Aide à la décision dans la gestion des parcs de compteurs d'eau potable

Alberto Pasanisi

ENGREF

Ecole Nationale du Génie Rural, des Eaux et des Forêts

Alberto Pasanisi

Aide à la décision dans la gestion des parcs de compteurs d'eau potable

Thèse présentée pour obtenir le grade de Docteur de l'ENGREF en Sciences de l'Eau (Option Statistique).

Soutenue le 6 février 2004 devant le jury suivant :

M. Gilles Celeux (INRIA Futurs, président)
M. Eric Parent (ENGREF, directeur de thèse)
M. Jacques Bernier (ENGREF, rapporteur)
M. Furio Cascetta (Seconda Università di Napoli, rapporteur)
M. Jean-Pierre Raoult (Université de Marne-la-Vallée, rapporteur)
M. Pascal Arnac (Veolia Water, examinateur)
M. Dominique Olivier (Veolia Water, examinateur)

M. Bernard Brémond (CEMAGREF Bordeaux, invité)
M. Lucien Duckstein (ENGREF, invité)

Février 2004

Table des matières

Table des figures

Résumé

La métrologie des compteurs d'eau se dégrade au long de leur vie opérationnelle, entraînant, pour la plupart des compteurs actuellement utilisés en France, une sous-estimation du volume d'eau facturé. Ce phénomène est source de problèmes pour les distributeurs d'eau : il se traduit en un manque à gagner non négligeable et détermine une situation d'inégalité entre les usagers. En outre, une réglementation, de plus en plus exigeante, obligera bientôt les distributeurs à limiter la proportion d'appareils à métrologie imparfaite en dessous d'une valeur fixée. La planification des renouvellements des compteurs est, par conséquent, un problème complexe qui demande la mise en place d'une stratégie optimale.

N'importe quelle méthode de planification nécessite la connaissance préliminaire de la métrologie des compteurs en conditions réelles d'exploitation. Le but de cette thèse est de fournir des éléments utiles à la mise en place des règles de gestion optimale adoptées par la Compagnie Générale des Eaux.

L'étude de la dégradation de la métrologie se fait avec un modèle dynamique (markovien) à quatre états discrets à métrologie de plus en plus dégradée. Du point de vue statistique la particularité du problème vient du mécanisme d'observation des données. Le manque d'observations répétées sur les mêmes individus ne permet pas la mise en œuvre des méthodes d'estimation normalement employées pour cette classe de modèles. Les calculs d'inférence sont réalisés dans un cadre bayésien avec des techniques MCMC (Markov Chain Monte Carlo). Cette méthode d'estimation est une alternative, plus que valide, aux procédures basées sur la recherche du maximum de la vraisemblance sous contraintes.

Ensuite, on se pose le problème de la recherche de nouvelles variables explicatives, autres que l'âge et, sur la base de la provenance géographique des compteurs étalonnés et des données relatives aux phénomènes de blocage, on partage les sites français d'exploitation en trois groupes à "agressivité" croissante.

Finalement, on montre que le modèle est capable de fournir des prévisions directement utilisables par les décideurs : l'estimation du sous-comptage et de la probabilité de non-conformité, en fonction de l'âge, de l'agressivité du site et de la consommation annuelle.

Mots-clés : Compteurs d'eau, métrologie, planification optimale, modèles dynamiques, modèles hiérarchiques, inférence bayésienne, méthodes MCMC.

Abstract

Water meters' metrology become more and more inaccurate during their operating life, which originates, in particular for meters actually used in France, the under-estimation of accounted water. That is a source of problems for water distribution companies : it gives rise to significant financial losses and to an unequal billing policy.

Furthermore, a more and more severe national standard will soon oblige water companies to keep the rate of inaccurate devices below a fixed value. Thus, the planning of meters' renewal is a complicate problem which needs the determination of an optimal strategy.

Every management method needs firstly the preliminary knowledge of meters' metrology in real operating conditions. The goal of this PhD thesis is to give indications useful to apply optimal management rules, currently used by "Générale des Eaux" water distribution company.

Meters' degradation is studied by a (markovian) dynamic model, based on four discrete states, each one of which characterises a more and more inaccurate metrology. By the statistical point of view, the problem is a little peculiar, due to data observation mechanism. The lack of repeated observations on individuals does not allow to use inference methods which are normally employed to estimate this class of models. Inference calculations are made in a Bayesian framework by MCMC (Markov Chain Monte Carlo) techniques. This estimation method is a very attractive alternative to usual procedures based on the constrained maximisation of likelihood.

After that, we look for new explanatory factors, in addition to age, and using geographical localisation of sampled meters and local "stuck" meters observations we divide French operating sites in three groups of increasing "aggressiveness".

Finally we show as the statistical model can give forecasts which can be directly use by decision-makers : the estimation of unaccounted-for water and non-conformity probability as a function of age, aggressiveness, and annual consumption.

Key-words : Water meters, metrology, optimal planning, dynamic models, hierarchical models, Bayesian inference, MCMC methods.

Glossaire

BMN	Base Métrologique Nationale.
BUGS	Bayesian inference Using Gibbs Sampler.
CGE	Compagnie Générale des Eaux.
COFRAC	COmité FRançais d'ACcréditation.
CIFRE	Convention Industrielle de Formation par la REcherche
DAG	Directed Acyclic Graph.
DN	Diamètre nominale.
EMT	Erreurs Maximales Tolérées.
ENGREF	Ecole Nationale du Génie Rural, des Eaux et des Forêts.
GRESE	Gestion du Risque En Sciences de l'Environnement
ICBC	Info centre Compteurs, Branchements et Clients.
LECE	Laboratoire d'Essai des Compteurs d'Eau.
MCMC	Markov Chain Monte Carlo.
MDEL	Modèle Dynamique à Etats Latents.
$\mathbf{q}_{max}$	Débit maximal.
$\mathbf{q}_{min}$	Débit minimal.
$\mathbf{q}_n$	Débit nominal.
$\mathbf{q}_t$	Débit de transition.
RCI	Réseaux, Comptage et Investissements.
SEDIF	Syndicat des Eaux d'Ile de France.

Prologue

Toute thèse a une petite histoire. La mienne est un peu spéciale et je crois qu'elle mérite d'être racontée.

En avril 1999, chargé de recherche de l'Université de Naples 2, je travaillais sur la modélisation floue de la propagation de polluants dans les cours d'eau et les estuaires, et je contactais le Prof. Lucien Duckstein à son adresse électronique de l'Université de Tucson (Arizona), pour lui demander des conseils sur comment choisir l'algorithme le plus adapté à mon cas d'étude.

Le lendemain en lisant la réponse, pratiquement immédiate, à mon courrier je découvrais qu'elle venait de l'ENGREF et que mon interlocuteur ne se trouvait pas au pays des cactus, mais à Paris, ce qui tombait bien, parce que j'avais déjà programmé d'y passer une période de vacances quelques jours plus tard. Alors, une semaine après mon premier courrier, un peu intimidé par son impressionnante liste de publications, je rencontrais le Prof. Duckstein. De cette rencontre est née une collaboration étroite, dont les résultats sont résumés en un article publié sur le JORBEL (Revue Belge de Recherche Opérationnelle), grâce à laquelle j'ai eu l'occasion de passer quelques temps à l'ENGREF, de connaître mon futur laboratoire d'accueil et son directeur, le Dr. Eric Parent.

Un soir de novembre 1999, Lucien Duckstein me demandait si j'étais intéressé à commencer un doctorat à l'ENGREF et de là est parti mon premier projet de thèse concernant la modélisation de la qualité des eaux superficielles avec des méthodes floues. Malgré nos efforts et l'opinion encourageante des nombreux chercheurs rencontrés pour mieux définir le cadre de la recherche, le projet échouait en mai 2000 par le désistement de l'organisme qui aurait dû le financer et je mettais de côté mes projets de thèse.

Le 13 juillet 2000, je recevais un courrier de Lucien me décrivant un projet de modélisation de la fiabilité des compteurs d'eau qui pouvait donner lieu à une thèse, et me priant de contacter dans les plus brefs délais Eric Parent.

Ensuite tout est allé très vite : la rencontre avec Messieurs Pascal Arnac et Dominique Olivier de la Compagnie Générale des Eaux, le dossier CIFRE et le démarrage de ma thèse sous la direction d'Eric Parent, un an après ma discussion avec Lucien.

Merci Lucien ! Merci de m'avoir convaincu que j'avais le potentiel pour faire une thèse, merci pour être à l'origine de cette aventure et merci pour ton sou-

tien dans les quelques moments difficiles que j'ai vécus dans ces trois ans. Tes observations et tes remarques tout au long de mon travail m'ont beaucoup aidé et cette thèse, même si très lointaine de notre projet original, t'appartient.

Je tiens à adresser ma profonde gratitude à mon directeur de thèse Eric Parent que je remercie d'avoir accepté de diriger ma thèse, me connaissant à peine, et d'avoir fait un pari si important sur la base du prior informatif de Lucien. Merci Eric, d'avoir perturbé mon esprit et d'avoir frappé durement ma vision d'ingénieur d'un monde déterministe et gouverné par des équations différentielles, et merci d'avoir été toujours à mon écoute, d'avoir partagé tes connaissances, et de m'avoir guidé dans un domaine scientifique dont j'ignorais le charme.

Au laboratoire GRESE, j'ai eu le privilège de travailler dans un environnement incroyablement riche d'idées. Les discussions avec les collègues thésards et anciens thésards Sandrine Micallef, Billy Amzal, Antoine Penciolelli, Etienne Rivot, ont toujours été toutes aussi intéressantes que sympathiques.

Environ la moitié de ce travail de thèse a été réalisée à la Direction Tehnique de la Compagnie Générale des Eaux, et précisément au sein du Service "Réseaux, Comptage et Investissements" (RCI). Je remercie Pascal Arnac, responsable de ce Service, de m'avoir accueilli dans son équipe et pour les fructueux échanges d'idées que nous avons eu dans ces 3 ans. Sa profonde connaissance de la problématique technique et opérationnelle des compteurs a été très importante pour mener à bien les tâches qu'on s'était fixées.

J'ai profité, tout au long de ce travail de l'aide précieuse des experts du comptage de l'équipe RCI : François Paquet (initialement), Serge Lamandé dont l'expérience de la pratique technique a rapproché mes études à la réalité du terrain, Hervé Cleraux, qui de son laboratoire à Nancy n'a jamais arrêté de m'alimenter en résultats expérimentaux, Laurent Chicot qui m'a fourni les données concernant les phénomènes de blocage des compteurs. Je remercie aussi Jean-Paul Guillaume, Alain Boireau et Laurence De Beir pour l'aide qu'ils m'ont apportée sur l'interprétation de la variable "agressivité" des sites d'exploitation et les correspondants régionaux qui ont mis en application mes plans d'échantillonnage et ont permis la réalisation de cette étude.

Je tiens à remercier les professeurs Jacques Bernier, Furio Cascetta et Jean-Pierre Raoult pour l'honneur qu'ils me font à être rapporteurs de mon travail et pour la lecture particulièrement attentive qu'ils ont fait de ma thèse.

Un grand merci aussi aux membres du comité de suivi dont l'avis et le conseil m'ont été très précieux : Messieurs Gilles Celeux, Bernard Brémond et Dominique Olivier (qui aura assisté au début et à l'achèvement de ce travail).

Ma reconnaissance s'adresse aussi à Madame Claude Pingeon et Mademoiselle Céline Guillemet qui, avec ma femme Sandrine, sont parties à la chasse des erreurs d'orthographe. Si j'ai trop maltraité la langue de Molière, je l'ai fait sans le vouloir.

Merci à Fabienne. Vous savez que votre aide a été décisive et je n'ai pas

besoin d'en dire davantage.

Dulcis in fundo, je tiens à dédier ma thèse à ma femme Sandrine, qui a vécu avec moi tous les états d'âme à travers lesquels je suis passé au long de ces trois années. C'est dur de partager sa vie avec un thésard qui rentre tard tous les soirs et qui parfois reste accroché des journées entières à un problème de calcul ... A toi, "victime" malgré toi de la statistique, des méthodes Monte Carlo, et des compteurs d'eau, qui mériterais un doctorat avec mention en "*Patience Appliquée*" mes remerciements les plus profonds.

Chapitre 1

Introduction

Le comptage de l'eau est la façon la plus équitable et correcte pour rémunérer le service de distribution d'eau potable, selon le principe élémentaire que plus on consomme, plus on paie.

Il s'agit aussi d'un système efficace de limitation des consommations et des gaspillages d'une ressource de plus en plus chère.

Toutes les études concordent : l'introduction du comptage entraînerait une baisse substantielle des consommations qui permettrait aux collectivités locales de dimensionner les infrastructures sur la base des exigences réelles des usagers. Il s'agit d'un problème vital dans les pays industrialisés mais encore plus dans les pays en voie de développement où la connaissance du besoin réel en eau devrait permettre la correcte allocation des ressources économiques, humaines et politiques nécessaires pour garantir l'accès à l'eau, un droit fondamental de l'homme dont on a beaucoup parlé en 2003, "année internationale de l'eau douce".

On reviendra sur ces arguments pour les discuter dans les chapitres suivants. Dans cette introduction je veux surtout parler des motivations de cette thèse.

Ce travail naît d'un partenariat CIFRE entre la Direction Technique de la Compagnie Générale des Eaux (CGE) et le laboratoire de Gestion du Risque en Sciences de l'Environnement (GRESE) de l'Ecole Nationale du Génie Rural, des Eaux et des Forêts (ENGREF). Les conventions CIFRE (Convention Industrielle de Formation par la Recherche) rapprochent une entreprise qui propose un sujet de recherche industrielle et un organisme de recherche qui se charge de l'aspect méthodologique du travail, finalisé à la réalisation d'une thèse.

La CGE, véritable institution dans la vie économique française avec ses 150 ans, est le premier distributeur d'eau potable en France : elle fournit environ 2.1

milliards de mètres cubes chaque année à 26 millions d'usagers.

Le problème technique au cœur de cette thèse est la dégradation de la métrologie des compteurs. Les erreurs de mesure augmentent avec l'âge des dispositifs et cette dégradation entraîne, dans l'écrasante majorité des cas, un sous-comptage : le compteur enregistre un volume d'eau inférieur à celui effectivement consommé.

Les conséquences pour le distributeur d'eau sont multiples. L'eau non comptée n'est pas facturée et donc constitue un manque à gagner qu'on peut chiffrer à plusieurs dizaines de millions d'euros par an, par rapport à une situation idéale de comptage parfait, mais elle est aussi source d'inégalité entre les usagers (qui à consommation égale et à tarif égal peuvent avoir des factures différentes).

En outre, la nouvelle réglementation française qui instaurera la vérification périodique des compteurs en service, obligera de plus en plus les distributeurs à améliorer la qualité de leur parcs de compteurs, en fixant une limite à la proportion tolérée d'appareils à métrologie imparfaite.

Face à ce problème technique, la CGE (comme d'autres distributeurs) a mis en place des techniques de gestion optimale des renouvellements de compteurs. La mise en œuvre de ces méthodes nécessite la connaissance du comportement des compteurs en conditions réelles d'exploitation. Par exemple le calcul de l'opportunité strictement économique de remplacer un compteur en service par un compteur neuf se base sur l'évaluation préliminaire du manque à gagner engendré par le sous comptage en fonction de l'âge du dispositif (et éventuellement d'autres variables explicatives).

Si les stratégies de gestion ne sont pas directement mises en cause et, par volonté de la CGE, n'ont pas fait l'objet de cette étude, en revanche le distributeur a souhaité améliorer sa connaissance du phénomène de dégradation et, par conséquent, se doter des moyens de fiabiliser les hypothèses techniques à la base des dites stratégies. Il s'agit d'un problème complexe dont l'étude dépasse le cadre des techniques de statistique et analyse des données normalement employées dans les entreprises. Pour entreprendre un véritable travail de modélisation, la CGE a sollicité la compétence d'un organisme de recherche appliquée : le laboratoire GRESE de l'ENGREF.

Au cœur des préoccupations du GRESE se trouve l'analyse et la modélisation de la composante aléatoire de phénomènes complexes dans différents domaines d'application.

La complexité d'un phénomène l'éloigne parfois du domaine de compétence

des experts "naturels" pour le rapprocher de problèmes très différents.

Revenons aux compteurs. En théorie, l'étude des erreurs de mesure est du ressort de la mécanique de précision et de la mécanique des fluides et justement les fabricants de compteurs font appel à cette catégorie de spécialistes pour la conception et la réalisation de leurs produits. En revanche quand un compteur est employé sur le terrain pour la facturation, il est soumis à une série de sollicitations pratiquement non reproductibles en laboratoire : surpressions, arrêts, démarrages, régimes de fonctionnement très variables d'un usager à l'autre, interaction chimique avec l'eau circulante, passages de particules solides entrées accidentellement dans le réseau ou qui se sont détachées des conduites etc. Dans ce cadre toute prévision déterministe est impossible et le seul moyen possible pour étudier le phénomène est l'analyse statistique des données recueillies sur le terrain. L'analyse et la modélisation statistique se basent sur des méthodologies qui ne sont pas strictement liées à un domaine d'application spécifique et qui d'une certaine façon rassemblent des problèmes différents. La même technique de modélisation dynamique peut servir pour expliquer la dégradation des compteurs mais aussi la croissance des arbres, le déroulement des études d'un jeune universitaire et des phénomènes de précipitations pluviales !

Cette thèse est volontairement recentrée sur le problème technique. Je suis ingénieur et dans le cadre d'un doctorat en "*Sciences de l'eau*" j'ai essayé d'écrire un document qui s'adresse essentiellement à un "public" d'ingénieurs, même si la partie méthodologique est prédominante et développe des concepts de statistique et de mathématiques appliquées. Pour cette raison, (par exemple) j'ai rappelé en annexe des notions sur les lois de probabilité employées dans ce travail et sur les méthodes d'estimation MCMC, outils peu habituels pour les ingénieurs.

Néanmoins, je pense que le lecteur statisticien pourra aussi être intéressé par le problème de modélisation qui montre, une fois de plus, que la statistique bayésienne fournit un cadre à la fois rigoureux et simple pour aborder de façon opérationnelle des modèles à structure mathématique assez complexe.

Cette thèse s'articule en 8 chapitres.

Le chapitre 2 décrit le principe de fonctionnement des principaux types de compteurs actuellement utilisés dans la pratique technique et donne des notions fondamentales pour la compréhension du problème technique (signature métrologique, histogramme de consommation, rendement). On insiste aussi sur les aspects réglementaires et en particulier sur la nouvelle norme qui entrera

bientôt en vigueur pour évaluer la conformité métrologique d'un parc de compteurs.

Dans le chapitre 3, on se focalise sur la gestion optimale des renouvellements des compteurs et sur les problèmes engendrés par un sous-comptage de plus en plus marqué en fonction de l'âge. Même si, par souci de confidentialité, on ne s'occupe pas de l'aspect décisionnel de la gestion, on souligne la nécessité d'une bonne connaissance du comportement des compteurs en service dans toute stratégie imaginable. On termine avec quelques (rares) exemples de stratégies de gestion proposées dans la bibliographie technique.

Le chapitre 4 marque la transition entre la partie technique et la partie mathématique de la thèse. Il s'agit d'un chapitre "*autobiographique*" parce qu'il reproduit, d'une certaine manière, le chemin intellectuel qui a conduit à la réalisation du modèle statistique de dégradation des compteurs, au cœur de ce travail. La conceptualisation naît de l'observation préliminaire des données qui suggèrent la présence simultanée dans un parc de compteurs de plusieurs populations, en proportions variables avec l'âge. Le découpage de quatre catégories de compteurs est réalisé ensuite sur la base de considérations techniques et réglementaires. La modélisation du phénomène se fait avec l'interprétation des dites catégories, à métrologie de plus en plus dégradée, comme des états à travers lesquels chaque dispositif passe le long de sa vie opérationnelle. C'est le point de départ du modèle dynamique de dégradation (qu'on suppose markovien) qui se présente comme mieux adapté à la structure stochastique des données, par rapport à d'autres techniques (notamment de régression) typiquement employées dans l'étude statistique des compteurs.

Dans le chapitre 5 nous abordons les aspects méthodologiques. Nous commençons avec les principes de base de l'inférence bayésienne et nous montrons que les techniques Monte Carlo par Chaînes de Markov permettent de surmonter assez facilement les difficultés d'un problème d'estimation peu usuel. En effet, l'inférence sur les modèles markoviens est normalement menée à partir de séries temporelles d'observations répétées des états de plusieurs individus à des pas de temps fixés. Dans notre cas, puisque l'échantillonnage d'un compteur est une mesure destructive, les données se présentent sous la forme d'observations ponctuelles et les calculs sont un peu plus complexes. En plus, un des quatre états n'est pas observé dans le processus d'échantillonnage et son occurrence doit être estimée sur la base d'une autre source d'information.

Dans le chapitre 6 nous nous posons le problème d'avoir une modélisation plus fine des facteurs d'influence de la dégradation des compteurs. La mise en œuvre d'un modèle hiérarchique met en évidence le rôle joué par la localisation géographique des compteurs dans le phénomène de dégradation. On calcule ainsi un paramètre de vitesse de dégradation qui traduit l'agressivité globale d'un certain site sur les compteurs installés. Ce paramètre est cohérent avec un autre indicateur technique de l'agressivité des conditions locales d'exploitation : le taux de blocage observé sur le terrain par les releveurs. Cette observation, d'une part confirme les résultats métrologiques expérimentaux, et d'autre part est très utile en pratique parce qu'elle autorise l'extrapolation de l'agressivité sur la base du seul taux de blocage, pour les sites pour lesquels n'existe pas de retour d'expérience, quant à la dégradation des compteurs. Nous terminons ce chapitre avec la description d'un autre facteur explicatif : la consommation annuelle. Comme on pouvait l'imaginer les compteurs soumis à des consommations importantes se dégradent plus vite que ceux qui ont des régimes de fonctionnement plus ordinaires.

Le chapitre 7 donne un exemple de la manière de prendre en compte âge, agressivité et consommation pour donner des prévisions sur un paramètre technique très important pour le distributeur : l'estimation du sous-comptage. Ces indications sont directement utilisables dans les outils de gestion optimale des parcs de compteurs.

Enfin le chapitre 8 trace un bilan du travail de thèse en soulignant les résultats obtenus et les perspectives.

> "Tous les événements, ceux mêmes qui par leur petitesse semblent ne pas tenir aux grandes lois de la nature, en sont une suite aussi nécessaire que les révolutions du soleil. Dans l'ignorance des liens qui les unissent au système entier de l'univers, on les a fait dépendre des causes finales, ou du hasard, suivant qu'ils arrivaient et se succédaient avec régularité ou sans ordre apparent ; mais ces causes imaginaires ont été successivement reculées avec les bornes de nos connaissances, et disparaissent entièrement devant la saine philosophie qui ne voit en elles que l'expression de l'ignorance où nous sommes des véritables causes."
>
> Pierre Simon de Laplace (1749-1827).

Première partie

Les compteurs d'eau

Chapitre 2

L'ABC des compteurs d'eau

2.1 Généralités

Les compteurs d'eau sont des machines hydrauliques qui permettent la mesure automatique du volume d'eau traversant en un temps donné une section déterminée d'un courant liquide. Les compteurs normalement utilisés pour la facturation de l'eau sont de type mécanique : l'organe de mesure est actionné directement par la force hydrodynamique de l'eau et met en rotation les mécanismes d'un dispositif d'affichage (totalisateur) grâce à l'énergie mécanique transmise par le courant d'eau. Les autres typologies de compteurs (notamment les débitmètres électromagnétiques) ne sont utilisées que dans des cas très particuliers et ne feront pas l'objet de cette étude.

Parmi les compteurs mécaniques on distingue deux grandes familles :

- les *compteurs volumétriques* dont l'organe de mesure (un ou plusieurs pistons à mouvement rotatif ou oscillant) se déplace sous l'effet d'une différence de pression dans le dispositif refoulant périodiquement un volume déterminé d'eau. Le nombre de refoulements donne donc une mesure du volume d'eau écoulé.
- les *compteurs de vitesse* dont l'organe mesurant est un rotor (turbine ou hélice) qui tourne sous l'effet de la poussée hydrodynamique de l'eau qui y transite. Le principe de mesure est la proportionnalité entre la vitesse angulaire du rotor et le débit de l'eau. C'est le nombre de tours du rotor sur une période de temps donnée qui indique le débit intégré, c'est-à-dire le volume d'eau écoulé.

La terminologie anglophone met bien en évidence le principe de mesure. Les

compteurs volumétriques sont désignés comme "*displacement meters*", alors que les compteurs de vitesse qui calculent "indirectement" le volume écoulé sont appelés "*inference meters*". Les plus anciens brevets de compteurs d'eau ont été déposés en Angleterre dans les années 1850 (Carlier, 1968) et concernent des compteurs volumétriques. L'utilisation courante de ces dispositifs date de 1860. Dans la suite nous donnons une description technique des principales typologies de compteurs actuellement utilisés.

2.1.1 Les compteurs volumétriques

Parmi les différents types de compteurs volumétriques, nous nous focalisons sur les compteurs dits *à piston rotatif*. Il s'agit des compteurs les plus précis et sensibles.

Le système de mesure, initialement breveté vers 1880 aux Etats Unis est représenté schématiquement dans la figure 2.1.

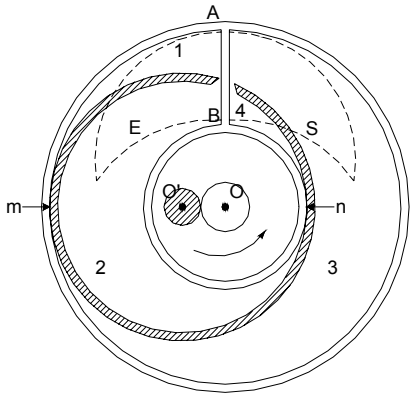


FIG. 2.1 – Principe de fonctionnement des compteurs volumétriques.

L'organe de mesure est formé de deux parties :

- Une boîte mesurante, constituée de deux cylindres coaxiaux (centre O) réunis par un diaphragme AB.
- Un piston cylindrique (de centre O') d'axe parallèle à celui de la boîte mesurante et constamment tangent, intérieurement au cylindre extérieur de la boîte (de rayon OA) et extérieurement au cylindre intérieur (de rayon OB). Les deux points de tangence (m et n) sont constamment alignés avec les deux centres O et O'.

Une fente est réalisée dans le piston pour pouvoir laisser passer le diaphragme AB. Pour permettre une rotation complète du piston à l'intérieur de la boîte, l'axe O' est guidé par un galet centré en O. Sur les plans inférieur et supérieur de la boîte se trouvent respectivement deux orifices symétriques par rapport au diaphragme (en pointillé) qui sont en communication l'un avec l'arrivée de l'eau (E), l'autre avec la sortie (S). Pour toute position du piston, la boîte mesurante est divisée en quatre compartiments (numérotés sur la figure) qui sont, à tour de rôle, en communication avec l'entrée E, puis isolés et ensuite en communication avec la sortie.

Un tour complet du piston correspond au passage d'un volume d'eau égal à la somme du volume des quatre compartiments. Le nombre de tours du piston est donc proportionnel au volume d'eau qui transite dans le compteur.

La figure 2.2 montre la section d'un compteur volumétrique actuellement disponible sur le marché (Altaïr V3) produit par le fabriquant français Sappel. Dans cette figure on observe bien le parcours de l'eau à l'intérieur du dispositif de mesure et la transmission du mouvement rotatif du piston à la minuterie du totalisateur, réalisée avec la juxtaposition de deux aimants. L'entraînement magnétique, dont l'utilisation courante date des années 70 a l'avantage d'isoler complètement le totalisateur de l'eau du réseau, évitant ainsi les phénomènes de blocage de la minuterie, causés par l'incrustation de particules solides dans les engrenages.

Cette technologie de mesure est largement la meilleure du point de vue métrologique, à la fois en termes de sensibilité (capacité de mesure à faible débit), et en termes d'exactitude (précision de la mesure). Elle a, en outre, l'avantage que la qualité de la mesure est indépendante de la position du compteur (horizontale ou verticale).

En revanche, ce type de compteur est relativement plus fragile que les compteurs de vitesse, supportant mal les débits excessifs, les mises en eau "musclées" et les passages d'air. En principe ces dispositifs sont aussi plus sensibles à la qualité de l'eau circulante. Le passage de particules solides peut provoquer des rayures sur l'ensemble boîte-piston qui, donnant lieu à des micro-fuites entre les compartiments de la figure 2.1, engendrent un sous-comptage, ou dans des cas plus graves causer un blocage du piston (aucun volume n'est alors enregistré).

Enfin, les compteurs volumétriques peuvent avoir un fonctionnement bruyant à débit élevé, ce qui oblige parfois les distributeurs d'eau à les remplacer, quand

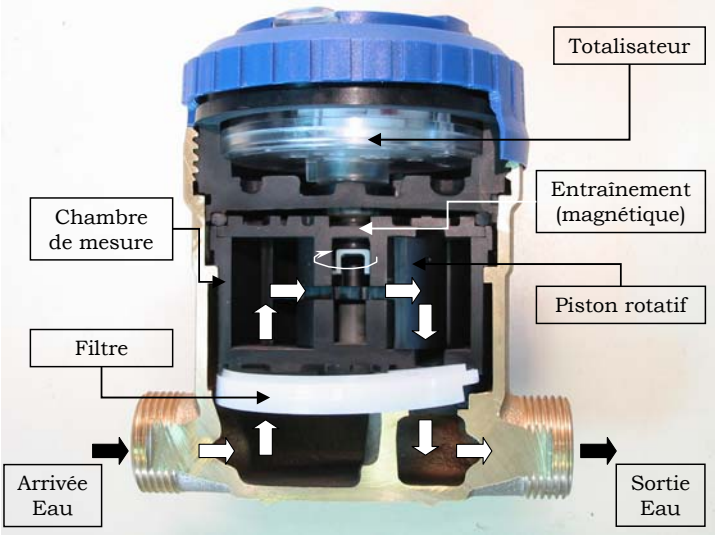


FIG. 2.2 – Circulation de l'eau dans un compteur volumétrique.

ils se trouvent à l'intérieur des habitations.

2.1.2 Les compteurs de vitesse

A cette grande famille appartiennent différents types de compteurs. L'organe de mesure peut être une turbine ou une hélice.

Les compteurs à turbine se divisent en *compteurs à jet unique* (l'eau attaque la turbine sous forme d'un seul jet, comme le montre la figure 2.3) et *compteurs à jets multiples* (figure 2.4) où l'organe de mesure est placé dans une *chambre d'injection* munie de plusieurs orifices obliques à travers lesquels l'eau rentre dans le mesureur et attaque la turbine sur toute sa périphérie. La sortie de l'eau se fait par d'autres orifices situés sur un plan parallèle à celui des ouvertures d'entrée (généralement dessus).

Dans les compteurs à jets multiples la section totale des ouvertures d'entrée n'est qu'une partie de la section de la veine d'eau à l'entrée du compteur et donc la vitesse périphérique de la turbine est sensiblement supérieure à celle de la veine fluide (contrairement aux compteurs à jet unique).

Un réglage, typiquement réalisé avec une vis (évidemment plombée) est nécessaire pour garantir le bon fonctionnement du compteur (figure 2.4). On signale ainsi que la formation de dépôt dans les orifices d'entrée peut, en réduisant leur section, augmenter la vitesse de rotation de la turbine et engendrer des

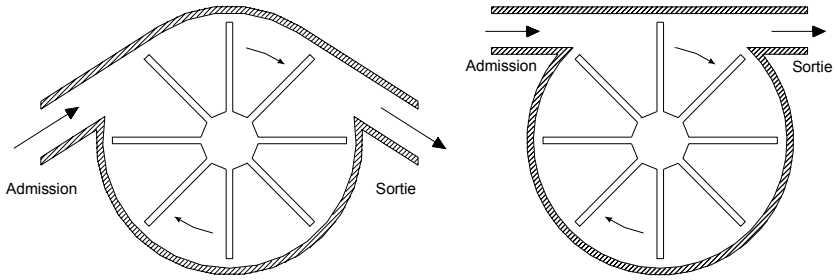


FIG. 2.3 – Principe de fonctionnement des compteurs de vitesse *à jet unique*.

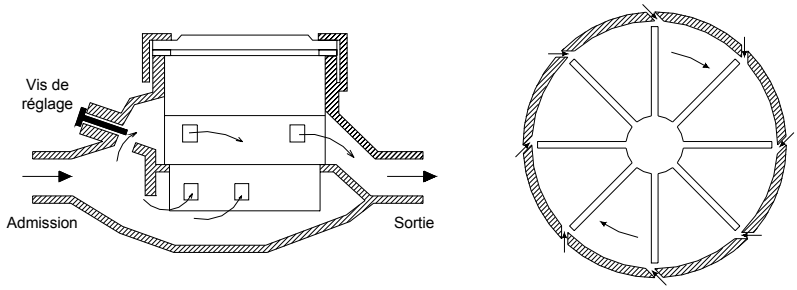


FIG. 2.4 – Principe de fonctionnement des compteurs de vitesse *à jets multiples*.

phénomènes de sur-comptage (Cascetta et Sepe, 1997). Les compteurs à jets multiples ne sont plus utilisés par la Compagnie Générale des Eaux (CGE).

Enfin on signale, parmi les compteurs de vitesse, les compteurs à hélice (verticale ou horizontale) dits aussi compteurs Woltman, utilisés en présence de débits très importants.

En général, les compteurs de vitesse ont des prestations métrologiques inférieures à celles des compteurs de volume en termes de sensibilité et exactitude.

En revanche, les compteurs de vitesse sont normalement réputés plus robustes que les compteurs volumétriques par rapport aux différentes agressions qui peuvent se manifester le long de leur vie opérationnelle (eaux dures et/ou chargées en particules solides, sur-débits) et sont moins encombrants et moins bruyants.

Par ailleurs, leur robustesse est à l'origine de leur diffusion (Fontana, 1996), qui a démarré à grande échelle lors de la reconstruction allemande de 1945. Les réseaux endommagés par la guerre étaient alors fortement ensablés et les compteurs volumétriques, inadaptés à des conditions si sévères, ont cédé la place à des compteurs de vitesse à jets multiples.

2.1.3 Compteurs actuellement utilisés en France par la CGE

Le tableau suivant montre le principaux types de compteurs utilisés aujourd'hui par la Compagnie Générale des Eaux en France métropolitaine.

Nom	Fabricant	Calibre	Techn. de mesure	Années de pose	% du parc compteurs
Aquadis	Actaris*	15 à 65 mm	Volumétrique	Depuis 1990	27.0%
Véga	Sappel	15 à 40 mm	Volumétrique	Depuis 1985	17.2%
Volumag	Schlumberger	15 à 65 mm	Volumétrique	1975 – 1990	12.9%
Altaïr	Sappel	15 et 20 mm	Volumétrique	Depuis 1995	12.2%
Flostar	Actaris	15 à 32 mm	Vitesse Jet Un.	Depuis 1990	12.1%
Marly 2	Wateau	15 et 20 mm	Volumétrique	Depuis 1999	2.6%
Microprecis MP2	Farnier	15 à 100 mm	Vitesse Jet Un.	Depuis 1977	2.0%

Compteurs utilisés en France par la CGE.

Les pourcentages ci-dessus sont calculés sur la base de la population connue de compteurs. On imagine donc que la répartition des différents types de compteurs est la même dans la population connue et celle inconnue (19.5% du total). Le nombre de compteurs installés est estimé à environ 6 millions.

Dans le langage technique courant ce qu'on appelle ici "type" de compteur est normalement appelé par les ingénieurs "modèle de compteur". On a préféré, dans cette thèse, utiliser ce terme exclusivement dans le sens attribué par le langage mathématique.

(*) **N.B.** Depuis novembre 2001, "*Actaris*" est le nouveau nom de la branche comptage du groupe Schlumberger.

2.2 Le débit maximal de fonctionnement dépend de la technologie de mesure

La norme internationale en vigueur en matière de compteurs d'eau froide (ISO 4064-1) définit le débit maximal de fonctionnement (débit de *surcharge* q_s) comme le débit le plus élevé auquel l'appareil doit pouvoir fonctionner, sans détérioration, pendant des durées limitées, en respectant les erreurs maximales tolérées (cette définition sera précisée dans la suite) et sans dépasser la valeur maximale admissible de la perte de pression provoquée par la présence de l'appareil de mesure dans la conduite. Les compteurs sont répartis en quatre groupes sur la base de cette valeur maximale de la perte de charge (0.1, 0.3, 0.6 et 1 bar).

Les compteurs d'eau sont caractérisés conventionnellement par leur "*débit permanent*" (q_p) défini comme le "débit auquel le compteur doit pouvoir fonctionner de manière satisfaisante en utilisation normale". Il est égal à la moitié du débit de surcharge.

Dans le langage technique courant les termes *débit permanent* et *débit de surcharge* sont normalement remplacés respectivement par les termes *débit nominal* (q_n) et *débit maximal* (q_{max}), utilisés dans le texte législatif français (Décret 76-130).

Au débit nominal est généralement associé un *diamètre nominal* (DN), relié approximativement aux dimensions de la tuyauterie du compteur.

Les compteurs volumétriques actuellement utilisés en pratique ont des débits nominaux compris entre 1.5 m^3/h (DN 15 mm) et 10 m^3/h (DN 40 mm), et donc des débits maximaux qui n'excèdent pas 20 m^3/h.

Les principales contraintes techniques qui limitent le débit de fonctionnement de ce type de compteur sont liées notamment aux pertes de charge et aux dimensions qui seraient nécessaires à un fonctionnement correct à fort débit.

Quand on dépasse ces valeurs, on utilise alors normalement des compteurs à turbine (à jet unique) pour des débits maximaux jusqu'à 100 m^3/h (DN 100 mm) et, pour des débits encore plus importants des compteurs à hélice, pouvant atteindre des valeurs jusqu'à 5 000 m^3/h (DN 500 mm).

