableau 17. Formalisation du modèle d'évaluation : dimensions de  $\mathcal{A}$ .

Les critères d'évaluations de  $\mathcal{A}$  ( $H$ ) vont être déterminés par les valeurs des variables de pilotage de  $D$ , à l'horizon  $h$  de prévision. Elles vont exprimer les hypothèses du décideur quant à la quantification prévisionnelle de l'activité de production. En 2009, la Direction Générale fournit une évaluation globale d'une hausse de l'activité de production d'un facteur 2,25 (passage de 2000wo/w à 4500wo/w). Les décideurs doivent désormais être capables de décliner cette information sur les variables de pilotage du modèle de prévision : si le nombre de *wafers-out* augmente, est-ce le fait d'une hausse du débit de la fabrication à volume d'en-cours de production constant, ou est-ce le fait d'une augmentation plus générale de l'activité de production ? *Quid* de l'automatisation de l'usine ? etc. Des hypothèses doivent donc être formalisées, dans le cadre de scénarios de prévision.

Avec l'aide des experts,  $j$  scénarios sont ainsi considérés pour caractériser les variations des variables  $X'_{i,b}$  à l'horizon  $b$  :

- *Scénario optimiste ( $j = O$ )* : la hausse du volume de wo/w correspond à un scénario de désengorgement de l'usine. Le nombre de plaques à produire reste stable, mais la cadence de production augmente. Le mix technologique ne change pas. Le mode automatique est déployé conformément aux objectifs du projet MFA (+30 points à l'issue de la hausse du volume de wo/w, puis stabilisation) ;
- *Scénario réaliste ( $j = R$ )* : la hausse du volume de wo/w correspond à une hausse du nombre de plaques à produire, à cadence de production constante. Le mix technologique ne change pas. Le mode automatique est déployé conformément aux objectifs du projet MFA ;
- *Scénario pessimiste ( $j = P$ )* : la hausse du volume de wo/w correspond à une hausse du nombre de plaques à produire, à cadence de production constante mais avec une complexification du mix technologique (+ 11% de *moves* environ pour réaliser une plaque). Le mode automatique n'est pas déployé.
- *Scénario historique implicite ( $j = I$ )* : ce scénario exprime les hypothèses qui furent implicitement utilisées par les experts lors de l'estimation de  $F_b$  sur la base d'une règle de trois : hausse du nombre de plaques à produire, à cadence de production constante et technologies identiques, sans être au fait d'un quelconque effet du déploiement du projet MFA sur le serveur.

Sur la base de ces scénarios, les variations des  $X'_{i,b}$  à l'horizon  $b$ , par rapport à l'instant  $t = 0$ , sont indiquées dans le Tableau 18.

<b>Scénario</b>	<b>Variation par rapport à <math>t = 0</math></b>		
	$X'_{1,b}$	$X'_{2,b}$	$X'_{3,b}$
O	stable	x 2,25	+ 30 points
R	x 2,25	x 2,25	+ 30 points
P	x 2,25	x 2,5	stable
I	x 2,25	x 2,25	stable

Tableau 18. Scénarios de prévision des variables de pilotage à l'horizon  $b$ .

Les incertitudes ( $U$ ) liées à ce modèle d'évaluation doivent être considérées :

- *Concernant  $H$ , la prise en compte de différents scénarios est une façon d'intégrer l'incertitude.* Néanmoins, aucune méthode mathématique n'est requise pour traiter cette incertitude. Définis qualitativement avec les experts, il est assumé que les valeurs des prédicteurs sont connues à l'avance<sup>58</sup> ;
- *Concernant  $D$ , l'incertitude reliant la variable  $Y_i$  à ses prédicteurs est directement exprimée par la forme du modèle de prévision utilisé.* En effet,  $F_b$  est calculé à partir du Modèle 4, qui est un modèle probabiliste.

Le dernier élément de la formalisation du modèle d'évaluation est la caractérisation de la procédure d'agrégation ( $R$ ). Elle tient compte des  $j$  scénarios, selon le Modèle 7 ci-après.

<sup>58</sup> Cf. Hyndman et Athanasopoulos (2012).

$$F_b^j = \frac{Y_b^j}{Y_0} = \frac{\hat{\alpha}_1 X_{1,b}^j + \hat{\alpha}_2 X_{2,b}^j + \hat{\alpha}_3 X_{3,b}^j + \varepsilon_b}{\hat{\alpha}_1 X_{1,0}^j + \hat{\alpha}_2 X_{2,0}^j + \hat{\alpha}_3 X_{3,0}^j + \varepsilon_0}$$

avec :

$F_b^j$  : facteur d'accroissement à l'horizon  $b$  pour le scénario  $j$ ,

$Y_0$  : estimation du besoin en ressource technologique au moment de réalisation de l'estimation, selon le Modèle 4,

$Y_b^j$  : estimation du besoin en ressource technologique à l'horizon  $b$  pour le scénario  $j$ , selon le Modèle 4,

$\varepsilon_b$  et  $\varepsilon_0$  sont i.i.d. de loi  $N(0, \sigma^2)$ ,

$Y_0$  et  $Y_b^j$  sont estimés en supposant les paramètres  $\alpha_i$  connus, c'est-à-dire en utilisant les estimations  $\hat{\alpha}_i$  (Application numérique 3)<sup>59</sup>

*Modèle 7. Procédure d'agrégation du modèle d'évaluation.*

Ces indications concernant le modèle d'évaluation vont permettre de fournir l'aide à la décision attendue. Avant cela, une précision méthodologique doit être apportée pour l'utilisation de la procédure d'agrégation proposée.

## 1.2. Considération méthodologique pour l'utilisation du modèle de prévision dans le cadre de la procédure d'agrégation du modèle d'évaluation

Le modèle de prévision MES (Modèle 4) est utilisé dans le cadre de la procédure d'agrégation de l'aide à la décision, pour fournir les estimations  $Y_0$  et  $Y_b$ , utilisées pour le calcul de  $F_b$  (selon  $j$  scénarios différents) Ces estimations sont fournies par un modèle linéaire classique, de forme générale :

$$Y_t = \sum_1^j \alpha_i X_{i,t} + \varepsilon_t$$

$\alpha_i$  sont les paramètres du modèle ;  $\varepsilon_t$  son terme d'erreur, réalisation d'un vecteur aléatoire de loi  $N(0, \sigma^2)$ . A partir de ce modèle, les managers de l'infrastructure TI-C2 peuvent calculer  $F_b$ , donné par le rapport  $\frac{Y_b}{Y_0}$  (Modèle 7). Ce calcul est basé sur le modèle linéaire, pour différentes valeurs de  $X_i$  en  $t = 0$  et  $t = b$ .

En conséquence, si l'on suppose  $\alpha_0$  et  $\alpha_i$  connus,  $F_b$  est égal en loi au ratio de deux lois normales : c'est une loi de type Cauchy (Cedilnik et al., 2004 ; Qiao et al., 2006, pour des détails au sujet de la loi de Cauchy). Dans le cadre de nos recherches, nous n'irons pas jusqu'à en évaluer les paramètres théoriques : les réalisations de  $F_b$ , pour différentes valeurs de  $X_i$ , seront analysées par simulation.

<sup>59</sup> On pourrait chercher à intégrer l'incertitude liée aux paramètres  $\alpha_i$ , mais cette possibilité ne sera pas développée.

## Section 2. Utilisation du modèle d'évaluation : résultat et recommandation pour l'aide à la décision.

### 2.1. Simulation et recommandation

Le modèle d'évaluation peut être utilisé pour produire un résultat ( $\Phi$ ). Pour cela, la procédure de simulation ci-après est appliquée :

- En  $t = 0$ , l'activité industrielle  $X_{i,0}$  est décrite avec les experts, à partir des données disponibles, telle que  $X_{i,0} = \{50000, 2000, 35\}$  ;
- Pour  $t = h$ , les valeurs  $X'_{i,h}$  sont évaluées pour chacun des scénarios, à partir de  $X_{i,0}$  et du Tableau 18. Ces valeurs numériques sont présentées dans le Tableau 19 ;

$j$	$X'_{1,h}$	$X'_{2,h}$	$X'_{3,h}$
O	50000	4500	65
R	112500	4500	65
P	112500	5000	35
I	112500	4500	35

Tableau 19. Valeurs numériques utilisées pour la simulation.

Les densités de probabilité des  $F_b$  sont évaluées par simulation, en répétant  $N = 10^6$  fois la procédure d'agrégation donnée par le Modèle 7. Le script R utilisé pour réaliser cette simulation est donné en annexe 4. Les résultats de cette simulation sont présentés en Figure 69.

Cette figure, dite « boîte à moustaches » ou boîte de Tukey, permet de résumer les densités de probabilité des simulations de  $F_b$ , pour chacun des  $j$  scénarios. Elle s'interprète de la façon suivante<sup>60</sup> :

- Les parties inférieures et supérieures des rectangles centraux, indiquent les valeurs des premiers et troisièmes quartiles des résultats des simulations ;
- Ces rectangles sont divisés par des traits épais, qui représentent les valeurs médianes des résultats des simulations ;
- Les traits fins supérieurs et inférieurs, en dehors des rectangles (les « moustaches »), correspondent à  $\pm 1,5$  fois les écarts interquartiles des densités de probabilités de chacune des simulations ;
- Les points hors des moustaches, peuvent être considérés comme des valeurs extrêmes et atypiques des résultats des simulations.

A titre de comparaison, l'estimation du facteur d'accroissement prévisionnel fourni par les experts en 2009 est également indiquée sur le graphique.

---

<sup>60</sup> Cf. Correa et González (2002).

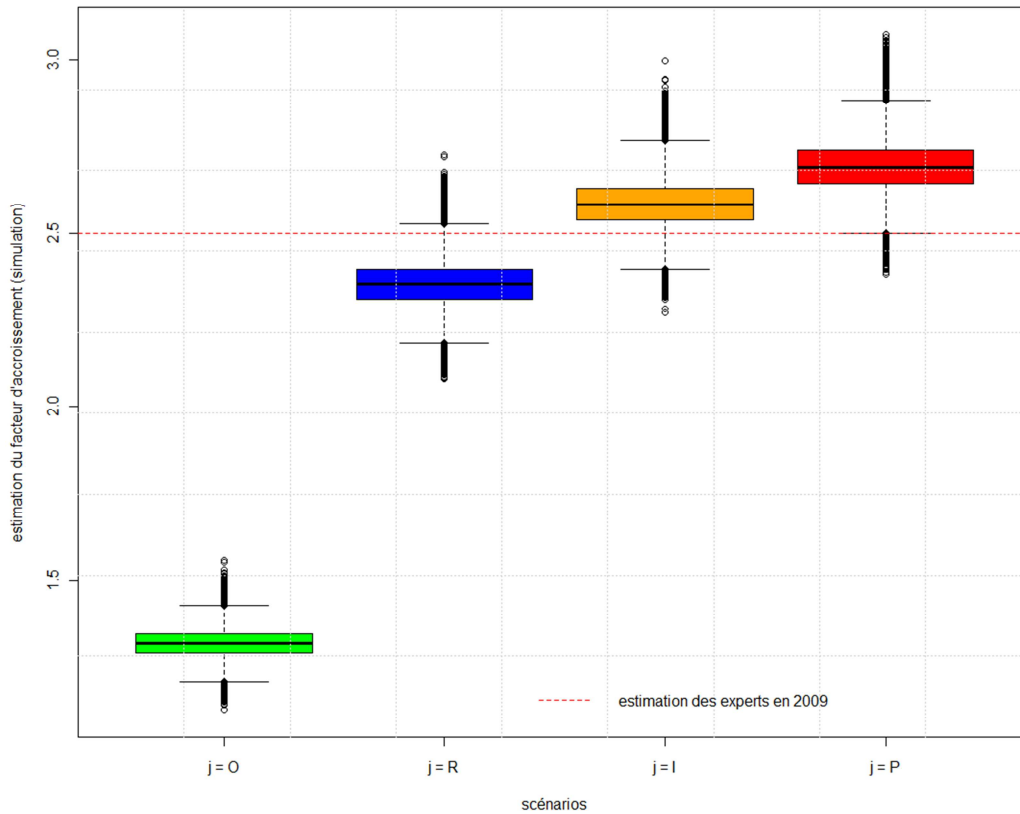


Figure 69. Densités de probabilité des résultats des simulations des  $F_b$  : boîte à moustaches ( $\Phi$ ).

Sur la base de ces résultats, la recommandation qui aurait pu être fournie en 2009, si ce modèle d'évaluation avait été utilisé, aurait été la suivante :

*Si l'on se place dans le cadre d'un scénario industriel réaliste, et en considérant les premier et troisième quartiles de la simulation,  $F_b$  devrait être compris dans l'intervalle  $[2,31 ; 2,40]$ .*

Cette recommandation peut être discutée au regard de l'estimation fournie par les experts lors de la prise de décision de 2009.

## 2.2. Discussion

Les scénarios O et P sont trop éloignés du scénario implicite I utilisé lors de la décision de 2009. Nous ne les discutons donc pas plus en détail. Ils permettent toutefois de montrer tout l'intérêt du modèle d'évaluation proposé, qui permet de caractériser des scénarios industriels selon plusieurs dimensions et d'aboutir à des résultats bien différents en fonction des scénarios. Il est ainsi possible de tester des combinaisons d'hypothèses variées, pour analyser le comportement du système et déterminer  $F_b$  sur la base d'une étude complète.

La recommandation basée sur le scénario R, fournie à partir du modèle d'évaluation, suggère une valeur de  $F_b$ , comprise dans l'intervalle  $[2,31 ; 2,40]$ . On peut donc dire que les experts, en proposant  $F_b = 2,5$  en 2009, surestimèrent légèrement le facteur prévisionnel d'accroissement. Néanmoins, la « moustache » supérieure de la boîte de

Tukey est presque égale (2,53) à la valeur fournie par l'expertise et, en ce sens, on pourrait penser que cette intuition fut bonne. Choisir 2,5, soit légèrement plus que la valeur médiane de la simulation (2,35), permettait également de limiter les risques de sous-capacités des futurs serveurs. En conséquence, l'utilisation de cette simulation aurait été un bon support pour appuyer leurs propos, dont le caractère informel fut critiqué alors (voir partie 2).

Cependant, ce résultat peut être modéré. En 2009, c'est en fait le scénario implicite I que les experts cherchèrent à estimer. Selon la simulation, les premier et troisième quartiles de  $F_b$  sont alors compris dans l'intervalle [2,54 ; 2,63], avec des valeurs pouvant s'élever jusqu'à 2,77 pour la moustache supérieure de la boîte de Tukey. Cela veut donc dire que, dans le cadre du scénario I, les experts sous-estimèrent plutôt le facteur d'accroissement.

Ainsi, l'anticipation de la hausse du besoin en ressource TI fut bonne, mais aurait pu être mieux maîtrisée. En effet, les experts fournirent une bonne prévision pour le scénario réaliste R, alors qu'ils pensaient anticiper le scénario implicite I. Si les managers de l'infrastructure TI-C2 avaient eu dès 2009 les résultats de ces simulations, peut-être auraient-ils orientée différemment leur décision d'anticipation de la hausse du besoin en capacité informatique. Effectivement, le scénario réaliste implique le succès du déploiement du projet MFA, ce qui n'est pas le cas du scénario implicite. En conséquence, si ce projet échoue, on pourrait aboutir à un risque de sous-capacité des nouveaux serveurs, puisque ce paramètre ne fut pas pris en compte dans la décision de 2009. On comprend tout l'intérêt de disposer d'un modèle quantitatif d'aide à la décision, qui permet de formaliser et de mettre à l'épreuve les propositions intuitives des experts.

## **Conclusion du chapitre 7**

Ce chapitre 7 marque la fin de l'application du processus socio-statistique d'aide à la décision. A partir d'un modèle statistique permettant de modéliser un système complexe, en tenant compte de son environnement métier, nous avons construit un modèle d'évaluation et proposé une recommandation décisionnelle. Nous avons montré que les managers de l'infrastructure TI-C2 auraient pu en tirer un bénéfice non négligeable, s'il avait eu un tel modèle d'évaluation à leur disposition, lors de la migration MES en 2009.

La recommandation apportée est très locale. Elle fournit en effet une réponse décisionnelle pour un problème exclusivement rencontré chez STMicroelectronics : la détermination d'un facteur d'accroissement, modélisé par un ratio de lois normales. Seule la phase contextualisation du processus socio-statistique d'aide à la détermination aura permis d'aboutir à une aide à la décision aussi adaptée aux besoins décisionnels des managers de l'infrastructure TI-C2.

Evidemment, l'utilisation de ce modèle repose sur une hypothèse de stabilité du système modélisé, c'est-à-dire de constance dans le temps du modèle statistique sous-jacent à l'aide à la décision. Il convient donc de mettre en place certains mécanismes permettant de s'assurer de cette stabilité et, si besoin, d'aider les managers de l'infrastructure TI-C2 à faire évoluer la connaissance formalisée dans le modèle : c'est l'objet du chapitre suivant.



## **Chapitre 8. Evolution de la connaissance organisationnelle.**

## Introduction

Le chapitre 6 a montré comment des relations de stabilité peuvent être identifiées dans les bases de données, puis modélisées ; le chapitre 7 comment cette modélisation peut être exploitée pour la prise de décision. Nous montrons désormais comment ces étapes stabilisatrices peuvent être combinées à une capacité à réagir à l'inattendu, selon les recommandations fournies par Morin (voir la première partie) pour l'appréhension d'un système complexe. Pour cela, le plan suivant sera suivi (Figure 70).

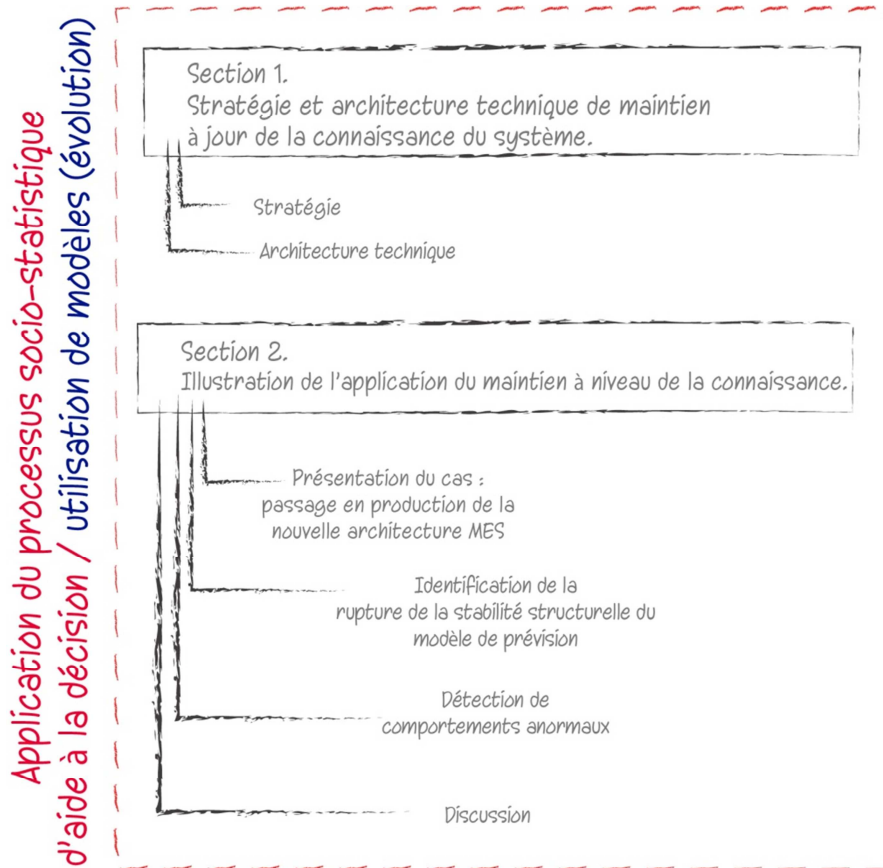


Figure 70. Plan du chapitre 8.

Dans un premier temps (section 1), l'ensemble des moyens mis en œuvre pour assurer cette capacité de réaction à l'inattendu est décrite : la mobilisation des méthodes quantitatives présentées précédemment, mais aussi l'architecture technique mise en place chez STMicroelectronics pour en permettre l'implémentation industrielle.

Dans un second temps (section 2), une illustration complète des résultats pouvant être apportés par ce maintien à niveau de la connaissance est fournie. Pour cela, le passage en production de l'architecture MES, déjà évoquée dans cette thèse, sera étudié. Après un rappel du cas, le comportement des mécanismes d'identification de la rupture de la stabilité structurelle et de détection des comportements anormaux sont présentés, puis discutés.

## **Section 1. Stratégie et architecture technique de maintien à jour de la connaissance du système.**

### **1.1. Stratégie**

Pour réagir l'inattendu, on peut tirer profit de la complémentarité entre certaines méthodes quantitatives développées dans le cadre des travaux de recherches présentés :

- *Le contrôle des modèles de prévision, dans le cadre de l'activité ITIL (Q) Modélisation et analyse des tendances* : les méthodes dynamiques de test de la stabilité structurelle, proposée par Zeileis et déjà présentée dans cette thèse peuvent être mise à contribution pour détecter tout changement de comportement du système, donc pour identifier une dérive de la réalité par rapport à la connaissance modélisée ;
- *L'activité ITIL (K) Contrôle* : en complément, on peut s'appuyer sur le besoin décisionnel de détection de comportements anormaux et sur les méthodes quantitatives développées en conséquences, telles que présentées dans cette thèse. L'identification de comportements inattendus de certaines variables permet effectivement de mettre en exergue des situations nouvelles, permettant d'enrichir ou de mettre à niveau les modèles prospectifs.

Une bonne stratégie de maintien à niveau de la connaissance peut donc s'appuyer sur la combinaison de ces deux approches. Par ailleurs, elle prend tout son sens lorsqu'elle est accomplie dynamiquement et automatiquement, c'est-à-dire lorsque les contrôles quantitatifs définis sont réalisés sans intervention humaine, et que les experts concernés sont systématiquement notifiés en cas de dérive potentielle du système, pour investigations complémentaires. Une telle approche a été implémentée à Crolles 2, grâce à l'architecture technique suivante.

### **1.2. Architecture technique**

*Remarque : nous adressons nos sincères remerciements à Pascal Palazon, dont l'expertise informatique fut une aide essentielle à l'implémentation de l'architecture technique présentée ci-dessous.*

Une architecture automatisant le contrôle du modèle de prévision et la détection de comportements anormaux est mise en œuvre chez STMicroelectronics. Elle permet de participer au maintien de la connaissance, de façon automatique et en « temps réel ». Les managers de l'infrastructure TI-C2 sont avertis de toute dérive de comportement du système, grâce aux méthodes statistiques déjà présentées. Pour ce faire, l'architecture technique illustrée par la Figure 71 a été définie et implémentée.

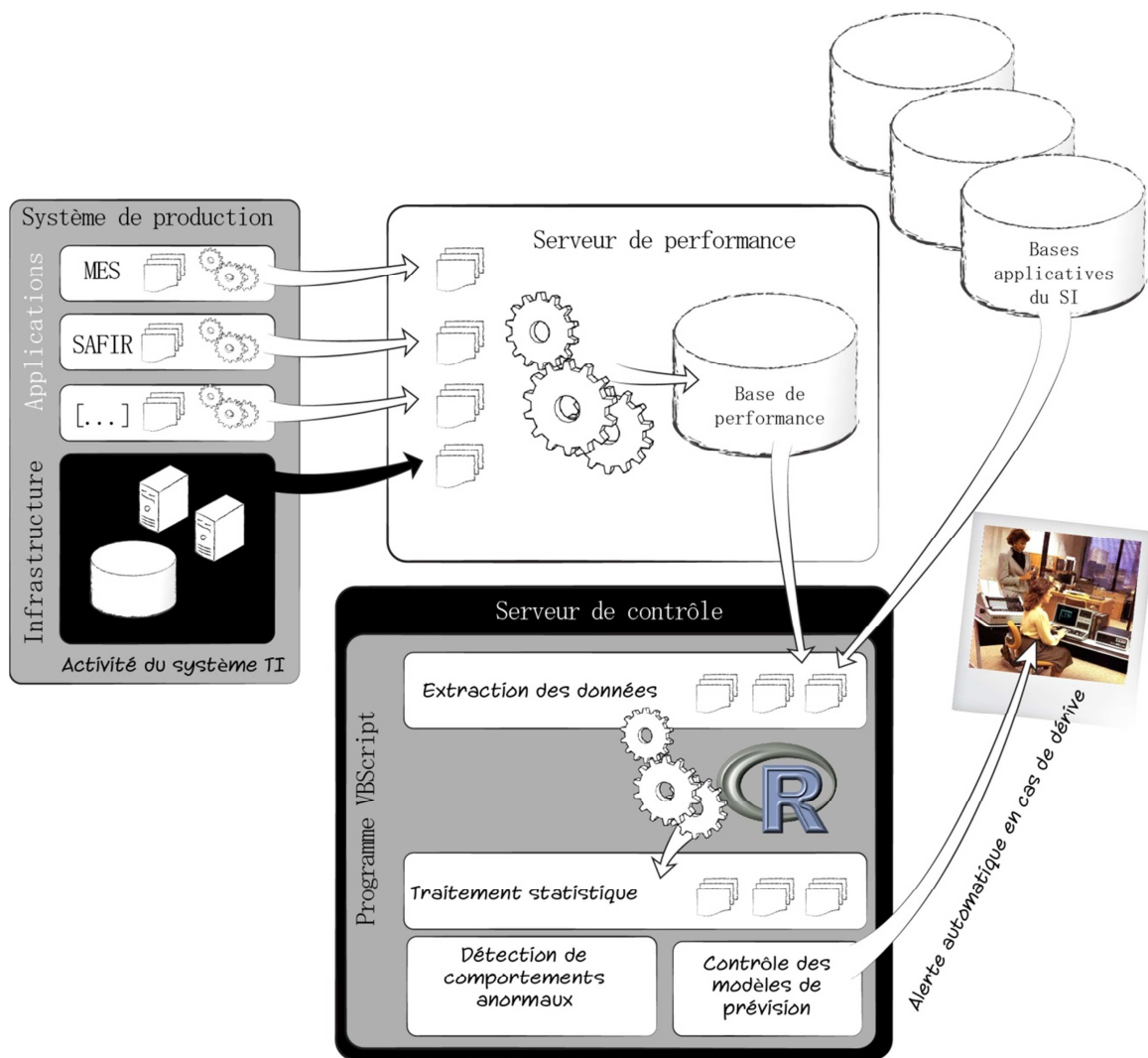


Figure 71. Architecture technique de maintien à jour de la connaissance.