A titre de curiosité, il est remarquable que la valeur limite de 100 m^3/h pour les compteurs à turbine était déjà indiquée dans la bibliographie technique des années 50 : Howe (1950) indique un chiffre de 10 ft^3/h, soit environ 102 m^3/h.

2.3 L'exactitude d'un compteur dépend du débit d'utilisation

Comme tout instrument de mesure, un compteur est susceptible d'erreur : le volume enregistré sur un branchement en une période donnée (v_e) est généralement différent du volume effectivement consommé (v_c). L'erreur relative de mesure est définie par la relation :

$$e = \frac{v_e - v_c}{v_c} \tag{2.1}$$

Quel que soit le principe de mesure, le fonctionnement d'un compteur mécanique se base sur le transfert d'énergie mécanique entre le courant d'eau et le dispositif mesurant.

Quand le débit est faible ce transfert d'énergie est plus difficile et, par conséquent, les erreurs relatives sont négatives et importantes en valeurs absolues, pouvant atteindre la totalité (-100%) quand l'énergie mécanique n'arrive pas à mettre en rotation les organes de mesure.

A mesure que le débit augmente, l'erreur diminue en valeur absolue jusqu'à atteindre des valeurs parfois légèrement positives.

La représentation graphique de la relation entre l'erreur relative de mesure et le débit circulant est dite *courbe métrologique* ou *signature métrologique* du compteur.

Tronskolanski (1963) propose des expressions théoriques de l'équation de la courbe métrologique des compteurs de vitesse basées sur *l'équation fondamentale d'Euler* (exprimant le couple de rotation exercé par un courant de débit donné sur la roue d'une turbine hydraulique).

En pratique les signatures métrologiques sont obtenues expérimentalement dans des laboratoires spécialisés en faisant circuler dans le compteur un volume connu d'eau à débit constant. La courbe est alors construite point par point avec les résultats des essais aux différents débits d'étalonnage.

Les méthodes et matériels d'essai font l'objet de la norme ISO 4064-3.

La figure 2.5 montre des exemples de courbes métrologiques de compteurs volumétriques neufs de diamètres nominaux égaux à 15, 20, 30 mm respectivement. Les débits sont toujours reportés sur l'axe des abscisses en échelle logarithmique.

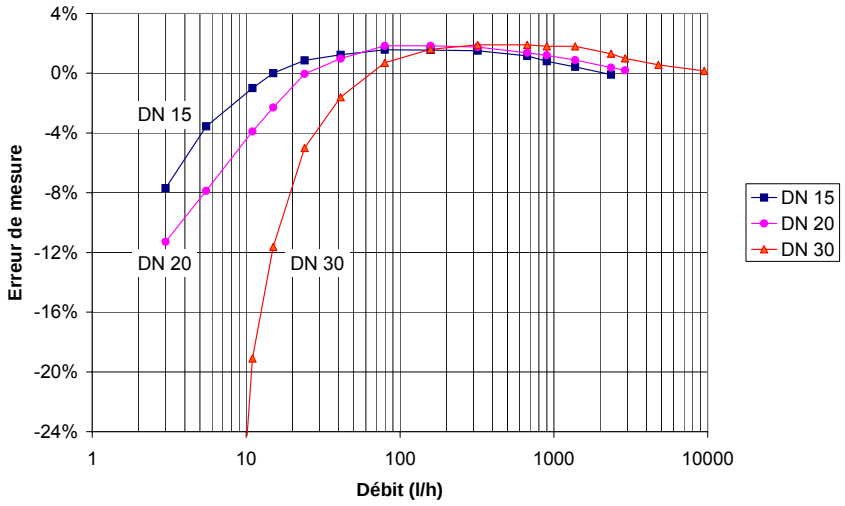


FIG. 2.5 – Exemple de courbes métrologiques de compteurs.

De l'examen de cette figure apparaît nettement que les compteurs ne fonctionnent correctement que dans une plage limitée de débit et aussi que la sensibilité d'un compteur (capacité d'enregistrer à faible débit) est inversement liée à ses dimensions. En effet, plus les organes de mesure sont grands, plus le compteur est "inerte" mais si on veut faire fonctionner le dispositif à haut débit, il est nécessaire d'augmenter la taille du mesureur (pour des questions de résistance aux sollicitations mécaniques et de limitation des pertes de charge). En définitive un compteur mécanique perd en sensibilité ce qu'il peut gagner en débit et inversement.

Le choix du compteur à installer sur un branchement est un problème délicat et un compromis s'impose entre le risque de surcharge hydraulique et la possibilité de mesurer à faibles débits

Une solution pour avoir une plage de fonctionnement très large, en termes de débit, pourrait, en principe, être représentée par les *compteurs combinés* (Betta et al., 2002), réalisés en disposant sur la conduite principale un gros compteur (par exemple à hélice) et en parallèle, sur une canalisation de plus faible diamètre, un petit compteur (par exemple volumétrique). Une valve automatique à l'amont ou à l'aval du gros compteur réglerait le fonctionnement du dispositif, en dirigeant les faibles débits vers le petit compteur et les hauts débits vers le gros compteur.

Cette solution, incontestablement attractive du point de vue théorique, s'est

révélée décevante, en termes de performance, en conditions réelles d'exploitation et n'est plus adoptée par la CGE. En particulier l'efficacité dans le temps du système de commutation des débits entre les deux compteurs, qui déjà au départ présente un comportement asymétrique, est mise en cause et des réglages fréquents seraient nécessaires, avec une évidente perte de rentabilité.

2.4 Le rendement dépend des modalités de puisage de l'eau

La répartition, en pourcentage, des volumes d'eau consommée par tranches de débit est un élément fondamental pour le choix du compteur à installer sur un branchement et pour l'évaluation des performances d'un compteur.

Cette répartition est normalement représentée sous forme d'histogramme (dit *histogramme de consommation*).

En principe chaque abonné a sa propre façon de puiser l'eau. En pratique, des études statistiques ont permis de définir des profils moyens de consommation pour des grandes catégories d'usagers (par exemple les consommations domestiques).

Les histogrammes types actuellement utilisés proviennent de campagnes de mesure réalisées dans les années 90 qui ont fait l'objet de plusieurs rapports de stages, par exemple (Girard, 1995).

La procédure pour obtenir l'histogramme de consommation d'un abonné s'articule en trois phases. D'abord, le compteur en service sur le branchement choisi est déposé et remplacé avec un compteur neuf (dont la courbe métrologique est connue et supposée invariante pendant la durée des mesures) équipé d'une tête émettrice qui envoie à distance une impulsion chaque fois qu'un volume d'eau donné (typiquement 0.01 l) est enregistré.

Ensuite, dans la phase d'enregistrement on recueille les données sur une période de temps fixée (une semaine pour les abonnés domestiques). Finalement, le traitement informatique des enregistrements temporels des impulsions permet le calcul du volume consommé dans chaque tranche de débit. Ces volumes sont corrigés sur la base de la courbe métrologique (connue) du compteur.

La figure 2.6 montre, à titre d'exemple, deux histogrammes de consommations enregistrés respectivement dans une maison individuelle et dans un immeuble de

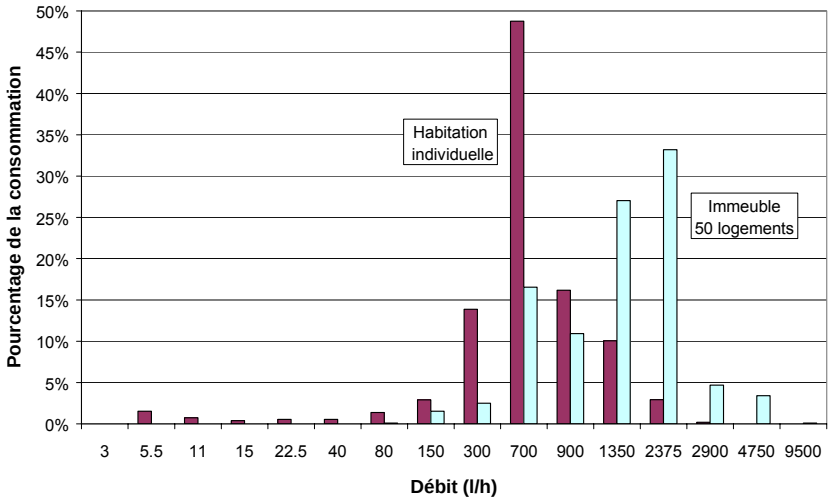


FIG. 2.6 – Exemple d'histogrammes de consommation.

50 appartements.

Dans l'histogramme relatif à la maison individuelle le pic dans la zone des faibles débits correspond à des fuites (robinets fuyants, chasses d'eau défectueuses etc.).

En superposant l'histogramme de consommation et la signature métrologique le long de l'axe des débits, on peut évaluer le *rendement* (R) d'un compteur, défini comme le rapport entre le volume enregistré et le volume consommé :

$$R = \frac{v_e}{v_c} = 1 + \sum_i e_i x_i \tag{2.2}$$

où e_i et x_i sont respectivement les erreurs de mesures et les proportions de la consommation en correspondance du débit q_i.

La quantité $\sum_i e_i x_i$ (moyenne pondérée des erreurs de mesure sur l'histogramme de consommation) est couramment désignée *Coefficient de Comptage Moyen* (CCM).

L'examen des figures 2.5 et 2.6 montre aussi l'importance de la connaissance du profil de consommation dans le choix du compteur à installer sur un branchement.

Dans l'exemple proposé, pour la maison individuelle on choisira un compteur de 15 ou 20 mm, alors que pour l'immeuble de 50 appartements le compteur le plus adapté au profil de consommation est celui de 30 mm.

La construction d'un profil type d'utilisation pour des grandes catégories d'utilisateurs est une opération délicate. Pour certains gros consommateurs (par exemple les industriels), compte tenu à la fois des importants enjeux économiques et de la grande variabilité des conditions d'utilisation strictement liées aux produits fabriqués et aux machines employées, il est envisageable d'avoir un histogramme personnalisé.

Pour les abonnés domestiques, la difficulté principale consiste en la détermination de la partie de consommation ayant lieu à faible débit, cette consommation parasite pouvant donner lieu à des volumes importants. Par exemple, un robinet qui goutte (env. 3.5 l/h) engendre une consommation annuelle d'environ 30 m^3/an soit environ 1/4 à 1/3 de la valeur annuelle moyenne pour cette catégorie d'usagers.

L'utilisation généralisée d'appareils de mesure de meilleure qualité (et notamment des compteurs volumétriques), capables de détecter (et de facturer) ces consommations a donc permis, par rapport au passé, une augmentation importante des rendements du comptage.

Par exemple Grau (1985), dans le cadre d'une étude réalisée dans les années 70 concernant des abonnés domestiques de la ville de Barcelone (Espagne) équipés de compteurs de vitesse inertes jusqu'à 10 l/h (à la sortie d'usine), a montré que le cumul des volumes facturés était inférieur d'environ 15% aux volumes effectivement consommés. En effet, 9.1% de la consommation avait lieu à des débits inférieurs à 9 l/h et donc les compteurs utilisés perdaient déjà à leur mise en service entre 9 et 10% des volumes consommés.

L'opinion courante des experts dans le domaine du comptage est que le fait que les distributeurs d'eau soient capables de facturer ces volumes a poussé les usagers à améliorer la qualité de leurs installations et donc la proportion de consommation à faible débit est en baisse par rapport au passé. Cette réduction des gaspillages d'eau potable s'inscrit dans un cadre plus général d'une baisse de la consommation d'eau potable (Detoc ; Grandjean et Jannin ; Stevenin et Jean-Marie, 2000), motivée d'une part par une prise de conscience de l'importance d'une ressource naturelle, imaginée comme de plus en plus *précieuse*, mais aussi par l'augmentation des prix : d'après une étude réalisée par le *Syndicat des Eaux d'Ile de France* (SEDIF), organisme qui regroupe 144 communes de la Région Parisienne, 70% des familles interpellées déclarent faire attention à leur consommation d'eau notamment en évitant les gaspillages (Francheteau, 2002).

Beaucoup d'encre a été versée sur l'importance de réduire les consommations d'eau et les ménages des pays industrialisés, soumis à un alarmisme excessif et non justifié par les données réelles (Lomborg, 2001) sur une possible "*Emergence Eau*" à l'échelle mondiale, se sont souvent sentis responsabilisés d'une tâche qui va bien au delà de leurs possibilités concrètes de modifier la demande en eau : dans l'Union Européenne (Margat, 2000) les consommations domestiques ne représentent que 17% des volumes totaux (8% sur échelle mondiale). D'autre part les économies d'eau sont importantes (Lee et al. 2001) parce qu'elles évitent le surdimensionnement des infrastructures : réseaux de distribution, stations de pompage, usines de production d'eau potable (de plus en plus chères à cause de l'augmentation de la pollution des nappes et des eaux superficielles), et donc des investissements inutiles de la part des collectivités. Eviter de gaspiller une ressource, qui en France et en Europe n'est pas *rare* mais *chère*, ne sauvera pas le monde d'une fantomatique menace de sécheresse planétaire mais est tout simplement ... une question de bon sens.

En pratique, le profil de consommation type des abonnés domestiques est obtenu par mélange de profils différents (avec fuites d'eau ou non) selon des proportions fixées (par exemple 25% avec fuites et 75% sans).

2.5 La réglementation fixe des limites aux erreurs de mesure

Le texte législatif français en matière de compteurs d'eau froide est le Décret n. 76-130 du 29/01/1976. Ce texte, qui reprend une directive CEE (75/33/CEE) est en accord avec la norme ISO 4064-1.

Tout d'abord le Décret définit le débit maximal q_{max} et le débit nominal q_n. Les définitions sont analogues à celles de débit de surcharge et débit permanent données par la norme internationale.

La valeur maximale de la perte de pression provoquée par la présence du compteur a été fixée dans un texte successif (arrêté du 19/07/1976) à 1 bar au débit maximal et 0.25 bar au débit nominal.

L'étendue des débits à l'intérieur de laquelle les compteurs doivent respecter les limites de loi est comprise entre un débit minimal (q_{min}) et le débit maximal. Cette plage est partagée par un *débit de transition* (q_t) en deux zones

dites *inférieure* et *supérieure* dans lesquelles les erreurs maximales tolérées sont différentes. Les débits q_{min} et q_t sont fixés en fonction du q_n et de la *classe métrologique* du compteur (A, B ou C en ordre décroissant d'exactitude) selon le tableau suivant :

	q_n <15 m^3/h		q_n ≥15 m^3/h	
Classe	$q_{\min}$	q_t	$q_{\min}$	q_t
A	0.04 q_n	0.10 q_n	0.08 q_n	0.30 q_n
B	0.02 q_n	0.08 q_n	0.03 q_n	0.20 q_n
C	0.01 q_n	0.015 q_n	0.006 q_n	0.015 q_n

Les *Erreurs Maximales Tolérées* (EMT) lors de la "*vérification primitive*" des compteurs (c'est à dire à la sortie d'usine) sont fixées de la manière suivante :

> "Cinq centièmes, en plus et en moins, du volume mesuré pour tout débit situé dans la zone inférieure comprise entre q_{min} inclus et q_t exclu."
>
> "Deux centièmes, en plus et en moins, du volume mesuré pour tout débit situé dans la zone supérieure comprise entre q_t inclus et q_{max} inclus."

Pour les compteurs en service, les valeurs des EMT sont le double de celles admises pour les compteurs neufs.

La réglementation fixe donc, en fonction de la classe métrologique et du débit nominal, des *canaux de tolérance* à l'intérieur desquels la courbe métrologique, construite à partir des erreurs de mesure à q_{min}, q_t, q_{max} doit être contenue. La figure 2.7 montre les canaux de tolérance pour des compteurs neufs (q_n<15 m^3/h). Les débits sont exprimés en pourcentages du q_n.

Des développements futurs de la réglementation font actuellement l'objet d'une étude à la Sous-Direction de la Métrologie du Ministère de l'Economie des Finances et de l'Industrie. Le projet d'arrêté définit les modalités de contrôle de la métrologie des compteurs en service (Costes et Pia, 2000).

La vérification périodique pourra être réalisée selon l'une des modalités suivantes :

- *Vérification unitaire*, avec une périodicité différente selon la classe métrologique.
- *Vérification statistique par lots* selon des plans d'échantillonnage dont la courbe d'efficacité (norme ISO 2859-0) présente une *probabilité indifférente*

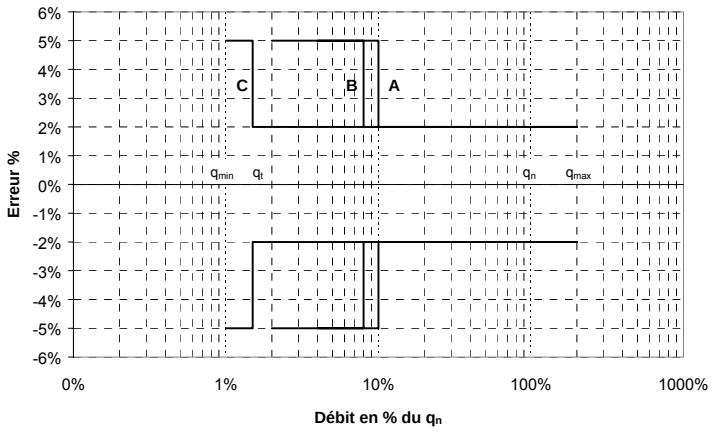


FIG. 2.7 – EMT pour compteurs neufs (Décret 76-130)..

de 50% d'accepter ou refuser un lot contenant une proportion de dispositifs non conformes inférieure ou égale à 10%.

Vraisemblablement, l'application de la réglementation sera graduelle : dans un premier temps le taux de non conformité acceptable sera de 15% pour passer progressivement à 12.5% et finalement à la valeur de régime de 10%. De cette manière, les distributeurs d'eau auront le temps de planifier la mise en conformité de leurs parcs.

Les distributeurs d'eau ayant mis en place un système d'assurance qualité équivalent à celui assuré par la vérification périodique (unitaire ou statistique) devraient être dispensés du contrôle.

Le contrôle de conformité prévoit, pour chaque compteur échantillonné, un essai d'exactitude aux deux débits suivants, dans l'ordre :

- Un débit compris entre 0.8 q_n et q_n.
- Un débit compris entre 0.1 q_n et 0.3 q_n pour les compteurs de q_n <10 m^3/h ou entre 0.3 q_n et 0.5 q_n pour les compteurs de $q_n \geq$10 $m^{3/}h$.

L'appareil est jugé conforme si les erreurs de mesure, évaluées en correspondance des deux débits d'essai sont inférieures ou égales (en valeur absolue) à 4%.

La réglementation en vigueur et ses développements probables ont inspiré le classement des compteurs en service en quatre catégories, qui forme la base du modèle statistique de dégradation présenté dans cette étude.

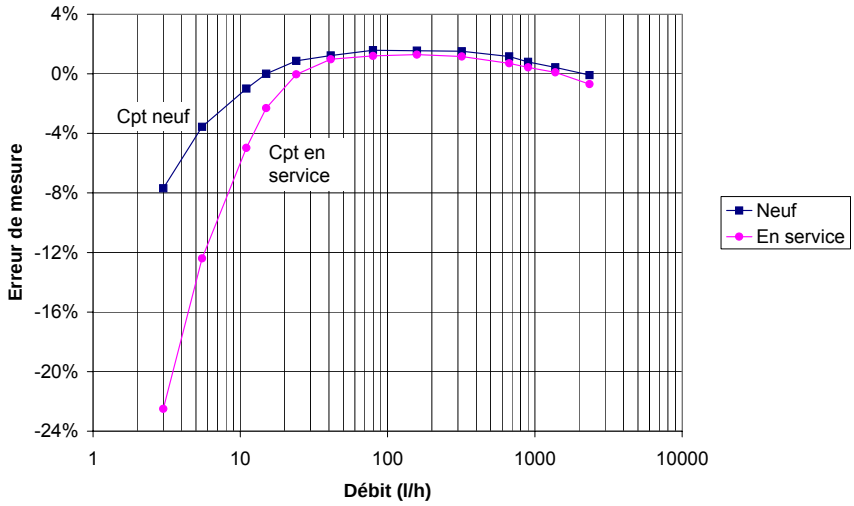


FIG. 2.8 – Evolution de la métrologie d'un compteur au fil du temps.

2.6 La performance des compteurs se dégrade au long de leur vie

L'observation des courbes métrologiques des compteurs en service et leur comparaison avec les performances des compteurs neufs, montrent que généralement la métrologie d'un compteur se dégrade au fil du temps.

Mis à part les compteurs de vitesse à jets multiples (qui ne représentent qu'une partie très marginale des compteurs de la CGE) la *déformation* de la courbe métrologique se manifeste normalement (figure 2.8) avec une réduction de la plage des débits de fonctionnement (le compteur devient moins sensible) et, éventuellement, une légère baisse dans la zone haute de la courbe (en correspondance des débits de tirage les plus fréquents).

Normalement, les facteurs explicatifs de la dégradation de la métrologie des compteurs, proposés depuis toujours par les experts, sont, outre les caractéristiques du compteur (marque, type, diamètre nominal), l'âge du dispositif, le volume enregistré, et la qualité de l'eau mesurée (notamment en termes de dureté et de particules en suspension).

Parmi ces facteurs, l'âge est reconnu comme le plus important et généralement les rares études disponibles dans la bibliographie technique ne prennent en compte explicitement que cette variable (AWWA, 1966), (Newman et Noss,

1982), (Tao, 1982), sous l'hypothèse que l'échantillon analysé est suffisamment homogène par rapport aux autres facteurs.

La notion de qualité de l'eau reste surtout liée à la localisation géographique des compteurs (Grau, 1985). En effet, si on peut imaginer que des eaux chargées en minéraux et/ou particules en suspension ont tendance à accélérer le phénomène de dégradation, il est très difficile d'isoler le rôle d'un facteur unique. Deux eaux ayant la même dureté peuvent, par exemple, être plus ou moins entartrantes en fonction de la température ou des conditions d'utilisation (compteur qui tourne continuellement ou soumis à des longues périodes d'arrêt). Il est donc plus simple d'imaginer que chaque site d'exploitation ait une certaine *agressivité* vis-à-vis des compteurs (résultant de l'effet de l'eau circulante et de l'état du réseau), plutôt que d'étudier un par un les différents facteurs.

Le rôle joué par le volume total enregistré ("*kilométrage*" du compteur) est souvent controversé et parfois nié (Sisco, 1967). Il s'agit d'une variable strictement dépendante de l'âge du dispositif et pour cette raison on a préféré introduire plutôt, dans cette étude, la variable "*consommation annuelle moyenne*", rapport entre l'index du compteur et son âge. En revenant à l'effet de cette variable sur le phénomène de dégradation, s'il est vrai qu'un compteur qui tourne plus vite a tendance à s'user plus vite, il est aussi vrai que les nombreux cycles d'arrêt et démarrage, qu'on peut observer sur des branchements à faible consommation peuvent être une source de stress tout aussi importante pour le dispositif.

L'opinion générale est que la consommation fait partie des facteurs aggravant la dégradation métrologique. En particulier, concernant les compteurs domestiques, on a pu observer une différence marquée entre les compteurs enregistrant plus ou moins de 200 m^3 par an, les premiers se dégradant sensiblement plus vite que les seconds.

Chapitre 3

Le retour d'expérience est à la base de la gestion optimale

3.1 Que signifie "gérer un parc de compteurs" ?

Les compteurs sont utilisés, dans la pratique technique, pour la facturation des volumes d'eau consommés par les abonnés. De leur précision dépend donc la rémunération (ou du moins une partie) du service de distribution d'eau potable.

Puisque, comme il a été déjà souligné dans le chapitre précédent, les compteurs ont tendance à sous-estimer les consommations, il est évident que leur dégradation se traduit par un manque à gagner pour les distributeurs d'eau.

Le fait que les erreurs de mesure s'accroissent le long de la vie opérationnelle des dispositifs, détermine, en outre, une situation d'inégalité entre consommateurs équipés de compteurs récents et consommateurs équipés de vieux compteurs, les premiers payant, à quantités égales d'eau consommée, plus que les seconds.

Enfin, la dégradation de la métrologie peut amener un compteur à dépasser les Erreurs Maximales Tolérées (EMT) fixées par la réglementation et rendre l'instrument non conforme et par conséquent inapte à la facturation.

Gérer un parc de compteurs d'eau signifie mettre en place une stratégie de remplacement des appareils de mesure afin d'améliorer globalement le rendement du parc, et poursuivre un ou plusieurs des objectifs suivants :

- Limiter les pertes économiques représentées par l'eau consommée et non facturée.

– Avoir un système de facturation équitable.
– Limiter le nombre d'appareils non conformes.

Dans la suite de ce chapitre chacune de ces problématiques sera brièvement explorée.

3.2 Le sous-comptage peut conduire à un manque à gagner important

Le volume d'eau consommé par les usagers sur une période donnée, mesuré par différence entre deux lectures successives de l'index du compteur, affecte de manière différente la rémunération du distributeur d'eau selon le principe de tarification choisi.

La figure 3.1 trace la relation entre la dépense de base, hors taxes et redevances, en eau pour le consommateur en fonction du volume d'eau consommé selon différents modes de tarification. Ces graphes montrent que le prix de base (rapport entre la dépense hors taxes et le volume consommé) dépend normalement de la consommation, selon la règle quasiment générale que le prix unitaire est fonction décroissante de la consommation (Erhard-Cassegrain et Margat, 1983).

Les tarifs normalement utilisés en France sont de type *binôme*, avec une partie fixe (dépendant du diamètre du branchement) et une partie variable dépendant du volume consommé (à prix constant ou variable par tranche dégressive).

En pratique, la rémunération du service d'alimentation en eau potable, couvrant les coûts de production et distribution de l'eau, mais aussi de l'abonnement, de la location du compteur, du raccordement au réseau etc., ne représente qu'une partie (entre 43 et 45%) de la facture payée par les usagers.

L'autre partie de la facture constitue la rémunération du service d'assainissement (si présent), couvrant la collecte et le traitement des eaux usées, des taxes (TVA de 5.5%), et des redevances au profit essentiellement des Agences de l'Eau, du FNDAE (*Fond National pour le Développement des Adductions d'Eau*) et, dans certains cas, des VNF (*Voies Navigables de France*). Ces différentes rubriques sont réparties de manière variable selon les collectivités desservies mais les valeurs ne s'éloignent guère de celles indicatives fournies dans la figure 3.2. Cette figure montre aussi, sur la base des indications fournies par Babillot et

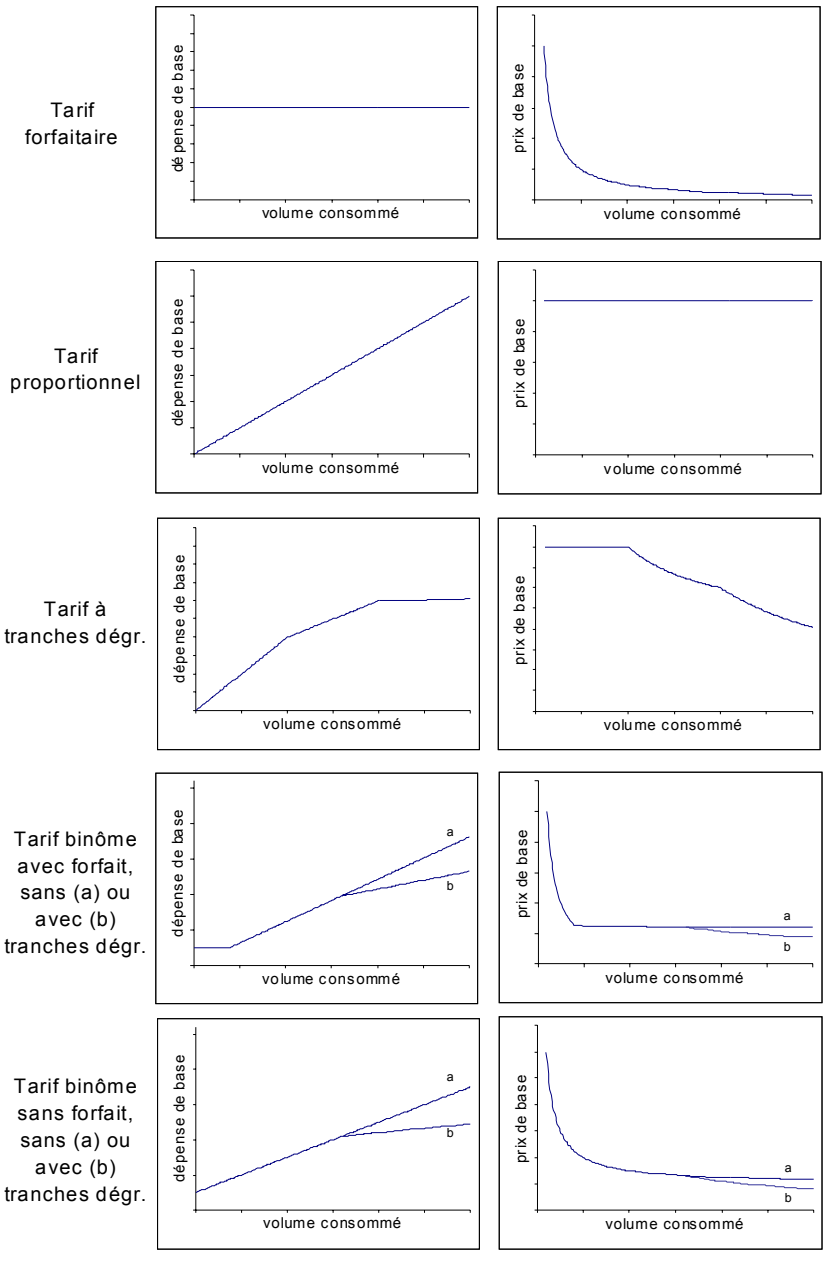


FIG. 3.1 – Différents principes de facturation de l'eau.

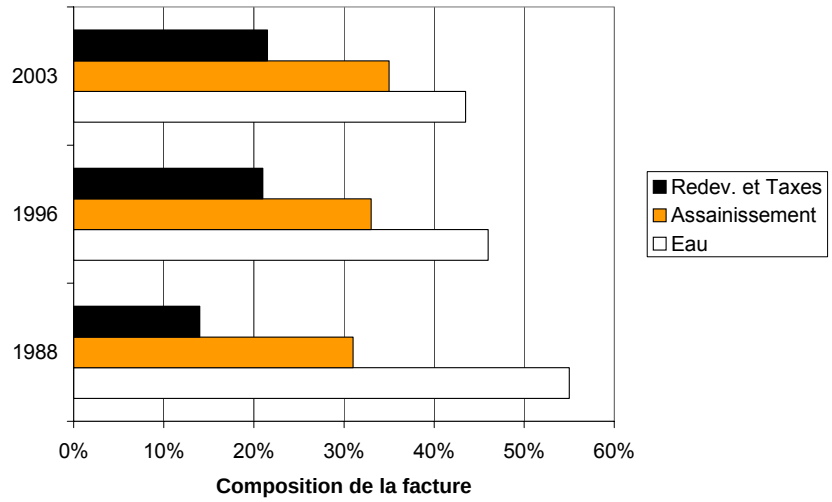


FIG. 3.2 – Structure de la facture d'eau potable en France.

Le Lourd (2000) et Chambolle (1988), l'évolution de la structure moyenne de la facture de l'eau dans les dernières années qui traduit l'effort national d'assainissement et l'augmentation des redevances.

Pour estimer le manque à gagner des distributeurs d'eau par effet du sous-comptage des volumes d'eau facturés, on doit d'abord calculer le prix moyen du m^3 d'eau distribuée. Le calcul est compliqué par effet des différentes règles de tarification.

La méthode généralement acceptée se base sur l'analyse des factures d'eau pour une consommation annuelle de référence[1] de 120 m^3.

Le *Centre d'Information sur l'Eau* (C.I.Eau) donne une fourchette de prix comprise entre 1.5 et 3 €/m^3, et la *Direction Générale de la Concurrence, de la Consommation et la Répression des Fraudes* (DGCCRF) calcule un prix moyen de 2.65 €/m^3.

Des valeurs un peu plus importantes (avec une moyenne de 2.9 €/m^3) sont fournies par le *Syndicat Professionnel des Distributeurs d'Eau* (SPDE), sur la base de son propre échantillon de services gérés en délégation[2]. L'écart de prix est imputable, d'une part au poids des charges sociales, plus lourdes pour les

[1] La valeur de 120 m^3/an a été retenue comme consommation de référence par L'Observatoire de l'eau créé par arrêté du 21 février 1996, ayant pour mission de développer l'information sur le prix de l'eau.

[2] Ces chiffres ont été fournis pat le C.I.Eau et sont disponibles sur le site web : www.cieau.com.

exploitants en délégation que pour ceux en régie, et d'autre part à la différence des prestations fournies.

En définitive on peut imaginer que la rétribution perçue par m^3 d'eau distribué vaut environ $2.65 \cdot 0.435 = 1.15$€.

Le manque à gagner peut s'évaluer par différence entre la valeur théorique du gain (calculée sur la base du volume effectivement consommé) et la valeur réelle calculée à partir du volume enregistré par le compteur.

Si on dénote r_{moy} le rendement moyen des compteurs installés, le manque à gagner (par m^3 effectivement distribué) peut facilement s'exprimer comme :

$$1.15 \cdot (1 - r_{moy}) \text{ €}/\text{m}^3 \tag{3.1}$$

ce qui donne, par exemple, pour des rendements compris entre 95% et 98% des manques à gagner entre 2.3 et 5.7 centimes d'Euros par m^3 consommé.

Ces chiffres, apparemment dérisoires, multipliés par les millions de m^3 distribués par des grands distributeurs, comme la CGE donnent lieu à des gisements potentiels importants.

Si, par exemple, on admet un rendement moyen des compteurs de 98%, compte tenu du fait que la CGE distribue, en France, environ 2.1 milliards de m^3 par an, il en resulte un manque à gagner annuel de 48 millions d'euros !

Il s'agit, évidemment, d'un calcul théorique, basé sur la différence avec les gains dans une situation idéale, et impossible en pratique, de comptage parfait (rendement égal à 1). Néanmoins ce calcul donne une idée de l'importance des enjeux économiques liés au comptage et de l'intérêt pour les grands distributeurs d'eau d'améliorer le rendement global des parcs de compteurs. Améliorer le rendement moyen de 0.1% signifie pour la CGE une augmentation des recettes annuelles de 2.4 millions d'euros, de quoi justifier (largement) les efforts de recherche dans le cadre desquels s'inscrit cette thèse.

3.3 L'équité du comptage : un engagement vis-à-vis des consommateurs

Dans le long débat sur l'opportunité d'utiliser ou non les compteurs dans les services de distribution d'eau potable, l'équité a toujours été un des deux arguments principaux (Guarino, 1976), (Lund, 1988).

L'opinion courant est qu'il s'agit de la seule façon "*logique et correcte*" (Hazen, 1918) de facturer l'eau et d'éviter inégalités et discriminations entre les usagers. L'autre argument majeur est la maîtrise des consommations d'eau[3], avec des baisses parfois supérieures à 50% par rapport à des systèmes de tarification forfaitaire.

Le débat n'est plus d'actualité aujourd'hui, surtout en France où le comptage est généralisé, mais l'argument de l'équité de la facturation reste très important.

La dégradation des compteurs, qui entraîne un sous-comptage de plus en plus marqué, pose, comme il a été déjà anticipé, un problème, dans la mesure où le rendement des compteurs installés n'est pas le même d'un usager à l'autre. Le phénomène met en cause un des principes fondamentaux du Service Public (Duroy, 1996) : l'égalité devant le tarif du service.

Le problème est délicat et, à vrai dire, la législation actuelle ne prévoit aucune mesure spécifique pour garantir un comptage équitable aux consommateurs, les seules obligations légales en la matière étant celles du Décret 76-130.

Par ailleurs, les collectivités locales qui délèguent le service à des sociétés privées, fixent parfois un âge limite à partir duquel tout compteur doit être remplacé. L'*Association des Maires de France* (2001) conseille de fixer ce délai en fonction de la "qualité" des appareils et du "degré d'agressivité de l'eau".

En définitive, aujourd'hui les distributeurs d'eau ont peu de contraintes en la matière et c'est à eux de définir leur politique de comptage.

Deux approches diamétralement différentes sont possibles. La première prévoit l'utilisation d'appareils de mesure très précis, avec des rendements peu dispersés et assez stables dans le temps, mais bien sûr plus chers ; l'investissement demandé pour la mise en œuvre des compteurs est compensé par un meilleur rendement (avec réduction des pertes dues à l'eau non comptabilisée), et une vie opérationnelle plus longue. Un autre choix pourrait être d'employer des compteurs moins performants, et moins chers, avec des rendements nettement inférieurs mais renouvelés plus souvent.

S'il est *a priori* difficile de se prononcer sur l'opportunité économique des deux différentes approches (tout dépend du prix de l'eau), il n'y aucun doute quant à celle des deux qui est la plus équitable : un comptage précis va davantage dans le sens des intérêts de l'usager qui paye effectivement pour ce qu'il

[3]La réduction des consommations d'eau par effet de l'introduction du comptage est l'objet de nombreuses études, dont, p. ex. celles de Shipman (1978), Lund (1987), Ditcham (1997), Austin et al., Guerquin et Grosjean, Richard et Huau, Varszegi (2000).

consomme.

En revenant au problème posé par la dégradation des compteurs, il est difficile d'évaluer l'équité d'une stratégie de gestion. L'*égalité* des usagers devant un service public ne doit pas être confondue avec l'*uniformité* de traitement (Chapus, 1993), et la règle simpliste de remplacer les compteurs, une fois atteint un âge limite, ne semble pas très adaptée à la gestion des parcs de compteurs où plusieurs types d'appareils de qualités différentes coexistent (et c'est souvent le cas dans la pratique).

Même dans le cas théorique où tous les appareils sont égaux, une stratégie de ce type, qui fixe en principe une limite inférieure aux rendements des compteurs, ne peut pas être considérée équitable. Il est évident que le même traitement ne peut pas être réservé aux gros et au petits consommateurs, étant donné que les montants non facturés par effet du sous-comptage sont bien différents dans les deux cas.

Changer plus souvent les compteurs sur les branchements à forte consommation est une règle logique qui va à la fois dans la direction de la rentabilité économique et de l'équité.

Uniformiser les typologies d'appareils utilisés est une mesure décisive pour améliorer l'équité du comptage. Mérite d'être signalé, notamment, l'effort de la CGE qui ne pose actuellement que des compteurs de classe C (cf. page 21), et systématiquement des compteurs volumétriques dans les diamètres de 15 à 40 mm (hors contraintes techniques). Les compteurs de classe C représentent aujourd'hui environ 96% des 6 millions de compteurs de la CGE (avec 79% de volumétriques). Une autre mesure prise par la CGE à l'encontre de cette exigence est la disparition progressive des compteurs de vitesse à jets multiples, les seuls dispositifs avec un risque non négligeable de sur-comptage. Il en reste quelques uns (un peu moins de 6% du parc) localisés surtout sur des contrats acquis récemment par la Compagnie et donc "hérités" par des exploitants précédents.

3.4 La nouvelle réglementation aura un impact sur les stratégies de gestion

Les dispositions normatives actuelles prévoient la "*vérification primitive*" des compteurs d'eau (chapitre 2) : à la sortie d'usine tout appareil doit respecter les

Erreurs Maximales Tolérées (EMT) en correspondance des trois débits q_{min}, q_t, q_{max}.

En revanche, aujourd'hui aucune norme n'impose la "*vérification périodique*" des compteurs en service. Le seul cas où cette éventualité se présente (en pratique) est à l'occasion d'un litige entre le distributeur et l'abonné. Le compteur est jugé conforme si ses erreurs de mesure ne dépassent pas les EMT des compteurs en service, soit le double de celles prévues pour les compteurs neufs, aux trois débits réglementaires.

La nouveauté principale du projet d'arrêté, concernant la nouvelle réglementation, est l'introduction de la vérification périodique des compteurs en service, selon les modalités décrites dans le chapitre 2. Du point de vue des grands distributeurs d'eau, qui à long terme opteront pour un contrôle statistique par lots, l'introduction de la nouvelle norme se traduira par l'intégration, dans les règles de gestion des parcs de compteurs, de mesures visant à limiter, à l'intérieur de chaque lot, le taux de dispositifs non conformes en dessous du seuil fixé (10% en conditions de régime).

La définition des lots et des plans d'échantillonnage étant à la charge du distributeur, les nouvelles règles de gestion seront visiblement plus compliquées et devront intégrer aussi le découpage en lots des parcs et le calcul du nombre de compteurs à étalonner dans chaque lot.

Pour répondre aux exigences de la nouvelle réglementation, il sera nécessaire d'estimer la probabilité de non-conformité d'un compteur en service, en fonction d'un certain nombre de variables explicatives qui traduisent ses conditions d'exploitation (âge, consommation, agressivité du site) mais aussi l'incertitude caractérisant cette estimation.

Le modèle statistique développé dans cette étude a été conçu pour répondre à ces questions, aujourd'hui intéressantes, demain inévitables.

3.5 Quelques exemples de formulation théorique des stratégies de gestion

Un des premiers articles où les critères de gestion des parcs de compteurs sont explicitement formulés en termes mathématiques est dû à la Section de Californie de l'*American Water Works Association* (AWWA, 1966).

Les auteurs déterminent la période optimale de remplacement (où plutôt de réparation) d'un compteur sur la base du calcul des pertes économiques annuelles dues à l'eau non facturée. La durée de service optimale est celle où le cumul des pertes est égal au coût de remplacement/réparation.

L'étude concerne les compteurs de petit calibre (diamètres de 5/8 et 3/4 de pouce soit environ 15 et 20 mm) en service par la *California Water & Telephone Company*. Le calcul économique, mené sur la base d'une perte moyenne de rendement de 0.21% par an, arrivait à la conclusion que la durée de vie optimale des compteurs, pour l'exploitation examinée, était de 20 ans (figure 3.3).

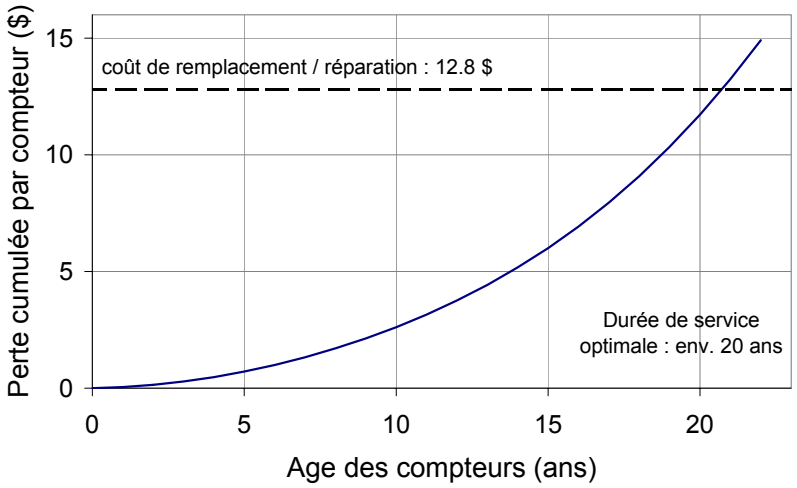


FIG. 3.3 – Calcul de la durée de service optimale d'un compteur (AWWA, 1966).

En plus du fait que cette étude semble surévaluer les performances des compteurs, compte tenu de la technologie de l'époque, on remarque que la méthode ne prend pas en compte la variabilité de la consommation entre les différents usagers, puisque le calcul économique est mené sur la base d'une consommation moyenne.

Une stratégie de gestion plus complexe est proposée par Noss et al. (1987). La durée de service optimale d'un compteur (au terme de laquelle l'appareil doit être remplacé ou réparé) est calculée en minimisant une fonction objectif, somme de trois termes :

1. Le coût annuel de la stratégie de gestion, fonction décroissante de la durée de service prévue T, étant le nombre de compteurs à remplacer chaque

année égal au rapport entre la taille du parc et la variable T. Le calcul se base donc sur l'hypothèse théorique que la pyramide des âges du parc de compteurs est plate.

2. La perte économique annuelle due aux volumes d'eau non facturés à cause du sous-comptage, calculés à partir d'un rendement moyen du parc et d'une consommation moyenne de référence. Le rendement moyen est évalué sur la base d'une loi de vieillissement linéaire : $A - B \cdot (T/2)$. Ce terme, est fonction croissante de l'âge moyen du parc et donc de la durée de service T.
3. La perte économique annuelle causée par l'eau non mesurée par les compteurs bloqués. Le calcul se base sur un taux de blocage annuel moyen et un temps moyen de permanence d'un compteur bloqué dans le parc (en pratique la période comprise entre deux relevés). Ce terme dépend en principe de la durée de vie des compteurs ; plus souvent on change les compteurs, plus on a de chance de remplacer un compteur bloqué et pas encore détecté par les releveurs. En réalité, ce terme, fonction décroissante de T, est négligeable par rapport aux deux premiers.

La recherche pratique du minimum de la fonction objectif se fait facilement et le calcul donne une expression exacte de la durée optimale (on omet volontairement cette formule ici).

Une analyse de sensibilité de cette formule montre une grande variabilité du résultat en fonction de la loi de vieillissement choisie : si on passe de –0.3% par an à –0.295% par an, on peut avoir des écarts dans la durée optimale de plus de 3 ans. A titre d'exemple, les calculs menés sur un exemple réel (ville de Taunton, Massachussets US) sur la base des indications de Newman et Noss (1982) concernant les pertes annuelles de rendement (environ –0.45% par an) et les taux de blocage (1.3% par an) donnent lieu à une durée optimale de service de 9 ans (figure 3.4).

Cette technique, comme la précédente, se base encore sur un calcul "*moyen*" valable pour tous les compteurs du parc.

On termine cette revue des principales études théoriques avec la méthode proposée par Lund (1988).

La stratégie se base sur le remplacement de tous les dispositifs ayant atteint une durée de service optimale T, calculée individuellement pour chaque compteur, et la détermination d'une règle de décision pour déposer les compteurs d'âge

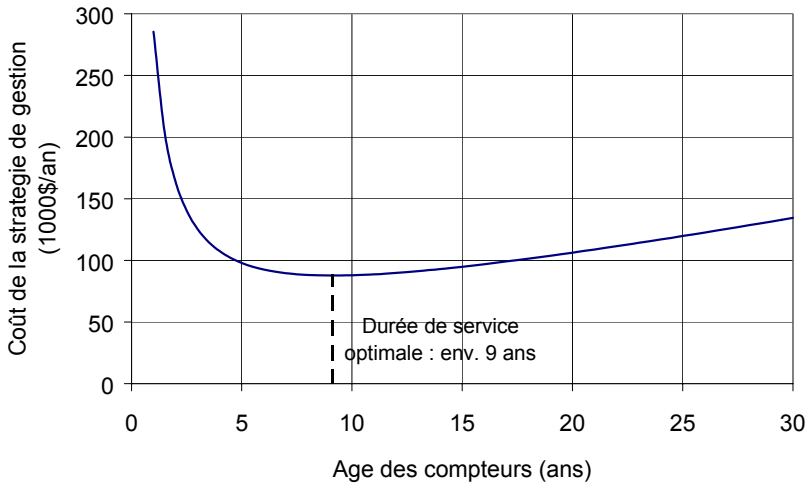


FIG. 3.4 – Exemple de calcul de la durée de service optimale d'un compteur (Noss et al. 1987).

inférieur à T, potentiellement bloqués. Les défaillances des compteurs sont parfois très facilement repérées sur le terrain ; c'est le cas quand l'appareil présente des signes évidents de détérioration (chocs mécaniques, gel, fraude). Dans d'autres cas la seule manière de détecter le blocage est la comparaison entre les deux index successifs, une différence anormalement faible étant imputable au blocage de l'appareil de mesure. L'auteur propose un méthode pour évaluer la probabilité[4] $[B|C]$ que le compteur se soit bloqué dans la période comprise entre l'$n$ème et l'$(n+1)$ème relevé, conditionnellement à la consommation C observée dans la même période (différence entre les deux derniers index). La probabilité $[B|C]$ peut être calculée avec la formule de Bayes :

$$[B|C] = \frac{[C|B] \cdot [B]}{[C|B] \cdot [B] + \left[C|\overline{B}\right] \cdot \left[\overline{B}\right]} \tag{3.2}$$

où $[B]$ est la probabilité de blocage dans la période entre les deux relevés et *vice versa* $\left[\overline{B}\right] = [not(B)] = 1 - [B]$, est la probabilité que le compteur ait fonctionné normalement dans cette période.

La probabilité $[B]$ peut être évaluée sur la base de la connaissance de la densité de probabilité $f(t)$ de la durée de vie du compteur, qu'on imagine déjà

[4]On utilise ici la notation *entre crochets* (Gelfand et Smith, 1990) des probabilités qui exprime avec $[A]$ la probabilité de l'évènement aléatoire A. On reviendra sur cette convention, et sur la formule de Bayes, dans le chapitre 4.

connue :

$$[B] = \int_{nT_L}^{(n+1)T_L} f(t)dt \tag{3.3}$$

où T_L est l'intervalle de temps entre deux relevés successifs.

Si le compteur a fonctionné normalement entre les deux relevés, alors la consommation observée C est égale (à l'erreur de mesure du compteur près) au volume réellement consommé C_r sur le branchement. La probabilité $\left[C|\overline{B}\right]$ est alors égale à la probabilité $[C_r]$ d'occurrence de la consommation C_r sur le branchement examiné. La loi de probabilité de C_r (qu'on imagine déjà connue) peut être, par exemple, modélisée en fonction de différents facteurs d'influence[5] parmi lesquels la catégorie d'usager, la saison etc.

En revanche, si le compteur s'est bloqué à l'instant u compris entre nT_L et $(n+1)T_L$ alors la consommation C n'est qu'une fraction de C_r qu'on peut estimer comme :

$$C = C_r \cdot \frac{u}{T_L} \tag{3.4}$$

La densité de probabilité de $C(u)$, qu'on peut considérer ici comme une fonction de l'instant de blocage u, s'obtient en multipliant la probabilité $[C_r] = [C \cdot T_L/u]$ par $f(u)$.

Par conséquent, la probabilité d'observer une consommation C, sachant que le compteur qui l'a enregistrée s'est bloqué dans la période comprise entre nT_L et $(n+1)T_L$ est exprimée par l'intégrale :

$$[C|B] = \int_{nT_L}^{(n+1)T_L} [C \cdot T_L/u]\, f(u)du \tag{3.5}$$

La mise en œuvre de la stratégie prévoit alors, sur la base du calcul de $[B|C]$, le choix parmi les alternatives possibles (remplacer le dispositif, procéder à une inspection *in situ*, ou ne rien faire) de celle qui minimise une fonction objectif,

[5] Parmi ces facteurs explicatifs on peut imaginer, à juste titre, d'inclure les caractéristiques (type, âge etc.) du compteur ayant enregistré les consommations dont on étudie la distribution de probabilité. De cette manière la consommation C_r serait interprétée comme " la consommation enregistrée par un compteur fonctionnant normalement dans la période comprise entre nT_L et $(n+1)T_L$ ".

somme des coûts d'intervention et des pertes économiques imputables à l'eau non facturée.

L'avantage principal de cette méthode est l'application des critères d'optimisation à chaque compteur. Cette approche est plus adaptée à la réalité opérationnelle où plusieurs types de compteurs sont présents dans le même parc et les consommations sont très variables entre les usagers.

3.6 La pratique technique de la gestion des parcs de compteurs

La définition d'une stratégie de gestion des parcs de compteurs a toujours été au cœur des préoccupations des distributeurs d'eau. Une étude réalisée aux Etats Unis dans les années 70 (Williams, 1976) montrait déjà que plus de 3/4 des distributeurs interpellés avaient des stratégies basées sur la dépose des appareils ayant atteint un âge limite ou, plus rarement un index maximal fixé, mais seulement 1/4 admettaient que la valeur de cette âge limite était issue d'un calcul économique réalisé *ad hoc*. Les valeurs des durées maximales de service étaient comprises entre 5 et 25 ans avec un pic dans la tranche 10-15 ans.

S'il a été déjà mis en évidence qu'une stratégie de ce type, basée sur un unique critère d'âge maximal pour tous les appareils, a des limites, il faut aussi observer que la mise en place de règles de gestion plus efficaces nécessite des moyens importants que souvent les petites collectivités n'ont pas. *A contrario*, les grands distributeurs d'eau français, et notamment la CGE, ont la capacité de définir des stratégies de comptage de plus en plus optimisées à leurs parcs de compteurs. L'approche suivie consiste en un calcul économique sur chaque compteur installé, sur la base du type d'appareil, de son rendement présumé (fonction d'un certain nombre de facteurs explicatifs et notamment de l'âge) et de la consommation enregistrée sur le branchement. Néanmoins dans la pratique technique la gestion des parcs est plus compliquée et les calculs économiques théoriques ne suffisent pas à définir un programme de renouvellement.

Tout d'abord, il y a les contraintes budgétaires, les investissements consacrés au comptage étant fixés année par année sur la base des exigences et de la disponibilité financière de l'entreprise. Ensuite il y a des contraintes locales et notamment les limitations d'âge fixées par certaines collectivités et, plus rarement,

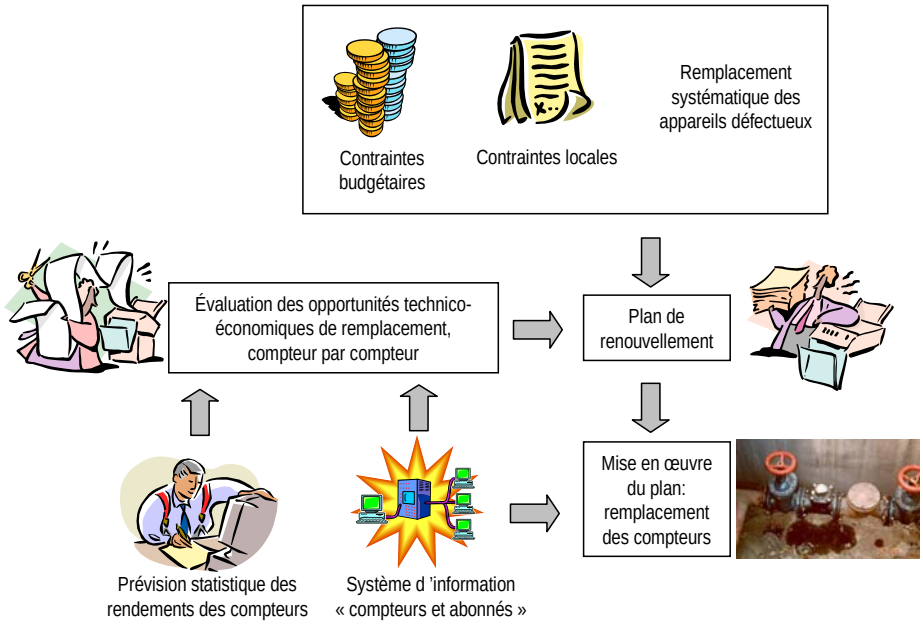


FIG. 3.5 – Gestion des parcs de compteurs d'eau en pratique.

des situations exceptionnelles (campagnes de dépose systématique de types de compteurs défectueux) qui peuvent complètement monopoliser les ressources (financières et humaines) destinées au renouvellement des compteurs. La réalité des grands distributeurs d'eau est très loin des règles théoriques de gestion suggérées dans la bibliographie technique (figure 3.5).

A l'échelle mondiale, les critères sont très variables d'un pays à l'autre et la stratégie utilisée dépend de facteurs locaux tels que la qualité de l'eau, les modalités de distribution (constante ou intermittente comme dans certains pays en voie de développement) mais aussi le prix de l'eau par rapport au prix des compteurs. Aujourd'hui encore l'opportunité économique du comptage est parfois mise en cause là où le prix de l'eau est très faible comme en Inde (Stedman, 2003).

Pour les gros consommateurs, compte tenu des enjeux économiques, de la grande variabilité des consommations et des modalités de puisage, un suivi au cas par cas peut être envisageable (Van der Linden, 1998), alors que pour les petits compteurs, les valeurs des variables utilisées pour le calcul économique, et notamment les rendements des dispositifs, sont déduites à partir d'études statistiques.

Des calculs économiques, compteur par compteur, sont aujourd'hui à la base de la politique de gestion de la CGE. Volontairement on ne rentre pas dans le détail du calcul (pour des raisons de confidentialité) mais on veut encore une fois souligner que, quelle que soit la règle de décision sur l'opportunité technico-économique de remplacement, il est indispensable de bien connaître le comportement des compteurs installés et notamment leur rendement, les critères économiques étant en général très sensibles à ce dernier paramètre. A ce propos les distributeurs d'eau procèdent à des étalonnages de compteurs en service, expressément déposés et testés dans leurs propres laboratoires d'essais. Ces données sont à la base des études statistiques mises en œuvre.

Le deuxième élément de la réussite d'une stratégie de gestion est un système d'information capable de fournir, pour chaque branchement, le maximum de renseignements sur le compteur (type, diamètre, âge, dernier index) et sur l'abonné (localisation, typologie, dernières consommations). Un tel outil est indispensable à la fois pour la définition d'un programme de remplacement, puisqu'il fournit les variables du calcul économique, et pour la mise en œuvre (programmation des interventions de déposes par zones géographiques). L'amélioration du comptage passe, plus que jamais, par la fiabilisation des systèmes d'information (Vlontakis, 2000).

Deuxième partie

Modélisation statistique de la dégradation des compteurs

Chapitre 4

De l'examen des données à un modèle conceptuel

4.1 Les données métrologiques

L'étalonnage des compteurs en service permet aux distributeurs d'eau de connaître le comportement des appareils de mesure en conditions réelles d'exploitation et constitue la base pour toute réflexion sur la gestion des parcs de compteurs. Les résultats des étalonnages sont des courbes métrologiques construites point par point en correspondance d'un nombre significatif de débits, choisis en fonction du calibre du compteur, sur une grille standard. Les histogrammes de consommation (représentations des proportions de la consommation ayant lieu dans une tranche donnée de débits) sont normalement référés à la même grille, ce qui rend plus facile le calcul du rendement du compteur étalonné.

Le Groupe Veolia Water, dont la Compagnie Générale des Eaux (CGE) fait partie, dispose actuellement de trois laboratoires d'essais : le LECE (*Laboratoire d'Essai des Compteurs d'Eau*) à Vandoeuvre-Lès-Nancy, les installations du Service Compteurs de la SADE à Ivry-sur-Seine et un troisième laboratoire, plus petit, à Lyon. Le LECE et le laboratoire de la SADE sont accrédités par le COFRAC (*Comité Français d'Accréditation*). L'accréditation COFRAC reconnaît la conformité aux critères généraux de qualité prévus par la norme ISO/CEI 17025 et l'aptitude à réaliser des étalonnages dans un domaine défini avec des incertitudes précisées dans le programme d'accréditation. Les critères de conformités pour un laboratoire d'essai de compteurs d'eau sont déterminés par le

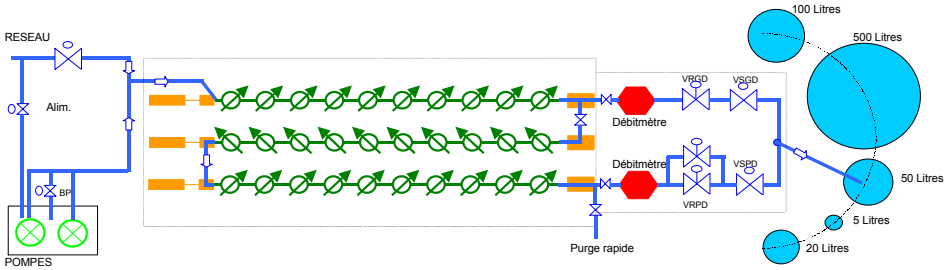


FIG. 4.1 – Schéma hydraulique du banc d'essai du LECE.

document COFRAC n. 2023, qui analyse toutes les sources possibles d'incertitude sur le calcul de l'erreur de mesure d'un compteur et en fixe les valeurs maximales acceptables.

Les données métrologiques utilisées dans le cadre de cette étude proviennent du LECE.

Au LECE sont étalonnés les compteurs de provenance de toutes les exploitations françaises (hors SEDIF, de compétence de la SADE) et occasionnellement étrangères de diamètre jusqu'à 40 mm.

La figure 4.1 montre un schéma hydraulique du banc d'essai. L'eau, pompée à débit constant à travers les compteurs disposés en série sur 3 rails est envoyée vers des jauges de différentes capacités (selon le débit d'étalonnage). Quand le volume d'eau prévu a rejoint la jauge, l'essai s'arrête et l'erreur de mesure est calculée en comparant la différence d'index après et avant le test (volume mesuré) avec le volume réellement transité.

Les débitmètres d'une part assurent la constance du débit et d'autre part mesurent le volume écoulé, déterminant l'arrêt automatique de l'essai.

Le banc de mesure peut accueillir simultanément un nombre maximal de 30 compteurs (volumétriques)[1] de DN 15 mm.

L'ensemble des étalonnages LECE, avec les essais des gros compteurs de la SADE, constitue une importante base de données, dite *Base Métrologique Nationale* (BMN). A l'heure actuelle la BMN comprend un peu plus de 32 000 résultats d'étalonnages dont environ 27 000 concernant des compteurs en service. Parmi eux un peu plus de 22 000 sont exploitables pour des études métrologiques.

[1] Les compteurs de vitesse sont sensibles aux perturbations hydrodynamiques engendrées par la présence en amont en en aval d'autres dispositifs. Par conséquent, le nombre d'étalonnages simultanés de ces appareils de mesure, limité par ce phénomène d'interférence plutôt que par les dimensions du banc, est inférieur à celui relatif aux compteurs volumétriques.

FIG. 4.2 – Le banc d'essai des compteurs du LECE.

Pour qu'un étalonnage soit exploitable il faut que les erreurs de mesure soient calculées au moins en correspondance de 5 débits d'essai. On estime que 5 points sont suffisants pour donner une appréciation globale de la métrologie du compteur et notamment pour déterminer son *état métrologique* (figure 4.7, page 51).

En revanche pour le calcul du rendement, seuls les étalonnages complets (courbe métrologique déterminée sur la base des erreurs de mesure en correspondance de tous les débits d'essai prévus par la grille standard) sont retenus.

Les 22 000 étalonnages exploitables sont répartis par diamètre nominal, selon le schéma de la figure 4.3. On peut remarquer que les compteurs de petit calibre (DN 15 mm) sont largement prédominants dans la base de données.

Concernant la répartition des étalonnages par type de compteur on retrouve dans la Base Métrologique tous les compteurs actuellement utilisés en France par la CGE (figure 4.4).

Dans le groupe des "*Autres Volumétriques*" on a regroupé tous les vieux types de compteurs volumétriques (notamment PSM, SOCAM 501 et 510, Stella J) qui présentent un intérêt limité dans le cadre de cette étude puisque ils ne sont plus installés et sont en voie de disparition dans les parcs de compteurs. Dans le même principe on a regroupé tous les compteurs à jets multiples (notamment Corona, SOCAM 401 et 410 et Doris SM). Enfin le groupe des "*Autres*" est formé par tous les compteurs qui n'ont pas d'intérêt pour des finalités métrologiques (vieux types

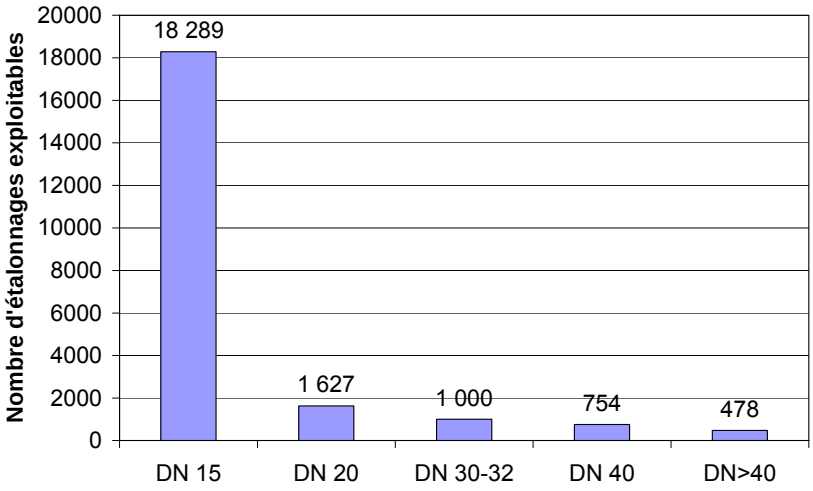


FIG. 4.3 – Répartition des essais métrologiques par diamètre.

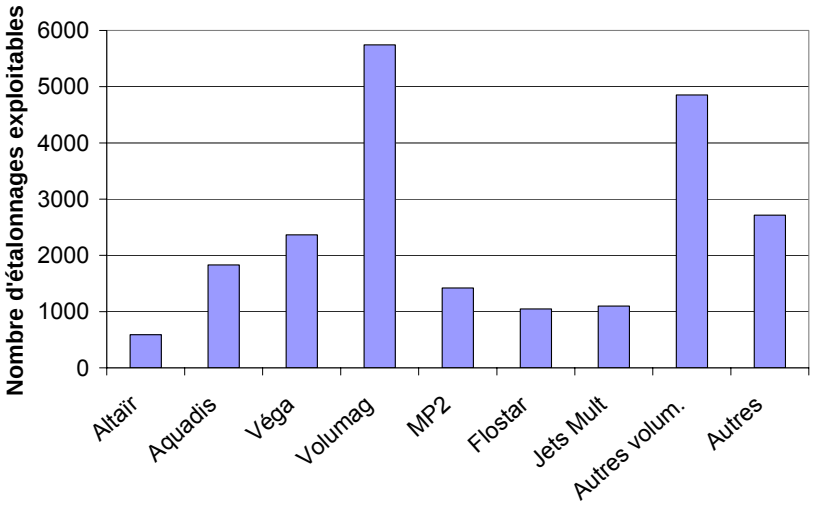


FIG. 4.4 – Répartition des essais métrologiques par type de compteur.

de compteurs pratiquement inexistants dans les parcs ou, *a contrario*, nouveaux compteurs, comme le Marly 2, dont l'apparition est trop récente pour pouvoir espérer un retour d'expérience).

4.2 Les données de facturation

L'ensemble des données utilisées pour la facturation des clients, stockées dans des bases informatisées, constitue un important réservoir d'informations sur les compteurs.

En particulier, dans cette étude on a utilisé des données issues d'une base appelée ICBC (*Info-centre Compteurs, Branchements et Clients*). Il s'agit d'un outil interne de gestion des parcs de compteurs obtenu par déversement périodique des informations des bases de facturation. Actuellement le périmètre géographique d'ICBC couvre la plupart des exploitations françaises de la CGE. La figure 4.5 montre les "*régions*" dont l'outil de facturation est compatible avec ICBC. Les *régions* sont des unités géographiques issues d'un découpage administratif du territoire français interne à l'entreprise. La France métropolitaine est aujourd'hui divisée en 11 régions dont 8 font partie du périmètre ICBC. Des détails ultérieurs sur l'organisation géographique de la CGE, seront donnés dans le chapitre 6, consacré à la caractérisation des différentes exploitations quant à leur "*agressivité*" vis à vis des compteurs.

Les informations de la base ICBC sont contenues en plusieurs tables dont les principales sont :

- La table "*Abonnés*" avec les indications relatives aux clients. Chaque abonné est détecté avec un code numérique à 16 chiffres. Les premiers 14 chiffres codifient le branchement et les 2 derniers sont un "*numéro de rang*" qui est augmenté d'une unité à chaque fois que sur le même branchement un abonnement est fermé et un autre est ouvert à sa place. Les informations principales contenues dans cette table sont : numéro et nom de l'abonné, adresse de consommation et facturation, indications géographiques sur la base du découpage du territoire de la CGE (agence, centre opérationnel, contrat), dernière consommation annuelle.
- La table "*Compteurs*" qui constitue un véritable *Registre d'Etat Civil* des compteurs. De chaque appareil on connaît le numéro de série, le numéro de l'abonné correspondant, le type, le diamètre l'année de fabrication, et

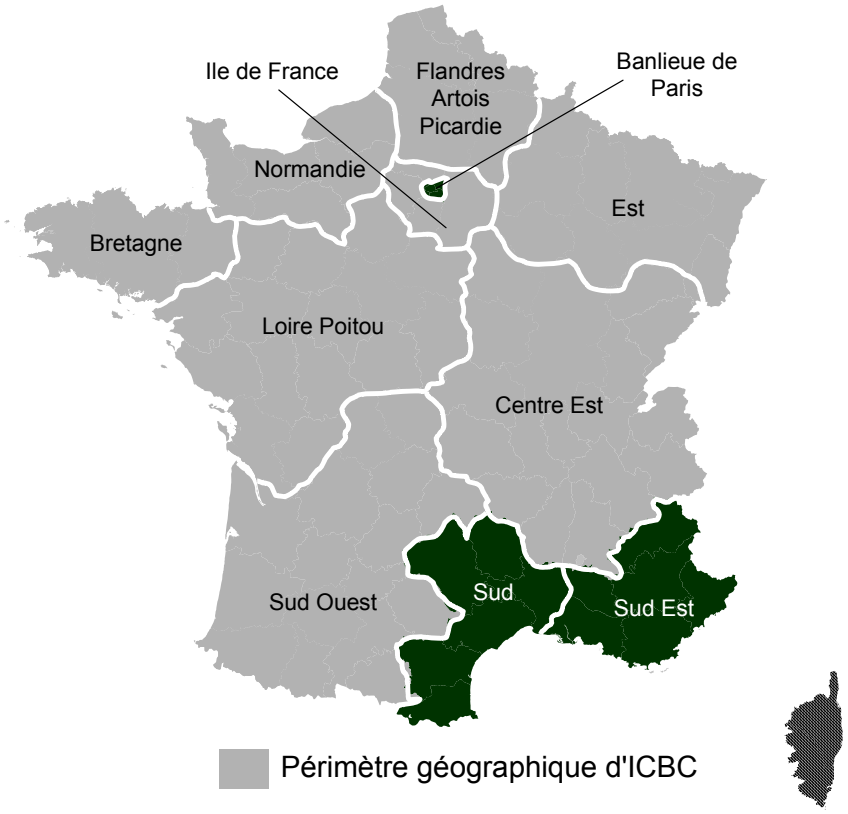


FIG. 4.5 – Couverture territoriale de la base de données ICBC.

des indications utiles à la politique de gestion (notamment le rendement estimé et un paramètre qui exprime l'intérêt technico-économique de son remplacement). Dans cette table on trouve les compteurs actuellement en service et les compteurs déjà déposés. Un paramètre dit "*Etat de Pose*" discrimine les compteurs opérationnels des compteurs qui n'existent plus dans les parcs ("*P*" pour les compteurs en service et "*D*" pour les compteurs déposés).

- La table "*Changement des compteurs*" où toutes les interventions de poses et déposes sont répertoriées. A chaque intervention correspond une ligne dans cette table et un code "*Intervention*" distingue les opérations de dépose de celles de pose. On y trouve notamment la date de l'intervention, le numéro de l'abonné, le numéro du compteur, l'index du compteur, et le motif de pose ou dépose. Parmi les causes de dépose on distingue trois grandes familles : les déposes préventives, curatives, administratives. Dans le premier groupe (75 à 80% des interventions) on trouve tous les remplacements réalisés dans le cadre de la politique de gestion des parcs de compteurs. L'appellation "*préventive*" met en évidence le fait que le compteur déposé fonctionne et il est renouvelé avant de pouvoir engendrer des problèmes. Le second groupe (10 à 15% du total) comprend les déposes irrévocables décidées suite à l'observation d'une défaillance du dispositif en place : blocage, choc mécanique, rupture due au gel, fraude. Enfin les déposes administratives se vérifient, parfois, lors d'un changement d'abonné sur le même branchement ou à l'occasion de la fermeture d'un branchement. Ces interventions représentent 5 à 10% des déposes totales.
- La table "*Historique des relevés*" qui contient les informations relatives à la lecture des index : numéro d'abonné, numéro du compteur, index, nature de l'index (lu par le releveur, estimé, communiqué par l'abonné, radio-relevé).

Du point de vue de l'étude du comportement des compteurs en service, ICBC donne des informations importantes concernant les défaillances des compteurs et notamment les phénomènes de blocage. Ces données seront utilisées pour caractériser l'agressivité des conditions d'exploitations des différents sites français. Le grand avantage d'ICBC est que cette base nous renseigne sur la quasi-totalité des compteurs français, alors que les données métrologiques sont forcement établies sur la base d'un échantillon de la population des compteurs en service.

La principale difficulté dans l'utilisation de ces informations est d'ordre pratique, la base de données n'ayant pas été conçue pour réaliser des études statistiques mais simplement comme un outil de gestion des parcs de compteurs. Par exemple, dans la table "*Changement des compteurs*" on n'a pas toutes les informations dont on aurait besoin concernant les appareils déposés ; la récupération de ces données dans la table "*Compteurs*" n'est pas immédiate parce que le numéro de série du compteur n'est pas un code unique à niveau national et le numéro de l'abonné peut avoir changé après la dépose de l'appareil (cf. chapitre 6). Il aurait fallu tout simplement prévoir dans la table "*Compteurs*" un champ "*Date de dépose*" pour simplifier énormément la phase de collecte de l'information.

L'autre question qu'on pourrait soulever concerne la fiabilité des données. Certaines informations non indispensables pour la facturation (comme le motif de dépose) pourraient être mal renseignées par les exploitants à l'acte de saisie dans l'outil qui permet d'établir les factures (source des données ICBC). Au cours de ces années on a constaté une amélioration de la qualité de l'information (réduction des compteurs de type ou âge inconnu, et des motifs de déposes inconnus) et on a raison de croire en la fiabilité de cette base de données, mais on sait que quand on mène des études à caractère strictement local, comme celle concernant les taux de blocage, on risque de prendre en compte des informations aberrantes sur un nombre (heureusement limité) de sites.

Un des résultats atteints dans ce travail de thèse est justement de montrer aux exploitants comment ces informations (apparemment complémentaires) peuvent être utilisées pour fournir des éléments importants dans la gestion des parcs de compteurs et les motiver à veiller sur leur bon renseignement dans les bases de données.

4.3 Un problème compliqué de modélisation

Parmi les paramètres techniques qui caractérisent un compteur, le rendement est sans doute le plus important parce qu'il donne immédiatement une estimation de la consommation non facturée et par conséquent des pertes économiques. D'ailleurs, les stratégies de gestion optimale des parcs de compteurs se basent normalement sur des calculs technico-économiques où intervient la prévision des rendements des compteurs en fonction de leurs caractéristiques d'exploitation et

notamment de l'âge.

La relation qui exprime comment le rendement d'un type de compteur dépend de l'âge est normalement appelée dans le langage technique "*loi de vieillissement*" du compteur. Dans cette appellation le terme "*vieillissement*" ne doit pas être interprété dans le sens communément attribué par les statisticiens dans le domaine de la fiabilité, mais simplement comme synonyme de "*dégradation*" de l'appareil.

Une approche simple au problème de l'étude des compteurs en service pourrait être la modélisation directe de la relation entre âge et rendement à l'aide d'un modèle linéaire, comme par exemple dans (Newman et Noss, 1982).