Cette architecture s'appuie sur les données de la base de performance, mais aussi sur celles des bases applicatives du SI. Les données devant faire l'objet d'une mise sous contrôle sont identifiées lors de l'application de la démarche de modélisation, ou par des groupes de travail complémentaires effectués avec les experts de STMicroelectronics. Pour exploiter ces données, un programme VBScript a été réalisé. Il assure principalement les fonctions suivantes :

- *Extraction des données devant être mises sous contrôle* : les variables analysées sont définies par plusieurs fichiers de configuration, définissant les requêtes SQL appropriées ;
- *Analyse statistique des données et récupération des résultats* : le programme R est appelé, associé à des scripts *ad hoc*, pour opérer les traitements de contrôle des modèles de prévision et de détection de comportements anormaux ;
- *Avertissement des utilisateurs concernés par courriel* : les experts sont automatiquement notifiés de toute dérive du système, suggérée par les traitements quantitatifs.

Le code du programme VBScript et les scripts R de traitement statistique ne peuvent être diffusés dans cette thèse, pour des raisons de confidentialité. Dans le cadre de notre

terrain de recherche, ce programme est automatiquement appelé quotidiennement, afin de réaliser l'ensemble des contrôles du système attendus. Il contribue ainsi de façon continue à la génétique du système de connaissances organisationnelles.

Nous illustrons ci-dessous les bénéfices que l'on peut tirer de cette stratégie et architecture de maintien à niveau de la connaissance.

## **Section 2. Illustration de l'application du maintien à niveau de la connaissance.**

### **2.1. Présentation du cas : passage en production de la nouvelle architecture MES**

Plaçons-nous dans la situation d'un observateur qui n'aurait pas été au fait du passage en production de la nouvelle architecture MES discutée dans cette thèse. Nous montrons que les mécanismes de contrôle mis en place lui auraient permis :

- *De détecter un changement de comportement du système*, grâce au contrôle du modèle de prévision (activité ITIL (Q) Modélisation et analyse des tendances) ;
- *De récolter un faisceau d'indices complémentaires*, qui lui auraient permis d'inférer les causes de ce changement (activité ITIL (K) Contrôle).

Tous ces éléments auraient sans doute incité l'observateur à réviser la connaissance formalisée au travers du modèle de prévision développé dans le Chapitre 6.

### **2.2. Identification de la rupture de la stabilité structurelle du modèle de prévision**

Le modèle de prévision journalier (Modèle 3) est considéré. On utilise la carte de contrôle déjà appliquée à ce modèle en période de stabilité (Figure 55). Elle est cette fois-ci initialisée sur l'ensemble de l'échantillon structurellement stable, c'est-à-dire les 170 observations utilisées pour l'estimation du Modèle 3. Les sept observations de la semaine qui suit sont contrôlées à partir de cette carte de contrôle. On obtient les résultats présentés dans la Figure 72 (script R dans l'annexe informatique n°5).

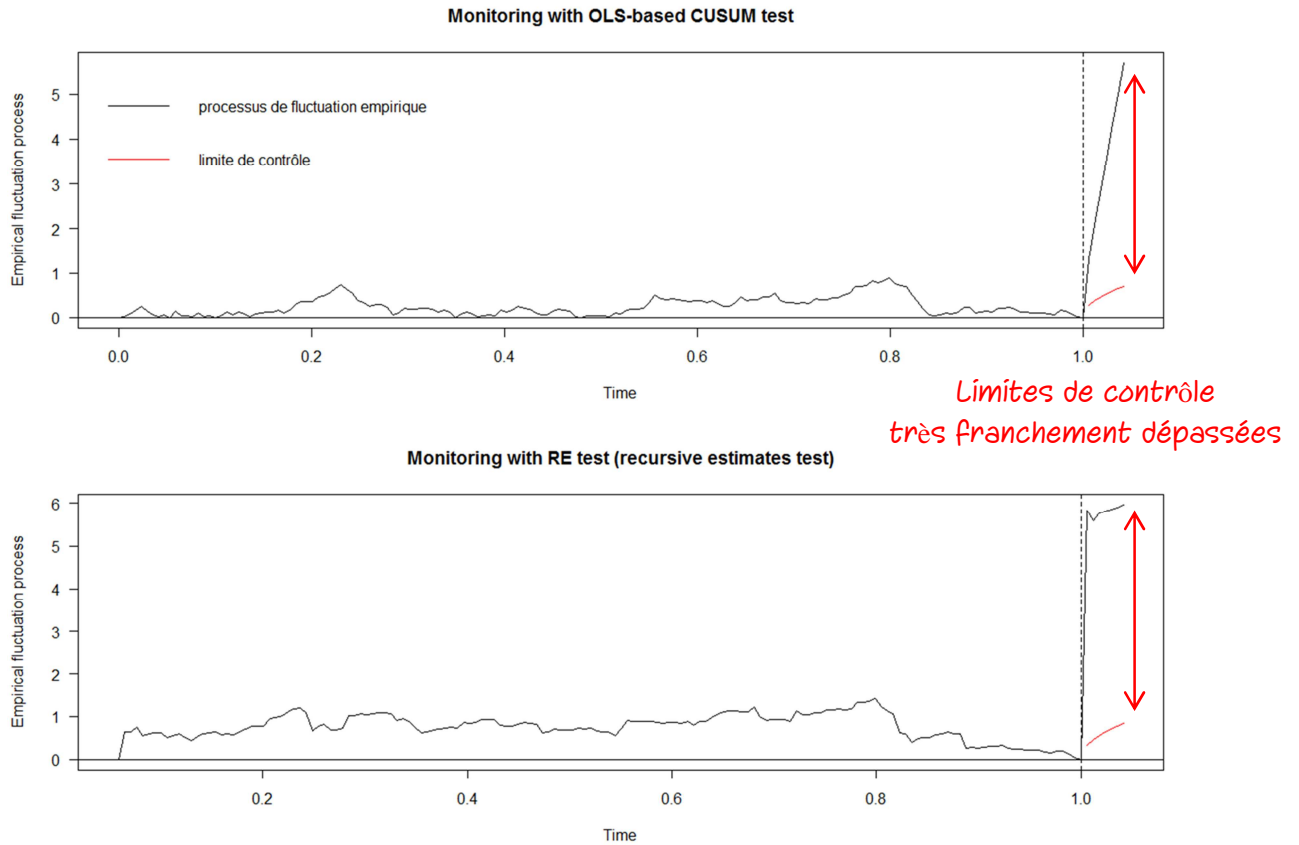


Figure 72. Carte de contrôle de la stabilité du modèle de prévision pour la semaine consécutive à la migration MES.

On constate une nette rupture de la stabilité du modèle. Dès la première nouvelle observation, la limite de contrôle est franchement dépassée. Le modèle de prévision MES ne peut plus être utilisé tel qu'il est connu. La détection de comportements anormaux permet de mieux comprendre les causes de cette dérive.

### 2.3. Détection de comportements anormaux

On présente ici le contrôle d'un certain nombre de variables représentatives de l'activité du système MES (ce cas, donné à titre d'illustration, est une simplification de la véritable stratégie de contrôle en œuvre chez STMicroelectronics, qui porte sur plusieurs centaines de variables). Le groupe contrôlé comprend les variables présentées dans le Tableau 20.

Variable	Description
$Y$	Nombre total de secondes CPU consommées par jour
$S$	Temps de réponse applicatif moyen journalier <i>300Works</i> pour la préparation de lots
$X_2$	Nombre moyen journalier de <i>moves</i>
$X_3$	Déploiement du projet MFA
$W$	Nombre moyen journalier de plaques produites ( <i>wafer-out</i> )

Tableau 20. Groupe de variables mises sous contrôle de détection des comportements anormaux.

Le contrôle est effectué sur une période de 72 observations consécutives, la dernière correspondant au premier jour de rupture de la stabilité structurelle détectée en Figure 72 (jour 171, D171). C'est les résultats du contrôle à cette date qui nous intéressent. La Figure 73 présente la carte de contrôle de la statistique globale du  $\chi^2$  obtenue<sup>61</sup>. Elle n'a pas de signification physique, mais permet de mesurer l'importance du comportement anormal, s'il y en a un.

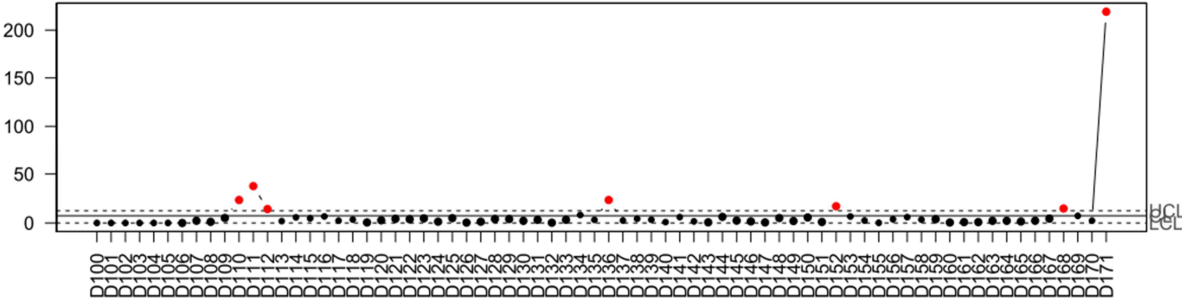


Figure 73. Carte de contrôle basée sur la statistique du  $\chi^2$ , pour le groupe de variables observées.

On voit que la dernière valeur est très grande. On peut clairement supposer un comportement anormal au sein du groupe de variables observé. Afin de savoir quelle variable peut être problématique, on s'intéresse alors au graphique des contributions partielles (Figure 74).

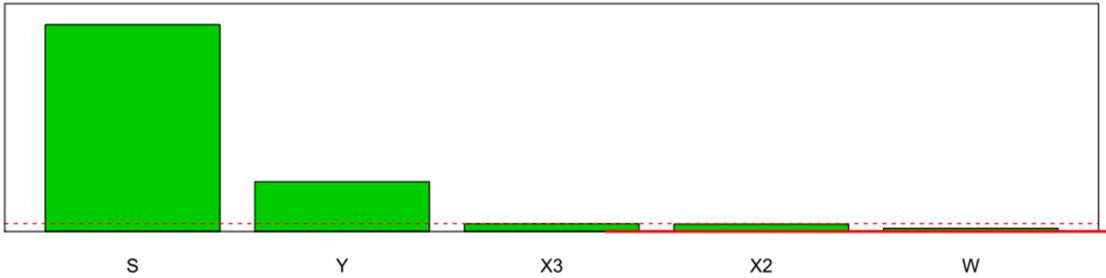


Figure 74. Graphique des contributions partielles.

Deux valeurs semblent aberrantes : au jour 171,  $S$  et  $Y$  dépassent nettement la limite de contrôle imposée par le groupe (pointillé rouge). Il est probable que ces valeurs soient hors-contrôle. Pour compléter cette analyse, des cartes de contrôle univariées sont également générées pour chacune des variables (Figure 75). Les traits pointillés verts correspondent à la modélisation Holt-Winters robuste de la série temporelle tracée en noir et les bandes vertes représentent les limites de contrôle. On remarque que dans ce cas, les alarmes données par le contrôle multivarié sont bien des alarmes dans le cas univarié.

<sup>61</sup> Les scripts R relatifs à la détection de comportements anormaux ne peuvent être fournis dans cette thèse, pour des raisons de confidentialité.



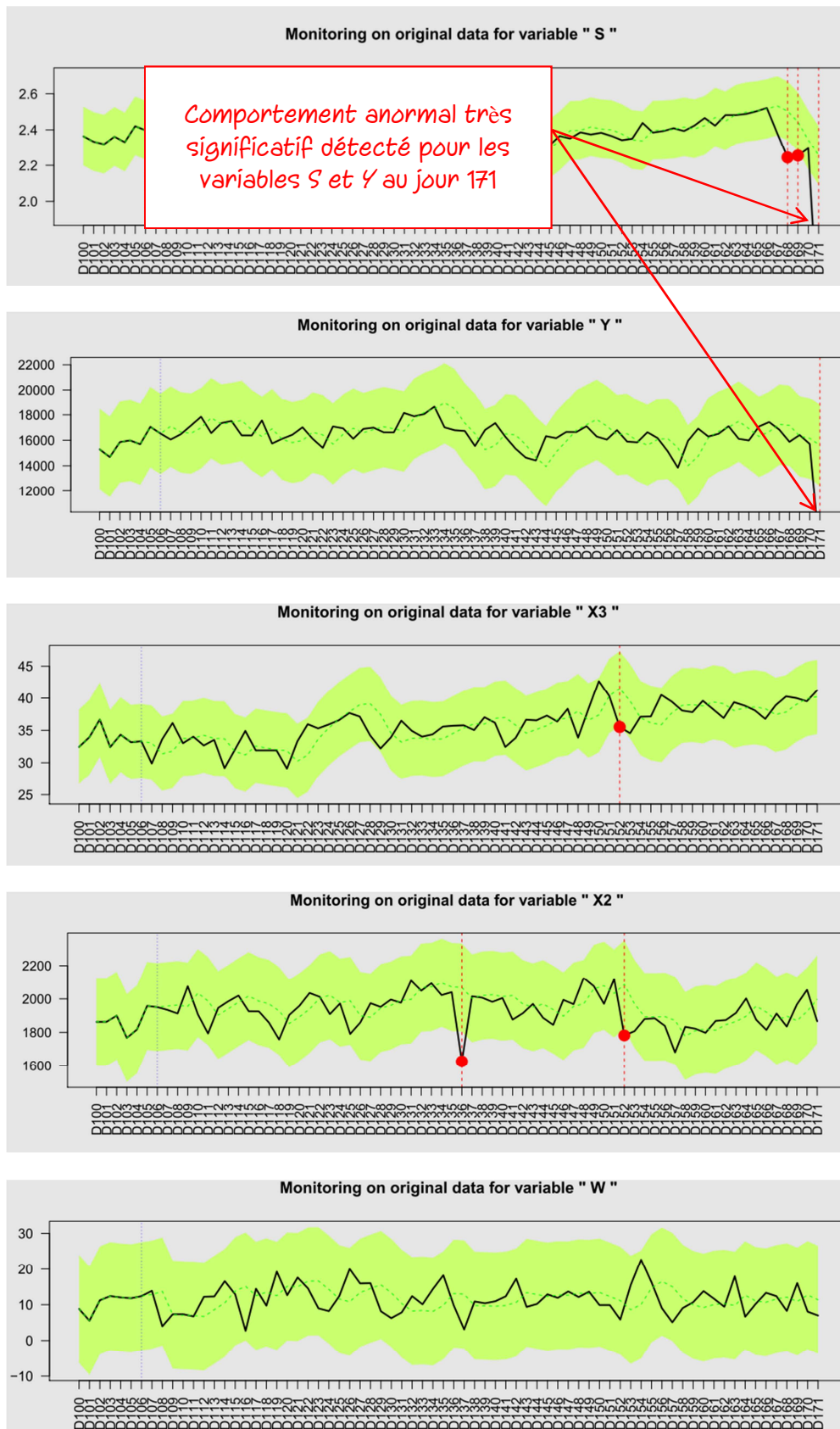


Figure 75. Cartes de contrôle univariées.

Les conclusions à tirer de ces contrôles peuvent être discutées.

## 2.4. Discussion

Les résultats présentés ci-dessus montrent que d'une part, il n'y a pas de perturbation en ce qui concerne l'activité métier : à la vue des cartes de contrôle présentées,  $X_2$ ,  $X_3$  et  $W$  ne présentent pas de valeurs aberrantes au jour 171. D'autre part, la consommation CPU ( $Y$ ) et le temps de réponse applicatif ( $J$ ) diminuent anormalement. Ces valeurs indiquent donc qu'à activité métier équivalente, la consommation CPU et la performance sont meilleures.

On peut donc penser que ces résultats sont la conséquence d'une modification au niveau de l'architecture des TI. Leur observateur pourra alors donc mener une enquête, et être rapidement mis au fait de la migration MES. Ces observations s'expliquent donc : baisse de consommation CPU et amélioration de la performance, correspondent simplement aux bénéfiques pouvant être attendus d'une telle migration de technique. La rupture de la stabilité structurelle du Modèle 3 est alors toute naturelle, la variable endogène n'étant plus la même (nouveau serveur). Le modèle de prévision devra donc être ré-estimé.

Ainsi, on peut dire que la conjonction des deux types de contrôle aura permis de comprendre :

- *Que le modèle d'aide à la décision en usage n'est plus valide* : il faudra le ré-estimer pour une prochaine prise de décision s'appuyant sur ce modèle ;
- *Que la migration de l'architecture technologique a engendré un certain nombre d'améliorations du comportement du système*, à savoir la baisse de la consommation CPU et la hausse de la performance applicative. Ce phénomène peut désormais être étudié plus en détail si nécessaire.

Les connaissances concernant le comportement du système considéré vont donc pouvoir évoluer en conséquence.

## Conclusion du chapitre 8

Cette application aura permis de montrer en quoi les méthodes de contrôle dynamique implémentées chez STMicroelectronics permettent d'apporter des réponses aux limites des approches usuelles de *capacity planning*, discutées dans cette thèse. En se basant sur le contrôle de modèles de prévision locaux, orientés métier, et en effectuant une sélection pertinentes des variables devant être mises sous contrôle de comportements anormaux, l'ensemble de l'activité des niveaux composant, ressource et métier peut être suivi au jour le jour.

Ce suivi est réalisé par l'emploi de méthodes statistiques, implémentées dans le quotidien de l'activité industrielle de STMicroelectronics. Elles permettent d'automatiser le contrôle du système, sans hypothèse *a priori* sur son comportement. C'est pourquoi nous pensons avoir proposé des méthodes de gestion de l'infrastructure des technologies de l'information, adaptées aux enjeux de gestion contemporains.

**Quatrième partie. Perspectives et conclusion générale.**

## **Plan détaillé de la quatrième partie.**

### **Section 1. Perspectives.**

#### 1.1. Perspectives méthodologiques.

*1.1.1. Méthodes quantitatives.*

*1.1.2. Exploitation quantitative de données non-structurées.*

#### 1.2. Perspectives décisionnelles.

*1.2.1. Réutilisation des connaissances et des méthodes de formalisation des connaissances.*

*1.2.2. Réutilisation et méthodes de décision.*

#### 1.3. Vers un système de support à la décision.

*1.3.1. Stockage des connaissances et des méthodes de formalisation des connaissances.*

*1.3.2. Les fonctions du système de support à la décision.*

#### 1.4. Travaux de recherche et *Big Data*.

*1.4.1. Présentation du concept.*

*1.4.2. Positionnement des travaux de recherches par rapport au *Big Data*.*

### **Section 2. Conclusion générale.**

\*  
\* \*

*En développant et appliquant un processus socio-statistique d'aide à la décision, les travaux présentés dans cette thèse auront permis de proposer :*

- *Des méthodes de formalisation de connaissances organisationnelles, pour constituer des artefacts utiles à l'aide à la décision. Ces apports constituent une contribution à la création de connaissances organisationnelles (ontologie du système de connaissances) ;*
- *Des applications concrètes de ces méthodes, ayant abouti à une utilisation opérationnelle de ces connaissances (phénoménologie), dans le cadre d'un modèle d'évaluation permettant de déterminer le facteur d'accroissement de la charge d'un serveur informatique ;*
- *Des méthodes permettant de maintenir à niveau la connaissance, en permettant le contrôle des modèles de prévision et la détection de comportements anormaux (génétique).*

*Dans la section 1 de cette partie, nous proposons ci-après quelques axes de développement de ces travaux. Tout d'abord, deux perspectives méthodologique et décisionnelle, sont envisagées. Nous montrons ensuite comment elles pourraient être intégrées dans le cadre d'un système de support à la décision. Pour finir, ces propositions nous amènent à considérer nos travaux dans le cadre d'un « buzzword » actuellement en vogue : le Big Data.*

*Ces considérations nous conduiront finalement à la conclusion générale de nos travaux de thèse, en section 2.*

\*  
\* \*

## Section 1. Perspectives.

### 1.1. Perspective méthodologiques

#### 1.1.1. Méthodes quantitatives

L'ensemble des méthodes quantitatives employées pourraient faire l'objet d'améliorations. Notamment, deux collaborations scientifiques ont été initiées à l'heure où cette thèse est rédigée :

- *Un projet de recherche est en cours de réflexion pour combiner les étapes d'exploration et d'explication/de prévision de la démarche de modélisation (création de connaissances).* Pour ce faire, la première étape de ce projet consiste à expérimenter et comparer des méthodes de régression, favorisant la sélection de variables en grandes dimensions ; Hastie et *al.* (2001) identifient les trois familles de méthodes suivantes :
  - \* *Subset selection* : parmi  $k$  prédicteurs potentiels, ces méthodes visent à sélectionner un sous-ensemble permettant de minimiser les erreurs du modèle (exemples de méthodes : *leaps and bounds*, *forward/backward stepwise selection*),
  - \* *Shrinkage* : dans ce cas, tous les  $k$  prédicteurs sont conservés, mais des pénalités sont appliquées aux coefficients de la régression (exemples : régression *ridge* ou *lasso*),
  - \* *Derived input selection* : cette famille est proche des méthodes déjà employées. Elle se base sur des combinaisons linéaires des  $k$  prédicteurs (régression à partir des composantes d'une ACP, régression PLS) ;
- *Un projet de recherche est en cours de développement pour améliorer la procédure de détection des comportements anormaux (évolution des connaissances).* Ce projet est actuellement essentiellement centré sur l'amélioration de l'intégration des changements structurels. Pour l'instant, la procédure en place s'appuie sur un test de Chow. Or, en présentant les travaux de Zeileis, nous avons montré que nombreuses autres possibilités de tests sont envisageables. Il convient donc de mener une réflexion plus approfondie sur ce sujet.

Par ailleurs, ces notions de changement structurel et d'évolution des connaissances ouvrent une autre question intéressante, qui pourrait faire l'objet de recherches ultérieures. Pour l'instant, que l'on soit en présence du contrôle des modèles de prévision ou de la détection de comportements anormaux, nous avons pris le parti de travailler en deux étapes : d'abord, la détection d'un changement structurel ; ensuite, la ré-estimation des paramètres des modèles. Une approche différente pourrait être explorée : elle intégrerait non pas la ré-estimation des paramètres, mais leur évolutivité. Concernant les modèles de prévision, des méthodes dynamiques d'estimation des paramètres, ajustant automatiquement les paramètres des modèles de régression pourraient être envisageables (pour une introduction à ces notions : Young, 2011). Au sujet de la détection de comportements anormaux, l'évolutivité dans le temps des

algorithmes de lissage exponentiel a été abordée par certains chercheurs (Williams, 1987 ; Taylor, 2004).

### 1.1.2. *Exploitation quantitative de données non-structurées*

Tout d’abord, pour la création de connaissances relatives à la contextualisation de l’aide à la décision, nous avons procédé à des exploitations entièrement manuelles des *verbatim* d’entretien. On pourrait imaginer aider l’analyste/le chercheur à constituer ces artefacts, par le recours à des méthodes statistiques d’exploitation du matériau récolté lors des enquêtes de type ethnographique. Lejeune (2010) suggère des procédés d’automatisation de l’activité de catégorisation, grâce à des méthodes de statistiques textuelles calculant les distributions des mots dans un corpus de textes. Par ailleurs, des procédés dits de *text mining* pourraient aussi être mobilisés. En effet, Feldman et Sanger (2006) en proposent la définition suivante :

*« text mining seeks to extract useful information from data sources [...] the data sources are document collections, and interesting patterns are found not among formalized database records but in the unstructured textual data in the documents in these collections. »*

Ces méthodes pourraient s’appliquer à l’analyse d’un ensemble de *verbatim* d’entretien. Cette proposition est toutefois relativement triviale dans le cadre de nos recherches, qui considèrent cet aspect du processus socio-statistique d’aide à la décision comme un prérequis, mais pas comme un élément central devant faire l’objet d’attentions particulières.