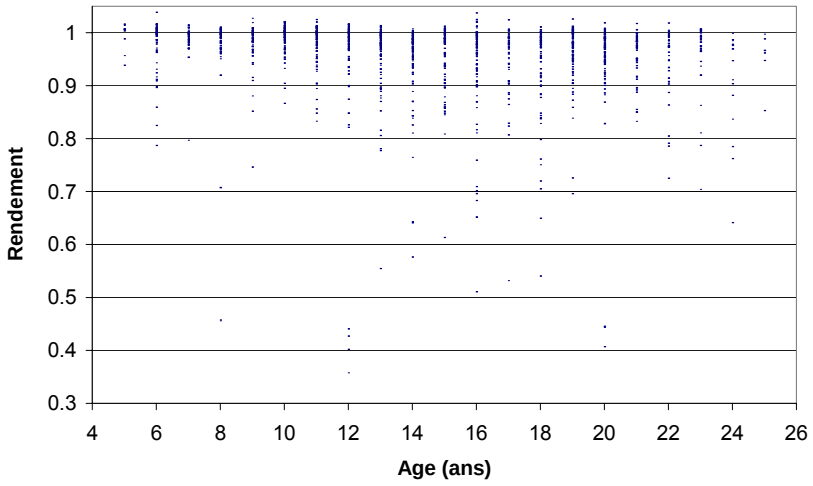


FIG. 4.6 – Exemple de données métrologiques. Visualisation de la relation rendement-âge pour un échantillon de 1 400 compteurs volumétriques.

Un coup d'œil aux données montre que la mise en œuvre de ces techniques de base dans le cadre de cette étude n'est pas possible. La figure 4.6 montre, sous forme de nuage de points les résultats d'environ 1 400 étalonnages de compteurs volumétriques de DN 15 mm, tous de même type, d'âges compris entre 5 et 25 ans. Les données montrent, de manière assez claire, la coexistence, à l'intérieur d'échantillons de même âge, de compteurs ayant des rendements très proches de l'unité, de compteurs à métrologie plus imprécise mais dont le rendement reste dans l'intervalle [0.8,1] et de compteurs nettement défaillants qui ont des rendements distribués de manière très dispersée entre 0.4 et 0.8. On remarque aussi, que, au fur et à mesure que l'âge augmente, on a plus de chances de trouver parmi

les individus échantillonnés, des compteurs à faible rendement, et par conséquent le nuage a tendance à se disperser. En d'autres termes les observations n'ont pas les mêmes variances (les statisticiens parlent de non respect de l'hypothèse de *homoscédasticité*).

Ce problème est bien connu par les experts du comptage qui ont toujours remarqué la présence, dans les parcs de compteurs, d'appareils particulièrement défectueux, présentant un sous-comptage important, mais qui ne sont pas bloqués. Puisque leurs index continuent de progresser, ces dispositifs défaillants sont pratiquement indécelables sur le terrain et pour cette raison les ingénieurs les appellent parfois "*compteurs sous-marins*".

Compte tenu des données, un modèle dynamique de mélange de plusieurs populations de compteurs, en proportions variables en fonction de l'âge, semble plus adapté au problème, qu'un modèle "classique" de régression qui cherche à décrire le comportement de l'individu moyen et quantifier les écarts par rapport à ce comportement.

4.4 Quatre groupes de compteurs

Sur la base des considérations inspirées par l'analyse exploratoire des données métrologiques on a imaginé de découper les compteurs en service en quatre groupes de qualité métrologique décroissante.

Le découpage est suggéré par la réglementation actuelle et les différents groupes sont ainsi définis (figure 4.7) :

- **Groupe 1** : compteurs dont la courbe métrologique est entièrement contenue dans les canaux fixés par les erreurs maximales tolérées (EMT) des compteurs en service (cf. chapitre 2). On rappelle que les valeurs des EMT sont $\pm 10\%$ à faible débit (entre q_{min} et q_t) et $\pm 4\%$ à moyen et fort débit (entre q_t et q_{max}). Quand on dit que la courbe métrologique doit être contenue "*entièrement*" cette assertion veut dire que les erreurs de mesure doivent être comprises entre les EMT en correspondance de tous les débits d'étalonnage prévus. Il n'est pas superflu de rappeler que cette condition est beaucoup plus restrictive que celle requise pour la conformité réglementaire, qui prévoit le respect des EMT en correspondance de seulement deux débits d'essai. Les compteurs appartenant à cette catégorie ont une métrologie excellente, proche de celle d'un compteur neuf, et des rende-

ments très proches de l'unité.

- **Groupe 2** : compteurs dont la courbe métrologique sort en quelques points des canaux de tolérance mais respecte les valeurs fixées par les EMT au moins en correspondance des débits : 0.2 q_n et 0.9 q_n (0.5 q_n si $q_n \geq 10$ m^3/h). Ces dernières restrictions sont les conditions minimales de conformité demandées par la nouvelle réglementation (cf. chapitre 2). En pratique appartiennent à cette catégorie les compteurs à métrologie imparfaite mais encore acceptable du point de vue réglementaire.
- **Groupe 3** : compteurs dont la métrologie ne respecte pas les conditions de conformité réglementaire, mais encore en état de marche. Ces dispositifs ont évidemment des rendements très faibles et leur présence dans le parc est doublement dangereuse parce que, en plus d'être non-conformes et de sous-estimer nettement la consommation, ils sont aussi difficilement décelables lors des relevés d'index. On peut identifier cette catégorie de compteurs comme les fameux "*sous-marins*" dont parlent les experts.
- **Groupe 4** : compteurs bloqués à tout débit et n'enregistrant aucune consommation. Ces appareils sont assez facilement détectés sur le terrain par les releveurs lors de la lecture périodique de l'index et aussitôt remplacés, dans un délai de quelques semaines. En pratique, il y a très peu de compteurs bloqués dans les parcs et la probabilité d'échantillonner aléatoirement un appareil bloqué est très faible. En plus les types de compteurs actuellement utilisés sont de moins en moins affectés par les phénomènes de blocage. Par conséquent, les rarissimes compteurs bloqués qu'on retrouve dans la base métrologique ne sont absolument pas représentatifs des défaillances observées sur le terrain et pour estimer l'occurrence de ces phénomènes il faudra faire appel à d'autres sources d'informations (notamment les données de la base ICBC). Une étude détaillée des phénomènes de blocages est développée dans le chapitre 6.

Concernant la relation entre les groupes ainsi définis et les rendements, la figure 4.8 montre clairement les différences entre les 3 premiers groupes de compteurs (le quatrième groupe n'est pas pris en compte parce que c'est celui des compteurs à rendement nul). Ces histogrammes ont été établis sur la base des mêmes données représentées dans la figure 4.6 classées en trois catégories selon les critères de découpage précédemment décrits.

Les répartitions des fréquences empiriques des rendements sont cohérentes

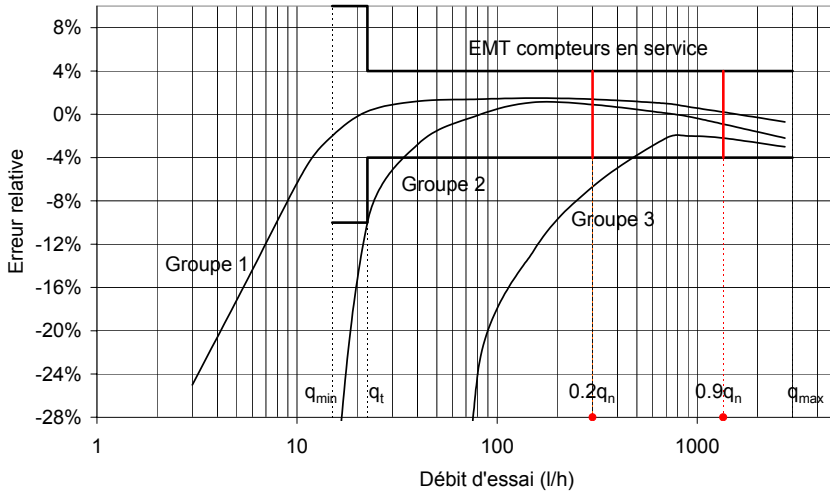


FIG. 4.7 – Découpage de groupes de compteurs sur la base de la signature métrologique (exemple relatif à DN = 15 mm).

avec la notion de qualité métrologique associée à chacune des trois catégories. Dans le premier groupe on trouve des compteurs avec des rendements très proches de l'unité et relativement peu dispersés et, au fur et à mesure qu'on passe dans les catégories les moins performantes, le rendement moyen baisse et la dispersion augmente.

L'analyse des répartitions par groupes des échantillons de même âge (figure 4.9) montre que la proportion de "*bons*" compteurs est une fonction décroissante de l'âge des dispositifs et que, inversement, au fur et à mesure que cette variable augmente la proportion de compteurs appartenant aux catégories moins performantes a tendance à augmenter.

4.5 Un modèle conceptuel de dégradation des compteurs

Le point de départ du développement d'un modèle conceptuel qui explique la dégradation des compteurs est l'interprétation des quatre catégories de compteurs comme des *états* ($X(0)$, $X(1)$, $X(t)$...) à travers lesquels chaque appareil passe au long de sa vie opérationnelle.

Concernant l'état initial, une hypothèse réaliste est que tout appareil dé-

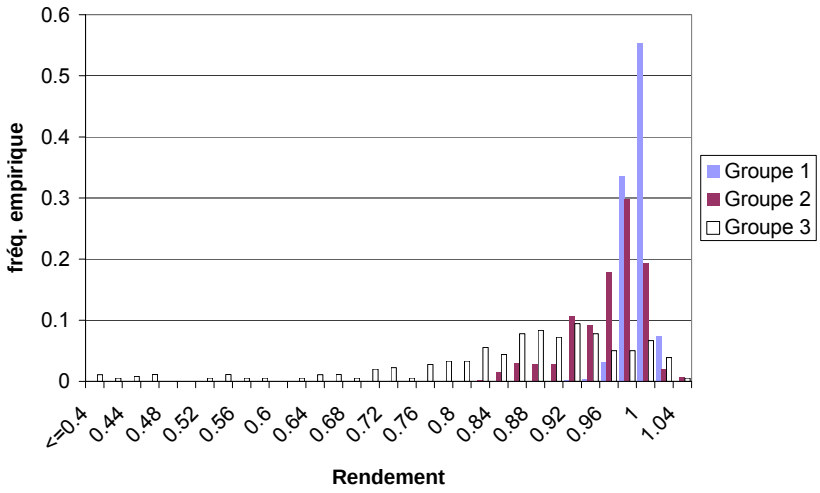


FIG. 4.8 – Distribution des rendements en fonction des "groupes" de compteurs.

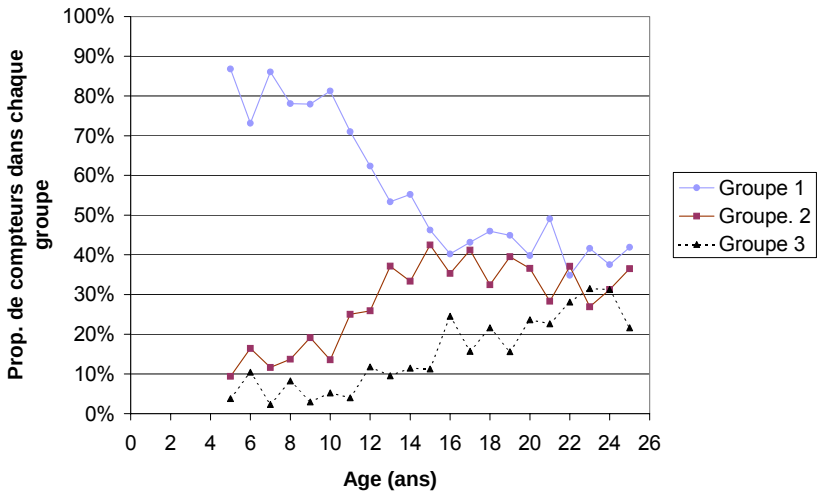


FIG. 4.9 – Evolution de la répartition par groupes des échantillons en fonction de l'âge.

marre sa vie opérationnelle dans l'état 1. Cette hypothèse est justifiée par le fait que tout compteur, à la sortie d'usine, doit respecter les conditions prévues par la *vérification primitive* (erreurs inférieures, en valeur absolue, aux EMT des compteurs neufs en correspondance des débits q_{min}, q_t, q_{max}). Compte tenu du fait que les EMT des compteurs neufs sont la moitié des EMT des compteurs en service on peut imaginer que la conformité réglementaire *à neuf* entraîne aussi le respect des conditions prévues par l'état 1.

N'ayant pas à disposition d'observations répétées sur les mêmes individus, on n'observe pas réellement les modalités des transitions mais on peut imaginer que la probabilité qu'un compteur se trouve dans un état donné à un âge t dépend uniquement de son état à l'âge t-1 et est indépendante de l'histoire du compteur dans la période [0, t-1[. Du point de vue mathématique cette hypothèse traduit le fait que la succession des états du système $X(t)$ est une chaîne de Markov[2]. Du point de vue de l'ingénieur, une dynamique de ce type exprime le fait que la dégradation d'un compteur est due à une succession d'événements aléatoires qui le font basculer d'une catégorie à une autre plutôt qu'à une dégradation continue due, par exemple, à des phénomènes d'usure progressive.

En réalité les deux dynamiques (continue et "*par accidents*") coexistent dans la dégradation des compteurs mais on peut imaginer que, en conditions normales d'utilisation, l'effet des accidents est prévalent, ce qui est confirmé par le fait que certains dispositifs démontés et analysés après plusieurs années de bons et loyaux services ne présentent pas d'usure apparente et ont des rendements pratiquements équivalents à ceux des compteurs neufs.

Cette conclusion pourrait sembler en contraste avec les résultats des essais d'endurance réalisés par les fabricants qui consistent à faire tourner, en eau claire, des compteurs neufs à débit constant très élevé (typiquement égal à q_{max}) pendant des cycles de durée fixée. Effectivement, si on étalonne les compteurs au terme de chaque cycle et on compare entre elles les courbes métrologiques, on observe une dégradation de la métrologie due à un phénomène d'usure, mais, si ces tests donnent des indications importantes sur l'endurance du dispositif, ils ne sont absolument pas représentatifs des conditions réelles d'exploitation. D'une part, des débits de fonctionnement si élevés se vérifient très rarement et d'autre part, dans la réalité, les compteurs sont sujets à des agressions difficilement

[2]En l'honneur du mathématicien russe Andrei Andreyevich Markov (1856-1922) à qui l'on doit les bases de la théorie des processus stochastiques.

imaginables et reproductibles en laboratoire : surpressions, arrêts et démarrages brutaux, passages de particules solides, dont l'effet sur la métrologie est loin d'être quantifié.

Revenons maintenant aux aspects mathématiques du problème.

Puisque on regroupe les compteurs de même âge pour étudier la probabilité d'occurrence des différents états, il est naturel de faire référence à une variable discrète pour exprimer le temps. On peut utiliser l'âge, exprimé en ans, mais aussi, quand les échantillons sont de taille plus petite, des classes d'âge (0 à 4 ans, 5 à 9 ans etc.). Le modèle statistique qu'on utilise pour décrire la vie d'un compteur est donc markovien à temps discrets.

Ces modèles, si on suppose connu l'état initial du système, sont paramètrés par les probabilités de transitions θ_{ij} d'un état i à un autre j. En fonction de l'état initial il est alors possible d'écrire les probabilités d'appartenance aux différents états pour toute valeur de la variable temporelle.

Concernant les probabilités de transition on imagine qu'elles ne dépendent pas du temps ; en langage statistique on dit que le modèle est *homogène*. Cette hypothèse signifie qu'on néglige le vieillissement du système dans la dynamique de dégradation et que les vieux appareils ont la même probabilité de subir un accident que les jeunes.

Une autre hypothèse importante qu'on fait sur les probabilités de transitions est que le processus de dégradation est irréversible, c'est à dire que le système ne peut pas revenir en arrière ou, de manière équivalente, que la métrologie d'un compteur ne peut jamais s'améliorer avec l'âge. Par conséquent l'état n. 4 (compteur bloqué) est sans retour (on dit qu'il est *absorbant* pour la chaîne) et pour $t \to +\infty$ la probabilité que le système se trouve dans l'état 4 tend vers 1.

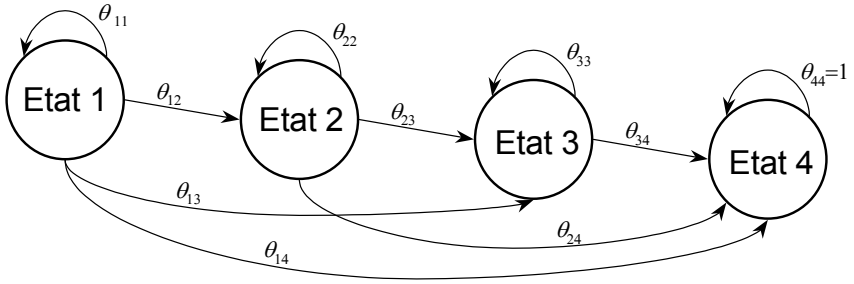


FIG. 4.10 – Mécanisme de dégradation des compteurs "par états" .

La figure (4.10) montre une représentation schématique du modèle de dégra-

dation. La formulation mathématique du problème et l'estimation des probabilités de transition seront l'objet du chapitre suivant.

4.5.1 Une schématisation bien générale

Les modèles markoviens sont très largement utilisés en statistique appliquée notamment quand on peut associer à un système complexe un certain nombre (fini ou infini) d'états (observables ou non) qui résument ses conditions de fonctionnement.

Les applications en fiabilité où chaque état exprime la notion de dégradation du système sont nombreuses.

Ang et Tang (1984) développent par exemple, un modèle irréversible de dégradation de la stabilité d'un immeuble, suite à des tremblements de terre répétés, basé sur 5 états d'endommagement de plus en plus graves (absence de dégâts, endommagement léger, modéré, important, effondrement de l'immeuble). Dans le même esprit, des schématisations markoviennes sont utilisées pour modéliser les conditions de santé d'individus atteints de maladies graves (notamment le SIDA), comme dans Guihenneuc-Jouyaux et al. (2000).

Enfin une autre application intéressante est donnée par Dhillon et Yang (1997) qui s'occupent de la fiabilité d'un système à plusieurs composantes. Dans cet exemple les états sont définis par le nombre de composantes défaillantes au même temps.

Dans un domaine d'application complètement différent, Abi-Zeid et Bobée (1994) fournissent un résumé des principales applications de cette classe de modèles en hydraulique et hydrologie, parmi lesquelles on signale la modélisation des apports fluviaux (Parent et al., 1991) et lacustres (Duckstein et Bogardi, 1979) dans un système hydraulique. Plus récemment, Jimoh et Webster (1999) ont fourni un exemple de modélisation des précipitations pluviales avec une chaîne de Markov à 2 états (pluie, beau temps).

Les processus markoviens sont aussi couramment employés en écologie. Un aperçu général de l'utilisation de ces méthodes dans ce domaine d'application est donné par Legendre et Legendre (1979). Des exemples typiques sont la modélisation des cycles de vie (Caswel, 1989) et en particulier de la croissance des arbres (Guédon, 1997) et de la succession des espèces végétales sur une parcelle de réserve naturelle (Coquillard et Hill, 1997).

Enfin les sciences sociales ont aussi fourni différents exemples d'application

de cette classe de modèles. Parmi eux (Bickenbach et Bode, 2001) la modélisation des incréments rélatifs des revenus annuels *pro capita* d'une population (processus à 5 états). Un résumé d'autres applications se trouve dans Berchtold (1998) : entre autres, l'obtention d'une maîtrise universitaire (modèle à 6 états correspondants aux 4 années d'inscription, plus les deux états absorbants "Obtention du diplôme" et "Abandon des études"), les états civils successifs d'une personne, le nombre de passagers dans un avion, le prix de l'or (avec 3 classes de prix).

4.6 Modélisation "indirecte" du rendement en fonction de l'âge

Le problème initial de trouver la relation entre rendement et âge des compteurs est abordé de manière indirecte. Si à chaque état de marche (1, 2, 3) on associe une distribution de probabilité des rendements (dont les paramètres sont éventuellement fonctions de l'âge), alors, sachant les probabilités d'occurrence des quatre états, il est simple de simuler par la méthode de Monte Carlo la distribution de probabilité d'un parc de compteurs d'âge donné t.

Pour réaliser un tirage aléatoire dans cette distribution de probabilité il faut dans un premier temps effectuer un premier tirage dans un modèle d'urne à 3 dimensions (annexe B) pour établir dans quel état se trouve le compteur. Les paramètres de ce modèle d'urne sont les probabilités conditionnelles des 3 états de marche, sachant que le dispositif n'est pas bloqué.

Ensuite on réalise un deuxième tirage dans la loi de probabilité des rendements correspondant à l'état obtenu. La valeur obtenue est issue de la distribution de probabilité des compteurs d'âge t. La répétition de cette procédure un grand nombre de fois permet de simuler cette loi (figure 4.11).

En réalité le problème est un peu plus complexe parce que les probabilités d'occurrence des 4 états sont, elles aussi, des variables aléatoires. Une description complète de la méthode est donnée dans le cadre de l'exemple développé dans le chapitre 7.

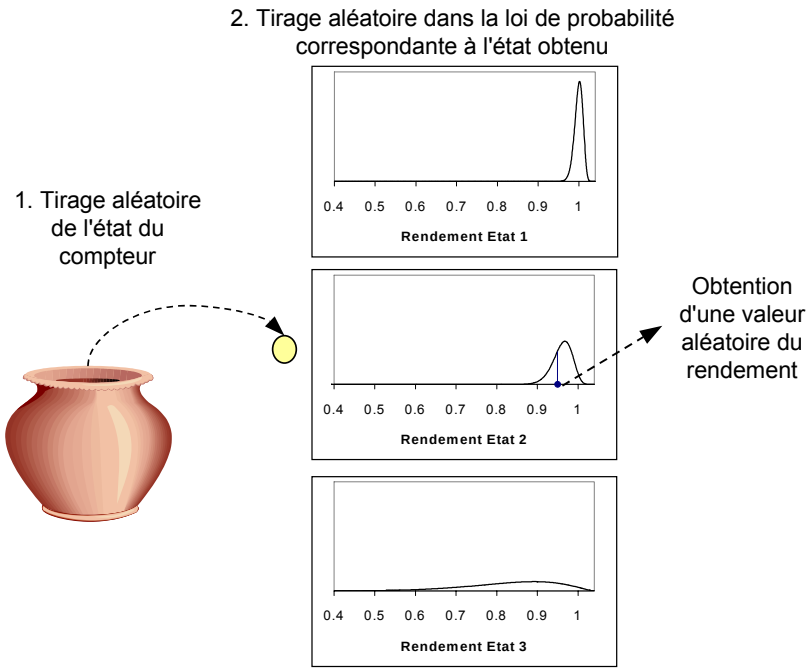


FIG. 4.11 – Simulation du rendement d'un compteur en fonction de son âge.

4.7 Pourquoi la modélisation "directe" du rendement en fonction de l'âge est problématique

L'originalité du modèle que nous proposons dans cette étude est que la qualité métrologique d'un compteur est décrite avec un paramètre "*objectif*" (l'état métrologique) établi sur la base uniquement de la signature métrologique. *A contrario*, le rendement est un paramètre non-objectif, puisque la pondération des erreurs de mesure en fonction des débits est faite sur la base d'histogrammes de consommation qui sont, par définition très variables. Parmi les deux éléments qui interviennent dans le calcul (compteur et usager) le "*facteur humain*" reste le plus incertain et le plus difficile à renseigner, les techniques d'obtention des profils de consommations étant coûteuses et assez difficiles à mettre en place. La diffusion du radio-relevé des compteurs, avec des lectures continues des index, permettra, dans un futur proche, de déterminer des histogrammes de consommation "sur mesure" pour chaque client équipé d'un dispositif émetteur et d'améliorer considérablement la connaissance, mais à l'heure actuelle on

a peu d'éléments pour confirmer la validité des hypothèses quant au choix des histogrammes types.

En pratique, les incertitudes les plus importantes dans le calcul du rendement concernent le pourcentage de la consommation à faibles débits qui est établi avec une pondération empirique entre les différents profils d'utilisateurs (fuyards ou non-fuyards) en fonction de données, acquises il y a quelques années, qui pourraient ne pas être représentatives de la situation actuelle. Aujourd'hui on a raison de croire que les usagers font davantage attention à leurs installations hydrauliques et que fuites et gaspillages ont diminués considérablement.

En outre, le rendement est sensible aux valeurs des erreurs de mesure dans la zone *haute* de la courbe métrologique, qui normalement corresponde aux débits de puisage les plus fréquents. Par exemple si on utilise l'histogramme type des compteurs de DN 15 mm on pondère par un coefficient 33% l'erreur de mesure à 300 l/h et 28% celle à 700 l/h. Ces erreurs dépendent du réglage à neuf des dispositifs et donc de facteurs aléatoires difficilement contrôlables. L'utilisation du rendement comme paramètre unique de la métrologie amplifie ces différences de comportement entre compteurs qui peuvent masquer le phénomène réel de dégradation.

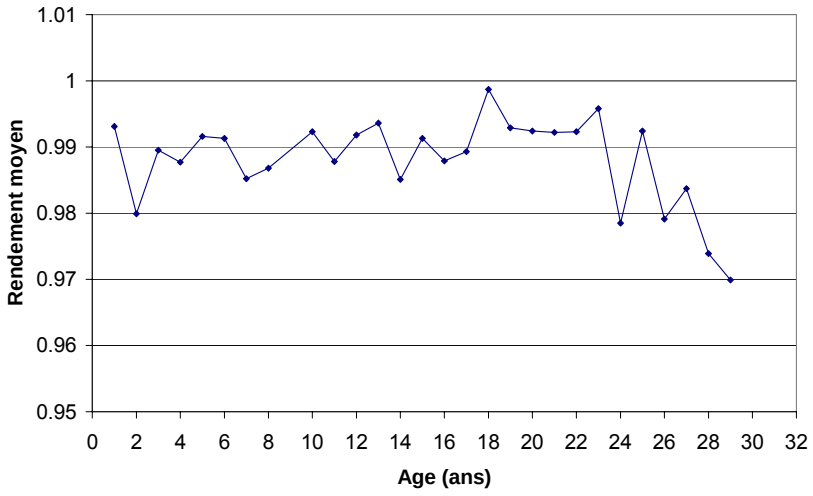


FIG. 4.12 – Evolution du rendement moyen en fonction de l'âge d'après Tao (1982).

Un exemple typique de la difficulté de cette approche "directe" est fourni par l'étude de Tao (1982) qui a examiné le rendement moyen en fonction de

l'âge de 3 200 compteurs volumétriques de petit diamètre (5/8 de pouce, soit environ 16 mm), au service de la collectivité locale de Hackensack (New Jersey). La prise en compte de l'âge comme seul facteur explicatif de la dégradation et l'analyse des rendements, calculés sur la base de signatures métrologiques en 3 points (144, 648 et 4 320 l/h) et en imaginant que 15% de la consommation avait lieu à des débits inférieurs à 215 l/h, 75% entre 215 et 2 150 l/h et 10% à des débits supérieurs à 2150 l/h conduisaient l'auteur à affirmer qu'aucune dégradation n'était observable, en fonction de l'âge, dans la métrologie des compteurs examinés (figure 4.12).

N'ayant pas à notre disposition les données sur la base desquelles cette conclusion a été établie, nous ne pouvons pas en dire plus ; mais l'exemple montre combien la modélisation du rendement en fonction de l'âge est une opération délicate, et l'utilisation de courbes métrologiques peu détaillées (uniquement 3 points de mesure, dont aucun dans la zone des faibles débits), associé au fait qu'on néglige des facteurs potentiellement importants (comme le type de compteur et le niveau de consommation), peut ne pas identifier le phénomène de dégradation.

4.8 Avantages de l'approche choisie

En définitive, les principaux points forts de la modélisation dynamique par mélange de populations peuvent être ainsi résumés :

- Du point de vue méthodologique, le modèle conceptuel proposé explique à la fois la présence de vieux compteurs extrêmement performants et de jeunes défaillants. L'aspect le plus évident de la dégradation est surtout l'évolution des proportions de bons et mauvais compteurs en fonction de l'âge (figure 4.9) plutôt qu'une baisse continue de la performance. D'autre part la structure même des données semble suggérer naturellement une approche de ce type : en conclusion ce modèle est, du point de vue conceptuel, plus correct.
- Du point de vue opérationnel, cette méthode permet d'utiliser aussi les résultats d'étalonnages partiellement incomplets. En pratique, un minimum de 5 points est suffisant pour établir la catégorie d'appartenance d'un compteur, alors que pour en calculer le rendement il faut beaucoup plus de précision, sachant qu'il est en général très dangereux de calculer des rendements

sur la base de signatures métrologiques obtenues par interpolation (ou, encore pire, par extrapolation) à partir des erreurs relatives à quelques débits d'essai. L'avantage de ne pas utiliser le rendement comme l'unique indicateur métrologique est évident : par exemple sur environ 18 500 étalonnages exploitables de compteurs de DN 15 mm à peine 7 100 sont des signatures métrologiques complètes sur la grille des débits d'essai et pour lesquelles le calcul du rendement est fiable.

– Du point de vue du distributeur d'eau, le modèle a l'avantage de donner une information très importante, c'est-à-dire la proportion en fonction de l'âge (et éventuellement d'autres facteurs explicatifs) des compteurs non conformes à la nouvelle réglementation. Ces indications seront de plus en plus précieuses puisque, dans un futur assez proche, les taux de non conformité des parcs ne devront pas dépasser des valeurs fixées (d'abord 15%, puis 12.5% pour arriver, en condition de régime à 10%) et le respect de ces contraintes sera un critère fondamental dans la gestion des renouvellements des compteurs.

En revanche, ce type de modèle est, du point de vue mathématique, nettement plus complexe que ceux normalement utilisés dans la pratique technique des ingénieurs et demande la mise en place de méthodes d'estimation plus sophistiquées qui feront l'objet du chapitre suivant.

Chapitre 5

Inférence bayésienne pour un problème complexe

5.1 La formule de Bayes, le Révérend Bayes et les "Bayésiens"

Soient A et B deux événements aléatoires. La probabilité de B, conditionnellement à la réalisation de A, est par définition exprimée (si la probabilité de l'événement A est non-nulle) par la relation suivante :

$$[B|A] = \frac{[B, A]}{[A]} \tag{5.1}$$

où $[B, A]$ est la probabilité que les deux événements A et B aient lieu simultanément.

Dans l'équation (5.1) on a utilisé la notation *entre crochets* (Gelfand et Smith, 1990) des probabilités, introduite déjà dans le chapitre 3, c'est-à-dire :

$$[A] \ : \text{ Probabilité de l'événement aléatoire } A$$

De manière analogue si X est une variable aléatoire à valeurs dans un sous-espace de $\mathbb{R}^n$ alors la notation simplifiée $[x]$ exprimera la probabilité $[X = x]$ que la variable X vaille x[1]. Dans la suite on se servira toujours de cette notation,

[1] On utilisera aussi la convention de représenter par des majuscules les variables aléatoires et par les minuscules correspondantes les valeurs de ces variables.

qu'on trouve très commode par rapport à d'autres notations d'usage courant comme $\Pr(A)$, ou $\mathbb{P}(A)$.

Revenons à l'équation (5.1). Puisque $[B, A] = [A, B]$ si dans l'expression de $[A|B]$:

$$[A|B] = \frac{[A, B]}{[B]}$$

on remplace $[A, B]$ par $[B, A] = [B|A] \cdot [A]$, on en déduit la relation entre les deux probabilités conditionnelles $[A|B]$ et $[B|A]$:

$$[A|B] = \frac{[A] \cdot [B|A]}{[B]} \tag{5.2}$$

L'équation (5.2), conséquence triviale de la définition de la probabilité conditionnelle, est appelée *Formule de Bayes* (ou aussi *Théorème de Bayes*) en l'honneur du Révérend Thomas Bayes (1702-1761) à qui l'on doit un des ouvrages qui sont à la base de l'inférence statistique : "*Essay Towards Solving a Problem in the Doctrine of Chances*" publié en 1763 dans les *Philosophical Transactions* de la *Royal Society* de Londres.

Thomas Bayes, pieux serviteur de Dieu[2] et grand mathématicien, n'est pas passé à l'histoire pour la découverte de la formule qui porte son nom, mais pour son interprétation et son application à un problème d'estimation.

Le problème du Rév. Bayes (Bernier et al., 2000) est le suivant : une balle est lancée sur une table parfaitement horizontale et s'arrête à un certain point $\mathtt{P_0}$. Ensuite, une deuxième balle est lancée n fois et on s'intéresse au nombre de fois y, qu'elle s'est arrêtée à la droite de la première (appelons ces événements "*succès*"). Comment estimer la probabilité de succès θ ?

Bayes imagine que θ est une variable aléatoire définie sur l'intervalle [0,1] et décrit avec une distribution de probabilité donnée $[\theta]$ sa connaissance concernant cette variable, préalablement à l'observation des données.

En utilisant la formule de Bayes, il est possible d'écrire l'expression de la distribution de probabilité de θ, conditionnellement à l'observation des données y :

$$[\theta|y] = \frac{[\theta] \cdot [y|\theta]}{[y]} \tag{5.3}$$

[2] Il est auteur en 1731 de l'essai : "*Divine Benevolence, or an Attempt to Prove That the Principal End of the Divine Providence and Government is the Happiness of His Creatures*".

où $[y]$ peut être calculé par marginalisation de la loi jointe $[y, \theta]$:

$$[y] = \int_{\Omega} [y, \theta]\, d\theta = \int_{\Omega} [y|\theta] \cdot [\theta]\, d\theta \tag{5.4}$$

Dans la dernière équation on a noté Ω l'ensemble des valeurs possibles de la variable aléatoire θ (l'intervalle [0,1] dans le problème de Bayes).

Les différents termes de la formule (5.3) peuvent être interprétés de la manière suivante :

- $[\theta]$ est la distribution de probabilité *a priori* du paramètre inconnu θ. L'appellation *"a priori"* exprime le fait qu'elle a été établie préalablement à l'observation des données y. Elle peut être issue de l'opinion personnelle du statisticien, ou établie sur la base de l'analyse d'autres données similaires ou de l'avis d'expert avec des techniques dite d'*élicitation de croyance*. Des exemples de mise en œuvre de ces méthodes sont fournis par Kadane et Wolfson (1998), O'Hagan (1998), Garthwaite et O'Hagan (2000) et Parent et Prevost (2003). En revanche, son indépendance des données est obligatoire pour éviter d'utiliser deux fois la même information (Berry, 1996).
- $[y|\theta]$ est la probabilité des observations conditionnellement à la valeur θ du paramètre du modèle statistique qu'on utilise pour leur description. Il s'agit de la *vraisemblance* des données, sous le modèle paramétré par θ.
- $[\theta|y]$ est la distribution de probabilité *a posteriori* du paramètre du modèle, sur la base de la connaissance *a priori* et de l'information apportée par les données. L'appellation *"a posteriori"* vient du fait que, logiquement, elle suit l'observation des données.
- Le dénominateur, indépendant de θ, est uniquement une constante de *normalisation*, telle que la distribution de probabilité $[\theta|y]$ respecte la propriété fondamentale : $\int_{\Omega}[\theta|y]d\theta = 1$.

Le passage de la distribution *a priori* à la distribution *a posteriori* des paramètres du modèle statistique, exprimé par la formule de Bayes, peut être alors interprété comme une mise à jour de la connaissance, sur la base des observations (figure 5.1).

Cette lecture de la formule de Bayes est à la base de l'origine de la distinction entre les statisticiens dits *fréquentistes* et les *bayésiens*.

La différence fondamentale entre les deux approches est que, pour les fréquentistes, toute technique statistique (inférence, test, choix de modèle ...) doit être fondée uniquement sur les données et aucune information externe à l'échantillon

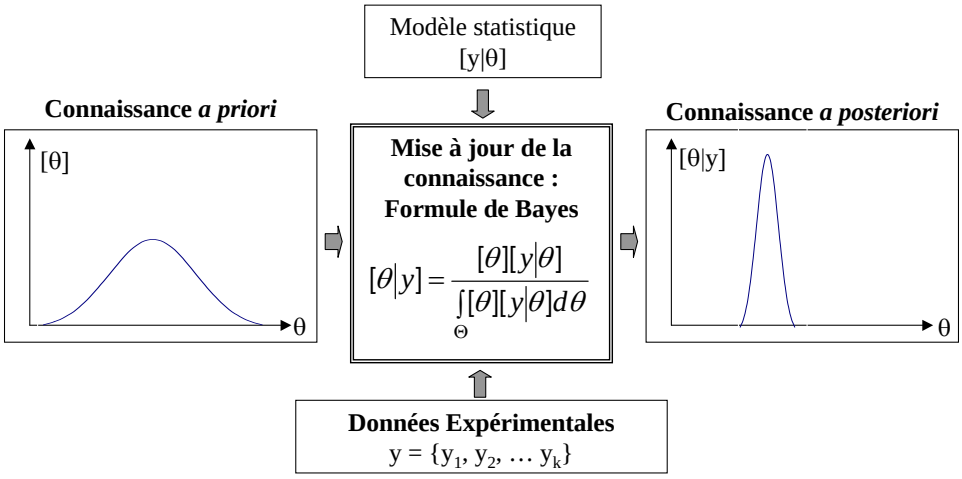


FIG. 5.1 – Mise à jour de la connaissance avec la formule de Bayes.

observé ne peut être introduite dans les calculs.

L'école fréquentiste, née au début du XXème siècle des travaux des fondateurs de la statistique moderne, notamment W.S. Gosset[3] "Student", K. Pearson, R.A. Fisher, N.J. Neyman, est aussi appelée *classique*, parce qu'elle a longtemps prévalu parmi les statisticiens même si historiquement l'approche bayésienne de T. Bayes, qui fut aussi, comme le fait remarquer Sivia (1996), celle de Jakob Bernoulli (1654-1705) et P. de Laplace (1749-1827) la précède d'environ un siècle et démi.

La raison de la mauvaise popularité des techniques bayésiennes a été longtemps la difficulté pratique des calculs d'inférence ; *a contrario*, le développement de techniques de simulation *Monte Carlo* des lois *a posteriori* (et la disponibilité d'ordinateurs capables de les mettre en œuvre) a marqué le début de la renaissance de l'intérêt pour la vision bayésienne dans la communauté statistique.

[3] W.S. Gosset (1856-1937) est universellement connu avec son pseudonyme "Student" qu'il a été contraint d'adopter parce que son employeur, le fabriquant de bière Guinness, ne permettait pas, à l'époque, à ses employés de publier.

5.2 La solution du problème de Bayes et ... le calcul des intégrales

Revenons au problème du Rév. Bayes. Si y est le nombre de succès observés et θ la probabilité de succès, le modèle naturel pour décrire le phénomène est le modèle *binomial* (annexe B). C'est à dire que, conditionnellement à θ, la probabilité d'observer y succès s'écrit :

$$[y|\theta] = \binom{n}{y} \theta^y (1-\theta)^{n-y} \tag{5.5}$$

Concernant la distribution *a priori* de θ, un bon choix est une loi de la famille *Bêta* (annexe B). Ces lois, à deux paramètres α et β, sont définies dans l'intervalle $[0, 1]$ et, selon les valeurs de α et β peuvent avoir des formes très différentes. Cette souplesse rend la famille *Bêta* très adaptée à la formalisation de la connaissance préliminaire sur une variable bornée. En vertu de ce choix, l'expression de la loi *a priori* de θ est :

$$[\theta] = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \tag{5.6}$$

D'après la formule de Bayes la loi *a posteriori* $[\theta|y]$, proportionnelle au produit entre loi *a priori* et vraisemblance :

$$[\theta|y] \propto \theta^{y+\alpha-1} (1-\theta)^{n-y+\beta-1} \tag{5.7}$$

est alors encore une loi *Bêta* de paramètres : $y + \alpha$ et $n - y + \beta$.

Dans ce cas l'expression du numérateur de la formule de Bayes nous a permis directement de reconnaître que la loi *a priori* et la loi *a posteriori* appartiennent à la même famille (*Bêta*). On exprime cette circonstance heureuse en disant que les lois *Bêta* et *binomiales* sont *conjuguées*.

Malheureusement, sauf quelques autres cas d'école, dans la pratique courante de la modélisation, la propriété de conjugaison n'est pas vérifiée et alors le calcul du dénominateur de la formule de Bayes s'impose.

Or, le calcul analytique de cette intégrale, souvent multidimensionnelle, est infaisable dans la plupart des cas. C'est pour cette raison que l'approche bayésienne a été longtemps mise à l'écart, à l'avantage de l'approche classique qui offre, elle, des solutions simples à de nombreux problèmes statistiques.

La découverte et la possibilité de mettre en œuvre des algorithmes de simulation capables d'obtenir des tirages aléatoires dans la loi *a posteriori* des paramètres a libéré les bayésiens des fardeaux du calcul intégral et a permis l'estimation de modèles à structures complexes. L'expression mathématique de la loi *a posteriori* restera inconnue à jamais mais, avec des échantillons aléatoires de cette loi de taille significative, on peut en calculer empiriquement la moyenne, la variance, les percentiles et toutes les autres grandeurs statistiques qui la décrivent, ce qui est d'ailleurs plus intéressant en pratique que d'avoir la formule mathématique de la loi jointe des paramètres du modèle (figure 5.2).

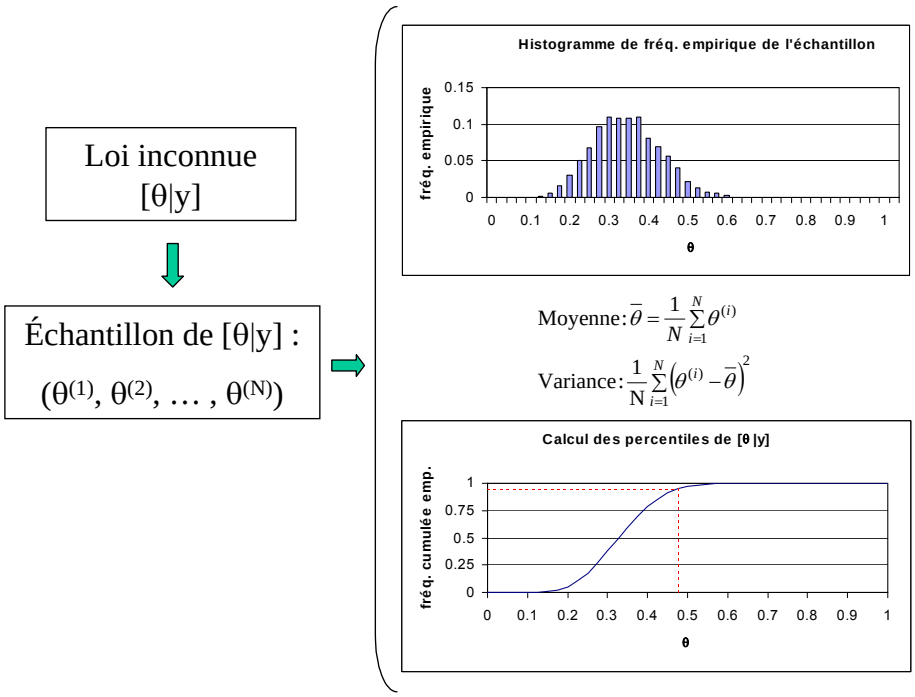


FIG. 5.2 – Approximation de la loi *a posteriori* avec des échantillons aléatoires.

Des procédures de simulation aujourd'hui très utilisées dans l'inférence bayésienne sont les méthodes dites MCMC (*Monte Carlo Markov Chain*), qui, à partir de l'expression d'une loi de probabilité $f(x)$, connue à une constante près, construisent une chaîne de Markov dont la distribution stationnaire est la loi visée $f(x)$. Le principe est alors de faire tourner de nombreuses fois l'algorithme de calcul et, une fois vérifié que la chaîne a atteint sa distribution asymptotique, utiliser les dernières valeurs obtenues comme échantillon de $f(x)$.

Les deux principales méthodes MCMC sont celle de *Metropolis-Hastings*

(Metropolis et al., 1953), (Hastings, 1970) et l'*échantillonneur de Gibbs* (Geman et Geman, 1984), (Gelfand et Smith, 1990), qui d'ailleurs peut être interprété comme un cas particulier de l'algorithme de Metropolis-Hastings.

Une brève présentation de ces deux algorithmes de calculs est donnée dans l'annexe C. Une description exhaustive des méthodes MCMC et de leur mise en œuvre peut être trouvée dans de nombreux textes et monographies de référence, parmi lesquels (Neal, 1993), (Gilks et al., 1996), (Robert, 1996), (Brooks, 1998), (Robert et Casella, 1999).

5.3 Avantages de l'approche bayésienne

Dans ce travail de thèse, j'ai décidé d'utiliser l'approche bayésienne pour les raisons suivantes :

- La vision *subjective* des probabilités, à la base de l'approche bayésienne, me semble plus cohérente que les théories fréquentistes qui interprètent la probabilité comme une proportion limite déterminée sur la base d'une séquence infinie d'expériences (Von Mises et Geiringer, 1964). D'ailleurs on associe souvent une probabilité à des événements qui par définition ne sont pas répétables, (cf. par exemple l'argumentaire de Hartigan (1983) autour de l'assertion, qui n'est heureusement plus d'actualité : "*Je pense que la probabilité qu'un conflit nucléaire éclate entre Etats Unis et Union Soviétique avant l'année 2000 est de 0.05*").

  L'interprétation bayésienne de la probabilité (De Finetti, 1937, 1974) est associée à une notion de *pari rationnel* : la probabilité attribuée à un événement est définie par les conditions auxquelles un individu rationnel est prêt à parier sur la réalisation de tel événement. La *rationalité* de l'individu, est nécessaire pour éviter que la définition de la probabilité soit arbitraire et est décrite par certaines règles de comportement, face à l'incertitude (Savage, 1972).
- La question de choisir une interprétation subjective ou fréquentiste de la probabilité n'est pas uniquement philosophique. Si l'on s'appuie à la définition subjective, alors il est possible de combiner les informations objectives, apportées par les données, avec des informations extérieures à l'échantillon observé et la formule de Bayes est l'instrument qui rend possible ce couplage dans un cadre cohérent basé uniquement sur le calcul des prob-

abilités (Bernardo et Smith, 1994). Le rôle de la formule de Bayes comme trait d'union entre la vision subjective du modélisateur et l'évidence des résultats expérimentaux est souligné, entre autres, par Box et Tiao (1973), Berry et al. (1996), Press et Tanur (2000).

Cette possibilité d'intégrer des éléments extérieurs aux échantillons est très adaptée à la pratique technique. Dans le monde réel, le statisticien est normalement confronté à des données peu représentatives ou incohérentes mais en revanche il dispose de l'avis technique des experts qui, sur la base de leur expérience et savoir-faire, sont capables de donner des informations complémentaires de grande utilité, qu'il serait dommage de ne pas prendre en compte (Cullen et Frey, 1999), (Bernier et al., 2000), (Perreault, 2000).

– L'analyse bayésienne fournit des résultats d'interprétation plus directe que ceux de la statistique classique. L'exemple le plus flagrant est la définition de l'*intervalle de confiance* du paramètre θ d'un modèle. Pour les bayésiens, qui préfèrent parler plutôt d'*intervalle de crédibilité*[4], il s'agit de l'intervalle qui contient le paramètre avec une probabilité donnée. Par exemple l'intervalle de crédibilité *a posteriori* à 95% est typiquement celui délimité inférieurement par le percentile d'ordre 2.5% et supérieurement par le percentile d'ordre 97.5%. Dans l'inférence classique cette assertion n'est plus vraie parce que le paramètre (inconnu) du modèle n'est pas une variable aléatoire mais une grandeur constante. L'interprétation correcte de l'intervalle de confiance est que, si on imagine l'ensemble des échantillons aléatoires pouvant être obtenus à partir du modèle, paramétré par θ, 95% des intervalles de confiance calculés (sur la base des différents échantillons) contiennent la vraie valeur du paramètre. L'interprétation bayésienne, décidément plus naturelle, est d'ailleurs celle de la plupart des praticiens qui font de l'inférence bayésienne ... sans le savoir (Lecoutre et Poitevineau, 1996).

– Les résultats de l'inférence bayésienne sont plus riches que les estimateurs fournis par les techniques classiques d'inférences (Berger, 1985). Les techniques bayésiennes permettent d'obtenir la loi jointe des paramètres du modèle et donc de prendre en compte simultanément l'effet de l'incertitude globale sur l'ensemble des paramètres inconnus sur les prévisions futures

[4] Autres terminologies d'usage moins courant : *Région à plus grande probabilité* ou *HDR*, (Lee, 1997), *intervalle bayésien de confiance* (Lindley, 1965).

du comportement du système étudié et sur les décisions suggérées par ce comportement (Krzysztofowicz, 1983).

– Enfin, les techniques d'inférence par méthode MCMC, relativement faciles à mettre en œuvre, sont très adaptées à l'estimation de modèles complexes à plusieurs paramètres ou à structure hiérarchique. L'estimation de ces modèles avec la technique usuelle du *Maximum de Vraisemblance* se révèle parfois nettement plus compliquée. L'argument de la commodité opérationnelle qui a longtemps joué contre l'approche bayésienne commence aujourd'hui à peser dans le sens opposé dans la vieille querelle entre fréquentistes et bayésiens (Robert, 1992).

5.4 Estimation bayésienne du modèle de dégradation des compteurs

La dégradation des compteurs est décrite avec un modèle markovien. Normalement l'estimation de ces modèles est réalisée à partir de données sous la forme d'observations répétées d'un certain nombre d'individus. Dans le cas présent, puisque l'essai d'un compteur peut être considéré comme une mesure destructive de l'appareil, les techniques d'estimation usuelles ne sont pas utilisables et le problème est mathématiquement plus complexe.

Dans la suite nous décrivons d'abord les techniques employées pour l'estimation de cette classe de modèles et ensuite nous montrons comment mettre en œuvre un algorithme de type Metropolis-Hastings pour mener les calculs d'inférence dans un cadre bayésien.

5.4.1 Un problème d'estimation peu usuel

Soit $X(t)$ une chaîne de Markov discrète et homogène avec un nombre fini (s) d'états : $1, 2...s$. Ce modèle est paramétré par une matrice carrée de transition $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_{11} & \theta_{12} & ... & \theta_{1s} \\ \theta_{21} & \theta_{22} & ... & \theta_{2s} \\ ... & & & \\ \theta_{s1} & \theta_{s2} & ... & \theta_{ss} \end{pmatrix} \tag{5.8}$$

dont l'élément θ_{ij} est la probabilité que le système soit dans l'état j au temps t, sachant qu'il se trouvait dans l'état i au temps $t-1$:

$$\theta_{ij} = [x(t) = j | x(t-1) = i] \tag{5.9}$$

Ces paramètres sont indépendants de t, en vertu de l'hypothèse d'homogénéité, et vérifient les relations :

$$0 \leq \theta_{ij} \leq 1 \qquad \sum_{j=1}^{s} \theta_{ij} = 1 \tag{5.10}$$

Si on dénote avec $[\underline{x}(t)]$ le vecteur ligne (à s composantes) des probabilités inconditionnelles d'appartenance à chacun des s états :

$$[\underline{x}(t)] = \{[X(t) = 1], ..., [X(t) = s]\} \tag{5.11}$$

l'évolution du système est décrite par l'équation dynamique :

$$[\underline{x}(t)] = [\underline{x}(t-1)] \cdot \boldsymbol{\theta} \tag{5.12}$$

qui peut s'écrire, en fonction des probabilités de l'état initial du système $[\underline{x}(0)]$:

$$[\underline{x}(t)] = [\underline{x}(0)] \cdot \boldsymbol{\theta}^t \tag{5.13}$$

L'estimation de ce modèle se fait classiquement sur la base des observations des états à travers lesquels passe un certain nombre z d'individus sur une période donnée t_{obs}. Les données se présentent donc sous la forme :

$$\{x_k(0), x_k(1), ..., x_k(t_{obs})\} \qquad k = 1, 2, ..., z \tag{5.14}$$

Dans ce cas des estimateurs simples par *Maximum de Vraisemblance* sont disponibles :

$$\widehat{\theta}_{ij} = \frac{m_{ij}}{\sum_{j=1}^{s} m_{ij}} \tag{5.15}$$

Dans cette dernière formule m_{ij} ést le nombre total de transitions de l'état i à l'état j (entre deux temps successifs) observées sur l'ensemble des individus pour tout t compris entre 0 et t_{obs} :

$$m_{ij} = \sum_{k=1}^{z} \sum_{t=1}^{t_{obs}} \mathbf{1}_{(x_k(t)=j, x_k(t-1)=i)} \tag{5.16}$$

où $\mathbf{1}_{(x_k(t)=j,x_k(t-1)=i)}$ est une fonction (dite indicatrice) qui vaut 1 si la condition entre parenthèses est vérifiée et 0 autrement.

Ces estimateurs (Anderson et Goodman, 1957), (Bickenbach et Bode, 2001) sont sans biais, leur distribution asymptotique est normale et leur écart type peut être estimé avec la formule :

$$\widehat{\sigma}_{\widehat{\theta}_{ij}} = \sqrt{\frac{\widehat{\theta}_{ij} \cdot (1 - \widehat{\theta}_{ij})}{\sum_{j=1}^{s} m_{ij}}} \tag{5.17}$$

Dans de nombreux problèmes réels, les données à disposition ne sont pas des séries temporelles des états, mais l'individu k est observé une seule fois à un instant u_k.

Pour fixer les idées on peut imaginer que les données sont issues de mesures *destructives* réalisées sur un certain nombre (z') d'individus :

Individu #	Instant de mesure	Etat observé
1	2	1
2	10	4
3	5	2
...		
z'	7	3

Ces données sont résumées de manière exhaustive par les statistiques :

$$\{y_1(t), y_2(t), ...y_s(t)\} \qquad t = 1, 2, ..., t_{obs} \tag{5.18}$$

$y_i(t)$ étant le nombre d'individus observés à l'instant t et qui s'y trouvent dans l'état i :

$$y_i(t) = \sum_{k=1}^{z'} \mathbf{1}_{(u_k=t, x_k=i)}$$

et $t_{obs} = \max(u_1, .., u_{z'})$.

Dans la suite on dénotera $\underline{y}(t)$ le vecteur ligne des observations au temps t :

$$\underline{y}(t) = \{y_1(t), y_2(t), ...y_s(t)\} \tag{5.19}$$

Les premières méthodes utilisées, dès les années 50, pour résoudre le problème de l'estimation de la matrice de transition $\boldsymbol{\theta}$ dans ce cas "peu usuel" étaient basées sur le remplacement des probabilités inconditionnelles $[\underline{x}(t)]$ par les proportions des individus observés de même âge, qui se présentent comme leurs estimateurs *naturels* :

$$[\widehat{\underline{x}}(t)] = \left\{ \frac{y_1(t)}{n(t)}, \frac{y_2(t)}{n(t)}, .., \frac{y_s(t)}{n(t)} \right\} \tag{5.20}$$

où $n(t) = \sum_{j=1}^{s} y_j(t)$ est la taille de l'échantillon d'âge t. Il n'est pas superflu d'observer que le nombre d'observations $n(t)$ n'est pas forcement le même pour toutes les valeurs de t. On peut imaginer, par exemple, d'étalonner 10 individus de 5 ans, aucun de 10 ans et 20 individus de 15 ans.

Le remplacement de $[\underline{x}(t)]$ par $[\widehat{\underline{x}}(t)]$ dans les équations (5.12), donne lieu à un système de $t_{obs} \cdot s$ équations en les s^2 inconnues θ_{ij}, qui, si $t_{obs} > s$, peut être résolu avec des techniques de *moindres carrés* (Miller, 1952), (Goodman, 1953), (Madansky, 1959). La difficulté principale dans la résolution de ce système est le respect des contraintes spécifiées par les équations (5.10). Cette méthode est sensible à la présence de données manquantes, parce que le manque d'observations pour un âge donné t^* entraîne la suppression de $2s$ équations (celles aux temps t^* et $t^* - 1$). Cette approche a été récemment revisitée dans un cadre bayésien par Congdon (2001).

Le problème a été reformulé successivement, dans les années 60, par Lee et al. (1968) qui ont proposé une technique d'estimation par *Maximum de Vraisemblance*. L'écriture de la vraisemblance se fait sur la base de l'observation que, conditionnellement au vecteur des probabilités d'appartenance aux s états $[\underline{x}(t)]$, le vecteur des observations $\underline{y}(t)$ est la réalisation d'un tirage multinomial (annexe B) de paramètres $[\underline{x}(t)]$ et $n(t)$. La vraisemblance est alors le produit de $t_{obs} + 1$ termes indépendants :

$$[\underline{y}|\boldsymbol{\theta}] = \prod_{t=0}^{t_{obs}} \frac{n(t)!}{\prod_{j=1}^{s} y_j(t)!} \prod_{j=1}^{s} [x_j(t)]^{y_j(t)} \tag{5.21}$$

dans cette expression $[x_j(t)]$ est la composante j du vecteur $[\underline{x}(t)]$: $[x_j(t)] = [X(t) = j]$.

Lee et al. (1968) obtiennent un estimateur de $\boldsymbol{\theta}$ en maximisant l'expression (5.21) sous les contraintes (5.10) avec des techniques de programmation quadratique.

Pour estimer ce modèle dans un cadre bayésien, il faut choisir préalablement une loi *a priori* pour $\boldsymbol{\theta}$. Une distribution de probabilité adaptée pour $\boldsymbol{\theta}$ est le produit de s distributions de *Dirichlet* indépendantes (cf. annexe B), une pour chaque ligne $\underline{\theta}_i$ de la matrice, chacune paramétrée par s constantes positives $\{\alpha_{i1,...}, \alpha_{is}\}$:

$$\underline{\theta}_i \sim \mathcal{D}(\underline{\alpha}_i) \tag{5.22}$$

avec $\underline{\alpha}_i = \{\alpha_{i1,...}, \alpha_{is}\}$.

Les distributions de *Dirichlet*, normalement utilisées en statistique bayésienne (Good, 1965 ; Gelman et al., 1995) comme lois *a priori* conjuguées des lois *multinomiales* (cf. demonstration page 178) sont des généralisations multidimensionnelles de lois *Bêta* et ont la propriété que leurs réalisations vérifient automatiquement les conditions (5.10).

La distribution *a priori* de $\boldsymbol{\theta}$ s'écrit alors[5] :

$$[\boldsymbol{\theta}] = \prod_{i=1}^{s} \frac{\Gamma(\alpha_{i1} \cdot ... \cdot \alpha_{is})}{\Gamma(\alpha_{i1}) \cdot \Gamma(\alpha_{is})} \theta_{i1}^{\alpha_{i1}-1} \cdot ... \cdot \theta_{is}^{\alpha_{is}-1} \tag{5.23}$$

Lee et al. (1968) se servent de la méthode utilisée pour la maximisation de la vraisemblance pour obtenir un estimateur bayésien ponctuel de $\boldsymbol{\theta}$ en maximisant le produit de la vraisemblance (5.21) et de la loi *a priori* (5.23). Cet estimateur, puisque dans la formule de Bayes le dénominateur ne joue aucun rôle, maximise aussi la loi *a posteriori* de $\boldsymbol{\theta}$ et représente le *mode a posteriori*.

L'expression exacte de la loi *a posteriori* est pratiquement incalculable : elle demanderait l'intégration d'une fonction de s^2 variables. En revanche, la mise en place d'un algorithme *MCMC* de type Metropolis-Hastings permet, avec un choix

[5] Le lecteur statisticien remarquera que le terme à droite du signe d'égalité dans la formule (5.23) n'est pas un produit de densités de probabilité "canoniques" (par rapport à la mesure de Lebesgue) de variables de $\mathbb{R}^s$. La densité de probabilité (formule B.22, page 177) de la loi de Dirichlet est (en réalité) définie dans l'espace $\mathbb{R}^{s-1}$. Pour plus de détails sur les lois de Dirichlet, cf. annexe B.

astucieux des "*fonctions de saut*" (annexe C), d'obtenir des tirages aléatoires de ladite loi, comme il sera exposé dans la suite.

5.4.2 Mise en œuvre de l'algorithme MCMC

L'algorithme de Metropolis-Hastings est une procédure itérative pour obtenir des tirages d'une distribution de probabilité, connue à une constante près. A chaque itération k, sur la base des valeurs obtenues lors de l'itération $k-1$, appelons-les $\boldsymbol{\theta}^{(k-1)}$, on tire au hasard, dans une certaine loi éventuellement dépendante de $\boldsymbol{\theta}^{(k-1)}$, dite *loi instrumentale*, un *candidat* $\boldsymbol{\theta}^*$. Les lois instrumentales sont aussi appelées *fonctions de saut* (*jumping functions*) parce que le passage à une nouvelle valeur de tentative, à partir de la dernière valeur retenue, peut être interprété comme un saut, dans l'espace des valeurs possibles, dont la longueur dépend des caractéristiques de la fonction choisie (notamment de sa variance). Le candidat est ensuite retenu avec une certaine probabilité (annexe C), fonction de $\boldsymbol{\theta}^*$ et de $\boldsymbol{\theta}^{(k-1)}$.

Dans le cas présent, un choix commode est de tirer indépendamment chaque ligne $\underline{\theta}_i^*$ de la matrice candidate $\boldsymbol{\theta}^*$ selon une loi de *Dirichlet* dont l'espérance est le vecteur ligne $\underline{\theta}_i^{(k-1)}$ de la matrice $\boldsymbol{\theta}^{(k-1)}$, de manière à ce que tout candidat vérifie automatiquement les contraintes (5.10).

Ces lois de *Dirichlet* appartiennent toutes à la famille de lois paramétrées par le vecteur $h_i \cdot \underline{\theta}_i^{(k-1)}$, $\forall h_i \in]0, +\infty[$:

$$\underline{\theta}_i^* \sim \mathcal{D}\left(h_i \cdot \underline{\theta}_i^{(k-1)}\right) \tag{5.24}$$

La démonstration de cette assertion est simple ; il faut juste expliciter le calcul de l'espérance des éléments θ_{ij}^* de la matrice $\boldsymbol{\theta}^*$:

$$\mathbb{E}\left(\theta_{ij}^*\right) = \frac{h_i \cdot \theta_{ij}^{(k-1)}}{h_i \cdot \left(\theta_{i1}^{(k-1)} + ... + \theta_{is}^{(k-1)}\right)} = \frac{h_i \cdot \theta_{ij}^{(k-1)}}{h_i} = \theta_{ij}^{(k-1)} \tag{5.25}$$

puisque $\theta_{i1}^{(k-1)} + ... + \theta_{is}^{(k-1)} = 1$.

La constante h_i peut être interprétée comme un paramètre de forme de la distribution de probabilité. En effet h_i affecte les variances des variables θ_{ij}^* :

$$\mathbb{V}\left(\theta_{ij}^*\right) = \frac{\theta_{ij}^{(k-1)} \cdot \left(1 - \theta_{ij}^{(k-1)}\right)}{h_i + 1} \tag{5.26}$$

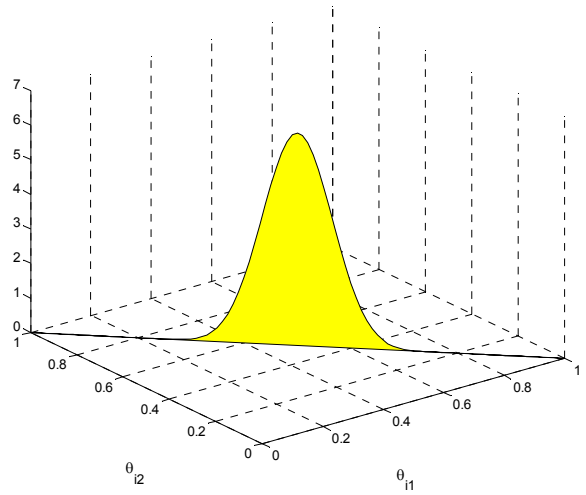


FIG. 5.3 – Exemple de fonction de saut dans un espace de dimension 2.

et plus h_i est grand, plus la loi de *Dirichlet* est "pointue" (et dégénère asymptotiquement vers une *Dirac* quand $h_i \to +\infty$).

Une visualisation graphique des lois instrumentales est possible quand s=2 et s=3 (pour des espaces de dimension supérieure à 3 il faut faire preuve d'imagination).

Dans un espace bidimensionnel tous les couples possibles des probabilités de transition θ_{i1} et θ_{i2} se trouvent sur la droite ℓ d'équation $\theta_{i1}+\theta_{i2} = 1$. La fonction de saut alors prend des valeurs positives sur ℓ et vaut 0 partout ailleurs dans $[0,1]^2$ (dans ce cas la *Dirichlet* dégénère en une loi de type *Bêta*).

Dans un espace de dimension 3, tous les triplets possibles de probabilités de transition, définis dans $[0,1]^3$, peuvent être décrits par un point dans un des trois plans engendrés par les couples $(\theta_{i1}, \theta_{i2})$, $(\theta_{i1}, \theta_{i3})$ ou $(\theta_{i2}, \theta_{i3})$. Puisque chaque composante du vecteur est connue, sachant les deux autres, cette représentation est unique et sans ambiguïté. Pour fixer les idées, supposons de représenter le triplet $(\theta_{i1}, \theta_{i2}, \theta_{i3})$ par ses deux premières composantes. Le point de coordonnées θ_{i1} et θ_{i2} se trouve à l'intérieur du triangle $\mathcal{T}$ défini par les semi-axes positifs de θ_{i1} et θ_{i2} et la droite ℓ. Tout point externe à ce triangle ne vérifie pas la condition $\theta_{i1} + \theta_{i2} + \theta_{i3} = 1$ et, vice-versa, pour tout point interne à $\mathcal{T}$, il existe un et un seul $\theta_{i3} = 1 - \theta_{i1} - \theta_{i2}$ qui est compris entre 0 et 1. Telle représentation de triplets de probabilités, très utilisée en théorie de la décision, a été introduite

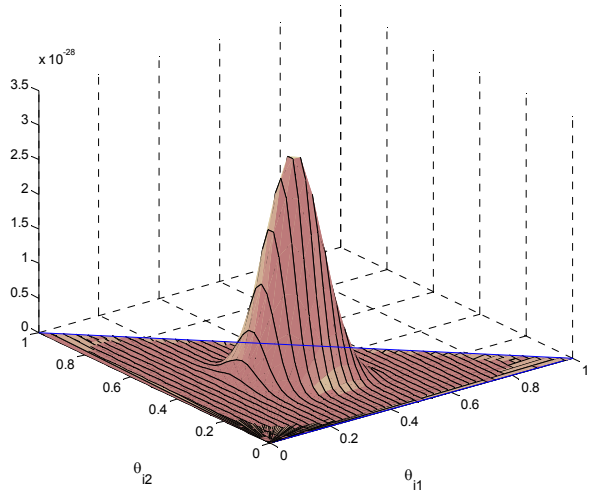


FIG. 5.4 – Exemple de fonction de saut dans un espace de dimension 3.

pour la première fois par Marshak (1950) et a acquis une grande popularité dans les années 80 après les travaux de Machina (1983).

Le rôle des constantes h_i dans l'estimation des θ_{ij} est important. Même si la convergence de la chaîne construite par l'algorithme de Metropolis-Hastings vers la distribution visée est assurée pour toute fonction de saut, en pratique des lois instrumentales trop plates génèrent des candidats avec des probabilités d'acceptation trop faibles et donc le nombre d'itérations requis pour atteindre la loi visée augmente considérablement. *A contrario*, si les fonctions de saut sont trop pointues, d'une part plusieurs itérations sont nécessaires pour *explorer* toute la distribution de probabilité visée (on dit aussi que les valeurs obtenues ne se *mélangent* pas) et d'autre part la chaîne construite risque de rester bloquée sur un maximum local de la loi *a posteriori*.

En pratique, on a procédé par réglages successifs de la variance des lois instrumentales en se fixant comme objectif d'atteindre la convergence de la chaîne après un nombre maximal de 5 000 itérations.

5.4.3 Une classe de modèles plus générale

Le modèle de dégradation des compteurs peut être interprété comme un cas particulier d'une importante classe de modèles : les *modèles dynamiques à états latents* (MDEL). Tous ces modèles décrivent un système qui évolue, au fil du

temps, selon un mécanisme aléatoire de transition, conditionnellement à ses états précédents. En particulier, dans un modèle markovien d'ordre 1 (ou markovien tout court) on fait l'hypothèse que seul l'état immédiatement précédent intervient dans le processus. L'autre point commun à tous ces modèles est le fait que la suite des états $\underline{X}=\{X(1), X(2), ..X(t_{obs})\}$ n'est pas observée (alors on dit qu'il s'agit de variables *latentes*) mais on dispose d'un certain nombre d'observations d'autres grandeurs $\underline{Y}=\{Y(1), Y(2), ..Y(t_{obs})\}$ liées aux états du système qu'on utilise pour reconstruire le processus de transition. Le terme θ dénotera encore l'ensemble des paramètres du modèle. Parmi eux un groupe, θ_x, affecte uniquement le mécanisme de transition entre les états, et un autre, θ_y uniquement le mécanisme d'observation des données.

Un MDEL est décrit par deux groupes d'équations

- *Equations de transition*, qui décrivent l'évolution de l'état du système entre deux pas de temps successifs (qu'on suppose discrets pour simplifier) :

$$X(t+1) = f\left(x(t), \theta_x, \epsilon_x(t)\right) \tag{5.27}$$

  Dans cette équation le facteur aléatoire qui perturbe (éventuellement) le mécanisme déterministe d'évolution est représenté par le terme $\epsilon_x(t)$.
- *Equations d'observation*, qui décrivent le mécanisme d'observation des données, en fonction des états inobservés et d'un bruit aléatoire d'observation $\epsilon_y(t)$:

$$Y(t) = g\left(x(t), \theta_y, \epsilon_y(t)\right) \tag{5.28}$$

Par exemple dans le modèle markovien du paragraphe précédent les équations de transition sont les équations dynamiques (5.12), alors que les équations d'observation sont les équations (5.21) qui expriment la dépendance (aléatoire) entre les probabilités des états du système et les données observables (dénombrement par état des individus de même âge). Dans ce cas on a :

$$\theta = \theta_x = \{\theta_{ij}\} \qquad i, j = 1...s \tag{5.29}$$

puisque les équations d'observation (5.21) n'introduisent aucun paramètre supplémentaire (les seuls paramètres du modèle dynamique sont les probabilités de transition).

Le modèle de dégradation des compteurs est juste un peu plus compliqué, puisqu'il faut rajouter l'équation d'observation des compteurs bloqués (cf. paragraphe suivant).

La représentation graphique des MDEL (figure 5.5) met bien en évidence la différence entre ces deux groupes d'équations. Pour cette raison nous présenterons dans le prochain paragraphe des généralités concernant la théorie des graphes appliquée à la modélisation statistique. Ensuite, nous utiliseront ces considérations pour décrire les algorithmes d'inférence bayésienne normalement employés pour l'estimation des MDEL.

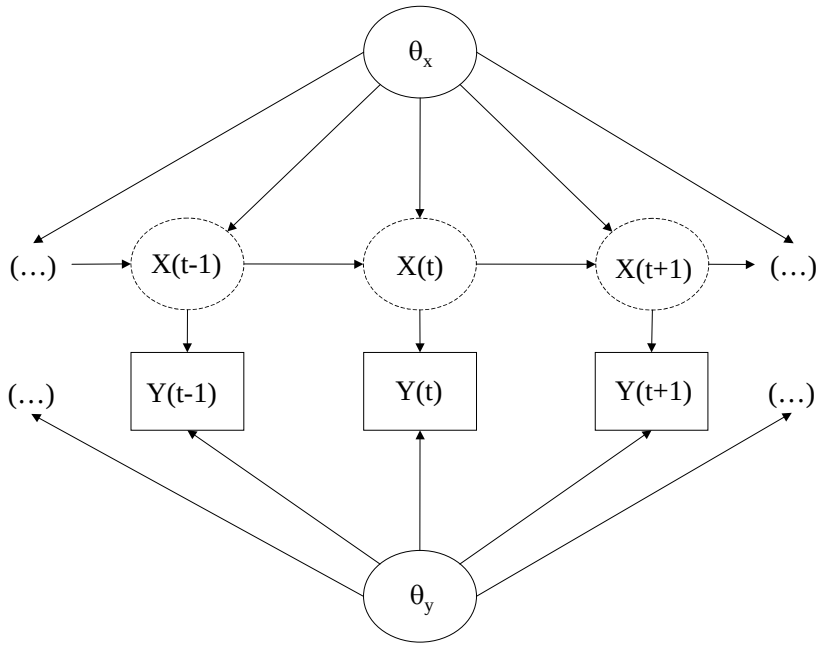


FIG. 5.5 – Représentation générale d'un MDEL sous forme de DAG.

Quelques éléments de modélisation graphique

Un modèle statistique peut être représenté par un graphe où toutes les variables sont représentées par des nœuds et les liens entre elles (déterministes ou probabilistes) par des connexions orientées ou non.

Les liens non-orientés traduisent l'idée de la corrélation (Spiegelhalter et al., 1996-a) entre les voisins proches d'un graphe (c'est-à-dire les variables liées entre elles).

Les liens orientés expriment la causalité entre deux variables. Ils sont représentés par une flèche orientée dans la direction de la variable dépendante. Cette dernière est appelée "*fils*" et la variable dont elle dépend en est un "*parent*". Pour une interprétations intuitive de la différence entre les deux types de liens, cf. le célèbre argumentaire de Cowell et al (1999) sur les deux représentations possibles du lien entre les variables *tabagisme* et *cancer des poumons*[6].

Dans un DAG les propriétés d'indépendance conditionnelle entre les variables (on reviendra sur ce concept) sont exprimées par le fait que, sachant ses parents, toute variable est conditionnellement indépendante de toutes celles qui ne sont pas ses fils ("*directed local Markov property*").

L'indépendance conditionnelle de deux variables A et B, sachant C, qu'on peut exprimer par la notation :

$$A \perp B|C \tag{5.30}$$

signifie que pour tout couple (b, c) de valeurs de B et de C respectivement on a :

$$[a|b, c] = [a|c] \tag{5.31}$$

En vertu de la propriété ci-dessus, la loi jointe de toutes les variables d'un DAG $(X_1, \ldots, X_n)$ peut être factorisée de la manière suivante :

$$[x_1, \ldots, x_n] = \prod_{i=1}^{n} [x_i|pa(x_i)] \tag{5.32}$$

où $pa(x_i)$ est l'ensemble des parents de x_i.

Dans un cadre bayésien, l'intérêt de la mise en forme graphique du modèle et de la factorisation de la loi jointe *a posteriori* des paramètres est particulièrement évident quand on utilise pour les calculs d'inférence la méthode MCMC dite de "*l'échantillonneur de Gibbs*" (Gelfand et Smith, 1990).

Cette méthode (cf. annexe C) permet de réaliser un tirage selon la loi jointe de plusieurs variables $[x_1, ..., x_n]$ avec des tirages séquentiels selon les lois conditionnelles de chaque variable, sachant toutes les autres :

[6] Un lien orienté entre les variables aléatoires *tabagisme* et *cancer des poumons* traduit l'hypothèse du modélisateur que fumer a une incidence sur la probabilité de contracter la maladie. Un lien non-orienté indique simplement qu'il y a une relation, non explicitée dans le modèle, entre les deux variables, par exemple qu'il existe un gène inconnu qui donne envie de fumer et, en même temps, prédispose aux maladies pulmonaires.

$$[x_i|(x_1, ..., x_n)\backslash x_i] \tag{5.33}$$

En d'autres termes, le problème de la simulation d'une variable aléatoire de dimension n est réduit à la simulation de n variables unidimensionnelles. La connaissance préliminaire des relations d'indépendance conditionnelle entre les variables simplifie la mise en œuvre de l'algorithme de Gibbs.