Un autre point concernant la construction de connaissances à partir de données non-structurées nous semble quant à lui plus important. Effectivement, l’ensemble des données quantitatives utilisées pour l’instant sont issues de bases de données structurées. Ces données nécessitent généralement un premier traitement pour être récoltées : comme nous l’avons indiqué, la base de performance de STMicroelectronics est souvent constituée à partir d’opérations de *parsing* des fichiers de *log* transactionnels. S’affranchir de cette étape de traitement intermédiaire permettrait d’optimiser le mécanisme de constitution de connaissances, en intégrant directement ces fichiers automatiquement générés par les systèmes d’information dans la syntaxique du système de connaissances. Un courant de recherche pourrait fournir un certain nombre de fondements théoriques pour ce faire : le *process mining* (Van der Aalst et Weijters, 2004 ; Song et Van der Aalst, 2008 ; Rozinat et al., 2009), qui se définit comme :

*« a tool to extract non-trivial and useful information from process execution logs. These so-called event log (also called audit trails, or transaction logs) are the starting point for various discovery and analysis techniques that help to gain insight into certain characteristics of the process. »* (Rozinat et al., 2009)

Toutes ces perspectives méthodologiques, ainsi que leurs conditions d’application au sein de notre terrain de recherche, doivent désormais être étudiées plus en détail. Elles méritent d’être approfondies, car elles permettraient de faciliter les conditions d’un déploiement industriel à plus grande échelle des réflexions conduites dans les travaux de cette thèse, en favorisant des solutions permettant de tenir compte de façon plus

automatique (adaptation aux changements structurels) d'un grand volume potentiel de données (grande dimension) et en étendant le périmètre du type de données exploitées (*text mining, process mining...*).

## 1.2. Perspectives décisionnelles

### 1.2.1. Réutilisation des connaissances et des méthodes de formalisation des connaissances

Les travaux proposés visent à développer une connaissance utile à la prise de décision, mais aussi des méthodes de formalisation de ces connaissances. Nous pensons que ces deux produits peuvent faire l'objet de réutilisations futures, en s'inscrivant dans le cadre général du cycle de vie de la gestion des connaissances, qui intègre des étapes de génération, stockage, partage et réutilisation (Reix et *al.*, 2011 ; Moalla et *al.*, 2012).

D'une part, les connaissances produites (artefacts du cadre de Tsoukiàs et modèles statistiques) peuvent être formellement documentées et stockées, pour des utilisations ultérieures.

D'autre part, les méthodes développées lors des applications du processus socio-statistique d'aide à la décision et destinées à créer des connaissances, pourraient être considérées comme génériques dans bon nombre de cas. Alors, ces démarches permettraient de créer ultérieurement des connaissances nouvelles :

- *La méthode de constitution d'étude de cas est générique et peut être utilisée dans tout cas de figure, pour préciser un contexte organisationnel et les besoins décisionnels qu'il implique ;*
- *Les méthodes quantitatives développées à partir d'un besoin décisionnel clairement identifié et formalisé, pourraient être réutilisées si ce besoin décisionnel est un jour identifié à nouveau. C'est ce que nous avons proposé de démontrer en parlant de la généricité de la démarche de modélisation constituée pour le besoin de décision de modélisation prospective (partie 3).*

Ainsi, ces méthodes elles-mêmes peuvent être formellement documentées et stockées, pour des utilisations ultérieures.

### 1.2.2. Réutilisation et méthodes de décision

Nous proposons d'envisager la réutilisation de connaissances organisationnelles et de leurs méthodes de formalisation selon la perspective suivante. Avant tout, il convient de distinguer différentes situations de décisions pouvant être rencontrés dans une



organisation. Une classification fréquente, selon les méthodes de décision, a été proposée par Gory et Scott Morton<sup>62</sup> (Le Moigne, 1974 ; Davenport, 2010, 2012) :

- *Les décisions programmables* : l'ensemble du raisonnement d'identification et de résolution d'un problème de décision peut être automatisé par un traitement informatique ;
- *Les décisions non-programmables structurées* : une fois que le problème a été posé, certaines structures génériques de raisonnement peuvent être utilisées pour aider le décideur dans la résolution de son problème ;
- *Les décisions non-programmables faiblement structurées* : dans ce cas, il est difficile de développer une aide formelle pour la prise de décision.

Nous pensons que l'application du processus socio-statistique d'aide à la décision tend à l'identification de décisions programmables et structurées au sein de l'organisation. En effet, la phase de contextualisation de l'aide à la décision permet de clairement identifier des espaces décisionnels, pouvant être complétés par des méthodes quantitatives. Nous suggérons alors que cette approche favorise la réutilisation des connaissances et de leurs méthodes de production :

- *Par exemple, dans le cas d'une décision programmable* : une connaissance formalisée pourrait très bien être utilisée pour prendre automatiquement un certain nombre de décisions de régulation du système. Supposons qu'un comportement anormal du système soit détecté. Certaines connaissances formalisées pourraient alors être automatiquement sollicités pour évaluer l'état du système informatique (par exemple, en se basant l'activité ITIL de gestion des seuils). Cette évaluation pourrait alors amener à un certain nombre de conclusions claires concernant le fonctionnement du système, qui pourraient engendrer automatiquement des décisions destinées à maintenir son intégrité (par exemple, *killer* un processus informatique, purger une base de données, etc.). On considère ici le point de vue phénoménologique du système de connaissances, c'est-à-dire d'utilisation d'une connaissance précédemment constituée ;
- *Autre exemple, concernant une décision cette fois-ci structurée* : Imaginons qu'un projet quelconque concernant l'infrastructure des TI soit initié. On peut constituer une étude de cas (contribution à l'ontologie du système de connaissances), selon la méthode présentée dans cette thèse, pour en identifier les problèmes de décisions :

---

<sup>62</sup> A ces « méthodes » de décision, les auteurs joignent également une typologie des « niveaux » de décision, constituée de trois types de décisions : régulation, pilotage et planification.

- \* *Il se peut alors que qu'un problème de décision déjà connu émerge* : dans ce cas, une connaissance déjà formalisée peut exister (par exemple, un modèle de prévision) et être directement exploitée (utilisation de connaissances),
- \* *Il se peut aussi qu'aucune connaissance à jour ne soit disponible* : c'est alors qu'une méthode de création de connaissance peut exister ; et elle pourra être mise à profit pour aider les décideurs,
- \* *Bien entendu, il est également possible que la constitution de l'étude de cas amène à considérer un problème décisionnel totalement nouveau* : une méthode et des connaissances entièrement nouvelles devront alors être produites, tel que développé dans les parties 2 et 3 de cette thèse.

Cette perspective décisionnelle pourrait être rendue opérationnelle par le développement d'un système de support d'aide à la décision.

### **1.3. Vers un système de support à la décision**

#### *1.3.1. Stockage des connaissances et des méthodes de formalisation des connaissances*

Comme l'indique Davenport (2010, 2012), décisions programmables et non-programmables impliquent des liens plus ou moins étroits avec l'information disponible. En intégrant les perspectives décisionnelles citées ci-dessus dans le cadre d'un système d'information, permettant d'exploiter au mieux les données disponibles concernant l'activité du système informatique dans son contexte industriel, un système d'aide à la décision pourrait être constitué pour aider les managers de l'infrastructure TI-C2. Un tel système d'aide à la décision devrait permettre le stockage des éléments suivants :

- *Une base de méthodes* : elle donnerait des trames méthodologiques pour la constitution des études de cas, et pour des traitements quantitatifs socio-statistiques (exemple : la démarche de modélisation présentée dans cette thèse). Ces méthodes proviennent de l'ensemble des étapes de constitution de connaissances des diverses applications du processus socio-statistique d'aide à la décision, pouvant être opérées dans l'organisation ;
- *Une base de connaissances* : elle formalise l'ensemble des connaissances produites par l'application du processus socio-statistique d'aide à la décision (problèmes de décision, modèles...), mais également d'autres connaissances pouvant être inférées par l'observation du système (seuils critiques d'utilisation de certaines ressources ou SLA, par exemple).

Ces éléments de stockage permettraient d'assurer un certain nombre de fonctions d'aide à la décision.

### 1.3.2. Les fonctions du système de support à la décision

Les éléments de stockage du système de support à la décision, matérialisés par les bases de méthodes et de connaissances pourraient être à la base d'un cercle vertueux d'alimentation mutuelle. Un certain nombre de fonctions, s'appuyant sur ces éléments de stockage et sur les données disponibles (base de performances, base de données métier), rendrait possible une création continue de connaissances et de méthodes :

- *Fonctions de création (ontologie du système de connaissances)* : les méthodes mises à dispositions permettraient de produire de nouvelles connaissances (exemple : application de la démarche de modélisation dans cette thèse), et les connaissances pourraient amener à produire de nouvelles méthodes (cas d'un nouveau problème décisionnel). La contribution à l'ontologie du système de connaissances serait ainsi continue. Toutes les perspectives méthodologiques évoquées plus haut pourraient être intégrées à ces fonctions (méthodes statistiques, *process mining*...) ;
- *Fonctions d'évolution (génétique du système de connaissances)* : les méthodes de maintien à niveau de la connaissance, telles que décrite dans cette thèse (contrôle du modèle prospectif et détection de comportements anormaux), seraient intégrées au système de support à la décision. Notons que cette perspective a déjà été partiellement implémentée industriellement au sein notre terrain de recherche (partie 3). On pourrait aussi envisager d'y inclure les approches évolutives évoquées précédemment dans les perspectives méthodologiques.
- *Fonction d'utilisation (phénoménologie du système de connaissances)* :
  - \* *Pour les décisions programmables* : un ensemble de connaissances formalisées, basées sur les données d'activité de l'infrastructure, pourraient automatiquement engendrer un certain nombre de décisions,
  - \* *Pour les décisions structurées* : face à un problème de décision bien identifiée, le système de support à la décision pourrait diriger le décideur vers une connaissance appropriée déjà existante ou vers une méthode socio-statistique adaptée.

L'automatisation de ces fonctions permettrait d'en assurer la généralisation à un plus grand volume de donnée, mais aussi à un niveau de granularité fin : un certain nombre de traitements (on pense principalement aux décisions programmables) pourraient être opérés en quasi-temps réel et contribuer ainsi activement au bon fonctionnement du système d'information. La Figure 76 illustre ce concept général de support d'aide à la décision, basé le processus socio-statistique d'aide à la décision.

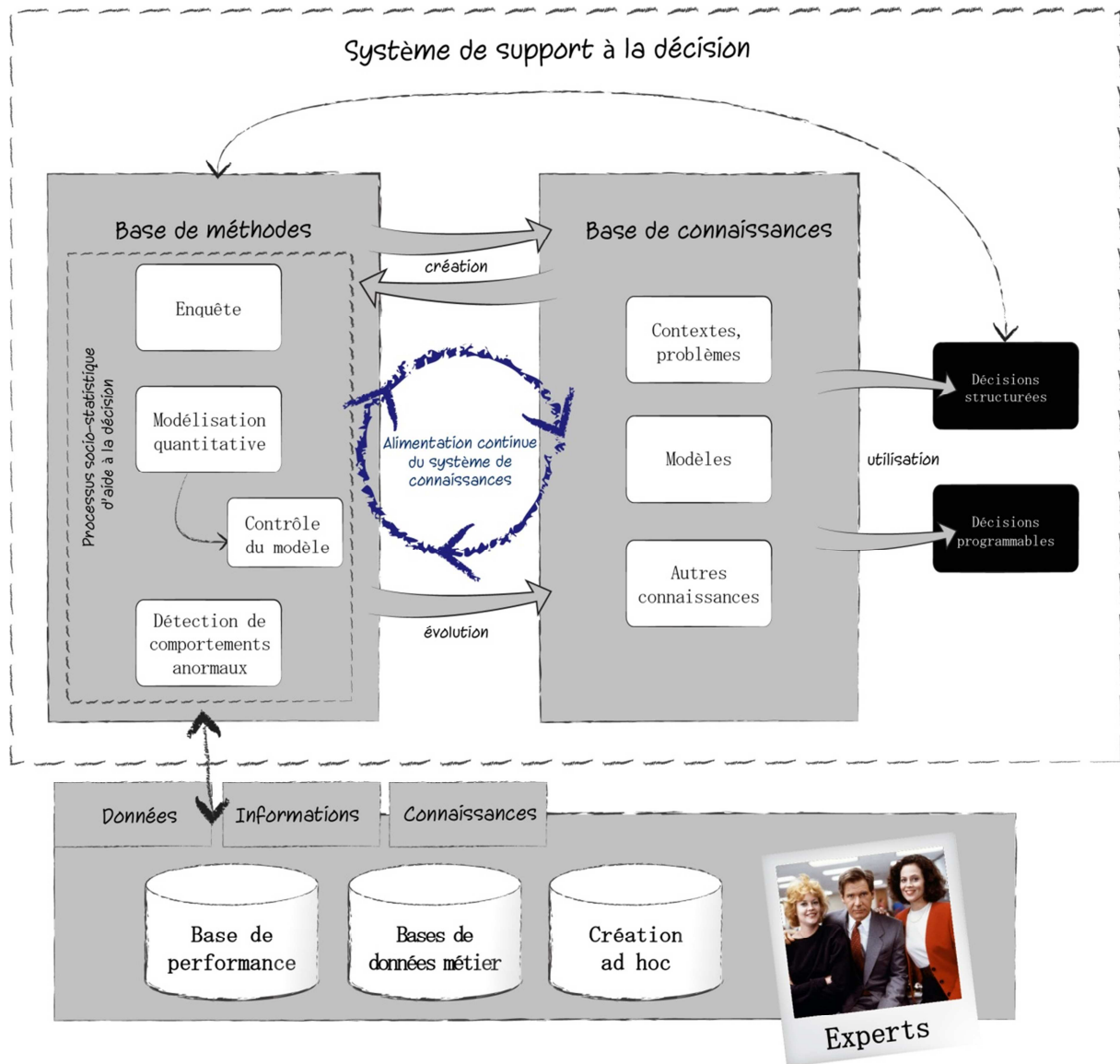


Figure 76. Vers un système de support à la décision : conceptualisation.

Nous avons d'ores et déjà identifié un certain nombre d'articles qui posent des bases conceptuelles et méthodologiques pertinentes pour la construction d'un tel système de support à la décision :

- *Power et Sharda (2007)* proposent une réflexion conceptuelle quant à la définition d'un système de support à la décision basé sur des modèles quantitatifs ;
- *Arinze et al. (1992)* suggèrent un système de support à la décision à base de connaissances, ayant pour but d'aider les décideurs dans leurs activités de gestion des capacités informatiques, en assurant un certain nombre fonctions clairement identifiées (évaluation de la performance informatique, prévision...) et basées sur des modèles quantitatifs ;

- *Chakraborty et Tab (2005)* montrent quant à eux comment des procédures de contrôle statistique peuvent être intégrées dans un système de support à la décision, en aidant au choix des méthodes statistiques et à l'interprétation des alertes grâce à une base de connaissances, formalisées par un ensemble de règles de décisions.

Selon le périmètre de ces travaux de thèse, développer un tel système de support à la décision permettrait de faire entrer le CMIS recommandé par ITIL dans une réelle dimension décisionnelle. La question des données relatives à la gestion des technologies de l'information, mais également leur exploitation à des fins d'aide à la décision seraient considérées dans le cadre d'un système d'information de gestion des capacités. Il contribuerait ainsi activement au bon fonctionnement de l'organisation.

Nos travaux de thèse et les perspectives présentées dans ce chapitre ont amené à évoquer conjointement des termes tels que « données », « prise de décision », « automatisation »... Cette conjonction renvoie à un concept actuellement en vogue : le *Big Data*. Nous proposons de conclure cette section en mettant ce concept en regard de nos travaux.

## **1.4. Travaux de recherche et *Big Data***

### *1.4.1. Présentation du concept*

Le *Big Data* est tout à fait le genre de « *buzzword* » qui alimente les discussions des services en charge de technologies de l'information et des cabinets de conseil. Il supplante les précédents engouements pour des concepts désormais établis et stabilisés, tel que le *data-mining* et le *knowledge discovery in database* (déjà évoqués dans cette thèse). Les quelques statistiques présentées dans la Figure 77, représentant les volumes de recherches Google depuis 2004 pour les trois mots-clés « *Data-mining* », « *Knowledge Discovery* » et « *Big Data* », en attestent.

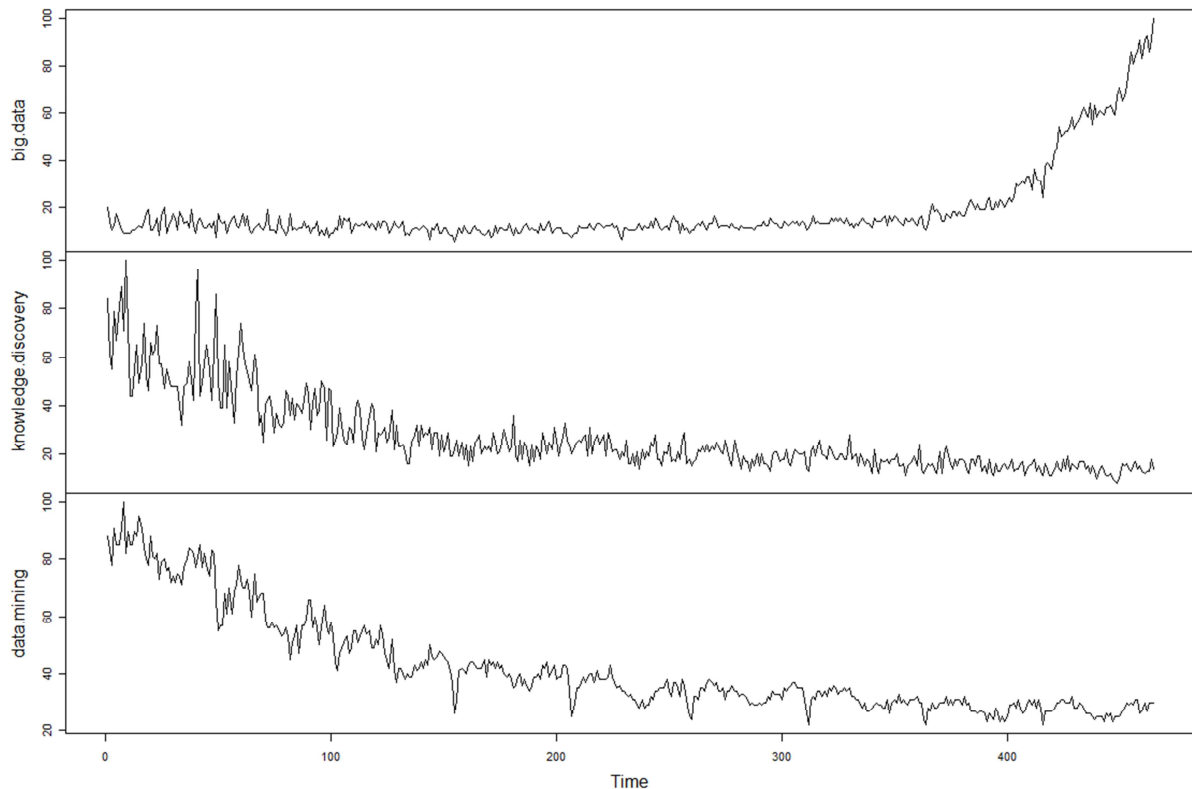


Figure 77. Volumes de recherche Google pour les mots-clés « Data-mining », « Knowledge Discovery » et « Big Data » (statistiques Google Trends : [www.google.com/trends](http://www.google.com/trends) ; données hebdomadaires, de janvier 2004 à décembre 2012)

Ces statistiques font référence à une base 100, qui correspond au plus haut volume de recherches Google pour le mot-clé considéré. On constate que les recherches liées aux termes désormais usuels « *Data-mining* » et « *Knowledge Discovery* » ont fortement diminué de 2004 à 2008, et se sont stabilisées depuis ; *a contrario*, les recherches du terme « *Big Data* » explosent littéralement depuis 2011.

Pourtant ces trois concepts visent à améliorer l'exploitation des données disponibles. Le « *Big Data* » est-il juste un produit marketing, faisant du neuf avec des problématiques déjà anciennes ? Pas tout à fait. Même s'il s'appuie toujours sur un noyau dur de méthodes statistiques (Lavalle et *al.*, 2011 ; Manyika et *al.*, 2011 ; Davenport, 2012), le *Big Data* se préoccupe explicitement d'un certain nombre de dimensions opérationnelles moins développées dans le cas du *data-mining* ou du *knowledge discovery in database* (IBM, 2012) :

- Capacité à traiter de très larges volumes de données ;
- Capacité à traiter rapidement les données ;
- Capacité à traiter des données variées, pas forcément structurées ;
- Capacité à traiter des données de façon à obtenir des éléments de réflexion pouvant être utilisés en toute confiance pour la prise de décision.

Ainsi, le *Big Data*, au-delà des aspects purement statistiques et algorithmiques d'exploitation de données, vise à adresser un certain nombre de contraintes concrètes de

déploiement des méthodes quantitatives dans les entreprises. Ces contraintes sont à la fois techniques (volume, vitesse et variété des données) et organisationnelles (intérêt décisionnel). Une réelle vision *Big Data* impose de prendre en considération l'ensemble de ces dimensions, avec une pareille importance (Shah et al. 2012, Davenport, 2010, 2012). Comment positionner les travaux présentés dans cette thèse par rapport à cette pratique en plein essor ?

#### 1.4.2. Positionnement des travaux de recherches par rapport au *Big Data*

Pour situer nos travaux par rapport au concept de *Big Data*, nous nous positionnons par rapport à ses dimensions de mise en œuvre, introduites ci-dessus. Tout d'abord, il est clair que nos recherches en ont pleinement intégré la contrainte organisationnelle. L'ensemble du processus socio-statistique d'aide à la décision a été proposé de façon à ce que tout travail quantitatif soit entièrement centré sur un problème décisionnel, distinctement identifié et formalisé. De plus, il intègre un questionnement complet sur sa place dans les processus cognitifs des acteurs de l'organisation et sur son rôle dans la création de connaissances organisationnelles. En ce sens, notre proposition implique des traitements quantitatifs étroitement associés à l'implication des acteurs de l'organisation (cf. les aspects syntaxiques et pragmatiques de l'ontologie du système de connaissances), de façon à ne jamais perdre de vue la valeur-ajoutée décisionnelle du travail effectué. Cette réflexion sur la place des statistiques dans l'organisation et cette focalisation sur l'amélioration de l'efficacité de la prise de décision, nous amènent à penser que nos travaux sont en phase avec l'approche *Big Data*.

Concernant les contraintes techniques, il est nécessaire de rester plus modeste :

- *Traitement de larges volumes de données* : le processus socio-statistique d'aide à la décision a fourni un guide permettant de s'orienter dans les vastes bases de données de STMicroelectronics (base de performance et bases de données métier), en impliquant les acteurs de l'organisation. Néanmoins nous n'avons pas développé de méthodes informatiques permettant d'optimiser cette étape (visualisation, analyse). De plus, il faut indiquer que nos recherches ont suscitées une prise de conscience de la valeur-ajoutée potentielle de l'exploitation des données stockées par l'organisation ; en conséquence, nous avons observé que certains managers ont désormais tendance à stocker de plus en plus de données (notamment en en créant de nouvelles, cf. la hiérarchie inversée du *continuum* données-informations-connaissances) et à stopper les mécanismes de purge des bases de données dans l'éventualité d'une exploitation future. Ainsi, le volume de données disponibles est en forte croissance et il conviendra tôt ou tard de mettre en place des moyens pour en faciliter le traitement. En ce sens, les perspectives méthodologiques visant à favoriser l'exploitation des données en grandes dimensions permettraient, entre autres possibilités, de tendre vers cet objectif. L'automatisation d'un certain nombre d'actions d'exploitation des données, *via* un système de support à la décision irait également dans cette direction ;

- *Rapidité de traitement des données* : cette question n'a pour l'instant pas été explorée. Les traitements réalisés sont essentiellement manuels, sauf pour les activités de maintien à niveau de la connaissance, qui sont réalisées une fois par jour. Toutefois, plus nous irons vers l'automatisation de l'exploitation des données dans le cadre d'un système de support à la décision, plus nous pourrions favoriser la rapidité du traitement. Tout un ensemble de fonctions de contrôles et de régulations du système informatique pourraient être implémentée en quasi-temps réel. Dans le cadre des perspectives méthodologiques, le développement de modèles évolutifs améliorerait également la rapidité du traitement des données. Par exemple, pour les modèles de prévision, plutôt que de se contenter d'avertir les utilisateurs d'une rupture de la stabilité du modèle de prévision nécessitant un travail quantitatif manuel de ré-estimation, on pourrait envisager un réajustement automatique du modèle en temps réel ;
- *Traitement de données variées* : pour l'instant, la question n'est pas posée, seules des bases de données classiques sont mobilisées. Toutefois, dans le cadre des perspectives méthodologiques, nous avons évoqué la possibilité d'intégrer des données non-structurées, notamment à partir de techniques de type *process mining*. Dans ce cas, pourraient directement être exploités des signaux émis par les équipements de production de l'usine ou par les applications du système d'information.

Nous pouvons donc dire que nos travaux ne peuvent se revendiquer comme la définition aboutie d'une solution *Big Data* chez STMicroelectronics. Néanmoins, ils ont permis de rassembler un certain nombre de prémisses propices au développement futur d'un projet *Big Data*. Ils s'inscrivent ainsi dans la « philosophie » de cette pratique émergente, en y apportant une contribution non négligeable : l'inscription de l'exploitation des données dans le cadre d'un processus socio-statistique d'aide à la décision, développé avec toute la rigueur méthodologique exigée par un projet de recherche académique.



## Section 2. Conclusion générale.

Construite à partir d'un contexte de recherche industriel, cette thèse a montré quelles sont les difficultés liées à la gestion des infrastructures des technologies de l'information contemporaines, à savoir : la gestion de la complexité, tout en ayant pour priorité de soutenir l'activité de l'organisation. Pour aider les managers à surmonter ces difficultés, un processus d'aide à la décision complet a été proposé. Un tel processus devait permettre de formalisation des connaissances organisationnelles, utilisables pour l'action et la prise de décision. Cette formalisation nécessitait un couplage entre méthodes quantitatives statistiques et méthodes qualitatives, pour impliquer les experts de l'organisation. Pour ce faire, il était nécessaire de développer une approche structurée. D'une part, elle devait permettre de contextualiser avec précision le besoin décisionnel. D'autre part, elle devait aboutir au développement d'une démarche de modélisation.

Une fois spécifiée et définie, cette démarche a été appliquée à un cas concret de STMicroelectronics. Deux besoins ont été clairement exprimé : la modélisation prospective (cœur des recherches) et la détection de comportements anormaux (travaux annexes). Une solution décisionnelle globale, couplant analyses quantitative et qualitative, a été proposée pour améliorer cette situation. En a résulté la formalisation d'une connaissance, sous la formalisation d'un modèle statistique. Nous avons montré comment cette connaissance pouvant être utilisée avec profit pour l'aide à la décision. Nous avons aussi suggéré certains mécanismes permettant d'en assurer le maintien à jour.

D'un point de vue scientifique, la valeur-ajoutée de ces travaux est la proposition d'un processus générique permettant une conduite complète de l'aide à la décision, basée sur une utilisation contextualisée de modèles quantitatifs formalisant une connaissance organisationnelle. De façon rigoureuse et argumentée, nous avons montré comment peut être contextualisée une situation décisionnelle et comment la gestion des connaissances industrielles peuvent permettre de construire les artefacts de cette contextualisation (constitution d'études de cas). Nous avons aussi montré comment l'utilisation de méthodes statistiques à l'état de l'art permettait de constituer des artefacts d'aide à la décision, basés sur des analyses quantitatives, situées dans une perspective cognitiviste. Ces analyses permettent à la fois de construire les connaissances utiles à l'aide à la décision, mais également de les maintenir à jour. Compte-tenu des volumes toujours croissants de données stockées par les organisations, nous pouvons dire que les travaux présentés offrent un ensemble de perspectives méthodologiques et décisionnelles, qui mériteraient d'être approfondies ultérieurement et confrontés à d'autres contextes de recherches.

D'un point de vue industriel, l'application de ce processus décisionnel a permis d'apporter des réponses concrètes à deux problématiques de gestion des capacités rencontrées chez STMicroelectronics : la modélisation prospective (cœur des recherches) et la détection de comportements anormaux (travaux annexes). Ces améliorations s'inscrivent comme une contribution directe à l'amélioration de la mise en œuvre de certaines activités du processus ITIL de gestion des capacités. Outre ces apports directs, nous estimons que les travaux de recherches menés chez STMicroelectronics ont favorisés le développement d'un terreau organisationnel, social et méthodologique favorable à une utilisation plus généralisée de méthodes statistiques, dans le cadre de la

gestion des connaissances et des mécanismes décisionnels de l'organisation. Cette évolution des mentalités et des pratiques pourrait, dans un second temps, être mise à profit dans le cadre d'un déploiement plus généralisé de nos propositions, voire d'une approche *Big Data* de plus grande ampleur. Dans cette optique, un prolongement de nos travaux consisterait à initier une réflexion approfondie, concernant les conditions d'industrialisation des résultats déjà obtenus et les perspectives suggérées. Cette réflexion permettrait une généralisation de nos travaux, à un nombre élargi de processus décisionnels au sein de Crolles 2. Elle permettrait également le déploiement vers d'autres usines, voir vers des mécanismes décisionnels inter-usines.

**Articles et communications issus des travaux présentés dans  
cette thèse**

## **Publications en revues scientifiques à comité de lecture**

M. Lutz, X. Boucher, O. Roustant, Methods and applications for decisions on IT capacity: bringing management frameworks into practice effectively (soumis à la revue Journal of Decision Systems en mars 2013).

E. Padonou, O. Roustant, M. Lutz, Robust monitoring of an industrial IT system in presence of structural change (soumis à la revue Quality and Reliability Engineering International en février 2013).

M. Lutz, X. Boucher, O. Roustant, Information Technologies capacity planning in manufacturing systems: proposition for a modelling process and application in the semiconductor industry, Computers in Industry vol. 63 (7) (2012) 659-669.

M. Lutz, X. Boucher, O. Roustant, M.A. Girard, Couplage entre système de production industriel et technologies de l'information : modélisation et création de connaissance, Ingénierie des Systèmes d'Information vol. 17 (4) (2012) 95-117.

## **Publication en revue industrielle**

M. Lutz, O. Roustant, X. Boucher, Monitoring an industrial information system using R with Holt-Winters filtering and  $Q_n$  robust standard deviation estimator, Journal of Computer Resource Management vol. 129 (2011) 44-51.

## **Conférences**

### *Internationales*

E. Padonou, M. Lutz, O. Roustant, Robust control chart to monitor the information system of semiconductor production plant, ENBIS 2012 (European Network for Business and Industrial Statistics), Ljubljana, Slovenia, September 2012.

M. Lutz, L. Mitchell, X. Boucher, Managing information technologies capacities through data analysis: data availability and exploitation, ENBIS 2012 (European Network for Business and Industrial Statistics), Ljubljana, Slovenia, September 2012.

M. Lutz, X. Boucher, O. Roustant, X. Ambrosioni, IT capacity forecasting: statistical modelling process and applications for semiconductor industry, INCOM 2012 (IFAC Symposium on Information Control Problems in Manufacturing), Bucharest, Romania, May 2012.

M. Lutz, O. Roustant, X. Boucher, Using statistics to monitor and model information technologies activities: case study in a semiconductor production plant, ENBIS 2011 (European Network for Business and Industrial Statistics), Coimbra, Portugal, September 2011.

M. Lutz, X. Boucher, O. Roustant, M.A. Girard, Capacity planning of a semi-conductor manufacturing information system, IESM 2011 (International Conference on Industrial Engineering and Systems Management), Metz, France, Mai 2011.

### *Nationales*

M. Lutz, R. Le Riche, X. Boucher, Informations limitées, dimensions et contradictions – application à la gestion des connaissances industrielles par l'emploi de méthodes statistiques, ASLC 2013 (Ateliers Sur La Contradiction), Saint-Etienne, France, Avril 2013.

M. Lutz, X. Boucher, O. Roustant, M.A. Girard, Capacity planning d'un système d'information industriel complexe, ROADEF 2011 (Conférence de la Société Française de Recherche Opérationnelle et d'Aide à la Décision), Saint-Etienne, France, Mars 2011.

### **Invitations à des conférences ou séminaires**

M. Lutz, Déploiement de méthodes analytiques pour le pilotage du système d'information d'une usine de production de semi-conducteurs – une première étape vers le Big Data, Congrès Big Data, CNIT Paris La Défense, Avril 2013.

M. Lutz, Monitoring et modélisation statistique d'un système d'information industriel de production de semi-conducteurs, Université de Rennes 1, Novembre 2011.

M. Lutz, Gestion de la capacité des ressources techniques d'un système d'information : une approche statistique orientée métier, Ecole Nationale Supérieure des Mines de Saint-Etienne - Laboratoire en Sciences et Technologies de l'Information, Juin 2011.

## Références

- H. Abdi, L.J. Williams, Principal component analysis, *WIREs Computational Statistics* 2 (4) (2010) 433-459.
- A.T.M. Aerts, J.B.M. Goossenaerts, D.K. Hammer, J.C. Wortmann, Architectures in context: on the evolution of business, application software, and ICT platform architectures, *Information and Management* 41 (2004) 781–794.
- N. Ahituv, M. Igarria, A model for predicting and evaluating computer resource consumption, *Communications of the ACM* 31 (12) (1988) 1467-1473
- M. Alavi, D.E. Leidner, Knowledge management and knowledge management systems: conceptual foundations and research issues, *MIS quarterly*, 25 (1) (2001) 107-136.
- J.R. Alcaras, P. Gianfaldoni, G. Paché (Dir.), *Décider dans les organisations – dialogues critiques entre économie et gestion*, L’Harmattan, Paris, 2004.
- S. Alford, R. Akker, R. Assaf, C. Bentley, J. van der Bent, P. Breslin, M. da Cruz, R. Donatz, R. Faber, E. Giesen, W. van Grembergen, S. De Haes, J.G. Hall, B. Hedeman, J. Hendriks, P. Hill, M. Iqbal, G. Koch, *Frameworks for IT management: an introduction*, van Haren, Zaltbommel, 2006.
- J. Allspaw, *The art of capacity planning*, O’Reilly, Sebastopol, 2008.
- L.C. Alwan, H.V. Roberts, Time-series modeling for statistical process control, *Journal of Business and Economic Statistics* 6 (1) (1988) 87-95.
- B. Ancori (Dir.), *Apprendre se souvenir décider – une nouvelle rationalité de l’organisation*, CNRS Editions, Paris, 1992.
- A. Appriou (Ed.), *Gestion de la complexité et de l’information dans les grands systèmes critiques*, CNRS Editions, Paris, 2009.
- C. Argyris, D.A. Schön, *Apprentissage organisationnel – théorie, méthode, pratique*, De Boeck, Bruxelles, 2001.
- B. Arinze, M. Igarria, L.F. Young, A knowledge based decision support system for computer performance management, *Decision Support Systems* 8 (1992) 501-515.
- M. Armatte, *La science économique comme ingénierie – quantification et modélisation*, Presses des Mines, 2010.
- M.J. Avenier, Pour quoi jeter le bébé avec l’eau du bain ? Méthodologie sans épistémologie n’est que ruine de la réflexion !, *Le Libellio d’AEGIS* 7 (1) (2011) 39-52.
- P. Azoulay, P. Ponthus, *Les modèles de décision dans l’entreprise*, Presses Universitaires de France, Paris, 1974.
- J. Bai, P. Perron, Computation and analysis of multiple structural change models, *Journal of Applied Econometrics* 18 (2003) 1-22.
- J. Bai, Estimation of a change point in multiple regression models, *The Review of Economics and Statistics* 79 (4) (1997) 551-563.
- S. Banerjee, M. Igarria, An empirical study of computer capacity planning in U.S. universities, *Information and Management* 24 (1993) 171-182.

- E. Barbin, J.P. Lamarche (Coord.), Histoire de probabilités et de statistiques, Ellipses, Paris, 2004.
- F. Barthelmé, J.L. Ermine, C. Rosenthal-Sabroux, An architecture for knowledge evolution in organisations, *European Journal of Operational Research* 109 (1998) 414-427.
- B. Baynat, Théorie des files d'attente – des chaînes de Markov aux réseaux à forme de produit, Hermès, Paris, 2000.
- E. Berghout, M. Nijland, P. Powell, Management of lifecycle costs and benefits: lessons from information systems practice, *Computers in Industry* 62 (7) (2011) 755–764.
- M. Bernardo, J. Hillston (Ed.), Lecture notes in computer science – proceedings of the 7<sup>th</sup> international conference on formal methods for performance evaluation, Springer, Berlin, 2007.
- A. Bharadwaj, A resource-based perspective on information technology capability and firm performance: an empirical investigation, *MIS Quarterly* 24 (1) (2000) 169-195.
- A. Bounfour (Ed.), De l'informatique aux systèmes d'information dans les grandes entreprises, *Entreprise et Histoire* 60 (3) (2010).
- D. Bouyssou, D. Dubois, M. Pirlot, H. Prade (Dir.), Concepts et méthodes pour l'aide à la décision 1 – outils de modélisation, Hermès, Paris, 2006a.
- D. Bouyssou, T. Marchant, M. Pirlot, A. Tsoukiàs, P. Vincke, Evaluation and decision models with multiple criteria – stepping stones for the analyst, Springer, Berlin, 2006b.
- J.P. Briffaut, Systèmes d'information en gestion industrielle, Hermès Sciences, Paris, 2000.
- L. Bronner, Overview of the capacity planning process for production data processing, *IBM System* 19 (1) (1980) 4-27.
- R.L. Brown, J. Durbin, J.M. Evans, Techniques for testing the constancy of regression relationships over time, *Journal of the Royal Statistical Society* 37 (1975) 149-163.
- T. Browning, Capacity planning for computer systems, AP Professional, Cambridge, 1994.
- C.P. Bruter, Les architectures du feu – considérations sur les modèles, Flammarion, Paris, 1982.
- J. Brutlag, Notes on RRDTOOL implementation of aberrant behavior detection (2000), [http://cricket.sourceforge.net/aberrant/rrd\\_hw.htm](http://cricket.sourceforge.net/aberrant/rrd_hw.htm), dernière consultation le 29 octobre 2012.
- R. Buyya, Economic-based distributed resource management and scheduling for grid computing, Thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy, School of Computer Science and Software Engineering, Monash University, Melbourne, 2002.
- J. Cady, B.R. Howarth, Computer systems performance management and capacity planning, Prentice Hall, New York, 1990.
- R. Cardinali, Information systems - a key ingredient to achieving organizational competitive strategy, *Computers in Industry* 18 (1992) 241–245.
- F. Carlier, A. Richard, Analyse stratégique de la décision, Presses Universitaires de Grenoble, Grenoble, 2002.



- A. Cedilnik, K. Košmelj, A. Blejec, The distribution of the ration of jointly normal variables, *Metodološki Zvezki* 1 (1) (2004) 99-108.
- S. Chakraborty, D. Tah, Real time statistical process advisor for effective quality control, *Decision Support Systems* 42 (2006) 700-711.
- A. Chakravarty, H.K. Jain, Distributed computer system capacity planning and capacity loading, *Decision Sciences* 21 (2) (1990) 253-262.
- J. Chapron, L'urbanisme organisationnel : méthode et aide à la décision pour piloter l'évolution du système d'information de l'entreprise, Thèse présentée en vue de l'obtention du titre de Docteur en Génie Industriel de l'Ecole Nationale Supérieure des Mines de Saint Etienne et de l'Université Jean Monnet, 2006.
- A. Charpentier, Différencier (indéfiniment) les séries temporelles ?, *Freakonometrics* (2009) [http://freakonometrics.blog.free.fr\\_index.php\\_tag\\_intégrer](http://freakonometrics.blog.free.fr_index.php_tag_intégrer), dernière consultation le 15 novembre 2012.
- J.B. Chatelain, Comment bien régresser – la statistique peut-elle se passer d'artefacts ?, *Prisme* 19 (2010).
- C.S.J. Chu, K. Hornik, C.M. Kuan, The moving-estimates test for parameter stability, *Econometric Theory* 11 (1995) 669–720.
- K.J. Cios, W. Pedrycz, R.W. Swiniarsky, L.A. Kurgan, *Data mining – a knowledge discovery approach*, Springer, Berlin, 2007.
- CMG, Computer Measurement Group, [www.cmg.org/](http://www.cmg.org/), dernière consultation le 18 septembre 2012.
- A. Cockroft, B. Walker, *Capacity planning for internet services*, Prentice Hall, Upper Saddle River, 2001.
- V. Cohen, *La recherche opérationnelle*, PUF, Paris, 1995.
- W.W. Cooper, K.K. Sinha, R.S. Sullivan, Measuring complexity in high-technology manufacturing: indexes for evaluation, *Interfaces* 22 (4) (1992) 38–48.
- J.C. Correa, N. González, *Gráficos Estadísticos con R*, Universidad Nacional-Sede Medellín, 2002.
- J.W. Cortada, *Managing DP hardware – capacity planning, cost justification, availability, and energy management*, Prentice Hall, Englewood Cliffs, 1983.
- P. Coutelle, *Introduction aux méthodes qualitatives en science de gestion*, Cours du CEFAG – séminaire d'études qualitatives, 2005.
- C. Croux, S. Gelper, K. Mahieu, Robust control charts for time series data, *Expert Systems with Applications* 38 (11) (2011) 13810-13815.
- M. Crozier, E. Friedberg, *L'acteur et le système*, Seuil, Paris, 2004.
- R.M. Dabbas, H.N. Chen, Mining semiconductor manufacturing data for productivity improvement - an integrated relational database approach, *Computers in Industry*, 45 (1) (2001) 29-44.
- T.H. Davenport (Ed.), *Enterprise analytics – optimize performance, process and decisions through big data*, FT Press, Upper Saddle River, 2012.

- T.H. Davenport, Business intelligence and organizational decisions, *International Journal of Business Intelligence Research* 1 (1) (2010) 1-12.
- A. David, Logique, épistémologie et méthodologie en sciences de gestion, Conférence de l'AIMS, 1999.
- J.G. De Gooijer, R.J. Hyndman, 25 years of time series forecasting, *International Journal of Forecasting* 22 (3) (2006) 443-473.
- T.S. Deisboeck, J.Y. Kresh, *Complex systems science in biomedicine*, Springer, New-York, 2006.
- S. Déjean, I. González, K.A. Lê Cao, X.Y. Chua, mixomics, <http://perso.math.univ-toulouse.fr/mixomics/>, dernière consultation le 09 octobre 2012.
- J. De Rosnay, *Le microscope – vers une vision globale*, Seuil, Paris, 1977.
- A. Desrosières, *La politique des grands nombres – histoire de la raison statistique*, La Découverte, Paris, 2010.
- A.C. Deuel, The benefit of a manufacturing execution system for plantwide automation, *ISA Transaction* 33 (2) (1994) 113–124.
- R. Dieng-Kuntz, N. Matta, *Knowledge management and organizational memories*, Kluwer, Dordrecht, 2002.
- B. Domanski, Distributed capacity planning, part 1, *Enterprise Systems Journal* (1999a), [http://esj.com/articles/1999/01/01/distributed-capacity-planning-part-1\\_633718635970687516.aspx?sc\\_lang=en](http://esj.com/articles/1999/01/01/distributed-capacity-planning-part-1_633718635970687516.aspx?sc_lang=en), dernière consultation le 12 septembre 2012.
- B. Domanski, Capacity planning in distributed environments, Part 2 (1999b), [http://esj.com/articles/1999/02/01/capacity-planning-in-distributed-environments-part-2\\_633718629354742660.aspx?sc\\_lang=en](http://esj.com/articles/1999/02/01/capacity-planning-in-distributed-environments-part-2_633718629354742660.aspx?sc_lang=en), dernière consultation le 12 septembre 2012.
- J. Dugmore, S. Taylor, ITIL V3 and ISO/IEC 20000 – alignment white paper, Office of Government Commerce, 2008.
- J. Dugmore, S. Lacy, *Capacity management*, British Standards Institution, London, 2005.
- D. Dupuy, C. Helbert, Construction and evaluation of metamodels, <http://cran.r-project.org/web/packages/DiceEval/DiceEval.pdf>, dernière consultation le 09 octobre 2012.
- M.J. Earl, *Management for information technology*, Prentice Hall, Herdfordhsire, 1989.
- M. Easterby-Smith, M.A. Lyles (Eds.), *Handbook of organizational learning and knowledge*, Blackwell, New-York, 2002.
- T.M. Edgington, T.S. Raghu, A.S. Vinze, Using process mining to identify coordination patterns in IT service management, *Decision Support System* 49 (2) (2010) 175-186.
- I. Englander, *The architecture of computer hardware and system software - an information technology approach*, Wiley, New-York, 2009.
- J.L. Ermine, *Les systèmes de connaissances*, Hermès, Paris, 1996.
- J.L. Ermine, *Systèmes experts – théorie et pratiques*, Lavoisier, Cachan, 1989.

- U. Fayyad, P. Stolorz, Data mining and KDD: promise and challenges, *Future Generation Computer Systems* 13 (1997) 99-115.
- U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery in databases, *AI Magazine* (1996a) 37-54.
- U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, The KDD process for extracting useful knowledge from volumes of data, *Communications of the ACM* 39 (11) (1996b) 27-24.
- R. Feldman, J. Sanger, *The text mining handbook: advanced approaches in analyzing unstructured data*, Cambridge University Press, Cambridge, 2006.
- D. Ferrari, *Computer systems performance evaluation*, Prentice Hall, Englewood Cliffs, 1978.
- G.A. Forgionne, Improving facility planning with decision technology systems, *Computer Science in Economic and Management* 4 (1991) 261-283.
- J.M. Fouet (Coord.), *Connaissance et savoir-faire en entreprise – intégration et capitalisation*, Hermès, Paris, 1997.
- J. Fox, S. Weisberg, *An R companion to applied regression*, Sage, Thousand Oaks, 2011.
- T. Foxon, M. Garth, P. Harrison, Capacity planning in client-server systems, *Distributed System Engineering* 3 (1996) 32-38.
- R. Fronheiser, ITIL capacity management – more than charts over coffee, CMG 2006, International Conference of the Computer Measurement Group (2006).
- R. Fronheiser, Too many servers, not enough eyes (or – where did all these servers come from?!?), Computer Measurement Group Internal Conference, Las Vegas (2004).
- E. Gardner, Exponential smoothing: the state of the art – part II, *International Journal of forecasting* 22 (3) (2006) 637-666.
- J.P. Gauchi, Utilisation de la régression PLS pour l'analyse des plans d'expériences en chimie de formulation, *Revue de Statistique Appliquée* 43 (1) (1995) 65-89.
- S. Gelper, R. Fried, C. Croux, Robust forecasting with exponential and Holt-Winters smoothing, *Journal of Forecasting* 29 (2009) 285-300.
- N. Geng, Z. Jiang, A review on strategic capacity planning for the semi-conductor manufacturing, *International Journal of Production Research* 47 (13) (2009) 3639–3655.
- J. Goo, C.D. Huang, Facilitating relational governance through service level agreement in IT outsourcing: an application of the commitment-trust theory, *Decision Support System* 46 (1) (2008) 216-332.
- M.B. Gordon, H. Paugam-Moisy (Dir.), *Sciences cognitives – diversités des approches*, Hermès, Paris, 1997.
- N.J. Gunther, 2007, *Guerrilla capacity planning*, Springer, Berlin, 2007.
- N.J. Gunther, *The practical performance analyst - performance-by-design techniques for distributed systems*, McGraw-Hill, New York, 1998.

- B. Guy, Ethique et épistémologie – convergence entre la démarche épistémologique (chercher le vrai) et la démarche éthique (chercher le bien) : point de vue des sciences de l'ingénieur, <http://hal.archives-ouvertes.fr/hal-00736247>, dernière consultation le 02 octobre 2012.
- J.A. Gowan, R.G. Mathieu, Critical factors in information system development for a flexible manufacturing system, *Computers in Industry* 28 (1996) 173–183.
- A. Grummitt, Capacity management – a practitioner guide, van Haren, Zaltbommel, 2009.
- D.N. Gujarati, *Econométrie*, De Boeck, Bruxelles, 2004.
- D.J. Hand, H. Mannila, P. Smyth, *Principles of data mining*, MIT Press, Cambridge, 2001.
- B.E. Hansen, The new econometrics of structural change: dating breaks in U.S. labor productivity, *Journal of Economic Perspectives* 15 (4) (2001) 117-128.
- M.A. Hardy, *Regression with dummy variables*, Sage, Newbury Park, 1993.
- T. Hastie, R. Tibshirani, J. Friedman, *The elements of statistical learning: data mining, inference and prediction*, Springer, New-York, 2001.
- J.L. Hellerstein, F. Zhang, P. Shahabuddin, Characterizing normal operation of a web server: application to workload forecasting and problem detection, *Proceedings of the Computer Measurement Group International Conference*, 1998.
- G.H. Hendriksen, *Method General Approach Performance Profiling (Method GAPP)* (2011), <http://method-gapp.com>, dernière consultation le 5 octobre 2011.
- M. Hubert, P.J. Rousseeuw, K. Vanden Branden, ROBPCA: a new approach to robust principal component analysis, *Technometrics* 47 (2005) 64–79.
- F. Husson, S. Lê, J. Pagès, *Analyse de données avec R*, Presses Universitaires de Rennes, Rennes, 2009.
- R.J. Hyndman, G. Athanasopoulos, *Forecasting: principles and practice – an online textbook*, <http://otexts.com/fpp/>, dernière consultation le 12 octobre 2012.
- R.J. Hyndman, Time series cross-validation: an R example, <http://robjhyndman.com/researchtips/tscvexample/>, dernière consultation le 06 décembre 2011.
- R.J. Hyndman, Y. Khandakar, Automatic time series forecasting: the forecast package for R, *Journal of Statistical Software* 27 (3) (2008).
- IBM, What is big data? Bringing big data to the enterprise, [www-01.ibm.com/software/data/bigdata/](http://www-01.ibm.com/software/data/bigdata/), dernière consultation le 21 octobre 2012.
- ISO-1, Norme Internationale ISO/CEI 20000-1:2005, Technologies de l'information – gestion des services – Partie 1 : spécifications, 2005.
- ISO-2, Norme Internationale ISO/CEI 20000-2:2005, Information technology – service management – Part 2: code of practice, 2005.
- ISO-3, ISO 20000 and ITIL, [20000.fwtk.org/20000-itil.htm](http://20000.fwtk.org/20000-itil.htm), dernière consultation le 5 septembre 2012.
- ITIL, Official ITIL website, [www.itil-officialsite.com](http://www.itil-officialsite.com), dernière consultation le 5 septembre 2012.