En effet, de la formule 5.32 on peut déduire (Cowell et al., 1999) que :

$$x_i \perp (x_1, ..., x_n)\backslash x_i|bl(x_i) \tag{5.34}$$

où l'ensemble $bl(x_i)$, dit "couverture markovienne" (Markov blanket) de x_i, est formé par les parents de x_i, par les fils $ch(x_i)$ de x_i, et par les éventuels co-parents des fils de x_i :

$$bl(x_i) = pa(x_i) \cup ch(x_i) \bigcup_{j \neq i} \{ch(x_j) \cap ch(x_i) \neq \emptyset\} \tag{5.35}$$

De la formule (5.34) on déduit immédiatement que :

$$[x_i|(x_1, ..., x_n)\backslash x_i] = [x_i|bl(x_i)] \tag{5.36}$$

et cette observation rend plus facile l'écriture des lois conditionnelles utilisées pour mettre en œuvre l'algorithme de Gibbs.

La détermination des couvertures markoviennes des différentes variables d'un DAG peut être faite à partir d'une procédure graphique très simple qui prévoit, dans un premier temps, le remplacement de tous les liens par des liens non-orientés et, ensuite, le rajout de liens fictifs entre les parents de fils communs ("*moralisation*" du graphe). La couverture markovienne de chaque variable est alors l'ensemble de ses voisins dans ce nouveau graphe, ainsi obtenu.

Retour sur les MDEL et sur leur estimation

Avec la convention de représenter les observations avec des nœuds carrés, et les variables latentes avec des nœuds pointillés, la figure 5.5 décrit tous les possibles MDEL dans l'hypothèse que le mécanisme de transition est markovien d'ordre 1. Dans cette figure on observe bien le mécanisme markovien d'évolution et le rôle joué par les variables latentes dans les équations d'observation.

L'approche bayésienne simplifie, du moins du point de vue conceptuel, le problème de l'estimation de ces modèles. Pour les bayésiens (West et Harrison, 1997) toutes les grandeurs non-observables (paramètres et variables latentes) jouent le même rôle dans le processus de mise à jour de l'information, exprimé par la formule de Bayes. Alors la loi jointe *a posteriori* des paramètres et des variables latentes s'exprime :

$$\left[\theta, \underline{x}|\underline{y}\right] \propto [\theta] \cdot [\underline{x}|\theta] \cdot \left[\underline{y}|\underline{x}, \theta\right] \tag{5.37}$$

et la loi *a posteriori* des paramètres s'obtient par marginalisation de la loi jointe (5.37) :

$$\left[\theta|\underline{y}\right] = \int \left[\theta, \underline{x}|\underline{y}\right] d\underline{x} \tag{5.38}$$

Pour obtenir des tirages aléatoires dans les lois (5.37) et (5.38), en pratique incalculables, on fait appel, évidemment, à des méthodes MCMC.

Normalement ces modèles sont estimés (dans un cadre bayésien) à l'aide de l'algorithme de Gibbs (annexe C). La mise en œuvre de cette méthode demande la connaissance préliminaire des lois conditionnelles de chaque variable, conditionnellement à toutes les autres. Pour les MDEL, en tenant compte des conditions (5.36) ces lois s'écrivent (Parent et Prevost, 2003) :

$$\left[\theta_x|\underline{y}, \underline{x}, \theta_y\right] \propto [\theta_x] \cdot [x\,(0)] \cdot \prod_{t=1}^{t_{obs}} [x\,(t)\,|x\,(t-1)\,, \theta_x] \tag{5.39}$$

$$\begin{aligned} \left[x(t)|\underline{y}, \underline{x}_{-(t)}, \theta_x, \theta_y\right] \quad \propto \quad & [x\,(t)\,|x\,(t-1)\,, \theta_x] \cdot \\ & \cdot [x\,(t+1)\,|x\,(t)\,, \theta_x] \cdot [y\,(t)\,|x\,(t)\,, \theta_y] \end{aligned} \tag{5.40}$$

$$\left[\theta_y|\underline{y}, \underline{x}, \theta_x\right] \propto [\theta_y] \cdot \prod_{t=1}^{t_{obs}} [y\,(t)\,|x\,(t)\,, \theta_y] \tag{5.41}$$

où :

$$\underline{x}_{-(t)} = \{x(0), ..., x(t-1), x(t+1), ...x(t_{obs})\} \tag{5.42}$$

Quand le développement des produits exprimés par ces formules, permet de reconnaître des lois de probabilités connues, la mise en place de l'échantillonneur

de Gibbs est très simple et cette méthode est nettement plus efficace que l'algorithme de Metropolis-Hastings. Dans le cas des modèles markoviens ici décrits, cette circonstance ne se vérifie pas. En fait l'équation (5.39), compte tenu du fait que $\theta = \theta_x$, devient :

$$\left[\theta_x|\underline{y},\underline{x}\right] \propto \prod_{i=1}^{i=s} \theta_{i1}^{\alpha_{i1}-1} \cdot \ldots \cdot \theta_{is}^{\alpha_{is}-1} \cdot \wp(\theta_{ij}) \tag{5.43}$$

$\wp(\theta_{ij})$ étant un polynôme d'ordre t_{obs} en θ_{ij} avec $i,j = 1...s$. La présence dans cette expression de termes du type $\theta_{ij}^{k_{ij}} \cdot \theta_{i'j'}^{k_{i'j'}}$ avec $ij \neq i'j'$ exclut toute propriété de conjugaison.

Alors, il faudrait mettre en place un algorithme de Gibbs "*hybride*" (Gilks, 1996 ; Brooks, 1998) dans lequel les tirages selon les lois conditionnelles dont l'expression est connue à une constante près sont réalisés à l'aide de la méthode de Metropolis-Hastings (sélection d'un candidat sur la base de la dernière valeur retenue et acceptation/rejet avec une certaine probabilité) et la procédure perd la simplicité de l'algorithme de Gibbs original.

C'est pour cette raison qu'on a utilisé pour toutes les grandeurs la méthode de Metropolis-Hastings. Cette démarche, d'une certaine manière, se rapproche de la méthode classique d'estimation des MDEL dans laquelle on fait abstraction du processus dynamique en écrivant la vraisemblance comme :

$$\left[\underline{y}|\theta\right] = \int \left[\underline{y}|\theta,\underline{x}\right] \left[\underline{x}|\theta\right] d\underline{x} \tag{5.44}$$

qui n'est rien d'autre que l'équation (5.21).

Dans Pasanisi et Parent (2003) et Pasanisi (2003) l'estimation du modèle de dégradation des compteurs était réalisé à l'aide du logiciel WinBUGS (annexe C) qui met en place l'algorithme hybride Gibbs/Metropolis, directement à partir du graphe de la figure 5.5. Cette méthode, pour un modèle à 4 états et pour des valeurs de t_{obs} supérieures à 10 se révèle assez lourde en termes de temps de calcul. L'utilisation de WinBUGS dans le cadre de ce modèle est donc faisable quand le nombre de pas temporels est limité.

L'implémentation d'un algorithme *ad hoc* de type Metropolis-Hastings a permis une augmentation spectaculaire de la vitesse de calcul (de plusieurs heures à quelques minutes pour réaliser 10 000 itérations MCMC dans le cas t_{obs}=20) et donc d'avoir un outil d'estimation plus général.

5.4.4 L'observation des compteurs bloqués

Dans la pratique technique, le blocage d'un compteur est détecté par les releveurs au moment de la lecture périodique de l'index. Si, par rapport à la dernière valeur connue, ce dernier n'a pas bougé, alors le compteur est jugé bloqué et une demande de remplacement est immédiatement établie. L'appareil défectueux est déposé dans les plus brefs délais.

Compte tenu aussi du fait que les compteurs actuellement utilisés ont tendance à se bloquer de moins en moins (les taux de blocages annuels moyens sont généralement inférieurs à 1%) la probabilité d'échantillonner un compteur bloqué est pratiquement nulle et leur faible nombre dans la base métrologique ne permet aucune estimation de leur proportion.

Par conséquent il n'est pas possible d'écrire une équation d'observation multinomiale du type (5.21) :

$$\left[\underline{y_1}, \underline{y_2}, \underline{y_3}, \underline{y_4} | \boldsymbol{\theta}\right] = \prod_{t=0}^{t_{obs}} \frac{n(t)!}{\prod_{j=1}^{4} y_j(t)!} \prod_{j=1}^{4} [x_j(t)]^{y_j(t)} \tag{5.45}$$

mais il faut distinguer entre l'observation des états "*observables*" (figure 4.7, page 51) parmi les données métrologiques (1, 2 et 3) et l'observation des compteurs bloqués.

Le nombre de compteurs, d'âge donné, qui se trouvent dans les états 1, 2 ou 3 est la réalisation d'un tirage multinomial de paramètres $n(t)$, taille de l'échantillon, et $\underline{q}(t)$, vecteur dont les composantes $q_j(t)$ sont les probabilités conditionnelles d'appartenance à l'état j, sachant que le compteur n'est pas bloqué :

$$\left[\underline{y_1}, \underline{y_2}, \underline{y_3} | \boldsymbol{\theta}\right] = \prod_{t=0}^{t_{obs}} \frac{n(t)!}{\prod_{j=1}^{3} y_j(t)!} \prod_{j=1}^{3} q_j(t)^{y_j(t)} \tag{5.46}$$

$$q_j(t) = \frac{[x_j(t)]}{\sum_{j=1}^{j=3} [x_j(t)]} \tag{5.47}$$

Concernant les compteurs bloqués, plusieurs solutions sont imaginables pour décrire leur observation.

Si on fait confiance aux informations de la base "*Branchements et Compteurs*" ICBC, alors on peut imaginer que le nombre $y_b(t)$ de compteurs bloqués

d'âge donné est la réalisation d'un tirage binomial de paramètres $p_b(t)$ et $m(t)$:

$$Y_b(t) \sim \mathcal{B}in\left(p_b(t), m(t)\right) \tag{5.48}$$

où $m(t)$ est le nombre de compteurs d'âge t répertoriés dans la base ICBC et $p_b(t)$ est la probabilité d'observer le blocage d'un compteur d'âge t. Ce dernier terme est, en première approximation, égal à la probabilité $b(t)$ que le blocage ait lieu dans la période comprise entre $t-1$ et t :

$$b(t) = [x_1(t-1)] \cdot \theta_{14} + [x_2(t-1)] \cdot \theta_{24} + [x_3(t-1)] \cdot \theta_{34} \tag{5.49}$$

dans l'hypothèse (réaliste) que le remplacement soit pratiquement immédiat et donc que les compteurs bloqués d'âge t soient ceux qui sont tombés en panne entre t-1 et t.

Eventuellement cette probabilité peut être multipliée par des termes correctifs qui prennent en compte l'incertitude liée à l'observation des blocages dans ICBC. Dans Pasanisi et Parent (2003) et Pasanisi (2003) la méthode d'extraction des données concernant les déposes de compteurs ne permettait d'avoir qu'une partie des informations. Ce phénomène était bien mis en évidence par un écart sensible entre le nombre de compteurs déposés et le nombre de compteurs posés sur une même année.

Le terme $p_b(t)$ était alors calculé comme :

$$p_b(t) = b(t) \cdot p_{ren} \tag{5.50}$$

où p_{ren} était une probabilité de renseignement de l'opération de dépose dans la base de données, rapport entre le nombre de compteurs déposés (issu d'ICBC) et le nombre réel de déposes, obtenu sur la base des informations concernant les achats de compteurs sur la même année. Ce rapport était égal à 0.58 pour les données de l'année 2001.

La méthode actuelle d'extraction (cf. chapitre 6) permet d'obtenir pratiquement la totalité des données concernant les déposes des compteurs dans le périmètre de la base ICBC ($p_{ren} = 1$).

Si l'on admet que les bases de données n'apportent pas d'information significative concernant les blocages on aborde le problème différemment. Ce cas peut se présenter en pratique quand on mène une étude sur des populations de petite taille ; alors le nombre de pannes observées se réduit à quelques unités et

le mécanisme binomial n'est plus très adapté pour décrire le phénomène. Une autre situation où les données ICBC ne sont pas significatives est quand les motifs des interventions de poses/déposes des compteurs ne sont pas suffisamment renseignés (cf. chapitre 6).

L'occurrence des blocages est alors estimée sur la base de l'avis d'expert qui, éventuellement s'appuyant sur le peu d'informations disponibles, donne un taux de blocage $u(t)$ et une incertitude sur son estimation (sous forme d'un écart type $\sigma(t)$). Une équation d'observation réaliste peut être alors une équation normale de type :

$$U(t) \sim \mathcal{N}\left(b(t), \sigma(t)^2\right) \tag{5.51}$$

où les taux de blocages fournis par l'expert remplacent les données (inexistantes ou peu fiables).

Enfin, une méthode encore plus pragmatique pour résoudre le problème est de faire complètement abstraction des blocages et de mener les calculs d'inférence et de prévision sur un modèle réduit dont l'état absorbant est l'état n. 3. Cette approche est justifiée par la faible occurrence des blocages et par le fait que, en pratique, ce qui intéresse le distributeur d'eau est surtout la prévision du comportement des compteurs survivants plutôt que l'estimation des probabilités de blocages.

5.5 Un exemple

5.5.1 Les données

Afin de montrer un exemple d'application des techniques d'inférence précédemment décrites à un cas réel, nous avons examiné les données reportées dans le tableau suivant concernant des compteurs volumétriques domestiques (DN 15 et 20 mm), d'âge compris entre 1 et 20 ans. On imagine que l'âge est le seul facteur explicatif de la dégradation.

	Données métrologiques			Données de facturation	
Age	y_1	y_2	y_3	Nombre de blocages	Taille de la pop.observée
1	43	5	2	136	63741
2	25	6	3	67	61983
3	78	16	7	82	41595
4	157	33	9	116	66475
5	237	50	7	149	82744
6	235	37	14	116	106898
7	243	72	11	64	45928
8	170	63	29	84	34452
9	174	61	21	95	33880
10	187	80	35	97	29209
11	179	94	27	134	32109
12	149	86	32	127	33132
13	150	125	51	176	26832
14	152	146	41	146	19203
15	157	137	50	394	64403
16	114	81	89	129	23155
17	80	65	66	85	10149
18	70	50	22	48	4676
19	79	62	25	107	10484
20	63	63	46	113	12010

Concernant les blocages, les données ICBC sont relatives à un échantillon significatif de l'effectif total de compteurs installés et on fait l'hypothèse que le taux de renseignement de l'information, au sein de la population examinée est de 100% : en pratique tous les blocages sont inclus dans les observations.

5.5.2 L'estimation

Les distributions *a priori* des probabilités de transition θ_{ij} sont des distributions de *Dirichlet* dont les paramètres sont tous égaux à 1. Ce choix donne lieu à des lois uniformes pour les 3 vecteurs des probabilités de transitions à partir des états 1, 2, 3 :

$$\begin{aligned} \underline{\theta}_1 &= \{\theta_{11}, \theta_{12}, \theta_{13}, \theta_{14}\} \sim \mathcal{D}(1,1,1,1) \\ \underline{\theta}_2 &= \{\theta_{22}, \theta_{23}, \theta_{24}\} \sim \mathcal{D}(1,1,1) \\ \underline{\theta}_3 &= \{\theta_{33}, \theta_{34}\} \sim \mathcal{D}(1,1) \end{aligned} \tag{5.52}$$

Le choix d'une loi *a priori* uniforme traduit le manque de connaissance préliminaire sur les probabilités de transition : toutes les valeurs possibles sont également probables *a priori*.

Pour obtenir des tirages aléatoires dans les loi *a posteriori* des θ_{ij} on a réalisé avec l'algorithme de Metropolis-Hastings 10 000 itérations de 5 chaînes

différentes, à partir de points initiaux très dispersés. L'utilisation de plusieurs chaînes permet de vérifier la convergence vers la loi visée, comme il sera plus clair dans la suite.

Les dernières 5 000 valeurs de chaque chaîne sont retenues pour simuler les loi *a posteriori*. Le tableau suivant montre les principales caractéristiques de ces lois, obtenues empiriquement, à partir des valeurs retenues.

					Percentiles	
Param.	Moyenne	Mode	Mediane	Ec. type	2.5%	97.5%
θ_{11}	0.9483	0.9485	0.9484	0.0011	0.9461	0.9504
θ_{12}	0.0492	0.0489	0.0491	0.0012	0.0469	0.0515
θ_{13}	0.0011	0.0008	0.0010	0.0005	0.0003	0.0022
θ_{14}	0.0013	0.0014	0.0013	0.0001	0.0012	0.0015
θ_{22}	0.9389	0.9395	0.9389	0.0026	0.9337	0.9444
θ_{23}	0.0610	0.0603	0.0611	0.0026	0.0556	0.0662
θ_{24}	0.0001	0.0000	0.0001	0.0001	0.0000	0.0004
θ_{33}	0.9667	0.9677	0.9669	0.0017	0.9633	0.9698
θ_{34}	0.0332	0.0322	0.0331	0.0017	0.0302	0.0367

Les percentiles d'ordre 2.5% et 97.5% représentent les bornes de *l'intervalle de crédibilité a posteriori* de niveau 95%, interprété par les bayésiens comme l'intervalle où se trouvent 95% des valeurs possibles des paramètres.

La figure 5.6 montre les histogrammes de fréquence empirique des valeurs simulées pour les probabilités de transition les plus intéressantes.

La convergence vers la loi *a posteriori* est vérifiée en comparant les "*parcours*" des 5 chaînes simulées pour chaque variable. La figure 5.8 montre, par exemple, les 5 chaînes relatives à la variable θ_{23}. Dans un premier temps les valeurs générées dépendent des valeurs initiales et les chaînes restent significativement distinctes. Au fur et à mesure que le nombre d'itérations augmente les chaînes se mélangent et deviennent pratiquement non distinguables, ce qui indique que les valeurs générées des 5 chaînes ont la même loi de probabilité. Il s'agit de la loi limite de la chaîne, soit la loi *a posteriori* qu'on se propose de simuler.

En pratique la convergence est vérifiée à l'aide de la méthode de Brooks-Gelman (1998), décrite dans l'annexe C. Cette technique s'appuie sur le calcul pour chaque variable d'une statistique, $\widehat{R}_{BG}$, qui fournit une mesure de la similitude des lois sous-jacentes des dernières valeurs simulées de chaque chaîne. Quand

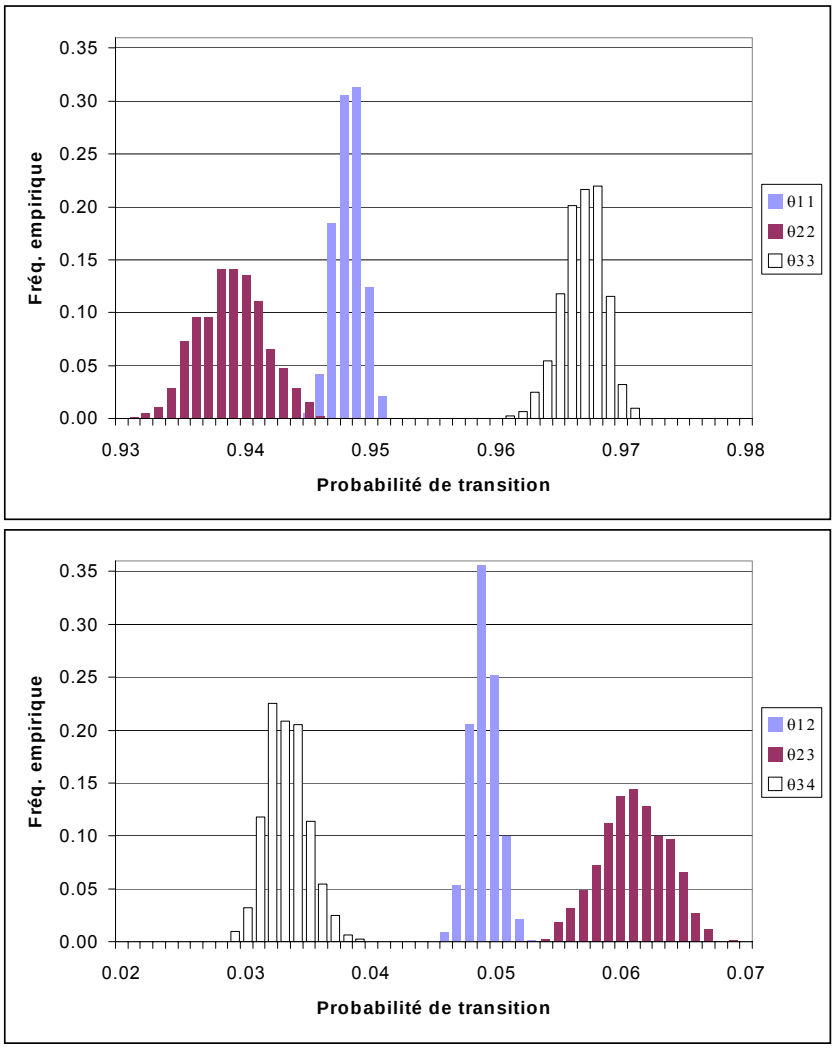


FIG. 5.6 – Histogrammes de fréquence empirique des probabilités de transition θ_{ii} et θ_{ii+1}.

pour toutes les variables simulées la valeur de $\widehat{R}_{BG}$ est suffisamment proche de 1, alors on peut affirmer que la convergence est atteinte.

La figure 5.7 montre que, pratiquement à partir de l'itération n. 3 000 toutes les chaînes ont atteint leur distribution limite et que les valeurs extraites peuvent être utilisées pour la simulation des lois *a posteriori*.

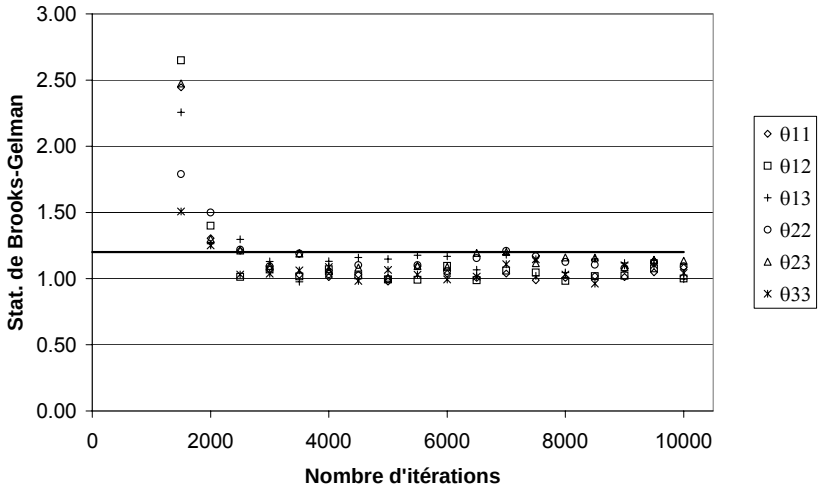


FIG. 5.7 – Contrôle de convergence de l'algorithme MCMC à l'aide de la statistique de Brooks-Gelman.

5.5.3 Prédictions et test d'adéquation

Sur la base des résultats de l'inférence, il est possible d'obtenir les distributions de probabilité *a posteriori* de toute grandeur W (observable ou non) dépendante des paramètres du modèle θ. Ils s'agit de la phase déductive (Girard et Parent, 2000) de l'inférence statistique qui suit la phase inductive de l'estimation.

Théoriquement, la distribution *a posteriori* $[w|y]$, que certains appelent "*prédictive*", de W peut être obtenue par marginalisation de la loi jointe $[w, \theta|y]$:

$$[w|y] = \int_{\Omega} [w, \theta|y] d\theta = \int_{\Omega} [w|\theta][\theta|y] d\theta \tag{5.53}$$

En pratique les lois prédictives sont simulées sur la base des valeurs extraites des distributions *a posteriori* des paramètres, obtenues avec l'algorithme

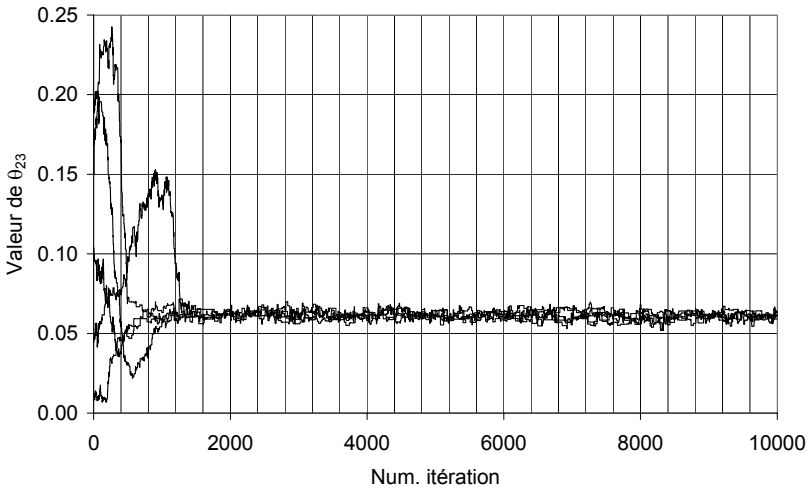


FIG. 5.8 – Parcours des 5 chaînes de valeurs de la variable θ_{23}, générées par l'algorithme de Metropolis-Hastings.

MCMC. En particulier on s'intéresse aux prédictions des grandeurs suivantes : probabilités d'appartenance aux différents états métrologiques et prédictions des grandeurs observables.

Probabilités d'appartenance aux différents états métrologiques

Il s'agit d'informations importantes du point de vue pratique, puisqu'elles permettent de reproduire la vie opérationnelle d'un compteur. La figure 5.9 montre, en fonction de l'âge, les intervalles de crédibilité à 95% et les moyennes *a posteriori* des probabilités des 4 états. Dans la partie inférieure de la figure, on a isolé les probabilités conditionnelles d'appartenance aux états de fonctionnement 1, 2 et 3, sachant que le compteur ne se trouve pas dans l'état absorbant de blocage. Ces probabilités permettent de prédire la composition d'un parc de compteurs opérationnels d'âge donné et, en particulier, le taux de non-conformité réglementaire, représenté par la probabilité conditionnelle d'appartenir à l'état 3

Prédictions des grandeurs observables

Il s'agit en pratique des simulations de nouvelles données $\widetilde{y}$, à l'aide du modèle, sur la base des lois *a posteriori* de ses paramètres. Cet exercice de retour

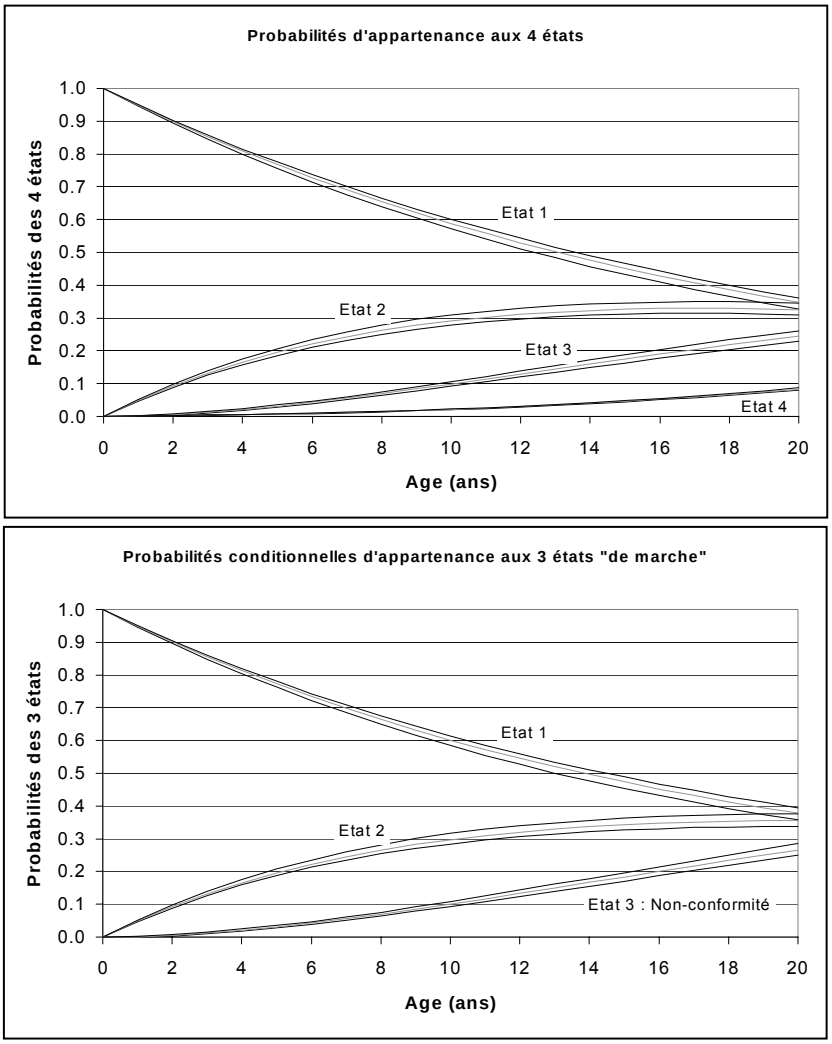


FIG. 5.9 – Intervalles de crédibilité prédictifs à 95% des probabilités d'appartenance aux différents états en fonction de l'âge.

du modèle à la réalité observable a pour but d'évaluer l'adéquation du modèle aux données. Une méthode intuitive pour répondre à cette question consiste en comparer les vraies données avec les intervalles prédictifs de leur répétitions. La figure 5.10 montre que les "tubes" prédictifs à 95% du nombre de compteurs appartenant à chacun des 4 états contiennent pratiquement toujours les observations et que donc le modèle est bien calé sur les données utilisées pour son estimation.

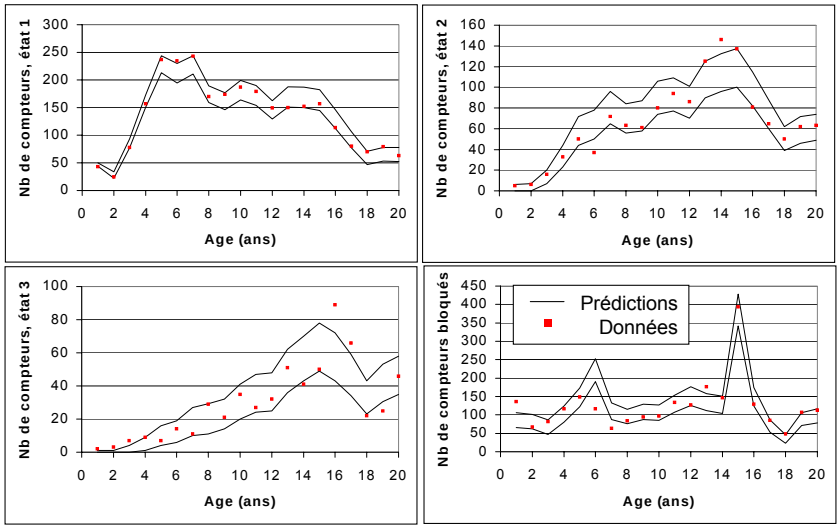


FIG. 5.10 – Intervalles de crédibilité prédictifs à 95% des répétitions des observations (modèle à 4 états).

Un jugement sur l'adéquation du modèle aux données peut être aussi établi avec la méthode des "*p-valeurs bayésiennes*". Cette technique se base sur le choix préalable d'une statistique de test T, fonction des grandeurs observables et des paramètres du modèle. Avec les mots de Gelman et al. (1995) la p-valeur bayésienne p est la probabilité que les données simulées soient "plus extrêmes" que les données réelles. Cette probabilité est calculée à l'aide de la statistique de test choisie :

$$p = [T(\widetilde{y}, \theta) > T(y, \theta) \,|y] \tag{5.54}$$

Des valeurs proches de 0.5 indiquent que le modèle est bien calé sur les données et, *a contrario* quand p est proche de 0 ou de 1, on peut conclure que le modèle proposé ne reproduit pas, de manière satisfaisante, la structure

stochastique des données. En théorie le calcul de cette grandeur se fait par le calcul de l'intégrale multidimensionnelle :

$$\int_{\Psi}\int_{\Omega} \mathbf{1}_{T(\widetilde{y},\theta)>T(y,\theta)}\,[\theta|y]\,[\widetilde{y}|\theta]\,d\theta d\widetilde{y} \tag{5.55}$$

où Ψ est l'espace de toutes les valeurs possibles de $\widetilde{y}$.

En pratique, si on a à disposition des échantillons de taille n des lois *a posteriori* de $\widetilde{y}$ et θ, le calcul se fait par méthode Monte Carlo en calculant le nombre de fois où $T(\widetilde{y},\theta) > T(y,\theta)$ divisé par n.

Dans le cas présent un calcul très simple peut être fait en utilisant comme statistique de test les données mêmes : on compte, pour $i = 1 \ldots n$, et pour $t = 1..t_{obs}$ le nombre de fois que deux parmi les trois valeurs simulées $\widetilde{y}_1^{(i)}(t)$, $\widetilde{y}_2^{(i)}(t)$, $\widetilde{y}_3^{(i)}(t)$ sont supérieures aux valeurs "vraies" correspondantes $y_1(t)$, $y_2(t)$, $y_3(t)$ et que la valeur de $\widetilde{y}_b^{(i)}(t)$ est supérieure à $y_b(t)$. La p-valeur est alors définie par :

$$p = \frac{\sum_{i=1}^{i=n}\sum_{t=1}^{t=t_{obs}}\left(\mathbf{1}_{\widetilde{y}_1^{(i)}(t)>y_1(t)} + \mathbf{1}_{\widetilde{y}_3^{(i)}(t)>y_3(t)} + \mathbf{1}_{\widetilde{y}_b^{(i)}(t)>y_b(t)}\right)}{3\cdot n\cdot t_{obs}} \tag{5.56}$$

dans l'hypothèse où l'on choisit pour les calculs les variables $\widetilde{y}_1^{(i)}(t)$ et $\widetilde{y}_3^{(i)}(t)$. Observons que puisque :

$$\widetilde{y}_1^{(i)}(t) + \widetilde{y}_2^{(i)}(t) + \widetilde{y}_3^{(i)}(t) = y_1(t) + y_2(t) + y_3(t) \qquad \forall i, \forall t \tag{5.57}$$

il convient de prendre en compte seulement deux des trois fonctions indicatrices :

$$\mathbf{1}_{\widetilde{y}_1^{(i)}(t)>y_1(t)}, \mathbf{1}_{\widetilde{y}_2^{(i)}(t)>y_2(t)}, \mathbf{1}_{\widetilde{y}_3^{(i)}(t)>y_3(t)}$$

qui ne sont pas indépendantes.

Si, par exemple $\widetilde{y}_1^{(i)}(t) > y_1(t)$ et aussi $\widetilde{y}_2^{(i)}(t) > y_2(t)$, alors forcement sera $\widetilde{y}_3^{(i)}(t) > y_3(t)$.

Pour l'exemple proposé la valeur de p ainsi calculée est de 0.47, ce qui confirme la bonne adéquation du modèle aux données.

En conclusion, du point de vue technique, les résultats montrent un bon comportement des dispositifs examinés : après 20 ans de service, de l'ordre de 40% des compteurs "survivants" ont encore une métrologie excellente et moins

de 30% ne respectent pas les conditions minimales d'exactitude requises par la future réglementation.

Ces résultats confirment l'excellente qualité des compteurs volumétriques et justifient le choix actuel de la CGE d'utiliser en prévalence ces types d'appareils, surtout en vue des prochaines contraintes réglementaires.

5.5.4 Et si on utilisait un modèle à 3 états ?

En pratique, les données relatives aux blocages sont parfois inconnues ou peu fiables. Les phénomènes de pannes sont aujourd'hui assez rares et les taux annuels normalement ne dépassent pas 1%. Si on prend le type de compteur et l'âge comme seuls facteurs explicatifs on arrive à avoir des données significatives sur les blocages (extraites sur la base d'une population de taille significative), mais, au fur et à mesure qu'on prend en compte d'autres facteurs (comme le site d'exploitation et la consommation annuelle) il devient de plus en plus difficile d'obtenir des informations exploitables de la base ICBC, puisque les populations concernées deviennent de taille trop petite.

Une façon d'aborder le problème est d'imaginer un modèle réduit à 3 états qui fait complètement abstraction du phénomène de blocage et ne considère que les compteurs en état de marche. Cette approche est adaptée à la pratique technique : les indications les plus intéressantes pour le gestionnaire concernent surtout les compteurs survivants et les pannes ne posent pas trop de problèmes parce qu'elles sont rares et en plus facilement décelables.

Dans un modèle réduit l'état 3 est absorbant et les transitions sont régies par la matrice 3×3 :

$$\begin{pmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ 0 & \theta_{22} & \theta_{23} \\ 0 & 0 & 1 \end{pmatrix}$$

Le nombre de paramètres indépendants passe donc de 6 à 3 : on dit que le modèle proposé est plus parcimonieux.

A titre d'exemple on peut comparer les résultats obtenus à l'aide du modèle réduit sur la base du jeu de données proposé au début de ce paragraphe avec ceux précédemment examinés. La figure 5.11 montre les moyennes prédictives *a posteriori* des probabilité d'appartenance aux 3 états de marche, calculées à l'aide du modèle réduit (en pointillé). Ces prévisions sont comparées avec les

moyennes des probabilités conditionnelles prédictives, évaluées avec le modèle à 4 états.