- S. Izza, *Intégration des systèmes d'information industriels – Une approche flexible basée sur les services sémantiques*, Thèse présentée en vue de l'obtention du titre de Docteur en Génie Industriel de l'Ecole Nationale Supérieure des Mines de Saint Etienne et de l'Université Jean Monnet, 2006.
- J.H. Jacot (Dir.), *Formes anciennes, formes nouvelles d'organisation*, Presses Universitaires de Lyon, Lyon, 1994.
- R. Jain, *The art of computer systems performance analysis - Techniques for experimental design, measurement, simulation, and modeling*, Wiley, New-York, 1991.
- G. Jiang, H. Chen, K. Yoshihira, *Profiling services for resource optimization and capacity planning in distributed systems*, *Cluster Computing* 11 (2008) 313–329.
- J. Johnston, J. Dinardo, *Méthodes économétriques*, Economica, Paris, 1999.
- T. Kelly, *Detecting performance anomalies in global applications*, Second Workshop on Real Large Distributed Systems, San Francisco, 2005.
- P. Kennedy, *A guide to econometrics*, Wiley, New-York, 2010.
- Y.G. Kim, S.H. Yu, J.H. Lee, *Knowledge strategy planning: methodology and case*, *Expert systems with applications* 24 (2003) 295-307.
- L. Klosterboer, *ITIL capacity management*, IBM Press – Pearson, Boston, 2011.
- G. Koenig, *L'apprentissage organisationnel: repérage des lieux*, *Revue Française de Gestion* 160 (2006) 293-328.
- K.D. Kumar, L. Karunamoorthy, H. Roth, T.T. Mirnalinee, *Computers in manufacturing: towards successful implementation of integrated automation system*, *Technovation* 25 (5) (2005) 477–488.
- M.C. Lacity, L.P. Willcocks, A. Subramanian, *A strategic client/server implementation: new technology, lessons from history*, *Journal of Strategic Information Systems* 6 (2) (1997) 95-128.
- S. Laflamme, *Analyse statistique linéaire et interprétation systémique*, *Nouvelles perspectives en sciences sociales : revue internationale de systémique complexe et d'études relationnelles* 4 (1) (2008) 141-159.
- A. Langley, H. Mintzberg, P. Pitcher, E. Posada, J. Saint-Macary, *Opening up decision making: the view from the black stool*, *Organization Science* 6 (3) (1995) 260-279.
- S. Lavalley, E. Lesser, R. Shockley, M.S. Hopkins, N. Kruschwitz, *Big data, analytics and the path from insights to value*, *MIT Sloan Management Review* 52 (2) (2011) 20-31.
- S.S. Lavenberg, *Computer performance modeling Handbook*, Academic Press, New York, 1983.
- E.D. Lazowska, J. Zahorjan, J.S. Graham, C.S. Kenneth, *Quantitative system performance - computer system analysis using queuing network models*, Prentice Hall, Englewood Cliffs, 1984.
- R.C. Leachman, D.A. Hodges, *Benchmarking semiconductor manufacturing*, *IEEE Transactions on Semiconductor Manufacturing* 9 (2) (1996) 158–169.
- L. Lebart, M. Piron, A. Morineau, *Statistique exploratoire multidimensionnelle*, Dunod, Paris, 2006.

- J.F. Lebraty, Technologies de l'information et décision : l'essor des datawarehouses, *Système d'Information et Management* 1 (3) (1998) 95-118.
- J.H. Lee, Y.G. Kim, A stage model of organizational knowledge management: a latent content analysis, *Expert Systems with Applications* 20 (2001) 299-311.
- A. Lefebvre, *L'architecture client-serveur*, Armand-Colin, Paris, 1994.
- C. Leikis, Consolidated capacity and performance reporting, *Proceedings of the Computer Measurement Group International Conference*, 2007.
- C. Lejeune, Montrer, calculer, explorer, analyser. Ce que l'informatique fait (faire) à l'analyse qualitative, *Recherches Qualitatives - Hors-Série – 9* (2010) 15-32.
- J.L. Le Moigne, *Les épistémologies constructivistes*, PUF, Paris, 2007.
- J.L. Le Moigne, *Le constructivisme – modéliser pour comprendre*, L'Harmattan, Paris, 2003.
- J.L. Le Moigne, *La modélisation des systèmes complexes*, Dunod, Paris, 1999.
- J.L. Le Moigne, *Les systèmes de décision dans les organisations*, Presses Universitaires de France, Paris, 1974.
- M. Lewis-Beck, A. Bryman, T. Futing (Eds.), *Encyclopedia of Social Sciences Research Methods*, Sage, Thousand Oaks, 2003.
- LINUX, linux.die.net, <http://linux.die.net/man/1/sar>, dernière consultation le 19 septembre 2012.
- J.D.C. Little, A proof of the queuing formula  $L = \lambda W$ , *Operations Research* 9 (1961) 383-387.
- V. Lloyd, C. Rudd, *Service Design*, The Stationery Office, Norwich, 2007.
- C. Lopez-Nicolas, P. Soto-Acosta, Analyzing ICT adoption and use effects on knowledge creation: an empirical investigation in SMEs, *International Journal of Information Management* 30 (2010) 521-528.
- G.S. Maddala, I.M. Kim, Structural change and unit roots, *Journal of Statistical Planning and Inference* 49 (1996) 73-103.
- S. Makridakis, S.C. Wheelwright, R.J. Hyndman, *Forecasting - methods and applications*, Wiley, New-York, 1998.
- J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. Hung Byers, *Big data: the next frontier for innovation, competition and productivity*, McKinsey Global Institute Report, 2011.
- J.G. March, *Décisions et organisations*, Editions d'Organisation, Paris, 1991.
- A. Martin, D. Dmitriev, J. Akeroyd, A resurgence of interest in information architecture, *International Journal of Information Management* 30 (1) (2010) 6-12.
- C.M. Mastrangelo, G.C. Runger, D.C. Montgomery, Statistical process monitoring with principal components, *Quality and Reliability Engineering International* 12 (1996) 203-210.
- D.A. Menascé, V.A.F. Almeida, *Capacity planning for web services – metrics, models, and methods*, 2002.

- D.A. Menascé, V.A.F. Almeida, *Scaling for e-business - technologies, models, performance and capacity planning*, Prentice Hall, Upper Saddle River, 2000.
- D.A. Menascé, V.A.F. Almeida, *Capacity planning for web performance - metrics, models, and methods*, Prentice Hall, Upper Saddle River, 1998.
- D.A. Menascé, V.A.F. Almeida, L.W. Dowdy, *Capacity planning and performance modeling: from mainframes to client-server systems*, Prentice Hall, Upper Saddle River, 1994.
- F.W. McFarlan (Ed.), *The information systems research challenge proceedings*, Harvard Business School, Boston, 1984
- H. Mintzberg, D. Raisinghani, A. Theoret, The structure of “unstructured” decision processes, *Administrative Science Quarterly* 21 (2) (1976) 246-275.
- E. Mitleton-Kelly, F. Land, *Complexity and information systems*, Published in the Second Edition of the Blackwell Encyclopaedia of Management, Volume on Management Information Systems, Wiley, New-York, 2011.
- B. McNaughton, P. Ray, L. Lewis, Designing an evaluation framework for IT service management, *Information & Management* 47 (4) (2010) 219-225.
- N. Moalla, H. Panetto, X. Boucher, Interopérabilité et partage de connaissances, *Ingénierie des Systèmes d'Information* 4 (2012) 7-16.
- D.C. Montgomery, E.A. Peck, G.G. Vining, *Introduction to linear regression analysis*, Wiley, Hoboken, 2006.
- D.C. Montgomery, *Introduction to statistical quality control*, Wiley, New-York, 2005.
- J.R. Montoya-Torres, Modélisation conceptuelle d'une unité de fabrication microélectronique, *Revista EIA – Escuela de Ingeniería de Antioquia* 7 (2007) 9-24.
- E. Morin, *Introduction à la pensée complexe*, Seuil, Paris, 2005.
- G. Neubert, *Contribution à la spécification d'un pilotage proactif et réactif pour la gestion des aléas*, Thèse de Doctorat présentée devant l'Institut National des Sciences Appliquées de Lyon, 1997.
- I. Nonaka, G. von Krogh, Tacit knowledge and knowledge conversion: controversy and advancement in organizational knowledge creation theory, *Organization Science* 20 (3) (2009) 635-652.
- I. Nonaka, N. Konno, The concept of “Ba”: building a foundation for knowledge creation, *California Management Review* 40 (3) (1998) 40-54.
- I. Nonaka, A dynamic theory of organizational knowledge creation, *Organization Science* 5 (1) (1994) 14-37.
- T.R. Norton, Don't predict applications when you should model the business, *Proceedings of the Computer Measurement Group International Conference*, 1998.
- P.C. Nutt, Framing strategic decision, *Organization Science* 9 (2) (1998) 195-216.
- P.C. Nutt, The formulation processes and tactics used in organizational decision making, *Organization Science* 4 (2) (1993) 226-251.

R.M. O'Brien, A caution regarding rules of thumb for variance inflation factors, *Quality and Quantity* 41 (2007) 673-690.

ONS, From Holt-Winters to ARIMA modelling: measuring the impact on forecasting errors for components of quarterly estimates of public service output, Office for National Statistics – UK Centre for the Measurement of Government Activity, 2008.

ORACLE, site de documentation Oracle, [http://docs.oracle.com/cd/B19306\\_01/server.102/b14237/stats002.htm#i28932](http://docs.oracle.com/cd/B19306_01/server.102/b14237/stats002.htm#i28932), dernière consultation le 19 septembre 2012.

A. Ostanello, A. Tsoukiàs, An explicative model of “public” interorganizational interactions, *European Journal of Operational Research* 70 (1993) 67-82.

E. Padonou, O. Roustant, M. Lutz, Robust monitoring of an industrial IT system in presence of structural change (2013 – soumis).

E. Padonou, M. Lutz, O. Roustant, Robust control chart to monitor the information system of a semiconductor production plant, ENBIS 2012, European Network for Business and Industrial Statistics, Ljubljana, 2012.

E. Paschetta, A. Tsoukiàs, A real world MCDA application: evaluating software, *Journal of Multi-criteria Decision Analysis* 9 (2000) 205-226.

A. Paschke, M. Bichler, Knowledge representation concepts for automated SLA management, *Decision Support System* 46 (1) (2008) 187-205.

J. Piaget, *Le structuralisme*, PUF, Paris, 1968.

G. Pison, S. Van Aelst, G. Willems, Small sample corrections for LTS and MCD, *Metrika* 55 (1) (2002) 111-123.

W. Ploberger, W. Krämer, K. Kontrus, A new test for structural stability in the linear regression model, *Journal of Econometrics* 40 (1989) 307-318.

J. Porra, R. Hirschheim, M.S. Parks, Forty years of the corporate information technology function at Texaco Inc. – a history, *Information & Organization* 16 (1) (2006) 82-107.

D.J. Power, R. Sharda, Model-driven decision support systems: concepts and research directions, *Decision Support Systems* 43 (2007) 1044-1061.

C.G. Qiao, G.R. Wood, C.D. Lai, D.W. Luo, Comparison of two common estimators of the ratio of the means if independent normal variables in agricultural research, *Journal of Applied Mathematics and Decision Sciences* (2006) 1-14.

A. Quaadgras, T. Edwards, Working smarter: leveraging information for group decision making, *CISR Research Briefing* 12 (4) (2012) 1-4.

M. Quinio Benamo, *Probabilités et statistique aujourd'hui*, L'Harmattan, Paris, 2009.

R Development Core Team, *R: a language and environment for statistical computing*, R Foundation for Statistical Computing, Vienne, 2011.



- R. Rakotomalala, TANAGRA data mining, <http://eric.univ-lyon2.fr/~ricco/tanagra/index.html>, dernière consultation le 09 octobre 2012.
- R. Reix, B. Fallery, M. Kalika, F. Rowe, *Systèmes d'information et management des organisations*, Vuibert, Paris, 2011.
- W. Revelle, *An introduction to psychometric theory with application in R*, <http://personality-project.org/r/book/>, dernière consultation le 16 octobre 2011.
- A. Revillard, *Fiches techniques – initiation à l'investigation empirique*, <http://annerevillard.com/enseignement/ressources-pedagogiques/initiation-investigation-empirique/fiches-techniques-initiation-investigation-empirique/>, dernière consultation le 02 octobre 2012.
- R.R. Rodriguez, J.J.A. Saiz, A.O. Bas, Quantitative relationships between key performance indicators for supporting decision-making processes, *Computers in Industry* 60 (2) (2009) 104-113.
- J.W. Ross, P. Weill, D.C. Robertson, *Enterprise architecture as strategy: creating a foundation for business execution*, Harvard Business School, Boston, 2006.
- P.J. Rousseeuw, C. Croux, Alternatives to the median absolute deviation, *Journal of the American Statistical Association* 88 (424) (1993) 1273-1283.
- P.J. Rousseeuw, A.M. Leroy, *Robust regression and outlier detection*, Wiley, New-York, 1987.
- B. Roy, Regard historique sur la place de la recherche opérationnelle et de l'aide à la décision en France, *Mathématiques et Sciences Humaines* 175 (3) (2006) 25-40.
- B. Roy, Decision science or decision-aid science?, *European Journal of Operational Research* 66 (1993) 184-203.
- B. Roy, Science de la décision ou science de l'aide à la décision ?, *Revue Internationale de Systémique* 6 (5) (1992a) 497-529.
- B. Roy, Recherche opérationnelle et aide à la décision, *AFCET Interflash* (1992b) 2-3.
- A. Rozinat, R.S. Mans, M. Song, W.M.P. Van der Aalst, Discovering simulation models, *Information Systems* 34 (2009) 305-327.
- G. Salaman, *Decision-making for business – a reader*, Sage, London, 2002.
- N. Salkind (Ed.), *Encyclopedia of measurement and statistics*, Sage, Thousand Oaks, 2007.
- T. Samad, P. McLaughlin, J. Lu, System architecture for process automation: review and trends, *Journal of Process Control* 17 (3) (2007) 191–201.
- G. Sanchez, Partial least squares, <http://www.plsmodeling.com/rpackages>, dernière consultation le 09 octobre 2012.
- R. Schiesser, *IT systems management*, Prentice Hall, Boston, 2010.
- L. Scrucca, qcc: An R package for quality control charting and statistical process control, *R News* 4 (1) (2004) 11-17.
- L. Segal, *Le rêve de la réalité*, Seuil, Paris, 1990.

- S. Shah, A. Horne, J. Capellá, Good data won't guarantee good decisions, *Harvard Business Review*, April 2012, 23-25.
- C.L. Sia, Y.S. Ho, Predictive capacity planning: a proactive approach, *Information and Software Technology* 39 (1997) 195-204.
- H.A. Simon, *Les sciences de l'artificiel*, Gallimard, Paris, 2004.
- H.A. Simon, De la rationalité substantive à la rationalité procédurale, *PISTES* 3 (1992), Disponible sur le site MCX-APC : [www.mcxapc.org/fileadmin/docs/lesintrouvables/simon5.pdf](http://www.mcxapc.org/fileadmin/docs/lesintrouvables/simon5.pdf).
- H.A. Smith, J.D. Mckeen, S. Singh, Developing information technology strategy for business value, *Journal of Information Technology Management* 18 (1) (2007) 49–58.
- C.U. Smith, L.G. Williams, *Performance solutions - a practical guide to creating responsive, scalable software*, Addison-Wesley, Boston, 2002.
- C.U. Smith, L.G. Williams, Software performance engineering: a case study including performance comparison with design alternatives, *IEEE Transactions on Software Engineering* 19 (7) (1993) 720-741.
- M. Song, W.M.P. van der Aalst, Towards comprehensive support for organizational mining, *Decision Support System* 46 (2008) 300-317.
- I. Stamelos, A. Tsoukiàs, Software evaluation problem situations, *European Journal of Operational Research* 145 (2003) 273-286.
- I. Stamelos, I. Vlahavas, I. Refanidis, A. Tsoukiàs, Knowledge based evaluation of software systems: a case study, *Information and Software Technology* 42 (2000) 333-345.
- J.H. Stock, M.W. Watson, Evidence on structural instability in macroeconomic time series relations, *Journal of business and economic statistics* 14 (1) (1996) 11-30.
- C. Sven, J. Björn, IS/ICT outsourcing decision project in large public organization – a case study, *Journal of Decision System* 20 (2) (2011) 137-162.
- J.W. Taylor, Smooth transition exponential smoothing, *Journal of Forecasting* 23 (2004) 385-394.
- M. Tenenhaus, *La régression PLS - Théorie et pratique*, Technip, Paris, 1998.
- M. Tenenhaus, J.P. Gauchi, C. Ménardo Régression PLS et applications, *Revue de statistique appliquée* 43 (1) (1995) 7-63.
- R.A. Thietart (et coll.), *Méthodes de recherche en management*, Dunod, Paris, 2007.
- G. Tiberghien (Dir.), *Dictionnaire des sciences cognitives*, Armand Colin, Paris, 2002.
- V. Todorov, P. Filzmoser, An object oriented framework for robust multivariate analysis, *Journal of Statistical Software* 32 (3) (2009) 1-47.
- H. Topi, J.S. Valacich, R.T. Wright, K.M. Kaiser, J.F. Nunamaker, J.C. Sipior, G.J. de Vreede, IS 2010 Curriculum guidelines for undergraduate degree programs in information systems, Association for Computing Machinery and Association for Information Systems, 2010.

- H. Tsoukas, E. Vladimirou, What is organizational knowledge?, *Journal of Management Studies*, 38 (7) (2001) 973-993.
- H. Tsoukas, The tyranny of light, *Futures* 29 (9) (1997) 827-843.
- H. Tsoukas, The firm as a distributed knowledge system: a constructionist approach, *Strategic Management Journal* 17 (1996) 11-25.
- A. Tsoukiàs, From decision theory to decision aiding methodology, *European Journal of Operational Research* 187 (2008) 138-161.
- A. Tsoukiàs, On the concept of decision aiding process, *Annals of Operations Research* 154 (2007) 3-27.
- I. Tuomi, Data is more than knowledge – implications of the reversed knowledge hierarchy for knowledge management and organizational memory, *Journal of Management Information Systems*, 16 (3) (1999) 107-121.
- M.C. Valiente, E. Garcia-Barriocanal, M.A. Sicilia, Applying an ontology approach to IT service management for business-IT integration, *Knowledge-Based Systems* 28 (1) (2012) 76-87.
- W.M.P. Van der Aalst, A.J.M.M. Weijters, Process mining: a research agenda, *Computers in Industry* 53 (2004) 231-244.
- W.N. Venables, B.D. Ripley, *Modern applied statistics with S*, Springer, New-York, 2002.
- L. Von Bertalanffy, *Théorie générale des systèmes*, Dunod, Paris, 1993.
- F. Wacheux, *Méthodes qualitatives et recherche en gestion*, Economica, Paris, 1996.
- B. Walliser, *Systèmes et modèles – introduction critique à l'analyse de systèmes*, Seuil, Paris, 1977.
- J. Wan, J.D. Jones, Managing IT service management implementation complexity: from the perspective of the Warfield version of systems science, *Enterprise Information Systems*, à paraître (2011) 1-33.
- J. Wan, Y. Wang, C. Zheng, Research on IT service management knowledge support structure, *WiCom 2007, International Conference on Wireless Communications, Networking and Mobile Computing* (2007) 6613-6616.
- P. Watzlawick (Dir.), *L'invention de la réalité – contributions au constructivisme*, Seuil, Paris, 1988.
- P. Weill, J.W. Ross, *IT governance: how top performers manage IT decision rights for superior results*, Harvard Business School, Boston, 2004.
- T.M. Williams, Adaptive Holt-Winters forecasting, *Journal of Operational Research Society* 38 (6) (1987) 553-560.
- S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, *Chemometrics and Intelligent Laboratory Systems* 58 (2001) 109-130.
- C.W. Yang, S.C. Fang, J.L. Lin, Organisational knowledge creation strategies: a conceptual framework, *International Journal of Information Management* 30 (2010) 231-238.
- P.C. Young, *Recursive estimation and time-series analysis: an introduction for the student and practitioner*, Springer, Berlin, 2011.

- A. Zeileis, A. Shah, I. Patnaik, Testing, monitoring, and dating structural changes in exchange rate regimes, *Computational Statistics and Data Analysis* 54 (6) (2010) 1696-1706.
- A. Zeileis, A unified approach to structural change tests based on F statistics, OLS residuals, and ML scores, Department of Statistics and Mathematics Wirtschaftsuniversität Wien – Research Report Series 13, 2005.
- A. Zeileis, C. Kleiber, Validating multiple structural change models – an extended case study, Department of Statistics and Mathematics Wirtschaftsuniversität Wien – Research Report Series 12, 2005.
- A. Zeileis, F. Leisch, C. Kleiber, K. Hornik, Monitoring structural change in dynamic econometric models, *Journal of Applied Econometrics* 20 (2005) 99-121.
- A. Zeileis, Implementing a class of structural change tests: an econometric computing approach, Department of Statistics and Mathematics Wirtschaftsuniversität Wien – Research Report Series 7, 2004.
- A. Zeileis, C. Kleiber, W. Krämer, K. Hornik, Testing and dating of structural changes in practice, *Computational Statistics and Data Analysis* 44 (2003) 109-123.
- A. Zeileis, F. Leisch, K. Hornik, C. Kleiber, strucchange: an R package for testing structural change in linear regression models, *Journal of Statistical Software* 7 (2) (2001) 1-38.

## **Annexes**

## Annexe 1

Chapitre 6. Application de la démarche de modélisation : création et formalisation de connaissances.

Section 1. Application à la migration MES : constitution d'un modèle de prévision.

1. Etape (A) intermédiaire d'observation n°1

1.2. Structuration des données - informations (aspect sémantique)

*Script R :*

```
#####
## Début de l'analyse
#####

#####
## Données
#####

## répertoire de travail
setwd("C:/XXX")

## données application
app300W <- read.csv2("acti_appli.csv")
rownames(app300W) <- app300W[,1]
jour <- app300W$JOUR
app300W$DATE <- NULL
app300W$JOUR <- NULL

## données serveur
cpu <- read.csv2("cpu.csv")
rownames(cpu) <- cpu[,1]
cpu$DATE <- NULL

#####
## ACP 1
#####

library(rrcov)

## acp :
acp1 <- PcaClassic(app300W, scale = TRUE)

## tracé de la figure du corps de thèse : saisonnalité de la 1ère composante
par(mfrow = c(1,2))
plot(acp1@scores[,"PC1"], type = 'l', main = "ACP1 - axe factoriel n°1",
     ylab = "ACP1 - axe factoriel n°1", xlab = "temps")
acf(ts(acp1@scores[,"PC1"]), main = "Corrélogramme")

## tracé de la figure du corps de thèse : corrélations entre composantes
population <- ifelse(jour%in%c("samedi","dimanche"), 2,16)

par(mfrow = c(1,1))
plot(acp1@scores[,"PC2"], acp1@scores[,"PC1"],
     pch = population,
     main = "ACP1 - axe factoriel n°1 et 2",
     xlab = "ACP1 - axe factoriel n°2", ylab = "ACP1 - axe factoriel n°1")
grid(9,lwd=1, col="1")

legend("topleft",
      c("Weekends","Jours de la semaine"),
      pch = c(2,16))
```

```

#####
## ACP 2
#####

library(psych)

## données sans weekends
app300W.semaine <- data.frame(app300W, jour)
app300W.semaine <- app300W.semaine[!app300W.semaine$jour%in%c("samedi", "dimanche"),]
app300W.semaine$jour <- NULL

## données serveur sans weekends
cpu.semaine <- cpu[!cpu$JOUR%in%c("samedi", "dimanche"),]

## acp :
acp2 <- PcaClassic(app300W.semaine, scale = TRUE, k = 4)

## on vérifie que les axes factoriels sont de bons prédicteurs :
acp.regression <- lm(cpu.semaine$CPU ~ acp2@scores[, "PC1"] +
                    acp2@scores[, "PC2"] +
                    acp2@scores[, "PC3"] + acp2@scores[, "PC4"])

summary(acp.regression)

## rotation varimax :
matrice.cor <- cor(app300W.semaine)

acp.varimax <- principal(r=matrice.cor, nfactors=4, rotate="varimax")

## tracé de la figure du corps de thèse : transactions et activité métier
par(mfrow = c(1,1))
plot(acp.varimax$loadings[,c("RC1", "RC3")], xlim=c(-1,1), ylim=c(-1,1),
     xlab = "ACP2 - axe factoriel n°1",
     ylab = "ACP2 - axe factoriel n°3")
abline(0,0)
abline(v=0)
grid(9,lwd=1, col="1")

varAx <- acp.varimax$loadings[,c("RC1", "RC3")]

text(varAx["CHECKPPIDS",1], varAx["CHECKPPIDS",2],
     labels="CHECKPPIDS", col = 'blue', cex = 0.8)

text(varAx["C2_IUIAPEXECUTE",1], varAx["C2_IUIAPEXECUTE",2],
     labels="C2_IUIAPEXECUTE", col = 'blue', cex = 0.8)

text(varAx["C2_EAPGETBESTPPID",1], varAx["C2_EAPGETBESTPPID",2],
     labels="C2_EAPGETBESTPPID", col = 'blue', cex = 0.8)

text(varAx["UA_CHECKLOTPROCESSABILITY",1], varAx["UA_CHECKLOTPROCESSABILITY",2],
     labels="UA_CHECKLOTPROCESSABILITY", col = 'blue', cex = 0.8)

text(varAx["C2_JOBPREPLOTTRULE",1], varAx["C2_JOBPREPLOTTRULE",2],
     labels="C2_JOBPREPLOTTRULE", col = 'blue', cex = 0.8)

text(varAx["C2_EAPMLJOBINLOTTRULE",1], varAx["C2_EAPMLJOBINLOTTRULE",2],
     labels="C2_EAPMLJOBINLOTTRULE", col = 'blue', cex = 0.8)

text(varAx["QRYUSERRULE",1], varAx["QRYUSERRULE",2],
     labels="QRYUSERRULE", col = 'blue', cex = 0.8)

text(varAx["UA_LOTPROCESSABILITYFOREQP",1],
     varAx["UA_LOTPROCESSABILITYFOREQP",2]-0.05,
     labels="UA_LOTPROCESSABILITYFOREQP", col = 'blue', cex = 0.8)

text(varAx["C2_EAPLOGPROCESSTIMERULE",1], varAx["C2_EAPLOGPROCESSTIMERULE",2]+0.02,

```

```

labels="C2_EAPLOGPROCESSTIMERULE", col = 'blue', cex = 0.8)

text(-0.75, -0.5, labels=bquote(bold(."Activités liées
au mode de production
automatique"))))

text(0.75, -0.5, labels=bquote(bold(."Activités liées
au mode de production
manuel"))))

text(-0.2, 0.9, labels=bquote(bold(."Transports et
traitements des lots"))))

legend("topleft", c("transaction 300Works", "activité industrielle sous-jacente"),
      text.col=c("blue", "black"),
      bg=NULL, bty = "n"
      )

### remarque : outil de visualition utilisé
### pour l'analyse des axes factoriels avec les experts :
var.load <- acp.varimax$loadings

score <- "RC1"
cormin <- 0.4

corel <- var.load[,score]
corel <- corel[corel > cormin | corel < -cormin]
corel <- sort(corel)

mycol <- colorRampPalette(c("green","red"))
par(mfrow = c(1,1))
par(oma=c(2,20, 0, 0), las = 2, cex = 1, las = 1)
barplot(corel, xlim = c(-1,1), col = mycol(length(corel)), horiz = TRUE, space = 0.4,
      main = "Loadings après rotation varimax")
box()
abline(v=c(-0.8, 0.8), col = 'red', lty = 2, lwd = 2)
abline(v=c(-0.5, 0.5), col = 'blue', lty = 2, lwd = 2)

#####
## Fin de l'analyse
#####

```



## Annexe 2

Chapitre 6. Application de la démarche de modélisation : création et formalisation de connaissances.