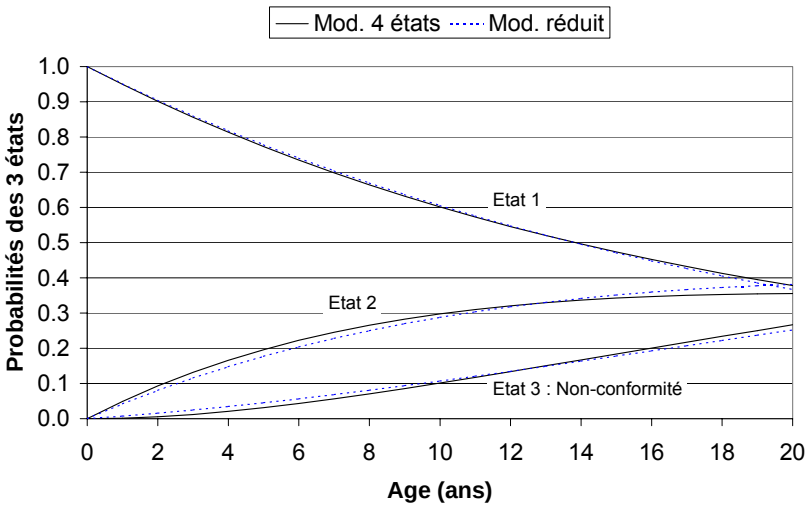


FIG. 5.11 – Moyennes prédictives des probabilités d'appartenance aux 3 états de marche, évaluées à l'aide des modèles à 3 et 4 états.

Les écarts entre les deux estimations sont très petits et, pratiquement nuls pour la probabilité d'appartenance au groupe des "bons" compteurs.

Concernant l'adéquation aux données, le modèle réduit se cale sur les observations aussi bien que le modèle à 4 états, comme il est évident dans la figure 5.12 qui montre, comme dans la figure 5.10, les vraies données et leur "*tubes*" prédictifs à 95%. Ces conclusions sont confirmées par un test d'adéquation, similaire à celui exposé dans le paragraphe précédent. La p-valeur calculée avec la formule :

$$p = \frac{\sum_{i=1}^{i=n} \sum_{t=1}^{t=t_{obs}} \left(\mathbf{1}_{\tilde{y}_1^i(t) > y_1(t)} + \mathbf{1}_{\tilde{y}_3^i(t) > y_3(t)} \right)}{2 \cdot n \cdot t_{obs}} \tag{5.58}$$

vaut 0.51, ce qui indique, en conclusion, que le modèle réduit est, lui aussi, bien adapté à la structure des données.

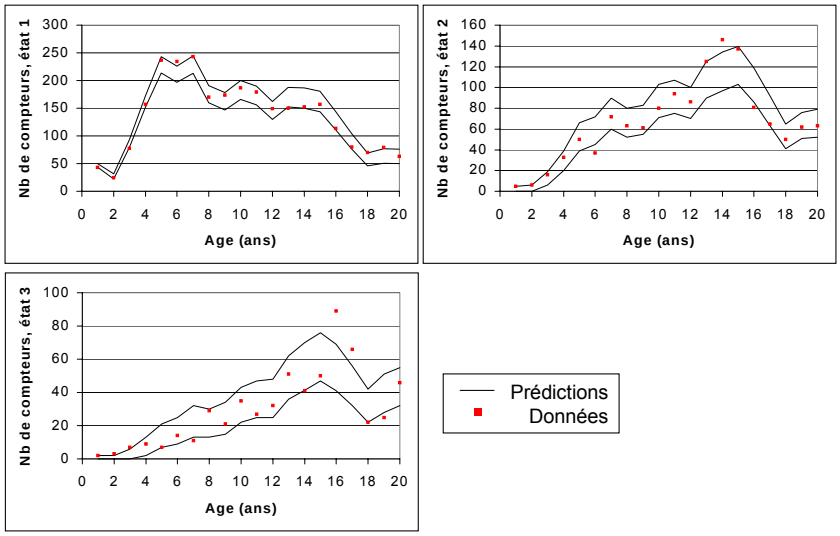


FIG. 5.12 – Intervalles de crédibilité prédictifs à 95% des répetitions des observations (modèle à 4 états).

Troisième partie

Amélioration et utilisation pratique du modèle de dégradation

Chapitre 6

A la recherche de nouvelles variables explicatives

6.1 Des milliers de sites d'exploitation

La Compagnie Générale des Eaux (CGE) est aujourd'hui le premier distributeur d'eau français. Quelques chiffres nous permettront de comprendre aisément la multiplicité des conditions d'exploitation et d'ici l'importance d'avoir des éléments d'aide à la gestion des parcs adaptés à chaque site.

Actuellement la CGE est partenaire d'environ 8 000 collectivités locales et dessert en eau potable 26 millions de consommateurs en France (hors DOM-TOM). L'eau distribuée, en provenance de 2 700 unités de production, est fournie aux usagers grâce à un réseau de 200 000 km de canalisations.

L'organisation géographique de la Compagnie prévoit *:*

- 11 *Directions Régionales :* Bretagne, Normandie, Flandres-Artois-Picardie, Est, Ile de France, Banlieue de Paris, Centre Est, Loire-Poitou, Sud Ouest, Sud, Sud Est.
- 48 *Centres Opérationnels*, chargés notamment des relations contractuelles avec les collectivités.
- 134 *Agences*, qui s'occupent principalement de la gestion quotidienne des contacts avec les usagers et les collectivités et exécutent les tâches techniques d'exploitation, entretien et travaux.

Un découpage encore plus fin et particulièrement intéressant en pratique peut être réalisé avec les *contrats*. Il s'agit littéralement des accords signés entre les

collectivités locales (communes ou regroupement de communes) et la CGE qui découpent donc des zones géographiques de plus petite échelle.

Les contrats sont des unités géographiques forcément inégales en extension et nombre d'usagers ; néanmoins cette répartition du territoire a une utilité opérationnelle, comme il sera précisé dans la suite. Le paramètre le plus significatif, dans le cadre de cette étude, pour décrire la taille d'un contrat est le nombre de compteurs, qui équivaut en pratique au nombre d'abonnés desservis.

Les deux contrats les plus importants sont la Banlieue de Paris (SEDIF) avec environ 520 000 compteurs et le Grand Lyon avec plus de 266 000 appareils. Parmi les autres grands contrats de la CGE on trouve notamment la Rive Droite de Paris, et des grandes villes dont Toulouse, Rennes, Brest, Ajaccio, Toulon, Metz, Nantes, Caen, Montpellier. A coté de ces gros contrats, avec des dizaines de milliers de branchements il existe des contrats de tailles nettement plus petites (de quelques centaines à quelques milliers de compteurs).

Dans la pratique technique il est bien reconnu par les experts et les exploitants que la dégradation des compteurs n'est pas la même sur tous les sites. Le principal facteur explicatif typiquement évoqué comme responsable de cette différence de comportement est la "*typologie*" d'eau distribuée. Les paramètres physico-chimiques classiquement proposés par les experts sont la dureté (responsable principale des phénomènes d'entartrage), la teneur en fer, le pH. En revanche aucune étude spécifique n'a jamais prouvé l'importance et le rôle joué par ces facteurs dans le mécanisme de dégradation des compteurs.

Par exemple, la formation de tartre dans un compteur ne dépend pas uniquement de la dureté mais aussi du pH, de la température de l'eau, et des modalités de puisage de l'eau, le dépôt ayant tendance à se former quand le compteur est sujet à des périodes d'arrêt. Maints consommateurs, obligés de détartrer régulièrement leurs machines à café, seront surpris de constater que leur compteurs au pied de l'immeuble, après expertise, ne sont pas du tout affectés par ce phénomène.

En outre, une différence (parfois significative) existe entre la composition de l'eau à la sortie d'usine et celle de l'eau qui arrive aux compteurs, après avoir transité dans les réseaux de distribution. L'altération des paramètres physico-chimiques de l'eau, par effet de la nature et de l'état des conduites, est actuellement un sujet de recherche active parmi les professionnels de l'eau (Boireau et al., 2000), (Mayadatchevski, 2000).

La présence de particules solides (notamment sables et micro-sables) dans l'eau distribuée peut être due à la réalisation de travaux d'entretien et réparation du réseau (après une coupure, au moment de la remise en service, une surcharge de particules solides peut se manifester temporairement), mais aussi à des phénomènes de corrosion (détachement de paillettes métalliques des canalisations). Ce facteur est très important dans la dégradation des compteurs, les particules solides pouvant rayer ou bloquer le mesureur d'un compteur volumétrique ou s'incruster dans le mécanisme de transmission.

Enfin, parmi les causes de cette différence de comportement des compteurs en service il y a des conditions d'utilisation particulières liées à l'utilisation de certains appareils domestiques. Un exemple typique est l'utilisation de chasses d'eau sans réservoirs qui entraîne des augmentations brusques du débit de puisage, pouvant dégrader les organes de mesure.

En bref, d'une part la méconnaissance de l'effet de chaque facteur sur le mécanisme de dégradation des compteurs et d'autre part la complexité des conditions réelles d'exploitation, difficilement traduisibles en un nombre réduit de paramètres, conseillent d'aborder le problème différemment. L'approche choisie consiste à associer à chaque site d'exploitation une notion d'*agressivité* qui traduit l'effet global de tous les facteurs énumérés (qu'on imagine "*locaux*"). Par ailleurs, les rares études de dégradation des compteurs disponibles en bibliographie ont toujours souligné la validité strictement locale des "*lois de vieillissement*" et des règles de gestion technique des parcs (AWWA, 1966), (Sisco, 1967), (Newman et Noss, 1982), (Tao,1982), (Grau, 1985), admettant implicitement la pertinence de cette démarche.

L'objectif principal de ce chapitre est de caractériser l'agressivité des sites d'exploitation de la CGE en utilisant les informations de la base de données de facturation et les résultats des essais métrologiques.

Comme il a été déjà anticipé, l'unité géographique choisie est le contrat. L'inconvénient de cette méthode est l'inégalité de ce découpage, les unités pouvant être très différentes en taille. En outre, à l'intérieur du même contrat on peut retrouver des conditions d'exploitations différentes en termes de qualité d'eau et/ou état du réseau de distribution.

En revanche, travailler avec les contrats permet d'obtenir plus facilement les données sur les compteurs en service par rapport à d'autres indicateurs plus classiques comme la ville. Le parc compteurs de la CGE est déjà découpé, pour

des raisons administratives internes, en contrats et dans les bases de données, cet indicateur est facilement accessible (à chaque contrat est associé un code alphanumérique de 5 chiffres, par exemple le code I4000 correspond à la ville de Toulouse).

En outre, dans la pratique technique, il est très intéressant d'avoir des indications sur la dégradation des compteurs au niveau des contrats parce que cette unité représente aussi (généralement) un périmètre géographique où les autres variables technico-économique utilisées pour la gestion des parcs de compteurs, et notamment le prix de l'eau, sont constantes (même si parfois on trouve des tarifs différents à l'intérieur du même contrat). D'autre part, les éventuelles contraintes relatives à l'âge maximal des appareils sont aussi fixées par les collectivités locales au niveau du contrat, ce qui impose, en pratique, des règles de gestion différentes de contrat à contrat, mais homogènes dans un même contrat.

6.2 Des sources différentes d'information

La source d'information privilégiée pour étudier la différence de comportement des compteurs en service, en fonction du site d'exploitation, est représentée par les résultats des essais métrologiques. Pour cette raison, dans la base métrologique on trouve aussi, parmi les autres informations relatives aux compteurs étalonnés, le contrat de provenance.

Puisque la dégradation des compteurs dépend de plusieurs facteurs, il est important dans l'analyse des données métrologiques de réduire, dans la mesure du possible, l'effet des variables qui ne sont pas objet de l'étude (et notamment le type de compteur). Pour cette raison on a isolé, parmi les résultats d'étalonnage, uniquement ceux concernant un groupe de compteurs domestiques volumétriques, à entraînement magnétique, techniquement assez similaires et qui présentent des résultats métrologiques peu différents. Les compteurs choisis pour l'analyse sont les Altaïr, Véga, Aquadis et Volumag dans les diamètres 15 et 20 mm. Par ailleurs ces compteurs représentent environ 70% du parc de la CGE.

On va chercher à mettre en relation l'âge et la probabilité d'appartenance à l'état métrologique n. 1 (courbe métrologique entièrement contenue dans les canaux de tolérance des compteurs en service sur toute l'étendue des débits d'essai). Un simple modèle "*hiérarchique*" de dégradation exponentielle permet, pour les contrats pour lesquels on dispose d'un nombre significatif d'étalonnages, le

calcul d'une *vitesse de dégradation métrologique* interpretée comme un indicateur de l'agressivité du site.

Malheureusement, le nombre de contrats pour lesquels on est capable d'estimer ce paramètre d'agressivité est limité (78 contrats sur un peu plus de 1 900). Ces sites représentent, néanmoins, environ 20% des compteurs français en terme d'effectif.

Une autre importante source d'information sur le comportement des compteurs en service est représentée par la base de données "*Branchements et Compteurs*" ICBC. Dans cette base on trouve trace des interventions de dépose des compteurs à partir de 1997 avec indication du motif de dépose. En d'autres termes on sait, pour chaque compteur remplacé, si l'opération a eu lieu à titre *préventif* ou *curatif* et notamment si l'appareil remplacé était bloqué.

L'hypothèse à la base de l'étude est que le taux de blocage observé sur un contrat peut être interprété comme un deuxième indicateur de l'agressivité du site. Comme pour l'étude de la dégradation métrologique, on s'intéresse aux taux de blocage d'un groupe de compteurs de référence qui sont en l'occurrence les mêmes étudiés dans le cadre de l'analyse des résultats métrologiques.

La méthode développée vise à détecter 3 groupes de contrats caractérisés par différents niveaux d'agressivité (A_1, A_2, A_3 en ordre croissant), et prévoit 3 phases :

1. Estimation d'un paramètre d'agressivité sur la base uniquement des résultats métrologiques (*vitesse de dégradation métrologique* λ). En s'appuyant sur la valeur de λ, les contrats sont découpés en trois groupes d'agressivité croissante.
2. Pour chacun des trois groupes ainsi obtenus, on calcule les taux de blocages des compteurs de référence, suivant les informations de la base de données ICBC. Les calculs montrent une cohérence entre les deux indicateurs d'agressivité (vitesse de dégradation et taux de blocage) : les sites où les compteurs se dégradent plus vite sont aussi ceux où l'on observe, en proportion, plus de blocages.
3. Estimation de l'agressivité des contrats pour lesquels uniquement les informations ICBC sont disponibles, grâce à l'extrapolation de la relation entre taux de blocage et agressivité. En pratique on assigne un contrat à un des trois groupes précédemment découpés en fonction de la valeur observée des taux de blocage des compteurs de référence. Si, par exemple, les taux

observés sur un site d'agressivité inconnue sont plus proches de ceux qui caractérisent le groupe A_1, alors on assigne le contrat examiné au groupe A_1 etc.

6.3 Un modèle hiérarchique de dégradation de la métrologie

6.3.1 Des vitesses de dégradation bien différentes

Les données métrologiques montrent clairement que la proportion de compteurs qui se trouvent dans l'état métrologique 1 décroît avec l'âge des appareils étalonnés. D'autre part, à âge égal, cette proportion est différente de contrat à contrat. La figure 6.1 montre, par exemple, la différence entre les proportions de "*bons*" compteurs sur trois contrats examinés : l'agglomération de Rennes, la ville d'Ajaccio et la commune de Rosières aux Salines (Lorraine).

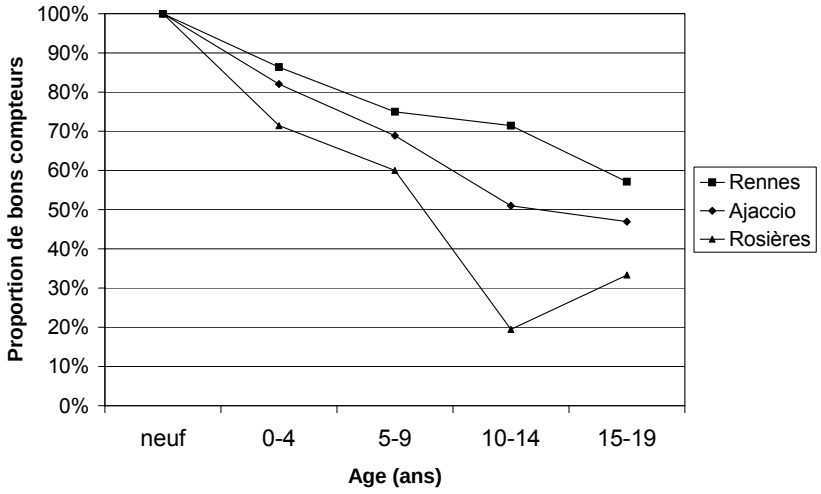


FIG. 6.1 – Proportion de bons compteurs sur 3 contrats différents.

Le découpage des données disponibles selon les deux variables contrat et âge donnant lieu à des échantillons de taille trop petite pour la plupart des contrats, dans cette étude nous avons préféré regrouper les compteurs par classes de 5 années d'âge : 0 à 4 ans, 5 à 9 ans etc.

Comme le montre la figure 6.1 les dynamiques de vieillissement peuvent être très différentes de contrat à contrat. Pour les trois exploitations examinées dans la figure, on observe que les compteurs se dégradent sensiblement plus vite à Ajaccio qu'à Rennes (ce qui concorde d'ailleurs avec l'avis des experts du comptage de la CGE).

L'application des modèles markoviens à 4 ou 3 états, décrits dans le chapitre 5, à chaque contrat est infaisable, compte tenu du fait que pour la majorité des contrats la taille des échantillons ne permet aucune estimation fiable de l'occurrence des états 2 et 3. En revanche on peut imaginer un modèle réduit, en regroupant dans une seule catégorie les états 2 et 3, et sans prendre en compte l'état 4 (en pratique on s'intéresse uniquement aux occurrences des états parmi les compteurs survivants). Le modèle qui en résulte a donc deux états, qu'on appellera b (*bon*) et d (*dégradé*), et la matrice de transition est :

$$\begin{pmatrix} \theta_{bb} & \theta_{bd} \\ 0 & 1 \end{pmatrix} \tag{6.1}$$

Puisque $\theta_{db} = 1 - \theta_{bb}$, les transitions sont régies par le seul paramètre θ_{bb}, probabilité de rester dans la catégorie des compteurs bons d'une unité de temps à une autre. Pour simplifier les notations on l'appelera simplement θ. Avec l'hypothèse que les compteurs neufs sont tous bons, un appareil d'âge donné, de provenance du contrat i, a une probabilité d'être bon exprimée par la formule :

$$P_i(t) = \theta_i^t \tag{6.2}$$

où t est une variable temporelle liée à l'âge qui vaut 1 si le compteur a un âge inférieur à 4 ans, 2 si l'âge est compris entre 5 et 9 ans etc.

Afin d'obtenir un paramètre qui soit une fonction croissante de l'agressivité du site on peut écrire la loi de décroissance exponentielle exprimée par la formule (6.2) de la manière suivante :

$$P_i(t) = \exp(-\lambda_i \cdot t) \tag{6.3}$$

On appelle le paramètre λ *vitesse de dégradation métrologique*. La détermination de la relation entre λ_i et θ_i, paramètre du modèle markovien à 2 états, est immédiate :

$$\lambda_i = \ln(1/\theta_i) \tag{6.4}$$

Conditionnellement à $P_i(t)$ le nombre observé $y_i(t)$ de compteurs bons d'âge donné est la réalisation d'un tirage binomial de paramètres $P_i(t)$ et $n_i(t)$, ce dernier étant la taille de l'échantillon de même *âge* t :

$$[y_i(t)|P_i(t)] = \binom{n_i(t)}{y_i(t)} P_i(t)^{y_i(t)} \left(1 - P_i(t)\right)^{n_i(t)-y_i(t)} \tag{6.5}$$

Une manière pour calculer les vitesses de dégradation des contrats examinés pourrait être d'écrire un modèle *ad hoc* pour chaque contrat et de mettre en œuvre individuellement les techniques usuelles d'inférence (classiques ou bayésiennes). L'inconvénient est que, pour les contrats pour lesquels les échantillons sont de tailles plus petites, les estimations risquent d'être affectées par une incertitude importante.

D'autre part, on peut imaginer que le comportement de compteurs issus de sites différents soit similaire, tout en restant quantitativement différent. Une technique plus efficace et élégante pour résoudre le problème d'estimation est de faire appel à un modèle dit *hiérarchique*. Le prochain paragraphe décrit brièvement l'esprit de cette technique de modélisation, sous une approche bayésienne.

6.3.2 Des compteurs et ... des rats

En statistique bayésienne on définit sur l'espace des paramètres d'un modèle $\underline{\varrho}=(\varrho_1, ..., \varrho_m)$ une loi *a priori* $\left[\underline{\varrho}|\underline{\eta}\right]$ paramétrée par des variables $\underline{\eta}$, qu'on appelle *hyper-paramètres*. Normalement le modélisateur imagine de connaître les valeurs des hyper-paramètres, de manière que la loi *a priori* est connue sans incertitude.

L'idée à la base de la modélisation "*hiérarchique*" (cf. par exemple Lee (1997)) est d'introduire un niveau supplémentaire dans les calculs d'inférence, en imaginant que les hyper-paramètres sont connus avec incertitude, une distribution de probabilité $\left[\underline{\eta}\right]$, dite "*hyper-prior*", exprimant leur connaissance *a priori* de la part du statisticien.

Un exemple célèbre pour expliquer les modèles hiérarchiques est fourni par Gelfand et al. (1990) et concerne l'évolution du poids de 30 jeunes rats, mesuré une fois par semaine durant 5 semaines. Les observations suggèrent qu'un modèle linéaire gaussien peut être approprié pour décrire l'augmentation du poids des animaux. Si l'on indique avec $y_{i,j}$ le poids du rat i mesuré le jour x_j on peut écrire :

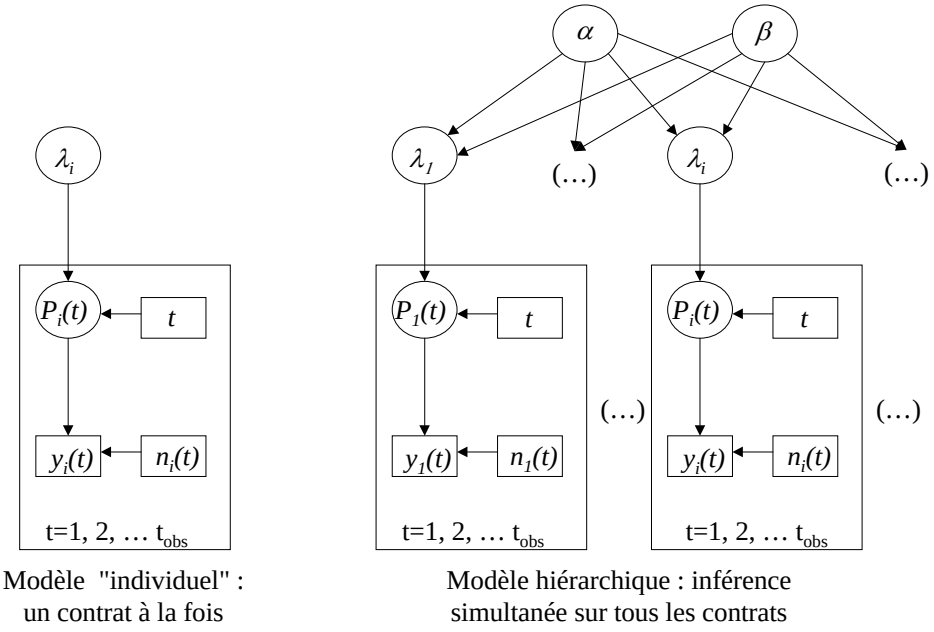


FIG. 6.2 – Hiérachisation du modèle de dégradation des compteurs.

$$y_{i,j} = (\kappa_i \cdot x_j + \nu_i) + \epsilon \tag{6.6}$$

où :

$$\epsilon \sim \mathcal{N}(0, \sigma) \tag{6.7}$$

Puisque tous les rats appartiennent à la même espèce on imagine qu'il existe un lien entre les paramètres de croissance. Cette similitude est décrite mathématiquement avec l'hypothèse que les κ_i et les ν_i ont tous les mêmes distributions de probabilités (normales) :

$$\kappa_i \sim \mathcal{N}(\mu_\kappa, \sigma_\kappa) \qquad \nu_i \sim \mathcal{N}(\mu_\nu, \sigma_\nu) \tag{6.8}$$

dont les paramètres sont issus de lois *a priori* appropriées (en l'occurrence des lois normales pour μ_κ et μ_ν et des lois *Gamma* pour σ_κ et σ_ν).

Dans le cas du modèle de dégradation des compteurs, le rôle des individus est joué par les contrats, et le paramètre qui décrit leur comportement est λ_i. On imagine que la loi commune des λ_i est une loi de type *Gamma* de paramètres α et β :

$$\lambda_i \sim \mathcal{G}(\alpha, \beta) \tag{6.9}$$

c'est-à-dire :

$$[\lambda_i|\alpha, \beta] = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_i^{(\alpha-1)} \exp(-\beta \cdot \lambda_i). \tag{6.10}$$

Le choix de cette loi de probabilité, définie uniquement pour des valeurs positives de λ_i, est justifié par l'hypothèse implicite que les compteurs se dégradent avec l'âge : si $\lambda_i > 0$, alors la fonction $\exp(-\lambda_i \cdot t)$ est strictement décroissante et sa limite pour $t \to \infty$ est 0.

Comme distribution *a priori* (*hyper-prior*) des paramètres α et β on choisit encore une *Gamma*. En particulier on fait appel à des lois "*peu informatives*", c'est-à-dire qu'elles n'apportent pas d'information significative sur la valeur des paramètres. Des lois *Gamma* peu informatives sont "classiquement" obtenues en choisissant la valeur 10^{-3} pour les 2 paramètres :

$$\alpha \sim \mathcal{G}(10^{-3}, 10^{-3}) \qquad \beta \sim \mathcal{G}(10^{-3}, 10^{-3}). \tag{6.11}$$

Cette paramétrisation donne lieu à des distributions de probabilité très aplaties sur $\mathbb{R}_+$ qui traduisent le manque d'information préliminaire sur ces paramètres. La répétition des calculs d'inférence en utilisant comme loi a priori des lois uniformes sur $\mathbb{R}_+$ (en pratique sur l'intervalle $]0, 10^5]$) donne lieu aux mêmes resultats.

La hiérarchisation du modèle, introduite grâce à l'hypothèse d'une loi commune pour les λ_i, peut être mise en évidence avec la représentation graphique du modèle sous forme de DAG (cf. chapitre 5).

La figure 6.2 montre clairement la différence entre le modèle "simple" relatif à un seul contrat et le modèle hiérarchique, caractérisé par la présence d'un niveau supplémentaire dans les calculs d'inférence. De cette manière l'inférence sur tous les λ_i est menée simultanément.

Dans cette figure, on a utilisé la technique de représentation des DAG du logiciel WinBUGS (cf. annexe C) : les calculs répétés de manière itérative (cycles `for ... next` régis par un indice k) sont visualisés à l'intérieur de rectangles (dits "*plateaux*") indexés par la variable k.

6.3.3 Résultats

Les résultats que nous présentons ici ont été obtenus en pratique à l'aide du logiciel WinBUGS (Spiegelhalter et al., 1996, 2000) sur la base des dernières 5 000 valeurs simulées avec l'algorithme de Gibbs sur un total de 10 000 itérations. Après vérification avec le diagnostic de convergence de Brooks et Gelman (1998), ces valeurs sont issues de la loi stationnaire de la chaîne. Plus de détails sur WinBUGS et le diagnostic de convergence se trouvent dans l'annexe C.

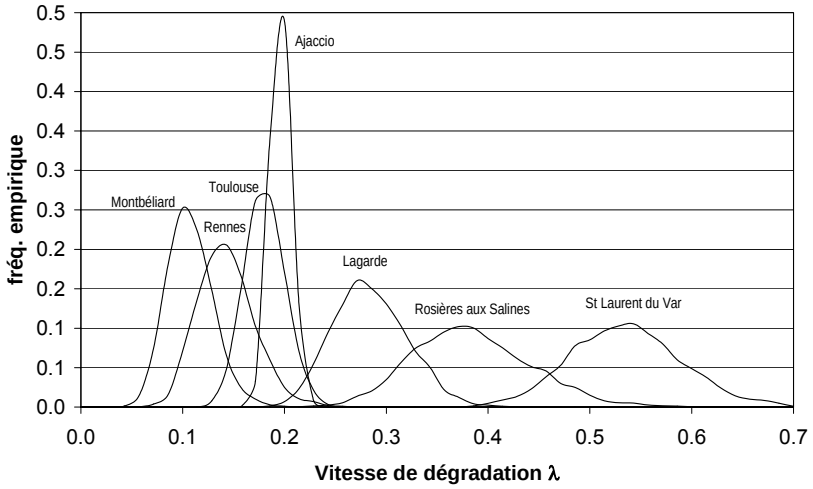


FIG. 6.3 – Visualisation des lois *a posteriori* des λ_i de quelques contrats.

Dans l'annexe A nous avons réporté le détail des résultats. On y trouve notamment les moyennes et les écarts type *a posteriori* des vitesses de dégradation λ_i et les intervalles de crédibilité à 95%. Les contrats examinés sont ceux pour lesquels au moins 25 étalonnages de compteurs de référence sont disponibles pour l'étude (78 contrats).

Une représentation graphique des lois *a posteriori* de quelques vitesses de dégradation est fournie en la figure 6.3. Ces courbes montrent bien la différence de comportement des sites d'exploitation (on retrouve notamment les différences entre Rennes, Ajaccio et Rosières aux Salines) mais aussi que l'incertitude sur les estimations des λ_i est très différente de contrat à contrat, l'écart type *a posteriori* dépendant principalement de la taille des échantillons examinés.

Pour examiner la pertinence de l'approche hiérarchique nous avons aussi estimé les vitesses de dégradation, une par une, pour tous les contrats.

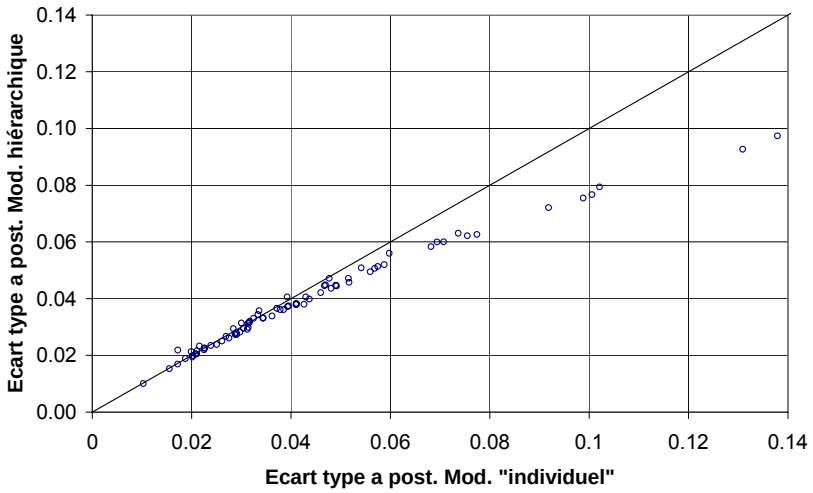


FIG. 6.4 – Ecarts type *a posteriori* des λ_i avec l'estimation hiérarchique et l'estimation individuelle.

L'hypothèse structurelle d'existence d'un lien entre les vitesses de dégradation entraîne, dans la quasi totalité des cas examinés, une réduction de l'écart type *a posteriori* de λ_i.

Si dans un graphe (figure 6.4) on représente sur l'axe des abscisses les écarts type pour l'estimation individuelle et sur l'axe des ordonnées ceux relatifs à l'estimation hiérarchique, de manière qu'à chaque point du plan corresponde un contrat, la plupart des points se disposent en dessous de la droite bissectrice du premier quadrant : c'est-à-dire que la valeur lue sur l'axe des ordonnées (modèle hiérarchique) est inférieure à celle en correspondance de l'axe des abscisses (modèle "individuel").

Cette réduction de l'écart type est plus marquée pour les contrats représentés par des échantillons de petite taille : pour moins de 45 unités la hiérarchisation du modèle entraîne une réduction de l'écart type d'environ 9% (en moyenne), alors que pour des échantillons de taille supérieure cette réduction est de l'ordre de 4%.

L'autre effet de cette hypothèse est de rapprocher le comportement des contrats. Le "*tassement*" vers la moyenne est un effet classique (*shrinkage effect*) de la modélisation hiérarchique (Gelman et al., 1995). Les écarts entre les moyennes *a posteriori* des λ_i ont tendance à se réduire, quand on passe de l'estimation individuelle à l'estimation simultanée, comme il est bien visible en la figure 6.5 :

le modèle hiérarchique donne des valeurs plus importantes des λ_i pour les exploitations à vitesse lente (en dessus de la bissectrice) et moins élevées pour les sites plus agressifs (en dessous de la bissectrice).

En définitif les résultats reflètent l'esprit de la modélisation hiérarchique : ce qu'on observe sur un individu ne nous renseigne pas seulement sur le comportement de l'individu mais aussi sur tous les autres et vice-versa.

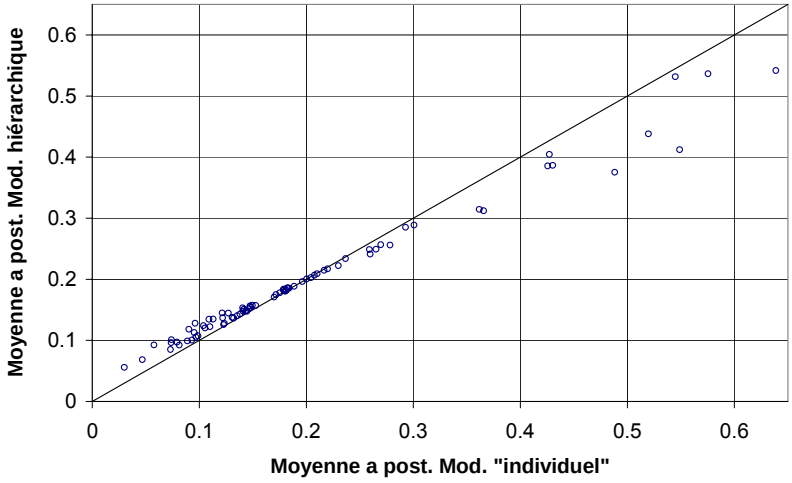


FIG. 6.5 – Moyennes *a posteriori* des λ_i avec l'estimation hiérarchique et l'estimation individuelle.

6.4 Un autre indicateur d'agressivité : le taux de blocage.

Comme il a été déjà souligné, les données métrologiques constituent la principale source d'information pour étudier le comportement des compteurs en service et aussi l'impact des conditions d'exploitation sur la dégradation de la qualité de la mesure. Néanmoins un autre indicateur d'agressivité d'un site pourrait être représenté par les taux de pannes des compteurs, notamment les taux de blocage, qui présentent des valeurs assez différentes d'un contrat à l'autre.

L'intérêt de l'analyse des taux de blocage repose sur les considérations suivantes :

- Puisque les essais métrologiques ne nous renseignent pas sur ces phénomènes, il s'agit de la seule manière d'estimer l'occurrence de l'état absorbant du modèle markovien de dégradation (chapitre 5).
- Il s'agit d'un paramètre intéressant pour la gestion des parcs de compteurs. Les compteurs bloqués étant évidemment remplacés systématiquement, il est important de savoir quelle partie du budget sera consacrée aux déposes inévitables et quelle partie aux déposes suggérées par les calculs technico-économiques.
- Si un lien existe entre la vitesse de la dégradation métrologique et le taux de blocage alors ce dernier, disponible "*gratuitement*" sur la majorité des exploitations, peut renseigner sur l'agressivité des contrats qui ne sont pas représentés dans la Base Métrologique.

6.4.1 La recherche de l'information

L'étude des taux de blocage est possible grâce à la disponibilité des informations relatives aux motifs de dépose des compteurs. Ces informations sont stockées dans l'outil de facturation et ensuite déversées périodiquement dans la base "*Branchement et Compteurs*" ICBC. On rappelle (chapitre 4) que cette base de données est formée par plusieurs tables dont celles intéressantes pour l'étude des blocages sont les tables "*Abonnés*", "*Compteurs*" et "*Changement des compteurs*".

En pratique, un compteur est jugé bloqué si, au moment du relevé, son index n'a pas bougé significativement par rapport au dernier index connu. Dans ce cas, le releveur constate le blocage et demande sa dépose. L'appareil est remplacé dans un bref délai et le blocage n'est pas normalement vérifié, ensuite, sur banc d'essai.

Quand le compteur est déposé et remplacé par un nouvel appareil de mesure, son *état de pose* passe de *posé* à *déposé* et deux nouvelles lignes apparaissent dans la table "*Changement des Compteurs*" : une opération de pose et une opération de dépose.

Les indications concernant le compteur déposé sont contenues dans la ligne relative à la dépose mais, malheureusement, au moment du déversement des données de la base de facturation à la base ICBC, une partie des opérations de déposes est perdue, alors que les informations les plus importantes sont effectivement celles concernant l'appareil remplacé (et pas le remplaçant).

Pour les récupérer on interroge la table "*Compteurs*" dans laquelle on rentre avec le numéro du branchement et à la condition que le compteur qu'on cherche soit déjà déposé (figure 6.6).

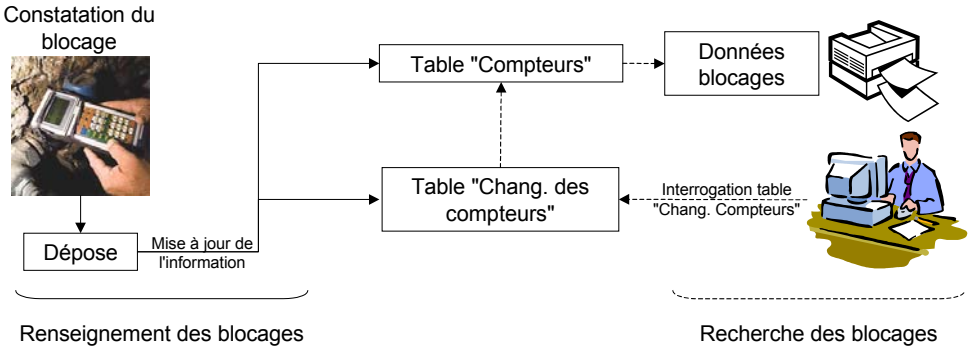


FIG. 6.6 – Extraction des données de blocage de la base ICBC.