Section 1. Application à la migration MES : constitution d'un modèle de prévision.

1. Etape (B) intermédiaire d'observation n°2

2.2. Structuration des données - informations (aspect sémantique)

*Script R :*

```
#####
## Début de l'analyse
#####

#####
## Données
#####

cpu.model <- read.csv2("modélisation_cpu_operations.csv", dec = ".")
rownames(cpu.model) <- cpu.model[,1]
cpu.model[,1] <- NULL

#####
## Tracé de Y contre tous les xi (fig. du corps de thèse)
#####
par(mfrow = c(2,2))
plot(cpu.model$CPU ~ cpu.model$X1, xlab = "x1", ylab = "Y")
plot(cpu.model$CPU ~ cpu.model$X2, xlab = "x2", ylab = "Y")
plot(cpu.model$CPU ~ cpu.model$X3, xlab = "x3", ylab = "Y")
plot(cpu.model$CPU ~ cpu.model$X4, xlab = "x4", ylab = "Y")

#####
## Traitement des risques de colinéarité
#####

## corrélations entre variables exogènes :
cor(cpu.model)
## Selon la règle de Klein, on pourrait conserver les variables originales
summary(lm(CPU~.,cpu.model))

## Nouvelle variable X1+X2 :
X1X2 <- cpu.model$X1+cpu.model$X2
## Quasiment <=> à X4
cor(X1X2, cpu.model$X4)

## Orthogonalisation :
ortho <- lm(cpu.model$X3 ~ X1X2 + cpu.model$X2)
summary(ortho)

## variables finales :
cpu.model.corrige <- data.frame(CPU=cpu.model$CPU,
                               X1X2=X1X2,
                               X2=cpu.model$X2,
                               R=residuals(ortho),
                               Xwe=cpu.model$Xwe)

## les corrélations ne sont plus significatives :
cor(cpu.model.corrige)

#####
## Modélisation - modèle provisoire avec violation des hypothèses du modèle linéaire
```

```

#####

## estimation du modèle (moindres carrés ordinaires)
modell1 <- lm(CPU ~., cpu.model.corrige)
summary(modell1)

## résidus non iid :
acf(modell1$residuals)
pacf(modell1$residuals)

#####
## Modélisation - modèle final avec modélisation ARIMA des erreurs
#####

## estimation du modèle (maximum de vraisemblance)
modell2 <- arima(cpu.model.corrige[,1], order = c(1,0,0),
               xreg = cpu.model.corrige[,2:5],
               method = "ML", include.mean = FALSE)

modell2

## Figure présente dans le corps de thèse : vérification des erreurs
par(mfrow = c(2,2))
plot(modell2$residuals, type = 'p', xlab="temps", ylab="résidus",
     main="Tracé dans le temps")
abline(h=c(-modell2$sigma2^0.5*qnorm(0.95), modell2$sigma2^0.5*qnorm(0.95)),
       col='red', lty = 2)
abline(h=0, col='blue', lwd=2)

qqnorm(modell2$residuals, main = "Droite de Henry",
       xlab = "quantiles théoriques", ylab = "quantiles empiriques")
qqline(modell2$residuals, col = "red")

acf(modell2$residuals, main = "Corrélogramme des résidus",
    xlab="décalage", ylab="corrélation")
pacf(modell2$residuals, main = "Corrélogramme partiel des résidus",
    xlab="décalage", ylab="corrélation")

## Figure présente dans le corps de thèse : résidus contre prédicteurs

par(mfrow = c(2,2))
plot(modell2$residuals ~ cpu.model.corrige[,2], ylab = "résidus", xlab = "(x1 + x2)")
plot(modell2$residuals ~ cpu.model.corrige[,3], ylab = "résidus", xlab = "x2")
plot(modell2$residuals ~ cpu.model.corrige[,4], ylab = "résidus", xlab = "R")
plot(modell2$residuals ~ cpu.model.corrige[,5], ylab = "résidus", xlab = "xWE (variable
indicatrice)")

## Figure présente dans le corps de thèse : données réelles vs. valeurs ajustées
library(forecast)

par(mfrow = c(1,1))

plot(fitted(modell2) ~ cpu.model.corrige[,1],
     xlab = 'Données réelles', ylab = 'Valeurs ajustées', main = "")

abline(a=0,b=1, lwd = 2, col = 'blue')
abline(a= modell2$sigma2^0.5*qnorm(0.95),b=1,
      lty = 2, col = 'red')
abline(a= -modell2$sigma2^0.5*qnorm(0.95),b=1,
      lty = 2, col = 'red')
grid(9,lwd=1, col="1")

legend("topleft", c("Prédiction parfaite", "Intervalles de prévision 5%"),
      lty=c(1,2), lwd = c(2,1), bty="n",
      col = c('blue', 'red'))

MAE.model <- mean(abs(fitted(modell2)-cpu.model.corrige[,1]))

```

```

#####
## Validation du modèle final : validation croisée
#####

## initialisation
D <- ts(cpu.model.corrige)

k <- 50 # taille de l'échantillon
n <- nrow(D)

model <- model2

### Cas 1 : augmentation de séquentielle de la période de modélisation (xshort) et prédiction
### sur la période suivante (xnext)

mae.arima <- matrix(NA,n-k-1)

for(i in 1:(n-k-1))
{
  xshort.G <- window(D, start = 1, end= k+i)
  xnext.G <- window(D, start= k+(i+1), end= n)
  fit2.G <- arima(xshort.G[,1], order = c(1,0,0), xreg = xshort.G[,2:5], method = "ML")
  fcast2.G <- forecast.Arima(fit2.G, xreg = xnext.G[,2:5])
  mae.arima[i] <- mean(abs(fcast2.G$mean-xnext.G[,1]))
}

### Cas 2 : modélisation sur une fenêtre mobile de taille fixe (xshort) et
### prédiction sur la période suivante (xnext)

maeMov.arima <- matrix(NA,n-k-1)

for(i in 1:(n-k-1))
{
  xshort.M <- window(ts(D), start= i, end= k+i)
  xnext.M <- window(ts(D), start= k+(i+1), end= n)
  fit2.M <- arima(xshort.M[,1], order = c(1,0,0), xreg = xshort.M[,2:5], method = "ML")
  fcast2.M <- forecast.Arima(fit2.M, xreg = xnext.M[,2:5])
  maeMov.arima[i] <- mean(abs(fcast2.M$mean-xnext.M[,1]))
}

## tracé de la figure dans le corps de thèse :
par(mfrow=c(1,2))
temp <- seq(from=1, by=10, to=(n-k-1))
lab <- seq(from=(k+1), by=10, to=(n-1))

plot(ts(mae.arima), col = '1', lwd = 1, axes = FALSE,
      ylab = "Erreur absolue moyenne de prévision",
      xlab = "Horizon de prévision",
      main = "Prévision avec estimation séquentielle")
axis(1, at = temp, labels = (n-1)-lab)
axis(2)
box()
grid(9,lwd=1, col="1")

plot(ts(maeMov.arima), col = '1', lwd = 1, axes = FALSE,
      ylab = "Erreur absolue moyenne de prévision",
      xlab = "Horizon de prévision",
      main = "Prévision avec estimation glissante")
axis(1, at = temp, labels = (n-1)-lab)
axis(2)
box()
grid(9,lwd=1, col="1")

#####
## Validation du modèle final : stabilité structurelle

```

```
#####

## initialisation
library(strucchange)

## redimensionnement des données pour plus de stabilité :
D <- transform(D,
               CPU = CPU/1e4,
               X1X2 = X1X2/1e2,
               X2 = X2/1e2,
               R = R/1e2)

## ajout des variables différés (strucchange ne fonctionne avec la fonction arima
## il faut donc calculer le modèle par moindres carrés ordinaires en intégrant
## les variables différées)
D$CPU1 <- c(NA, head(D$CPU, -1))
D$X1X21 <- c(NA, head(D$X1X2, -1))
D$X21 <- c(NA, head(D$X2, -1))
D$R1 <- c(NA, head(D$R, -1))

D <- ts(D)

## Figure présente dans le corps de thèse : test avec 2 statistiques
par(mfrow = c(1,2))

## calcul processus de fluctuation empirique :
sc <- gefp(CPU~., fit = lm, data = window(D, start = 2))

## test de Nyblom-Hansen :
sc1 <- sctest(sc, functional = meanL2BB)
plot(sc, functional = meanL2BB,
     main=paste("Test de Nyblom-Hansen, ", "p-value =", round(sc1$p.value, 2)),
     ylab="Processus de fluctuation empirique", xlab="Temps")

## test SupLM :
sc2 <- sctest(sc, functional = supLM(0.1))
plot(sc, functional = supLM(0.1),
     main=paste("Test SupLM, ", "p-value =", round(sc2$p.value, 2)),
     ylab="Processus de fluctuation empirique", xlab="Temps")

legend("topleft", c("seuil de rejet de H0"),
      lty = 1, col = 'red', bty="n")

#####
## Fin de l'analyse
#####
```

## Annexe 3

Chapitre 6. Application de la démarche de modélisation : création et formalisation de connaissances.

Section 1. Application à la migration MES : constitution d'un modèle de prévision.

1. Etape (C) d'aide à la décision

3.2. Structuration des données - informations (aspect sémantique)

*Script R :*

```
#####
## Début de l'analyse
#####

#####
## Données
#####

cpu.modelD <- read.csv2("modélisation_cpu_décision.csv", dec = ",")
rownames(cpu.modelD) <- cpu.modelD[,1]
cpu.modelD[,1] <- NULL

#####
## Tracé de Y contre tous les Xi (fig. du corps de thèse)
#####
par(mfrow = c(2,2))
plot(cpu.modelD$CPU ~ cpu.modelD$X1, xlab = "X1", ylab = "Y")
plot(cpu.modelD$CPU ~ cpu.modelD$X2, xlab = "X2", ylab = "Y")
plot(cpu.modelD$CPU ~ cpu.modelD$X3, xlab = "X3", ylab = "Y")

#####
## Modélisation
#####

model.D <- lm(CPU~.-1, cpu.modelD)
summary(model.D)

## Figure présente dans le corps de thèse : vérification des erreurs
par(mfrow = c(2,2))
plot(model.D$residuals, type = 'p', xlab="temps", ylab="résidus",
      main="Tracé dans le temps")
abline(h=c(-summary(model.D)$sigma*qnorm(0.95),
           summary(model.D)$sigma*qnorm(0.95)),
       col='red', lty = 2)
abline(h=0, col='blue', lwd=2)

qqnorm(model.D$residuals, main = "Droite de Henry",
       xlab = "quantiles théoriques", ylab = "quantiles empiriques")
qqline(model.D$residuals, col = "red")

acf(model.D$residuals, main = "Corrélogramme des résidus",
     xlab="décalage", ylab="corrélation")
pacf(model.D$residuals, main = "Corrélogramme partiel des résidus",
     xlab="décalage", ylab="corrélation")

## Figure présente dans le corps de thèse : résidus contre prédicteurs
par(mfrow = c(2,2))
plot(model.D$residuals ~ cpu.modelD$X1, ylab = "résidus", xlab = "X1")
plot(model.D$residuals ~ cpu.modelD$X2, ylab = "résidus", xlab = "X2")
plot(model.D$residuals ~ cpu.modelD$X3, ylab = "résidus", xlab = "X3")
```

```

#####
## Risque de colinéarité
#####

## corrélation des variables explicatives :
cor(cpu.modelD[,2:ncol(cpu.modelD)])

## Analyse du risque colinéarité :
library(car)
mod.analyse <- lm(CPU~.,cpu.modelD) ## l'intercept est nécessaire pour le calcul
VIF <- vif(mod.analyse) ## OK, <4, et <10

## Régressions axiliaire : part de la variabilité (Ri) de chacune des variables
## explicatives expliquée par les autres variables explicatives
## Problématique si  $R_i^2 > R^2$  global
TOL <- 1/VIF
Ri <- 1-TOL ## OK, <0.85

## Régression PLS : voir TANAGRA (résultats numériques vérifiés)

#####
## Visualisation du modèle
#####

## Figure présente dans le corps de thèse : valeurs réelles, prédites et intervalles
## de prévision
p <- predict(model.D, interval = "pred")
low <- p[,2]
high <- p[,3]
fitted <- model.D$fitted.values

N <- 1:nrow(model.D$model)

par(mfrow=c(1,1), las = 1)
matplot(N,cbind(low, high), type = "n",xlab = "",
ylab = names(model.D$model[1]), axes = TRUE)

grid(9,lwd=1, col="1")

polygon(c(N,rev(N)),c(low,rev(high)), col = "grey70")
lines(model.D$model[,1], lty = 2)
lines(fitted, lwd = 2)

legend("topleft", fill = "grey70",
c("intervalles de prévision"),
xjust=1)

legend("topright",
c("données réelles","valeurs prédites"),
lty = c(2,1), lwd = c(1,2),
xjust=1)

#####
## Validation du modèle d'aide à la décision : validation croisée
#####

## Initialisation
library(DiceEval)

Y <- model.D$model[,1]

X <- data.frame(model.D$model[,2:ncol(model.D$model)])
colnames(X) <- c(paste( "X", c(1:ncol(X)), sep=""))

formula <- as.formula(paste("Y~", paste (paste("X", c(1:ncol(X)), sep=""),
collapse="+"),sep=""))

```

```

model.CV <- modelFit(X,Y, type = "Linear", formula = formula)

## Nombre de groupes :
K = 3

## Validation croisée :
CV <- crossValidation(model.CV ,K=K)

## Figure présente dans le corps de la thèse : visualisation de la validation croisée
par(mfrow = c(1,2))

plot(as.ts(Y), lty = 2, xlab = "", ylab = "", axes = FALSE,
main = paste("Valeurs prédites avec validation croisée ",K," groupes",sep=""))
lines(as.ts(model.CV$model$fitted.values), lwd = 2, col = 'red')
lines(as.ts(CV$Ypred), type = 'b', lwd = 2, pch = 2, col = 'blue')
axis(1,at = N, labels = N)
mtext(side = 1, las = 1, text = "Temps", line = 3.5)
axis(2)
box()
grid(9,lwd=1, col="1")

legend("bottomleft",
      c("données réelles", "valeurs prédites", "valeurs prédites par validation croisée"),
      lwd = c(1,2,2), lty = c(2,1,1), pch = c(NA,NA,2), col = c('black', 'red', 'blue'),
      cex = 1, bty="n")

robCV <- testCrossValidation(model.CV)
grid(9,lwd=1, col="1")
mean(robCV)

#####
## Validation du modèle d'aide à la décision : stabilité structurelle
#####

## initialisation
library(strucchange)

## redimensionnement des données pour plus de stabilité :
D <- transform(cpu.modelD,
              CPU = CPU/1e5,
              X1 = X1/1e4,
              X2 = X2/1e3,
              X3 = X3/1e1)
D <- ts(D)

## Figure présente dans le corps de thèse : test avec 2 statistiques

par(mfrow = c(1,2))

## calcul processus de fluctuation empirique :
sc <- gefp(CPU~., fit = lm, data = window(D, start = 2))

## test de Nyblom-Hansen :
sc1 <- sctest(sc, functional = meanL2BB)
plot(sc, functional = meanL2BB,
      main=paste("Test de Nyblom-Hansen, ", "p-value =",round(sc1$p.value,2)),
      ylab="Processus de fluctuation empirique",xlab="Temps")

## test SupLM :
sc2 <- sctest(sc, functional = supLM(0.1))
plot(sc, functional = supLM(0.1),
      main=paste("Test SupLM, ", "p-value =",round(sc2$p.value,2)),
      ylab="Processus de fluctuation empirique",xlab="Temps")

legend("topleft", c("seuil de rejet de H0"),
      lty = 1, col = 'red', bty="n")

```

```
#####  
## Fin de l'analyse  
#####
```



## Annexe 4

### Chapitre 7. Usage du modèle de prévision pour l'aide à la décision

#### Section 2. Utilisation du modèle d'évaluation : résultat et recommandation pour l'aide à la décision.

##### 1. Simulation et recommandation

*Script R :*

```
#####
## Début de l'analyse
#####

#####
## Données
#####

## Données de modélisation
cpu.modelD <- read.csv2("modélisation_cpu_décision.csv", dec = ",")
rownames(cpu.modelD) <- cpu.modelD[,1]
cpu.modelD[,1] <- NULL

## données de simulation
simuD <- read.csv2("simulation_2.csv")

#####
## Initialisation
#####

## Constitution du modèle et récupération de la variance résiduelle :
model.D <- lm(CPU~.-1, cpu.modelD)

sigma.model <- summary(model.D)$sigma

## Nombre de répétitions
N <- 100000

#####
## Simulations
#####

## Initialisation
init <- predict(model.D, simuD[1,])

## Scénario optimiste
sc.O <- predict(model.D, simuD[2,])

simul.O <- matrix(ncol = 1, nrow = N)

simul.O[c(1:N),] <- (sc.O + sigma.model * rnorm(N))/
                  (init + sigma.model * rnorm(N))

## Scénario réaliste
sc.R <- predict(model.D, simuD[3,])

simul.R <- matrix(ncol = 1, nrow = N)

simul.R[c(1:N),] <- (sc.R + sigma.model * rnorm(N))/
                  (init + sigma.model * rnorm(N))

## Scénario pessimiste
sc.P <- predict(model.D, simuD[4,])
```

```

simul.P <- matrix(ncol = 1, nrow = N)

simul.P[c(1:N),] <- (sc.P + sigma.model * rnorm(N))/
                    (init + sigma.model * rnorm(N))

## Scénario implicite
sc.I <- predict(model.D, simuD[5,])

simul.I <- matrix(ncol = 1, nrow = N)

simul.I[c(1:N),] <- (sc.I + sigma.model * rnorm(N))/
                    (init + sigma.model * rnorm(N))

#####
## Figure du corps de thèse : résultat des simulations
#####

fig <- boxplot(data.frame(simul.O,simul.R,simul.I,simul.P), range = 1.5,
               col = c("green", "blue", "orange","red"),
               names = c("j = O", "j = R", "j = I", "j = P"),
               xlab = "scénarios", ylab = "estimation du facteur d'accroissement (simulation)")
abline(h=2.5, col = "red", lty = 2)
grid(9)
legend("bottomright", "estimation des experts en 2009", col = "red", lty = 2, bg = NULL,
      bty = "n")

#####
## Fin de l'analyse
#####

```

## Annexe 5

Chapitre 6. Application de la démarche de modélisation : création et formalisation de connaissances.

Section 1. Application à la migration MES : constitution d'un modèle de prévision.

1. Etape (B) intermédiaire d'observation n°2

2.2. Structuration des données - informations (aspect sémantique)

et

Chapitre 8. Evolution de la connaissance organisationnelle.

Section 2. Illustration de l'application du maintien à niveau de la connaissance.

Identification de la rupture de la stabilité structurelle du modèle de prévision

*Script R :*

```
#####
## Début de l'analyse
#####

#####
## Données
#####

cpu.model <- read.csv2("modelisation_cpu_operations_monitor.csv", dec = ".")
rownames(cpu.model) <- cpu.model[,1]
cpu.model[,1] <- NULL

## ajout des variables différés (strucchange ne fonctionne pas avec la fonction arima
## il faut donc calculer le modèle par moindres carrés ordinaires en intégrant
## les variables différées)
cpu.model$CPU1 <- c(NA, head(cpu.model$CPU, -1))
cpu.model$X11 <- c(NA, head(cpu.model$X1, -1))
cpu.model$X21 <- c(NA, head(cpu.model$X2, -1))
cpu.model$X31 <- c(NA, head(cpu.model$X3, -1))
cpu.model$X41 <- c(NA, head(cpu.model$X4, -1))

#####
## Contrôle dynamique / 1 (contrôle OK période de stabilité structurelle, chapitre 6
#####

## modèle
model <- CPU ~ X1 + X2 + X3 + X4 + Xwe + CPU1 + X11 + X21 + X31 + X41

library(strucchange)

## Initialisation du test sur la période historique
## avec la statistique OLS-CUSUM :
init <- mefp(model, type="OLS-CUSUM",
             data = window(ts(cpu.model), start = 1, end = 163))

## Contrôle des 7 prochains jours
moni <- monitor(init, data = window(ts(cpu.model), start = 1, end = 170))

## Figure du corps de thèse : carte de contrôle
par(mfrow = c(2,1))
```

```

plot(moni)

## Initialisation du test sur la période historique
## avec la statistique estimation récursive ces coefficients :
init <- mefp(model, type="RE",
             data = window(ts(cpu.model), start = 1, end = 163))

## Contrôle des 7 prochains jours
moni <- monitor(init, data = window(ts(cpu.model), start = 1, end = 170))

## Figure du corps de thèse : carte de contrôle
plot(moni)

#####
## Contrôle dynamique / 2 (contrôle NOK : instabilité structurelle, chapitre 8
#####

# Initialisation du test sur la période historique + 7 jours
## avec la statistique OLS-CUSUM :
init <- mefp(model, type="OLS-CUSUM",
             data = window(ts(cpu.model), start = 1, end = 170))

## Contrôle des 7 prochains jours
moni <- monitor(init, data = window(ts(cpu.model), start = 1, end = 177))

## Figure du corps de thèse : carte de contrôle
par(mfrow = c(2,1))
plot(moni)

## Initialisation du test sur la période historique + 7 jours
## avec la statistique estimation récursive ces coefficients :
init <- mefp(model, type="RE",
             data = window(ts(cpu.model), start = 1, end = 170))

## Contrôle des 7 prochains jours
moni <- monitor(init, data = window(ts(cpu.model), start = 1, end = 177))

## Figure du corps de thèse : carte de contrôle
plot(moni)

#####
## Fin de l'analyse
#####

```

## **Annexe 6**

*pp. suivantes* : version de prépublication de E. Padonou, O. Roustant, M. Lutz, Robust monitoring of an industrial IT system in presence of structural change (2013 – soumis).

# Robust monitoring of an industrial IT system in presence of structural change

Esperan Padonou\*, Olivier Roustant\*, Michel Lutz†

February 22, 2013

## Abstract

This paper presents an original research initiated by the monitoring needs of a semiconductor production plant. The industrial operations rely on an Information Technology (IT) system, and several time series data are controlled statistically. Unfortunately, these variables often contain outliers, as well as structural changes due to external decisions in the IT activity. As a consequence, it has been observed that the monitoring results obtained with standard techniques could be severely biased.

This paper presents some contributions to overcome such difficulties. A new monitoring method is proposed, based on robust Holt-Winters smoothing algorithm, and coupled with a relearning procedure for structural breaks detection. Such a method is flexible enough for a large-scale industrial application. We evaluate its performances through simulations studies, and show its usefulness in industrial real applications for univariate and multivariate time series. The scope of application deals with IT activity monitoring, but the introduced statistical methods are generic enough for being used in other industrial fields.

*Keywords:* Monitoring, Control charts, Holt-Winters, Robust statistics, Structural changes

---

\*Ecole Nationale Supérieure des Mines, EMSE-FAYOL, CNRS UMR6158, LIMOS, F-42023 Saint-Etienne, France

†STMicroelectronics - Ecole Nationale Supérieure des Mines, PIESO/UMR 5600

# 1 Introduction

Manufacturing activities efficiency relies increasingly on Information Technologies (IT). This is clearly true for wafer fabrication plants, where this research is ongoing through a partnership with the company STMicroelectronics. Several activities are useful for a proper IT management (see eg. Rudd *et al.*<sup>18</sup>). This paper is focused on the monitoring activity. A careful monitoring is based on a cautious observation of the IT system. The objective is to have a close look on all data recorded and stored to track the activity of a plant IT system.

As the size and complexity of IT systems strongly increase, the viability of monitoring can only be ensured by employing automated procedures as mentioned by Dugmore *et al.*<sup>5</sup> Usual monitoring is based on threshold exceeding detection (Rudd *et al.*<sup>18</sup>). However, facing the enormous diversity and expanse of the components of modern IT systems, IT experts do not have always enough knowledge or time to determine a priori the critical thresholds. Consequently, automated monitoring procedures that do not require any preliminary hypothesis about the monitored variables, are required. In this case, the monitoring activity requires the ability to distinguish normal system operations from exceptional events.

To quantitatively detect such abnormal behaviors in the fluctuation of a time series, a standard procedure is the following: 1) Model the time series to be monitored; 2) Monitor model residuals, through an appropriate control chart (see eg. Montgomery<sup>13</sup>, Croux *et al.*<sup>4</sup>). This procedure is known in IT management studies. By way of example, Hellerstein<sup>8</sup> has developed an approach based on ARIMA modelling, whereas Brutlag<sup>3</sup> and Leikis<sup>11</sup> have recourse to Holt-Winters smoothing. This paper focuses on the latter. Holt-Winters smoothing is well adapted in industrial contexts, where time series have to be modeled without any preliminary analysis. Moreover, it is a fairly good approximation for many kinds of time series, accurate enough for industrial applications (see eg. Makridakis *et al.*<sup>12</sup>, Croux *et al.*<sup>4</sup>). Furthermore, we observed that the Holt-Winters model is validated for most of the time series encountered in our research field.

However, a monitoring based on the usual Holt-Winters algorithm revealed several limitations at STMicroelectronics. That is why we developed some methods to overcome these difficulties. This paper provides further details about their implementation in such a real industrial background. First of all, we observed that the usual Holt-Winters monitoring procedure could be severely biased by the presence of

outlier observations. Consequently, we implemented a robust approach, as proposed by Gelper *et al.*<sup>6</sup> and Croux *et al.*<sup>4</sup> Nonetheless, we noticed that the efficiency of this robust Holt-Winters control is seriously affected when structural changes occur. That is why we developed a dynamical procedure, which detects structural breaks in a time series and reinitializes the robust smoothing parameters accordingly. This dynamical procedure is an original contribution to References<sup>4,6</sup>, contributing to improve the flexibility for applications in industrial fields. Finally, it should be mentioned that an industrial monitoring implies a simultaneous control of several dozens of variables. There are numerous approaches in multivariate process control (see e.g. Bersimis *et al.*<sup>2</sup> for a review), and a specific investigation of robust techniques in this context is beyond the scope of the paper. Alternatively, we show how the robust controlling procedure can be adapted to the framework of multivariate process control.

This paper is organized as follows. In Section 2, we introduce the main statistical methods that have been deployed to control the STMicroelectronics IT system: The robust Holt-Winters monitoring proposed by Gelper *et al.*<sup>6</sup> and Croux *et al.*<sup>4</sup>; The dynamical contribution developed through our researches, improving its flexibility in changing environments. In Section 3, some performance tests are introduced. They are grounded on quantitative simulations to: 1) Compare the usual and robust Holt-Winters monitoring procedures; 2) Evaluate the structural break detection capacity of the new dynamical procedure. Lastly, Section 4 introduces some example from real industrial cases: Univariate and multivariate applications are considered.

## 2 Investigation of a robust monitoring for trended time series with structural changes

We recall here the main principles of Holt-Winters based monitoring, and its robust version introduced by References<sup>4,6</sup>. Then, we propose a new methodology in presence of structural changes.

### 2.1 Monitoring based on Holt-Winters smoothing

The Holt-Winters (HW) algorithm is a popular technique used to provide short-term forecasts of a given time-series (see e.g. Makridakis *et al.*<sup>12</sup>). The predictions are built iteratively as a linear combination of the observed values and the prediction



obtained at last step.

For illustration, let us consider a time series  $y$ , observed at dates  $1, 2, \dots, n - 1$ . For the sake of simplicity, we assume that  $y$  is non-seasonal, though the methodology is similar. The HW algorithm is based on the assumption that  $y$  is a sum of two time-series  $\alpha$  and  $\beta$  corresponding respectively to a local level (order of magnitude) and a trend. These auxiliary time series are estimated iteratively as averages of the last observation and the last predictions, weighted by two parameters  $\lambda_1$  and  $\lambda_2$ :

$$\hat{\alpha}_t = \lambda_1 y_t + (1 - \lambda_1) \hat{y}_{t|t-1}, \quad t = 1, \dots, n - 1 \quad (1)$$

$$\hat{\beta}_t = \lambda_2 (\hat{\alpha}_t - \hat{\alpha}_{t-1}) + (1 - \lambda_2) \hat{\beta}_{t-1}, \quad t = 1, \dots, n - 1 \quad (2)$$

Logically, the one-step-ahead forecast done at date  $t - 1$  for date  $t$  is then given by:

$$\hat{y}_{t|t-1} = \hat{\alpha}_{t-1} + \hat{\beta}_{t-1}, \quad t = 1, \dots, n \quad (3)$$

which gives in particular the prediction at date  $n$ . In practice,  $\lambda_1$  and  $\lambda_2$  are estimated by minimizing a criterion (often the least-square criterion) based on the forecast errors:

$$E_t = y_t - \hat{y}_{t|t-1}, \quad t = 1, \dots, n - 1 \quad (4)$$

and the algorithm is initialized by a linear regression on the first  $m$  values.

As a second step, monitoring can be performed. While applying a control chart to  $y$  is not recommended, due to the violations of the usual assumption (*identically and independently distributed data*) especially if  $y$  is trended, the forecast errors  $E_t$  may be close to satisfy it. Then, assuming furthermore that  $E_t$  are normally distributed  $N(0, S^2)$  the upper and lower control limits for  $E_t$  are then given by:

$$UCL = +q_{\alpha/2} * \hat{S} \quad (5)$$

$$LCL = -q_{\alpha/2} * \hat{S} \quad (6)$$

where  $q_{\alpha/2}$  is the quantile of a Student distribution at level  $\alpha/2$ , and  $\hat{S}^2$  is the usual variance estimator:

$$\hat{S}^2 = \frac{1}{n - m - 1} \sum_{t=m+1}^{n-1} E_t^2 \quad (7)$$

These limits intend to detect the dates that correspond to an anomaly: A value of  $E_t$  outside the interval [UCL, LCL] should be a strong indication of an abnormal behavior (for a given confidence level  $\alpha$ ). However, the limits themselves are sensitive to outliers, since the variance estimator overestimates the true variance in presence

of outliers. Furthermore, the predicted value  $\hat{y}_{t|t-1}$  depends linearly on past values that may contain outliers. These problems are solved by the robust version of the Holt-Winters monitoring.

## 2.2 Robust monitoring based on Holt-Winters smoothing

The Robust Holt-Winters algorithm (RHW) introduced by Gelper *et al.*<sup>6</sup> and Croux *et al.*<sup>4</sup> considers two additional auxiliary time series:  $y^*$ , representing a *cleaned* proxy of  $y$  after outliers treatment, and  $\sigma$  the expected prediction error, representing a robust estimate of the forecast error  $E_t$ . To obtain a robust algorithm, large values are truncated when larger than a given threshold. More precisely, the expected errors  $\sigma_t$  are computed recursively by:

$$\hat{\sigma}_t^2 = \lambda_\sigma \left[ \psi_k \left( \frac{E_t}{\hat{\sigma}_{t-1}} \right) \right]^2 \hat{\sigma}_{t-1}^2 + (1 - \lambda_\sigma) \hat{\sigma}_{t-1}^2 \quad (8)$$

where  $\lambda_\sigma$  is a given weight, and  $\psi_k$  is the Huber function with boundary value  $k$ :

$$\psi_k(x) = \begin{cases} x & \text{if } |x| \leq k, \\ \text{sign}(x) \times k & \text{if } |x| > k \end{cases} \quad (9)$$

The error  $E_t$  is still given by  $E_t = y_t - \hat{y}_{t|t-1}$ , with  $\hat{y}_{t|t-1} = \hat{\alpha}_{t-1} + \hat{\beta}_{t-1}$ , but the local level and trend are now estimated by using the cleaned time series  $y^*$ :

$$\hat{\alpha}_t = \lambda_1 y_t^* + (1 - \lambda_1) \hat{y}_{t|t-1} \quad (10)$$

$$\hat{\beta}_t = \lambda_2 (\hat{\alpha}_t - \hat{\alpha}_{t-1}) + (1 - \lambda_2) \hat{\beta}_{t-1} \quad (11)$$

where  $y_t^*$  is given by:

$$y_t^* = \psi_k \left( \frac{E_t}{\hat{\sigma}_t} \right) \times \hat{\sigma}_t + \hat{y}_{t|t-1} \quad (12)$$

Notice that the role of the Huber function  $\psi$  is to truncate the forecast errors  $E_t$  when larger than  $k$  times the expected prediction error  $\hat{\sigma}_t$ :

$$\psi_k \left( \frac{E_t}{\hat{\sigma}_t} \right) \times \hat{\sigma}_t = \begin{cases} E_t & \text{if } |E_t| < k \hat{\sigma}_t, \\ \text{sign}(E_t) \times k \hat{\sigma}_t & \text{if } |E_t| > k \hat{\sigma}_t \end{cases} \quad (13)$$

For instance, Equation (12) can be rewritten in a simpler way:

$$y_t^* = \begin{cases} y_t & \text{if } |E_t| < k \hat{\sigma}_t, \\ \text{sign}(E_t) \times k \hat{\sigma}_t + \hat{y}_{t|t-1} & \text{if } |E_t| > k \hat{\sigma}_t \end{cases} \quad (14)$$

Finally the parameters  $\lambda_1$  and  $\lambda_2$ , and the standard deviation  $S$  of the forecast errors are computed with robust procedures, while  $\lambda_\sigma$  is fixed to 0.3 (see References<sup>4,6</sup> for more details). The algorithm initialization is also done in a robust way by repeated median regression. For that, we recommend using a period of length 7: This is short enough to assume a local linear trend and long enough to be resistant to 2 outliers.

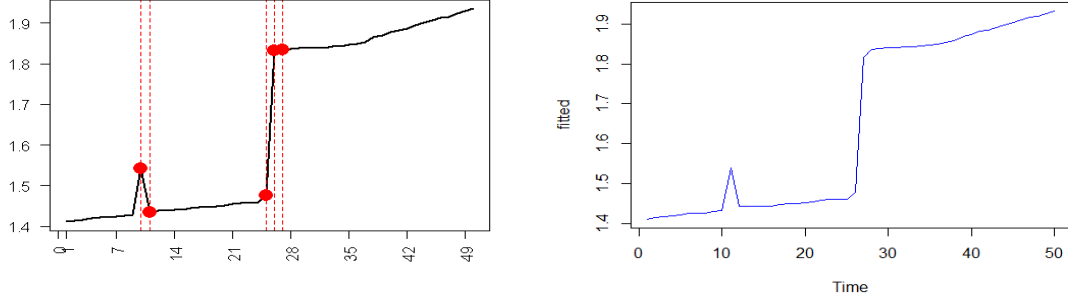
Since the RHW smoothing is a full robust procedure, the control charts based on the errors  $E_t$  (see Section 2.1) are now resistant to outliers, which is a clear improvement to the (non-robust) HW-based monitoring. However, an adaptation is necessary in presence of a structural change.

### 2.3 Robust monitoring with structural changes

In this section, we focus on time series that possibly contain outliers and structural changes. To detect structural changes in parametric models, three main classes exist (Zeileis<sup>22</sup>): F statistics, fluctuation tests and maximum likelihood scores. We choose to use a common and simple F statistics, the Chow test, since it can be easily adapted to a robust framework. It consists in splitting the sample into 2 groups: The first one before the break date and the second one after. The model parameters are estimated for both of them so that an F test be performed to judge whether they are equal or not. The Chow test is easy to use but restricted by 2 limitations. The first one is mentioned by Hansen<sup>7</sup>: "An important limitation of the Chow test is that the break date must be known a priori". Moreover, the exact number of changes is unknown. The second problem is a question of robustness: A break may be missed or falsely detected because of outliers. In this section, we show how the robust Holt-Winters smoothing allows to deal with these problems and deduce a strategy for structural change monitoring.

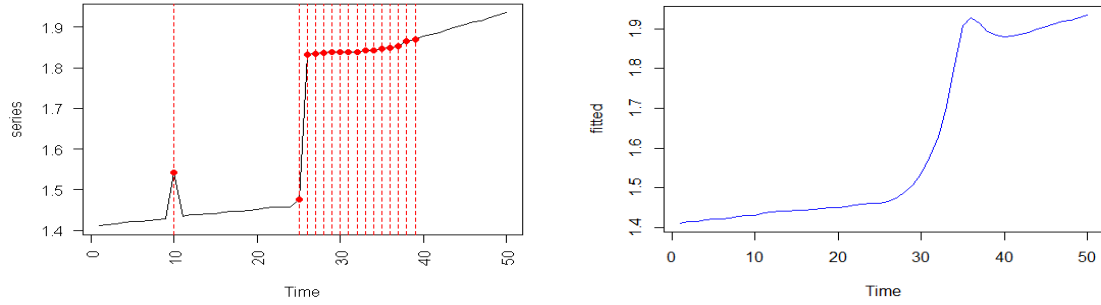
**An introducing example** Consider the time series in Figure 1 with an outlier at date 10 and a break at date 26. As expected, the robust algorithm is not sensitive to the outlier at date 10 contrarily to the classical smoothing: This is its main advantage. But after that, the level changes suddenly at date 26. The robust algorithm does not admit this modification quickly and many false alarms follow. Nevertheless, this specificity can be used to detect break dates. Indeed, when a structural change happens, there is a quite long period (here 26—34) of successive false alarms corresponding to successive large errors in the RHW smoothing. During this period, the predicted values  $\hat{y}_{t|t-1}$  seem to exhibit a deterministic pattern (Figure 1, d): Actually, we show below (Figure 2 and Proposition 1) that they match exactly with an

exponential function. These two facts strongly suggest that the periods of successive large errors given by the robust Holt-Winters smoothing are useful for break dates detection.



(a) Alarms given by the classical HW forecasting

(b) Predicted values by the classical HW smoothing



(c) Alarms given by the robust HW forecasting

(d) Predicted values by the robust HW smoothing

Figure 1: A real time series with outlier and structural change

### 2.3.1 Structural change and consecutive large errors of the robust Holt-Winters smoothing

When a structural break occurs in practice, there is a sequence of consecutive false alarms due to a succession of large errors. The robust smoothing enables to quantify the importance of these errors by comparing them to their predicted values. Let us call *relative error* the ratio  $\frac{E_t}{\hat{\sigma}_t}$ . A succession of large values of this ratio is a

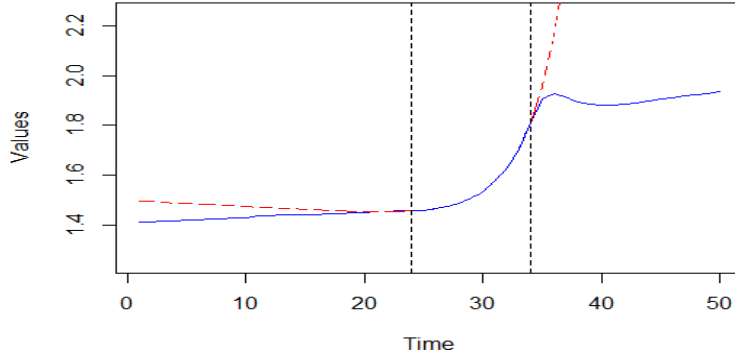


Figure 2: Zoom on Figure 1 (d): The predicted values (solid line) coincide with a function  $f$  of the form  $f(t) = ar^t + bt + c$  (dotted lines) during the period  $[26-34]$ .

forewarning sign of structural change. More formally, we call  $t_1$  a *suspicious date* for structural change if there exists an integer  $p \geq 3$  such that:

$$\psi_k \left( \frac{E_{t_1}}{\hat{\sigma}_{t_1}} \right) = \psi_k \left( \frac{E_{t_1+1}}{\hat{\sigma}_{t_1+1}} \right) = \dots = \psi_k \left( \frac{E_{t_1+p-1}}{\hat{\sigma}_{t_1+p-1}} \right) = k \quad (15)$$

or

$$\psi_k \left( \frac{E_{t_1}}{\hat{\sigma}_{t_1}} \right) = \psi_k \left( \frac{E_{t_1+1}}{\hat{\sigma}_{t_1+1}} \right) = \dots = \psi_k \left( \frac{E_{t_1+p-1}}{\hat{\sigma}_{t_1+p-1}} \right) = -k \quad (16)$$

The period  $[t_1, \dots, t_1 + p - 1]$  is a *suspicious period*: a period when forecasting errors remain  $k$  times higher than their expected values.

**Proposition 1** During a suspicious period  $[t_1, \dots, t_1 + p - 1]$ , the predictions  $\hat{y}_{t|t-1}$  of the robust HW smoothing are given by a deterministic and monotonic function  $f$  which does not depend on any observation posterior to date  $t_1$ . Its form is:

$$f(t) = \frac{r^{(t+1)}}{(r-1)^2} + a.t + b \quad (17)$$

with

$$r = \sqrt{\lambda_\sigma(k^2 - 1) + 1} = 1.378405 \quad (18)$$

$a$  and  $b$  are given by  $\hat{y}_{t_1}$  and  $\hat{y}_{t_1-1}$

**Proof** See Appendix 1.

### 2.3.2 A new methodology for structural change monitoring

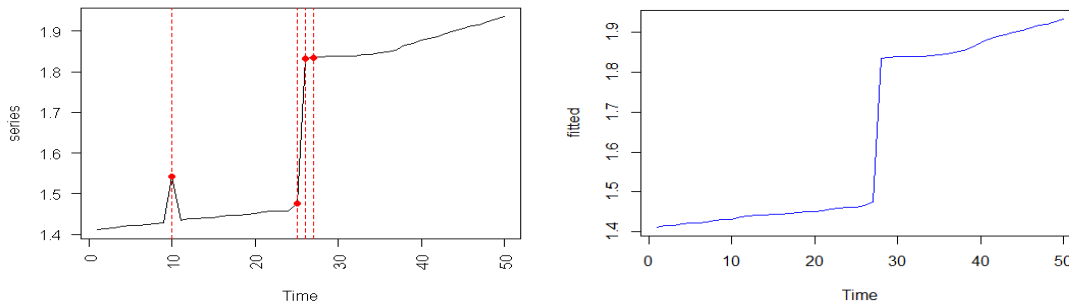
The previous observations and result about the RHW smoothing in case of a structural change suggest the following methodology that we call RHW-SC:

1. Find the suspicious dates by looking at consecutive relative errors given by the robust HW smoothing (Equations 15 and 16)
2. Apply a robust version of the Chow test to the suspicious dates detected in 1. One robust version of the Chow test consists in replacing the usual linear regression by the repeated median regression (Siegel<sup>19</sup>, Rousseeuw *et al.*<sup>17</sup>) and using a robust estimator of the residuals' variance (Croux *et al.*<sup>4</sup>) to compute the F statistic.

This strategy tackles the two main issues mentioned at the beginning of this section: All the possible break dates (number and locations) are automatically detected by the algorithm itself, and for a given date the statistical test for structural change is done in a robust way. In practice the methodology is applied dynamically, and when the robust Chow test is positive, the robust HW smoothing is reinitialized.

#### **Performance of the RHW-SC methodology on the introducing example**

To solve the problem raised by the introducing example, we perform dynamically the RHW-SC methodology with  $p=3$ . The methodology has detected only one suspicious date, namely day 26, which indeed corresponds to the structural change. Notice that only one Chow test has been used contrarily to exhaustive methods such as Quandt-Anderson that systematically test all dates. We reinitialized the robust HW smoothing at date 26, by using the repeated median regression coefficients. As a result, RHW-SC methodology remains resistant to the outlier at date 10 without generating a long sequence of false alarms after the break.



(a) Alarms given by the RHW-SC methodology (b) Predicted values with the RHW-SC methodology

Figure 3: Alarms and predicted values by the robust version with structural change detection

### 3 Global performance tests

Some ARIMA models correspond to exponential smoothing methods (see e.g. Hyndman *et al.*<sup>10</sup>). In particular, a Holt-Winters forecasting with smoothing parameters  $(\lambda_1, \lambda_2)$  is equivalent to an ARIMA  $(0, 2, 2)$  model of parameters  $(\theta_1, \theta_2)$  if:  $-1 \leq \theta_2 \leq 1$ ,  $\theta_2 - \theta_1 \leq 1$  and  $\theta_2 + \theta_1 \leq 1$ , with :

$$\lambda_1 = 1 + \theta_1 \quad (19)$$

$$\lambda_2 = \frac{1 - \theta_1 - \theta_2}{1 + \theta_2} \quad (20)$$

Thus, we first use ARIMA  $(0, 1, 1)$  time series with parameter  $\theta = 0.5$ . Indeed, they correspond to the special case where  $\theta_2 = 0$ . Next, we perform our simulations on ARIMA  $(0, 2, 2)$  models with  $(\theta_1, \theta_2) = (1, -0.25)$ . These parameters have been estimated from an industrial time series by assuming that it comes from an ARIMA  $(0, 2, 2)$  model. In this section, we compare first the performances of the RHW and the HW smoothing statically and dynamically. Afterwards, we evaluate the ability of the RHW-SC method to detect a structural change.

#### 3.1 Comparison tests in a static setting

We are interested in comparing 2 characteristics of the HW and the RHW methods:

1. The *power*: The probability that an outlying observation is detected
2. The *false detection rate*: The probability that a normal observation is detected. This risk is called *size* of the control chart when there is no outlier. Below, we use 0.95 as a confidence level, so the false detection rate is expected to be 0.05.

### 3.1.1 Simulation results for ARIMA (0, 1, 1) time series

We use the following strategy:

1. Generation of a time series: simulate an ARIMA (0, 1, 1) time series of length 160 with parameters  $\theta = 0.5$ . The first 60 values serve as a training sample (including  $m = 7$  values for HW and RHW initializing).
2. Generation of outliers: For a fixed contamination rate  $R$ , choose randomly  $160R$  dates among the 160 dates. Contaminate these observations by adding or subtracting to them a value  $e$ . Whether it is an addition or a subtraction is chosen at random. Two cases are considered for  $e$ :
  - (a)  $e$  is a fixed value among 10, 20 or 30
  - (b) "Mix":  $e$  is chosen at random uniformly among 10, 20, 30
3. Perform the HW and the RHW methods and estimate:
  - (a) Their *power*: as the percentage of the contaminated observations that are really detected
  - (b) Their *false detection rate*: as the percentage of non contaminated observations that are detected.
4. Repeat 100 times the steps 1 to 3.

The RHW and HW monitoring are equivalent when there is no outlier (Reference<sup>4</sup>). Their differences become significant when the data is contaminated.

On the one hand, Table 1 and Table 2 below show that the HW control chart is subject to two effects. The first one is the widening of its control limits by outliers, which tends to reduce abnormally its false alarms rate. The second one concerns the dates following these outliers: the corresponding predictions are biased and this fact tends to raise the false alarms rate. Here, the first effect is predominant. This explains the too low false alarms risk and the poor detection rate of the HW method.



On the other hand, Table 3 and Table 4 show that the RHW methodology outperforms the HW smoothing especially when the number of outliers raises and when the magnitude of these outliers is not fixed: which is a realistic case. In fact, the RHW control chart doesn't lose neither its power, nor the stability of its false alarms rate that remains stable around the theoretical value 5%.

Table 1: HW detection rate

	R = 2%	R = 5%	R = 10%
e = 10	93.5	70.0	31.0
e = 20	94	74	35
e = 30	91	72	36
Mix	72	55	33

Table 2: HW false alarms rate

	R = 2%	R = 5%	R = 10%
e = 10	3.3	3.6	2.1
e = 20	1.7	3.0	7.5
e = 30	1.9	6.0	10
Mix	3	5.3	4

Table 3: RHW detection rate

	R = 2%	R = 5%	R = 10%
e = 10	100	100	99.2
e = 20	100	100	99.9
e = 30	100	100	100
Mix	100	100	99.8

Table 4: RHW false alarms rate

	R = 2%	R = 5%	R = 10%
e = 10	5.2	4.8	6.1
e = 20	5.2	5.2	6.4
e = 30	5.2	4.8	7.1
Mix	5.1	4.9	6.1

### 3.1.2 Simulation results for ARIMA (0, 2, 2) time series

The same strategy as in the precedent section is used. Only the first step is modified to generate ARIMA (0, 2, 2) models with  $(\theta_1, \theta_2) = (1, -0.25)$ . The results are summarized below (Tables 5, 6, 7 and 8). They show that the RHW control chart remains better in term of detection. Nevertheless, its false risk becomes higher than expected even if it remains stable enough for a fixed contamination rate contrarily to the HW smoothing.

Table 5: HW detection rate

	R = 2%	R = 5%	R = 10%
e = 10	97.6	91.3	67.0
e = 20	100	92.2	74.4
e = 30	97.1	91.0	73.3
Mix	92.7	77.3	57.5

Table 6: HW false alarms rate

	R = 2%	R = 5%	R = 10%
e = 10	3.3	2.7	3.7
e = 20	2.1	4.1	8.0
e = 30	5.2	1.5	4.0
Mix	2.2	3.9	6.4

Table 7: RHW detection rate

	R = 2%	R = 5%	R = 10%
e = 10	99.8	99.0	83.0
e = 20	100	99.9	93.5
e = 30	100	99.9	95.0
Mix	99.7	99.3	91.5

Table 8: RHW false alarms rate

	R = 2%	R = 5%	R = 10%
e = 10	6.4	7.2	9.0
e = 20	7.1	7.1	8.5
e = 30	6.2	7.2	9.7
Mix	7.0	7.4	9.0

### 3.2 Comparison tests in a dynamical setting for ARIMA (0, 2, 2) time series

In practice, industrial variables are tracked daily; the smoothing parameters and control limits are re-estimated every day. So, we perform dynamical simulations that consist in using the same procedure as for the static setting but with updating the smoothing parameters and controls limits at each iteration.

Table 9: HW detection rate

	R = 2%	R = 5%	R = 10%
e = 10	99.6	95.4	80
e = 20	100	93.6	80
e = 30	100	97	78.6
Mix	95	70	52

Table 10: HW false alarms rate

	R = 2%	R = 5%	R = 10%
e = 10	4.6	4.3	4.2
e = 20	3.3	2.2	2
e = 30	2.5	2.2	2.4
Mix	3.0	3.2	2.0

Table 11: RHW detection rate

	R = 2%	R = 5%	R = 10%
e = 10	100	100	99.7
e = 20	100	100	100
e = 30	100	100	100
Mix	100	100	100

Table 12: RHW false alarms rate

	R = 2%	R = 5%	R = 10%
e = 10	6.3	6.8	8.6
e = 20	6.0	7.1	9.1
e = 30	6.2	6.9	9.3
Mix	6.1	7.0	7.2

Tables 11 and 12 show how suitable the HW monitoring is to update its smoothing parameters and control limits contrarily to the RHW method. Indeed, the HW detection rate is drastically improved even if the power of the RHW smoothing remains the better. Let's remark that the HW false alarm risk has raised and approaches better the theoretical value 5%. As for the RHW false alarms risk, it has decreased

to approach this same theoretical value; but this later improvement is less obvious. However, the RHW false alarms risk has become very stable and non dependant on the magnitude of outliers, which is not the case for the HW control chart. Anyway, the RHW control chart outperforms the HW one when the scale of the outliers is more realistically (randomly) chosen.

### 3.3 Structural change detection performance tests

To evaluate the RHW-SC methodology, we use the following strategy:

1. Generation of a time series: simulate an ARIMA (0, 2, 2) time series of length 160 with parameters  $(\theta_1, \theta_2) = (1, -0.25)$ . The first 60 values serve as a training sample (including  $m = 7$  values for RHW initializing).
2. Generation of outliers: For a fixed contamination rate  $R = 0.05$ , choose randomly  $160R$  dates among the 160 dates. Contaminate these observations by adding or subtracting to them a fixed value  $e = 5$ . Whether it is an addition or a subtraction is chosen at random.
3. Generation of structural changes: choose randomly 1 date among the 100 last values. From this date to the last one, add a linear function:  $A * DATE + B$  to the time series.
4. Perform the RHW-SC methods with 0.95 as confidence level for the robust Chow test and compute:
  - (a) The detection rate for structural changes
  - (b) The false detection rate for structural changes.
5. Repeat 1000 times the steps 1 to 4.

Table 13: Detection rate

	B = 0	B = $\pm 50$
A = 0	-	95
A = $\pm 50$	86	95

Table 14: False detection rate

	B = 0	B = $\pm 50$
A = 0	-	6
A = $\pm 50$	4	5

The RHW-SC methodology is efficient enough for structural change detection especially when a shift happens. It loses in efficiency when there is no shift. But in

these latest cases, the RHW period of successive false alarms is short. The RHW-SC false detection rate corresponds to the theoretical value expected since the confidence level used for the Chow test is 95%.

## 4 Applications

The RHW and the RHW-SC methods are performed daily on many indicators of the information system of a company. In general, the results are satisfactory. Here, we present some of these examples and a multivariate case.

### 4.1 Examples of univariate time series

This first example is illustrative of structural changes detection by the RHW-SC methodology.

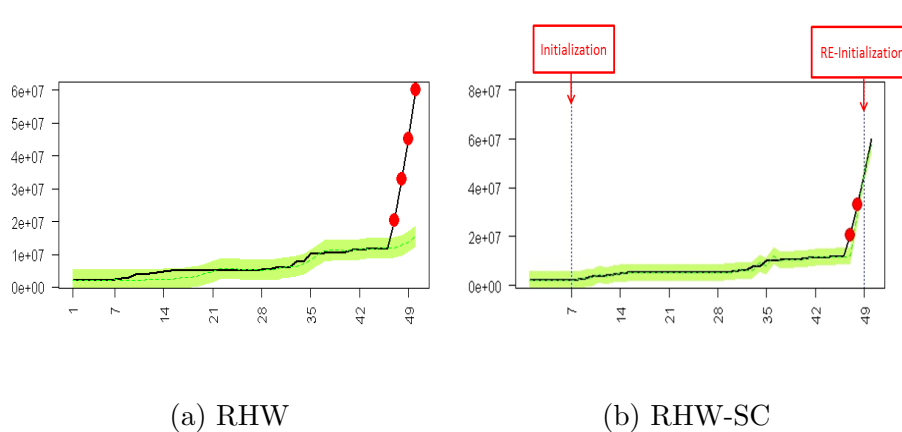


Figure 4: An industrial time series with a structural change

The time series of Figure 4 is subject to one slope change at date 47, which leads to four false alarms when the RHW methodology is performed (4a). This change is detected 3 days later by the RHW-SC monitoring, whence a reinitialization and reduction of the number of false alarms. The second example shows an extension of the RHW smoothing to seasonal time series (Reference<sup>6</sup>)

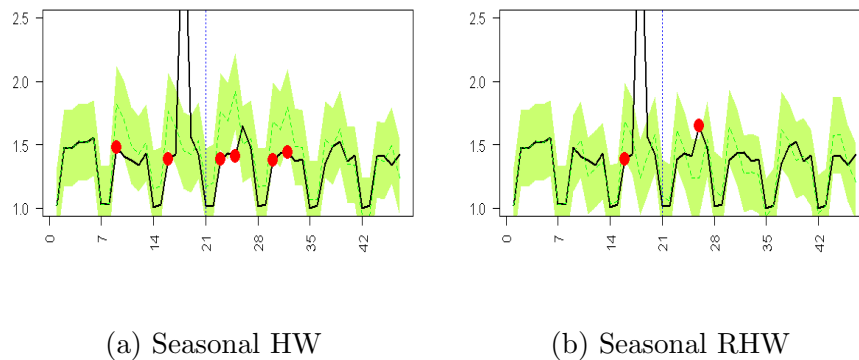


Figure 5: An industrial and seasonal time series

Figure 5 points out the sensitivity of the HW method to outliers. The abnormal observation of the third week has deformed the seasonal component. Because of the overestimation of this seasonality, the 26th observation that is really aberrant is not detected. Moreover, it leads to four false alarms the two following weeks. These problems no longer exist with the seasonal RHW smoothing.

## 4.2 The multivariate case

In our company, several indicators are tracked daily. So, monitoring them separately leads to false alarms every day. This is foreseeable given that the theoretical false alarms risk tends to 100% when the number of independent variables approaches 100: Whence the necessity to perform a multivariate monitoring. Notice that there are a lot of possible approaches in the multivariate framework (see e.g. Bersimis *et al.*<sup>2</sup>), and our aim is neither to do a comparative study nor to propose a best one. Rather, we show how the new robust methodology can be adapted to the multivariate case. Thus, as an example, we constitute groups of variables. For each group of  $p$  variables we use the following strategy :

1. Perform the RHW smoothing for each variable of the group.
2. Use a robust Hotelling T2 control chart to analyze simultaneously the  $p$  vectors of residuals given by step 1.

Among existing robust Hotelling T2 control charts (see e.g. Rousseeuw,<sup>15</sup> Alfaro *et al.*<sup>1</sup>) we consider here the computation of the confidence ellipsoid with the Minimum Covariance Determinant criterion introduced by Rousseeuw *et al.*<sup>15,16</sup> . There

are two identified difficulties: A poor orientation of the confidence ellipsoid and an underestimation of its size. To face these problems, we use recent solutions found in the literature. Firstly, the orientation is improved by choosing a subset of the 75 % *best points* for the MCD criterion instead of 50% (Huber *et al.*<sup>9</sup>). Secondly, the size estimation is improved by using two correcting factors: One asymptotic factor of consistency to the chi square distribution (Rousseeuw *et al.*<sup>17</sup>) and one empirical result for small samples (Pison *et al.*<sup>14</sup>).

Now, we present an example with a group of  $p = 4$  variables. Each day, the chi square statistic is computed for this group. The result is plotted in Figure 6. This control chart has successfully detected the 6 most important outliers but does not say where the problems come from. This is the well known problem of multidimensional *out of controls* interpretation. As an example, let us focus on the alarm on date 56. Its cause can be known by looking at each variable separately. Then, Figure 7 shows that the variable BIN2 is mainly responsible for it.

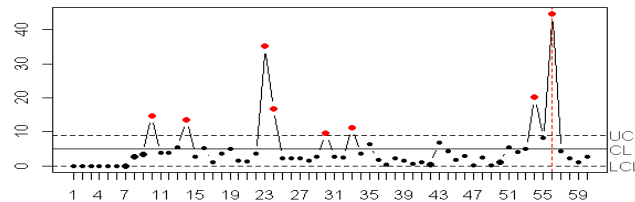
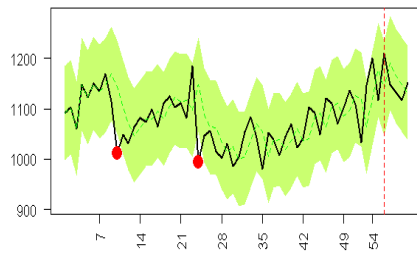
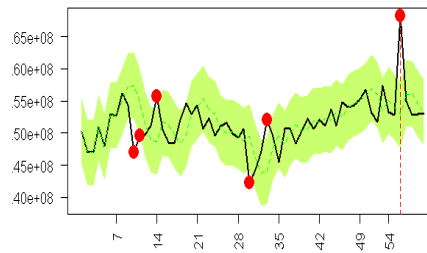


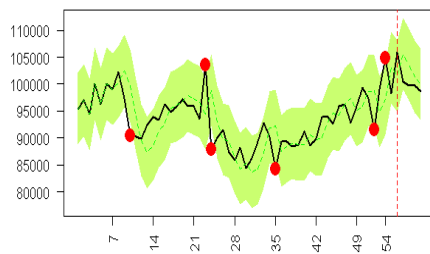
Figure 6: Chi squared daily statistics for a group of 4 variables



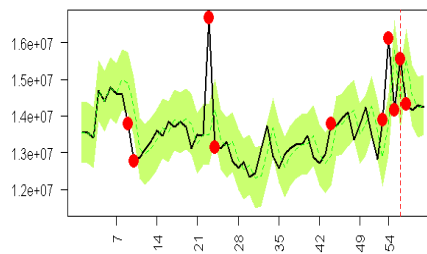
(a) Variable BIN1



(b) Variable BIN2



(c) Variable BIN3



(d) Variable BIN4

Figure 7: RHW smoothing for the four variables of the group

Nevertheless, this solution is not realistic for an industrial use because it produces too many graphs that need to be examined. Among the numerous existing criteria to interpret multivariate signals, the partial relative contributions mentioned by Montgomery<sup>13</sup> are very popular for their efficiency. In Figure 7, these relative contributions at date 56 confirm the influence of BIN2.

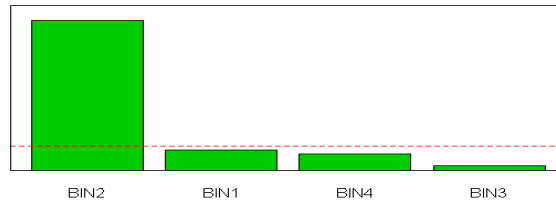


Figure 8: Partial contributions to variability at date 56

## 5 Conclusion

In this paper, we have studied a monitoring based on robust Holt-Winters smoothing, as proposed by Gelper *et al.*<sup>6</sup> and Croux *et al.*<sup>4</sup> Based on an industrial application of this method at STMicroelectronics and simulation studies, we can confirm its high robustness. However, as poor results were noticed in case of structural changes, we have also proposed and evaluated an improved dynamical approach, for a better integration of changing environments. The efficiency of this robust and dynamical method has been demonstrated on real univariate and multivariate STMicroelectronics case studies. This contribution has been developed as an improvement for the STMicroelectronics IT system monitoring, but it could be applied to any other industrial applications, where time dependent variables has to be statistically controlled.

For further research, there are several interesting outstanding questions. Firstly, some tuning choices could be investigated. Hence, the truncation parameter was set to  $k = 2$  and the smoothing parameter  $\lambda_\sigma$  to 0.3 (still proposed by References<sup>4,6</sup>). These values could be discussed: Do they always provide optimal results or should they be contextually adapted? Secondly, when structural changes occur, we decided to reinitialize the Holt-Winters parameters. An alternative could be to use dynamical smoothing parameters that may change over time, as proposed by Williams<sup>21</sup> or Taylor<sup>20</sup>. Finally, in our methodology, we chose a Chow test to detect structural breakpoints, since it is easily adapted to the robust framework. Nonetheless, as mentioned earlier, other approaches exist (other F statistics, fluctuation tests, maximum likelihood scores). Their potential of application in our monitoring procedure and a comparative study or their performance may bring to interesting further insights.



## References

1. J.L. Alfaro and J.F. Ortega. A robust alternative to Hotelling's T2 control chart using trimmed estimators. *Quality and Reliability Engineering International*, 24(5):601–611, 2008.
2. S. Bersimis, S. Psarakis, and J. Panaretos. Multivariate statistical process control charts: An overview. *Quality and Reliability Engineering International*, 23(5):517–543, 2007.
3. J.D. Brutlag. Aberrant Behavior Detection in Time Series for Network Monitoring. *Proceedings of the 14th Systems Administration Conference (LISA 2000)*, 2000.
4. C. Croux, S. Gelper, and K. Mahieu. Robust control chart for time series data. *Expert Systems with Applications*, 38(11):13810–13815, 2011.
5. J. Dugmore and S. Lacy. *Capacity management*. British Standards Institution, 2005.
6. S. Gelper, R. Fried, and C. Croux. Robust forecasting with exponential and Holt-Winters smoothing. *Journal of Forecasting*, 29(3):285–300, 2010.
7. B. Hansen. The new econometrics of structural change: dating breaks ins US. Labor Productivity. *Journal of Economic Perspectives*, 15:117–128, 2001.
8. J.M. Hellerstein. Quantitative Data Cleaning for Large Databases. United Nations Economic Commission for Europe (UNECE), 2008.
9. M. Hubert and K. Van Driessen. Fast and robust discriminant analysis. *Computational Statistics and Data Analysis*, 45:301–320, 2004.
10. R.J. Hyndman, M.L. King, I. Pitrun, and B. Billah. Local Linear Forecasts Using Cubic Smoothing Splines. *Australian and New Zealand Journal of Statistics*, 47(1):87–99, 2005.
11. C.A. Leikis. Consolidated Capacity and Performance Reporting. In *Int. CMG Conference*, pages 527–534, 2007.
12. S. Makridakis, S. Wheelwright, and R.J. Hyndman. *Forecasting, methods and applications*. Wiley, 3rd edition, 1998.

13. D.C. Montgomery. *Statistical Quality Control: A Modern Introduction*. John Wiley & Sons, 2012.
14. G. Pison, S. Van Aelst, and G. Willems. Small Sample Corrections for LTS and MCD. *Metrika*, 55:111–123, 2002.
15. P.J. Rousseeuw. Least Median of Squares Regression. *Journal of American Statistical Association*, pages 871–880, 1984.
16. P.J. Rousseeuw and K. Van Driessen. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, 41:212–223, 1998.
17. P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outlier Detection*. Wiley, 1987.
18. C. Rudd and V. Lloyd. *Service Design, Itil, Version 3*. Stationery Office Books, 2007.
19. A.F. Siegel. Robust regression using repeated medians. *Biometrika* 69, pages 242–244, 1982.
20. J.W. Taylor. Smooth transition exponential smoothing. *Journal of Forecasting*, 23:385–394, 2004.
21. T.M. Williams. Adaptive Holt-Winters forecasting. *Journal of Operational Research Society*, 38(6):553–560, 1987.
22. A. Zeileis. A Unified Approach to Structural Change Tests Based on ML Scores,  $F$  Statistics, and OLS Residuals. *Econometric Reviews*, 24:445–466, 2005.

## Acknowledgements

The authors would like to thank the company STMicroelectronics for giving the opportunity to work on this research project, and especially X. Ambrosioni and P. Palazon, for their support. We also acknowledge the participants of the ENBIS-12 conference (Ljubljana, 9-13 September 2012), for their relevant remarks on our first oral presentation of this work.

## Appendix

We prove the Proposition presented in Section 2.3.1: If  $t_1, t_1 + 2, \dots, t_1 + p - 1$  is a suspicious period, then the sequence  $\hat{y}_{t_1}, \hat{y}_{t_1+1}, \dots, \hat{y}_{t_1+p-1}$  is given by a deterministic and monotonic function  $f$  which does not depend on any observation posterior to date  $t_1$ .

### Proof

The period  $t_1, t_1 + 1, \dots, t_1 + p - 1$  is defined by either Equation 15 or Equation 16. Without loss of generality, consider Equation 15. By combining this definition 15 and Equation 8, we obtain the equation:

$$\hat{\sigma}_t^2 = (1 + \lambda_\sigma(k^2 - 1))\hat{\sigma}_{t-1}^2 \quad t_1 + 1 \leq t \leq t_1 + p \quad (21)$$

which shows that the predicted errors  $\hat{\sigma}_t$  follow a geometric progression with common ratio,

$$r = \sqrt{1 + \lambda_\sigma(k^2 - 1)} \quad (22)$$

Consequently, by denoting  $t_0 := t_1 - 1$ , the date before  $t_1$ , the predicted errors are given by:

$$\hat{\sigma}_t = r^{t-t_0}\hat{\sigma}_{t_0} \quad t_1 + 1 \leq t \leq t_1 + p \quad (23)$$

With the common choice  $k = 2, \lambda_\sigma = 0.3$ , we have  $r \approx 1.378 > 1$ . Thus, the expected error goes increasing exponentially. Furthermore, the cleaned time series  $y_t^*$  becomes :

$$y_t^* = k.r^{t-t_0}\hat{\sigma}_{t_0} + \hat{y}_t \quad (24)$$

Now, relying on the equivalence with ARIMA(0,2,2) model (see e.g. Hyndman<sup>10</sup>), the forecast values of the Holt-Winters smoothing with parameters  $(\lambda_1, \lambda_2)$  follow the recursive scheme:

$$\hat{y}_{t+1} = (2 - \theta_1) y_t^* + \theta_1 \hat{y}_t - (1 + \theta_2) y_{t-1}^* + \theta_2 \hat{y}_{t-1} \quad (25)$$

where  $\theta_1$  and  $\theta_2$  are the parameters of the corresponding ARIMA(0, 2, 2) model (See Equations 19 and 20). Given Equation 24, this scheme becomes:

$$\hat{y}_{t+1} - 2\hat{y}_t + \hat{y}_{t-1} = P(t) \quad (26)$$

with:

$$P(t) = k[(2 - \theta_1)r - (1 + \theta_2)]r^{(t-1-t_0)}\hat{\sigma}_{t_0} \quad (27)$$

This is a linear equation, whose solutions are given by the sum of the solutions of the homogeneous linear equation and a particular solution:

$$\hat{y}_t = f(t) = \frac{r^{t+1}}{(r-1)^2} + a.t + b \quad (28)$$

The values of the constants  $a$  and  $b$  are imposed by  $\hat{y}_{t_1}$  and  $\hat{y}_{t_1+1}$ . This shows indeed that the predicted values are purely deterministic and increasing exponentially during the suspicious period.



NNT : 2013 EMSE 0688

Michel LUTZ

INDUSTRIAL DECISION-AID SOCIO-STATISTICAL METHODS – Applied to the capacity management of an IS in the microelectronics industry

Speciality: Industrial Engineering

Keywords: decision-aiding, knowledge management, statistics and data analysis, information system and information technologies, capacity management, microelectronic industry

Abstract:

The exploitation of huge volumes of data stored in industrial systems is studied. A proper analysis of these data can provide valuable material for decision making. The research work presented deals with the following question: how can one convert raw data into useable information, to contribute to the knowledge management of an organization and improve its dynamic decision making? A structured decision-aid process, entitled socio-statistical, is proposed. It implies actors of the organization and use of formal methods. Firstly, the objective is to analyse and formalize decisional problems. Then, an appropriate decision-aid should be developed on the basis of exploratory and inferential statistics. Our generic methodology is applied to a specific management issue: capacity management of information technologies of a STMicroelectronics semiconductor production plant. This application raises a decision issue: managers have to ensure the right balance between infrastructure cost and service level offered to information system users. The goal is to demonstrate that the socio-statistical process may provide relevant support. This negates two managerial dilemmas, usually encountered when managing capacity: complexity of technological systems and incorporation of the business activity. Our application has been developed in the scope of the ITIL framework, well-known by industrial information systems experts. It will be shown how the process can build predictive models, which link the activity of hardware servers to the STMicroelectronics industrial activity. Methods are also proposed, to monitor daily the quality of these models, as well as the overall activity of the information system. This work helps at formalizing quantitatively organizational knowledge, facilitating its use in decisional processes, but also ensuring its positive change over time. We hope this research is laying some foundations for a broader exploitation of the data stored in modern manufacturing systems, through future development of decision-support systems and Big Data initiatives.

NNT : 2013 EMSE 0688

Michel LUTZ

METHODES SOCIO-STATISTIQUES POUR L'AIDE A LA DECISION EN MILIEU INDUSTRIEL – Application à la gestion des capacités d'un SI en industrie micro-électronique

Spécialité: Génie Industriel

Mots clefs : aide à la décision, gestion des connaissances, statistiques et analyse de données, système d'information et technologies de l'information, gestion des capacités, industrie microélectronique

Résumé :

Cette thèse pose la question de l'exploitation des larges volumes de données stockés par les systèmes industriels. Ces données offrent un matériau précieux pour soutenir les prises de décision. Il convient de pouvoir convertir ces données en connaissances, utilisables dans le cadre des processus décisionnels des acteurs d'une organisation. La thématique générale des travaux de recherches présentés est alors : comment opérer des transformations de données brutes en connaissances, pour contribuer au système de connaissances d'une organisation et améliorer son système décisionnel ? Un processus structuré d'aide à la décision, qualifié de socio-statistique, est proposé. Il implique les acteurs de l'organisation et l'emploi de méthodes formelles. Dans un premier temps, il vise à analyser et formaliser les problématiques décisionnelles. Ensuite, il permet de construire une aide la décision quantitative adaptée, sur la base de méthodes statistiques exploratoires et inférentielles. Cette méthodologie générique est appliquée dans cette thèse à un problème de gestion particulier : la gestion des capacités de l'infrastructure des technologies de l'information d'une usine de production de semi-conducteurs de la société STMicroelectronics. Ce cadre applicatif pose un problème décisionnel : les managers doivent assurer un juste équilibre entre le coût de l'infrastructure technologique et le niveau de service offert aux utilisateurs du système informatique. Cette thèse montre que le processus socio-statistique offre une aide pertinente. Il permet de surmonter deux enjeux de gestion, fréquents lors de la gestion des capacités : la complexité des systèmes technologiques et la prise en compte de l'activité métier de l'organisation. Situant ces travaux applicatifs dans le cadre du référentiel ITIL, bien connu dans le monde de l'informatique industrielle, il est montré comment l'application du processus permet de constituer des modèles prédictifs, qui mettent en relation l'activité des serveurs informatiques et l'activité industrielle de STMicroelectronics. Cette application permet aussi de contrôler dynamiquement la validité des modèles, ainsi que l'activité quotidienne du système de d'information. Ces travaux permettent de formaliser quantitativement des connaissances organisationnelles, d'en favoriser l'utilisation dans les processus décisionnels, mais aussi d'en assurer l'évolution dans le temps. Ces recherches nous semblent poser des fondations pour un recours plus généralisé à l'exploitation des données stockées par les systèmes de production contemporains, dans le cadre du développement de systèmes de support à la décision et de perspectives *Big Data*.