Cette méthode d'extraction des données permet de récupérer le maximum d'informations concernant les déposes des compteurs à l'intérieur du périmètre de la base ICBC. Actuellement ce périmètre est constitué par 8 Régions sur 11 (toutes sauf Banlieue de Paris, Sud et Sud-Est), à l'exclusion, pour la Région Ile de France, de la ville de Paris.

L'analyse réalisée porte, en particulier, sur les contrats avec plus de 5000 compteurs. Ce choix, qui limite le périmètre d'étude, est nécessaire parce que les valeurs des taux de blocage sont, en général, assez faibles (de l'ordre de 0.5 à 1.5% par an) et, pour avoir des estimations fiables il faut que ces taux soient calculés sur la base d'une population de taille importante.

Ces contrats constituent un échantillon important parce qu'ils représentent plus de la moitié de la base ICBC avec 2.7 millions de compteurs, soit 45% de tous les parcs français.

La figure 6.7 montre la répartition par taille des contrats ICBC et le nombre de compteurs correspondants. On observe qu'un peu plus de la moitié des compteurs se trouvent sur des contrats de taille comprise entre 1000 et 10 000, et 16% sur les 14 contrats avec plus de 30 000 compteurs.

6.4.2 Quelles données de blocage examiner ?

Les blocages des compteurs en service sont, en général assez rares et les taux de blocage ont aujourd'hui baissé de manière sensible par rapport au passé. La

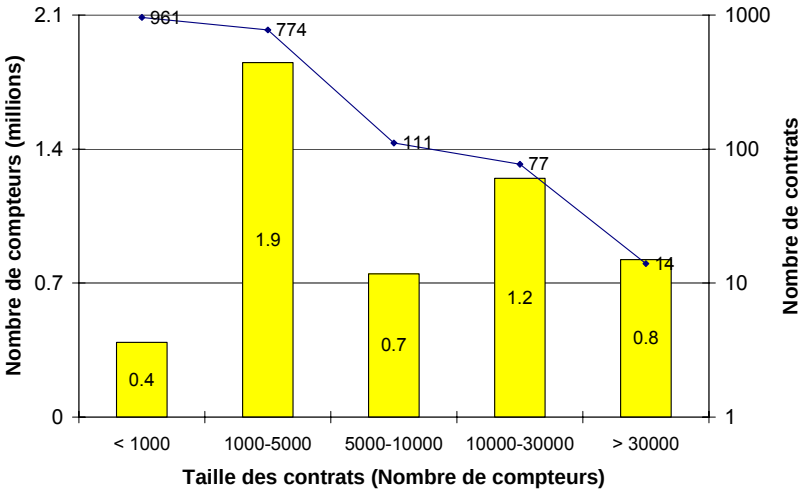


FIG. 6.7 – Répartition du nombre de contrats et des compteurs selon la taille des contrats.

principale cause de ce phénomène est l'amélioration technique des compteurs. A ce propos, une étape fondamentale a été marquée par l'introduction, dans les années 70, de l'entraînement magnétique (chapitre 2). L'isolement parfait du totalisateur de la matrice liquide, obtenu à moindre coût en termes de perte de sensibilité, a pour effet d'éviter les blocages du totalisateur causés par l'incrustation de particules solides dans les engrenages (en principe un simple grain de sable pourrait suffire). La meilleure qualité de ces compteurs est universellement reconnue et a motivé d'importantes campagnes de remplacement d'appareils avec des mécanismes de transmission moins performants, secs ou humides (Orr et al., 1977).

Les études et les essais réalisés par les distributeurs d'eau, de plus en plus sensibles au problème de la performance de leurs parcs de compteurs, ont en outre poussé les fabriquants à améliorer constamment leur produits en termes d'exactitude mais aussi de résistance aux agressions externes (passage de particules solides, surpressions, chocs).

Le type de compteur est donc un facteur explicatif important des blocages des compteurs et, si on veut isoler l'effet du site d'exploitation, il faut examiner uniquement un groupe de compteurs qui présente une certaine homogénéité technologique. En particulier, on s'intéresse aux quatre types de compteurs "de référence" qui ont déjà été examinés pour l'estimation de la vitesse de dégrada-

tion métrologique.

Un autre facteur explicatif, parfois controversé est le rôle de l'âge. Dans quelques études disponibles (Newman et Noss, 1982) le taux de blocage est imaginé constant en fonction de l'âge et telle est l'opinion de certains experts. La difficulté d'isoler ce facteur de l'effet du site d'exploitation et du type de compteur est accrue par la rareté du phénomène, de manière que, si on découpe trop les données, les taux très proches de 0 calculés sur la base de l'observation de quelques unités ne permettent pas d'obtenir des conclusions significatives.

L'autre facteur explicatif, que nous avons déjà implicitement évoqué est évidemment le site d'exploitation. On imagine que les blocages ne soient pas uniformément répartis sur le territoire nationale et qu'un taux de blocage plus ou moins important puisse nous renseigner sur l'agressivité de l'exploitation.

Afin d'améliorer la qualité des résultats, on tient compte, dans l'analyse des déposes, des interventions dont la cause n'est pas répertoriée en divisant chaque taux de blocage par la proportion de déposes bien renseignées dans le contrat. Cette démarche suppose qu'à l'intérieur d'un contrat le taux de non-renseignement soit uniforme et que la répartition des motifs de dépose parmi les interventions codifiées soit la même que celle parmi les non codifiées.

6.5 Trois groupes de contrats à agressivités différentes

Sur la base des estimations des vitesses de dégradation métrologique on peut réaliser un premier découpage des contrats examinés en 3 groupes. Cette partition est réalisée dans le but d'isoler des groupes qui sont sensiblement différents les uns des autres en termes de comportement métrologique et de taux de blocage.

En particulier on propose d'utiliser comme estimateur $\hat{\lambda}_i$ de la vitesse de dégradation métrologique la moyenne *a posteriori* (annexe A) et de définir ainsi trois groupes de la manière suivante :

$$\begin{array}{ll} \text{Groupe } A_1 & \hat{\lambda}_i \leq 0.145 \\ \text{Groupe } A_2 & \hat{\lambda}_i \in]0.145, 0.21[\\ \text{Groupe } A_3 & \hat{\lambda}_i \geq 0.21 \end{array}$$

Avec ce choix, des compteurs d'âge compris entre 5 et 9 ans (environ à la moitié de leur vie opérationnelle) ont une probabilité d'être très performants

supérieure à 3/4 s'ils se trouvent dans un environnement peu agressif, comprise entre 3/4 et 2/3 dans un site moyennement agressif et inférieure à 2/3 si exploités dans des conditions particulièrement agressives (figure 6.8).

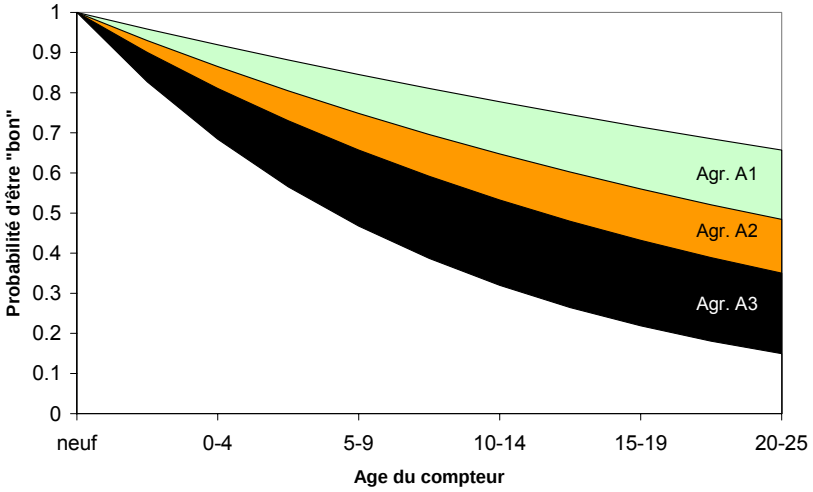


FIG. 6.8 – *Faisceaux* contenant les courbes de dégradation des compteurs en fonction de l'agressivité du contrat.

L'analyse des taux de blocage observés dans les trois groupes ainsi obtenus, montre une certaine concordance entre l'agressivité métrologique et les blocages : les contrats où les compteurs se dégradent plus vite sont aussi ceux où les pannes sont plus fréquentes (figure 6.9).

Parmi les contrats examinés du point de vue métrologique, il y en a 36 pour lesquels on a aussi les données de blocage, dont 16 appartenant au groupe A_1, 16 au groupe A_2 et 4 au groupe A_3.

La figure 6.9 montre les taux de blocage pour les 3 groupes de contrats des 4 types de compteurs de référence. Il s'agit des taux moyens[1] calculés sur les 3 années 2000, 2001 et 2002, en tenant compte du taux de non-renseignement des causes de déposes, selon la modalité décrite précédemment :

$$x_{i,j} = \frac{1}{3}\frac{s_{i,j}}{m_{i,j}}\frac{1}{1-u_i} \tag{6.12}$$

où $s_{i,j}$ est le nombre de blocages de compteurs de type j observés sur le contrat i dans les 3 années, $m_{i,j}$ est la taille de la population de compteurs de

[1] Taux de blocage sur la période des 3 années d'observation, divisé par 3.

type j présents dans le parc du contrat i et u_i est le taux de déposes, dans le contrat i, dont la cause n'est pas spécifiée. Sont exclus de l'analyse tous les contrats présentant sur une des trois années, un taux de non-renseignement des motifs de déposes supérieur à 50%.

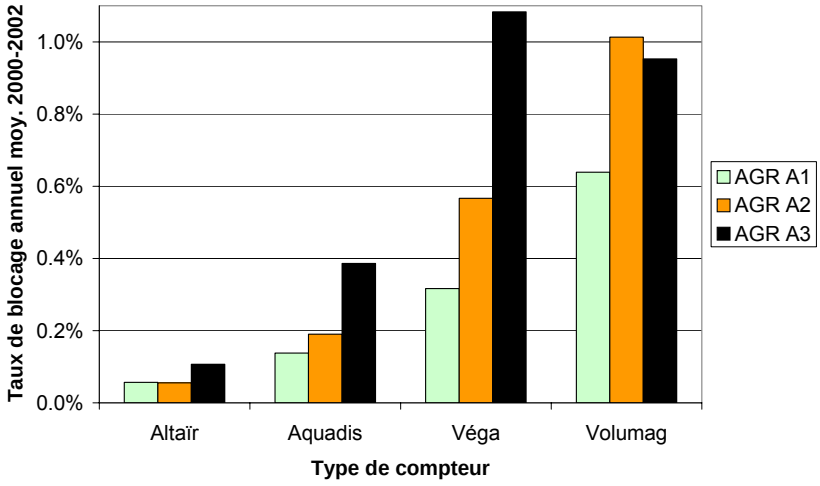


FIG. 6.9 – Taux de blocage annuels dans la période 2000-2002 en fonction de l'agressivité et du type de compteur.

L'analyse (figure 6.9) montre aussi que les blocages sont plus fréquents pour les Véga et le Volumag, par rapport aux Altaïr et aux Aquadis. Cela peut être dû aux différences entre les typologies de compteurs. L'Altaïr et l'Aquadis sont des appareils plus récents et la meilleure robustesse, en particulier de l'Altaïr, est confirmée par des essais d'usure accélérée réalisés en faisant circuler dans les appareils une eau expressément chargée de particules solides.

La figure 6.10 montre les mêmes taux de blocage moyens par classes d'âge et confirme encore la cohérence des deux indicateurs. Cette figure montre aussi que le taux de blocage obtenu en mélangeant les 4 types de compteurs est fonction croissante de l'âge, mais cette dépendance peut être dûe au fait que les jeunes compteurs sont essentiellement des Altaïr et des Aquadis et les plus âgés sont des Véga et des Volumag.

Une analyse plus fine avec prise en compte des trois facteurs : groupe d'agressivité, type, âge, n'est pas possible. Le découpage des données selon plusieurs facteurs explicatifs donne lieu à des échantillons de tailles trop faibles et les proportions calculées n'ont plus aucun sens.

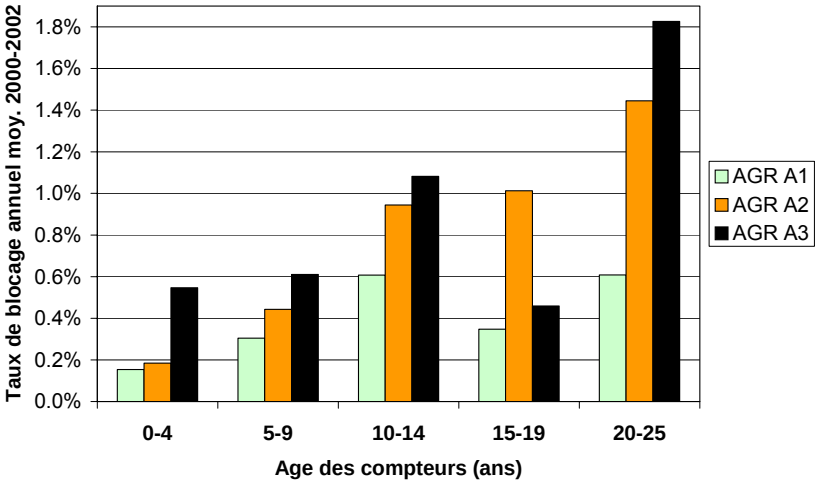


FIG. 6.10 – Taux de blocage annuels dans la période 2000-2002 en fonction de l'agressivité et de l'âge.

6.6 Attribution de l'agressivité à partir des taux de blocage

6.6.1 L'idée à la base de la méthode

Quand il n'y a pas assez d'informations dans la Base Métrologique pour caractériser l'agressivité d'un site d'exploitation, on utilise les taux de blocages pour assigner un contrat donné à un des 3 groupes précédemment identifiés. L'idée à la base de la méthode est que le comportement de chaque contrat peut être décrit avec un certain nombre de paramètres de blocage qui sont, en l'occurrence, les probabilités de blocage des 4 types de compteurs "de référence", plus le taux global de l'ensemble de ces compteurs. A chaque exploitation i on associe donc un point $\mathtt{C}_i$ dans l'espace $\mathbb{R}^5$ dont les coordonnées sont, dans l'ordre, les taux de blocages des Altaïr, Aquadis, Véga, Volumag et des 4 types confondus.

De même, les 3 groupes d'agressivité découpés précédemment, sont décrits par les valeurs de ces 5 paramètres, et donc par 3 points ($\mathtt{A}_1$, $\mathtt{A}_2$ et $\mathtt{A}_3$) de $\mathbb{R}^5$. L'attribution d'un contrat i à un groupe pourrait être faite alors sur la base de la comparaison des distances (euclidiennes) entre le point $\mathtt{C}_i$ et les points $\mathtt{A}_1$, $\mathtt{A}_2$ et $\mathtt{A}_3$, en assignant à chaque contrat l'agressivité du groupe le plus *proche* (figure 6.11).

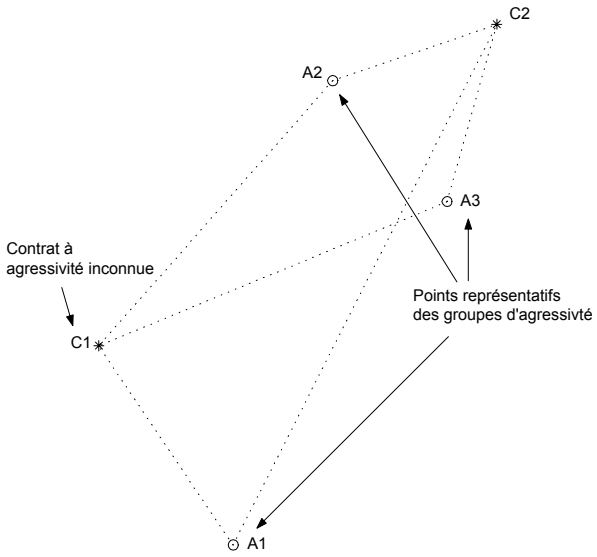


FIG. 6.11 – Assignation des contrats aux 3 groupes sur la base des taux de blocage.

Le défaut de cette méthode est d'utiliser la proportion "brute" de compteurs bloqués, par rapport à la population observée, pour décrire de manière déterministe le comportement du contrat. En réalité, les probabilités réelles de blocage sont connues avec incertitude et donc à chaque contrat on devrait plutôt associer un nuage de points, à la place d'un seul point, chacun d'entre eux ayant une certaine probabilité. La prise en compte de l'incertitude sur la détermination des probabilités de blocage permet, en outre, de donner un jugement probabiliste sur l'assignation d'un contrat à un groupe donné. Le calcul d'incertitude peut être réalisé de manière très simple dans un cadre bayésien, comme il sera exposé dans le paragraphe suivant.

6.6.2 Du taux de blocage à la probabilité de blocage

Si note τ la probabilité qu'un compteur se bloque en une période donnée, le nombre y_b de blocages observés dans ce laps de temps est une réalisation de la variable aléatoire binomiale Y_b de paramètres τ et m, étant m la taille de la population observée :

$$Y_b \sim \mathcal{B}in\,(\tau, m) \tag{6.13}$$

Si l'on choisit comme loi *a priori* pour τ des lois *Bêta* de paramètres α et β, alors on sait déjà (cf. chapitre 5) que la loi *a posteriori* $[\tau|y_b]$ sera encore une *Bêta* de paramètres $\alpha + y_b$ et $\beta + m - y_b$:

$$\tau \sim \mathcal{B}(\alpha + y_b, \beta + m - y_b) \tag{6.14}$$

Sur la base de ces simples considérations il est facile de déterminer les lois *a posteriori* des probabilités de blocages $\tau_{i,j}$ des compteurs examinés. Dans la notation $\tau_{i,j}$ l'indice i dénote le contrat (ou le groupe) et l'indice j le type de compteur (1-Altaïr, 2-Aquadis, 3-Véga, 4-Volumag, 5-Ensemble des 4 modèles de référence). La figure 6.12 montre les distributions *a posteriori* des probabilités de blocage des 3 groupes d'agressivité, obtenues en utilisant des lois *a priori* uniformes sur l'intervalle [0,1] ($\alpha = \beta = 1$).

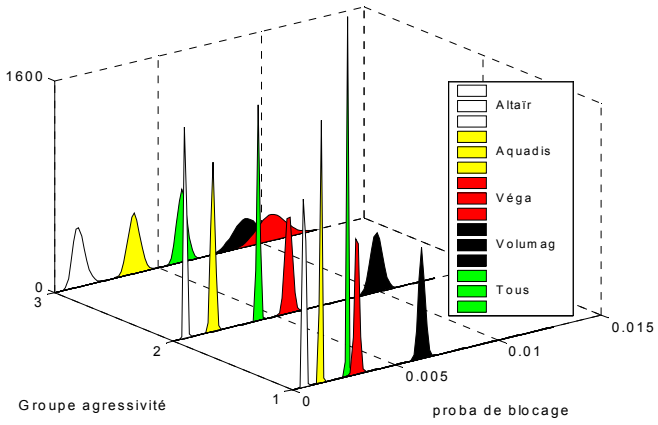


FIG. 6.12 – Distributions de probabilités *a posteriori* des taux de blocages.

Le même calcul, réalisé pour les contrats desquels on veut déterminer l'agressivité, permet de définir les distributions de probabilités des coordonnées $\tau_{i,j}$ des points ($\mathtt{A}_1$, $\mathtt{A}_2$ et $\mathtt{A}_3$)qui les représentent dans l'espace $\mathbb{R}^5$.

6.6.3 Détermination du groupe d'agressivité pour chaque contrat

Une fois déterminées les lois *a posteriori* des $\tau_{i,j}$ pour les 3 groupes et pour les contrats à examiner, la détermination de l'agressivité est réalisée de la manière suivante :

- Tirage aléatoire pour un contrat donné i d'une valeur de chacun des paramètres $\tau_{i,1},\ \tau_{i,2},\ \tau_{i,3},\ \tau_{i,4},\ \tau_{i,5}$ des distributions de probabilité respectives *a posteriori*, exprimées par la formule (6.14). Ces valeurs déterminent un point $\mathtt{C}_i$ de $\mathbb{R}^5$.
- Tirage aléatoire pour les trois groupes d'une valeur des paramètres $\tau_{A_1,j}$, $\tau_{A_2,j},\ \tau_{A_3,j}$ avec j=1 à 5. Ces 15 valeurs déterminent 3 points de $\mathbb{R}^5$: $\mathtt{A}_1$, $\mathtt{A}_2$, $\mathtt{A}_3$.
- Calcul des 3 distances du point $\mathtt{C}_i$ des 3 points $\mathtt{A}_1$, $\mathtt{A}_2$, $\mathtt{A}_3$:

$$\delta\left(\mathtt{C}_i, \mathtt{A}_1\right) = \sqrt{\sum_{j=1}^{5}\left(\tau_{i,j} - \tau_{A_1,j}\right)^2} \tag{6.15}$$

$$\delta\left(\mathtt{C}_i, \mathtt{A}_2\right) = \sqrt{\sum_{j=1}^{5}\left(\tau_{i,j} - \tau_{A_2,j}\right)^2} \tag{6.16}$$

$$\delta\left(\mathtt{C}_i, \mathtt{A}_3\right) = \sqrt{\sum_{j=1}^{5}\left(\tau_{i,j} - \tau_{A_3,j}\right)^2} \tag{6.17}$$

- Attribution du contrat i au groupe d'agressivité, dont le point représentatif est le plus proche du point $\mathtt{C}_i$.

On répète plusieurs fois ces opérations et finalement le groupe d'appartenance le plus probable est celui qui a été le plus proche, dans la plupart des itérations. Le nombre de fois où chaque groupe a été choisi par l'algorithme, divisé par le nombre d'itérations réalisées, donne aussi une estimation de la probabilité d'appartenance du contrat à chaque groupe. Les résultats des calculs se trouvent dans l'annexe A.

6.7 Résultats et commentaires

La méthode proposée, basée sur l'analyse des taux de blocage, permet de définir le groupe d'agressivité de 137 contrats ultérieurs : on arrive donc à un total de 215 contrats. Parmi eux, 199 se trouvent dans le périmètre ICBC et la somme de leurs effectifs (environ 2.6 millions) représente 52% des parcs compteurs d'ICBC.

64% des contrats examinés appartient au groupe d'agressivité A_1, 19% au groupe A_2 et 17% au groupe A_3. Si on pondère le nombre de contrats par leurs tailles, on peut estimer que 73% des compteurs se trouvent dans des zones à

faible agressivité, 18% à agressivité moyenne et à peine 9% en exploitations particulièrement agressives. Cette conclusion doit être prise avec précaution puisque l'analyse porte essentiellement sur les gros contrats.

D'autre part, les résultats montrent qu'un lien semble exister entre la taille du contrat et son agressivité. La figure 6.13 montre la répartition par groupes d'agressivité des contrats (202 sur 215 dont on connaît l'effectif de compteurs), regroupés par taille *:* 80% des contrats avec plus de 30 000 compteurs sont peu agressifs contre 45% des contrats de taille inférieure à 10 000 unités.

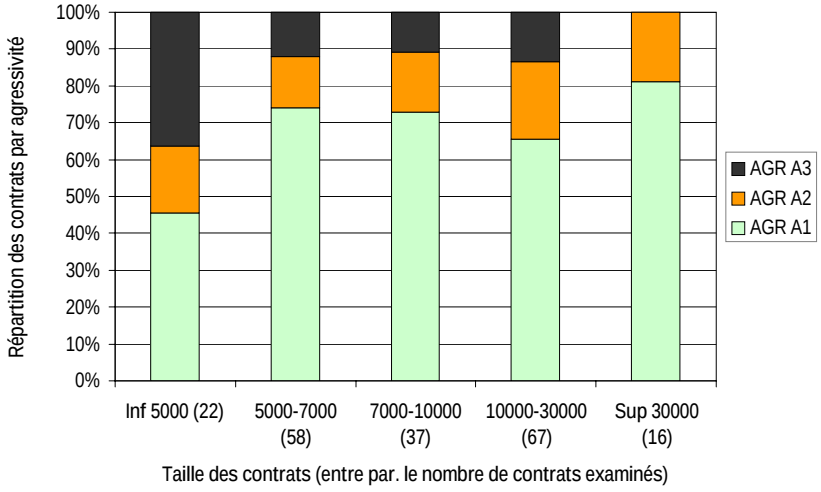


FIG. 6.13 – Répartition par agressivité des contrats examinés en fonction de leur taille.

Une autre analyse intéressante peut être menée sur la base d'une classification des réseaux de distribution typiquement utilisée par les gestionnaires. Le regroupement se base sur la valeur d'un paramètre technique dit "*Indice Linéaire de Consommation*" (ILC), défini de la manière suivante :

$$\text{ILC} = \frac{V}{L} \tag{6.18}$$

où V est le volume d'eau facturé journalier moyen (volume annuel *compté* divisé par 365) et L est la longueur du réseau de distribution. V est normalement exprimé en m^3/jour et L en km et donc ILC est mesuré en m^3/jour·km.

Sur la base des valeurs de l'ILC les réseaux sont usuellement classés en trois catégories selon le schéma suivant :

ILC$\leq$ 10	Rural
10 <ILC< 30	Semi-Rural
ILC$\geq$ 30	Urbain

Cette classification est basée sur la considération qu'en zone rurale les usagers sont plus éparpillés et donc, à nombre égal de branchements, les réseaux, devant couvrir un territoire plus vaste, ont des longueurs plus importantes qu'en zone urbaine.

Les critères de classification indiqués dans le tableau en fonction de la valeur de l'ILC sont ceux suggérés par la pratique technique et conventionnellement utilisés par les professionnels de l'eau. Sur les contrats examinés, il y en a 195 donc l'ILC est connu.

L'analyse (figure 6.14) montre que les contrats *ruraux* sont généralement plus agressifs que les contrats *semi-ruraux* et *urbains* : 70% des contrats *urbains* sont peu agressifs, alors que pour les contrats *ruraux* cette proportion descend à 57%. De même, parmi les contrats *ruraux* la proportion de sites très agressifs est de 22% contre 12 à 13% pour les *semi-ruraux* et les *urbains*.

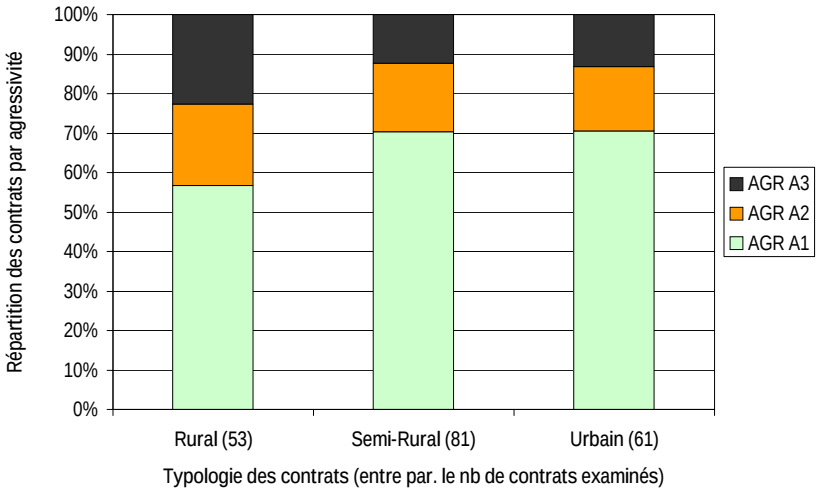


FIG. 6.14 – Répartition par agressivité des contrats examinés en fonction de leur typologie (*ILC*).

Cette différence de comportement pourrait être due au fait que l'eau qui transite dans un réseau plus étendu a plus de chances de se charger de particules

solides telles que sables ou microsables et paillettes métalliques qui peuvent se détacher des conduites et qui ne sont pas (évidemment) présentes dans l'eau mise en circulation "à la sortie d'usine". En d'autre termes, ce n'est pas tellement la composition chimique des eaux qui déterminerait l'agressivité d'un site vis-à-vis des compteurs mais plutôt ce qui se passe entre la mise en circulation et le compteur.

D'ailleurs, parmi les deux facteurs *eau* et *réseau* l'effet du premier est mieux maîtrisé puisque les paramètres physico-chimiques de l'eau mise en circulation doivent être compris entre les limites fixées par la réglementation en vigueur (Décret 2001-1220) et les matériaux utilisés par les fabriquants de compteurs sont en général compatibles avec les typologies d'eau les plus communes.

Néanmoins, si on analyse la composition de l'eau des contrats considérés, on trouve une relation intéressante entre la dureté, exprimée par le *titre hydrotimétrique* (TH), et l'agressivité. La figure (6.15) montre que quand le TH sort de l'intervalle usuel, compris entre 5 et 35 °F, la probabilité qu'un contrat soit très agressif double, passant de 14% à 27%. Ces données ont été obtenues sur la base de la moyenne annuelle des résultats des contrôles de qualité (CGE-DT, 2003) et sont relatives à l'année 2002. Le nombre de contrats pour lesquels ces informations sont disponibles est de 174.

Même si le nombre de contrats dans les deux regroupements est nettement différent, les différences entre leurs répartitions par agressivité est évidente. Du point de vue physique, en outre, cette différence s'explique par la considération que des eaux très dures (TH $>$ 35 °F) ou très douces (TH $<$ 5°F) favorisent la formation de dépôts.

Une dernière analyse des résultats a été menée sur base régionale (pour les 206 contrats pour lesquels la région est connue). Les résultats ne sont pas concluants, compte tenu du fait que le faible nombre de sites examinés dans chaque région ne peut pas (spécialement pour les régions Sud et Sud-Est) être considéré comme représentatif de l'ensemble des contrats régionaux. En outre, compte tenu de leur extension et des différentes conditions d'exploitation, les régions sont des unités territoriales forcement non-homogènes (eaux et réseaux très différents) et l'extrapolation d'une sorte d'agressivité locale à si large échelle est un exercice très délicat.

Néanmoins, les résultats (figure 6.16) sont en accord avec les opinions répandues parmi les experts du comptage. On retrouve notamment le meilleur

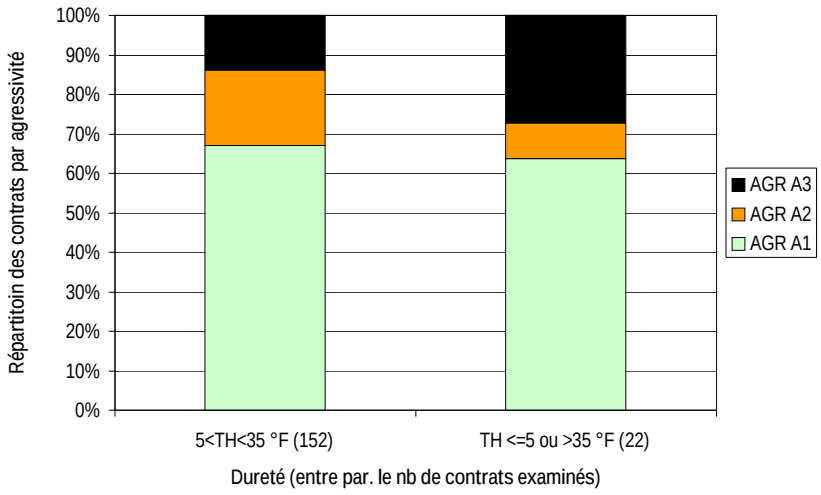


FIG. 6.15 – Répartition par agressivité des contrats examinés en fonction de la dureté de l'eau distribuée.

comportement général des compteurs de la région Centre-Est par rapport à ceux de provenance de régions, normalement réputées "à problème" comme Sud-Est et Est.

Une analyse plus pertinente peut être faite en pondérant les nombres de contrats par leurs effectifs (en terme de nombre de compteurs). Dans la figure 6.17 on montre la répartition, sur base régionale, des compteurs par agressivité. Pour chaque région on spécifie la proportion du parc régional examiné (valeurs proches de 50% pour la plupart des régions du périmètre ICBC). Ce graphe confirme les tendances exprimées par la figure 6.16, même si les différences entre les régions se font moins marquées. On observe aussi que la région Sud-Ouest perd des positions par rapport à l'analyse précédente à cause notamment du fait que quelques gros contrats (Toulouse, Ville d'Agen, Sud d'Agen) ne se trouvent pas parmi les "peu agressifs".

6.8 Une nouvelle variable explicative : la consommation

Un autre facteur explicatif de la dégradation, souvent mis en cause par les ingénieurs, est le volume total enregistré, affiché par l'index du compteur. A vrai

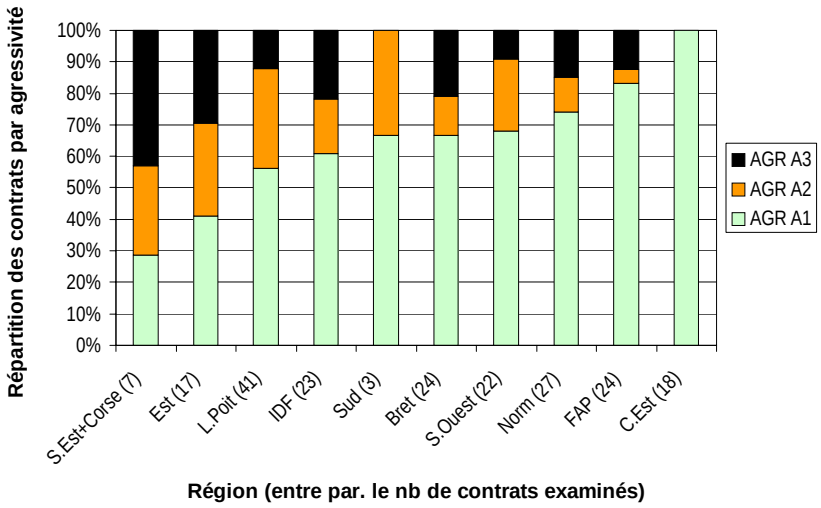


FIG. 6.16 – Répartition par agressivité des contrats examinés selon la région de provenance.

dire l'importance de cette variable dans le processus de dégradation est parfois niée. Sisco (1967) estime que, par rapport à l'âge, le volume mesuré a un effet significatif uniquement pour les compteurs réparés. Dans son étude déjà présentée dans le chapitre 4, Tao (1982) insiste surtout sur la corrélation entre index et âge et se sert de l'index pour calculer "indirectement" l'âge de compteurs non suffisamment renseignés (sur la base d'une consommation annuelle moyenne), laissant entendre que l'une des deux variables est redondante, étant donnée l'autre.

En revanche, Tort et al. (1988) soulignent l'importance de la prise en compte de ce facteur dans la planification des étalonnages. C'est aussi l'avis courant de beaucoup de praticiens, même si, à ce jour, aucune étude où les deux facteurs sont explicitement pris en compte n'est disponible.

Dans ce travail, compte tenu de la dépendance forte entre index et âge, on a préféré prendre en compte ce dernier avec l'introduction d'une nouvelle variable explicative, qu'on a appelé "*Consommation moyenne*". Il s'agit tout simplement du rapport entre le dernier index du compteur (en pratique l'index de dépose) et l'âge du dispositif.

Ce rapport est donc une estimation de la consommation annuelle moyenne enregistré par le compteur le long de sa vie opérationnelle.

Un jugement *a priori* concernant le rôle de cette variable dans le processus

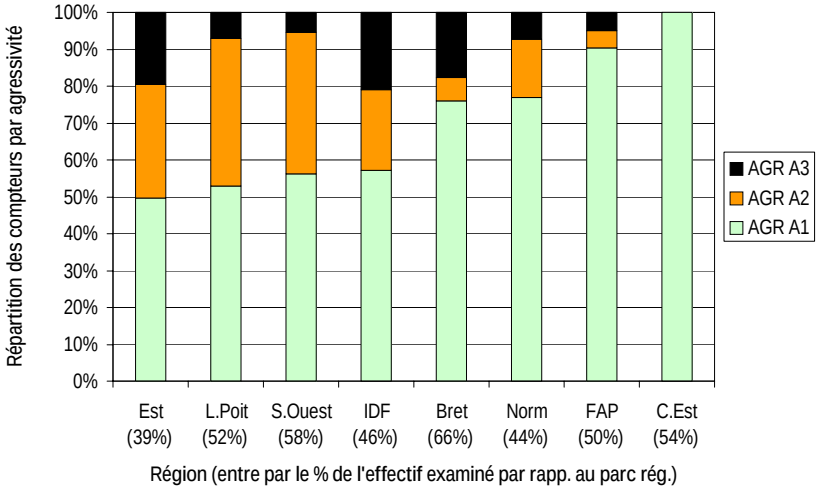


FIG. 6.17 – Répartition par agressivité (pondérée sur les effectifs) des contrats examinés selon la région de provenance.

de dégradation est difficile. D'une part, il est évident qu'un compteur qui tourne plus vite s'use aussi plus vite, comme le montrent les essais d'endurance réalisés par les fabricants ; le problème est de savoir à partir de quelle consommation le phénomène se manifeste. D'autre part, les arrêts et démarrages fréquents qu'on observe sur des branchements à faible consommation ne sont pas bénéfiques pour le dispositif de mesure et, en plus, les périodes prolongées d'arrêt, provoquant la stagnation de l'eau dans le compteur, favorisent la formation de dépôts.

Une réponse à cette question a été fournie indirectement par des études sur l'évolution des consommations domestiques, mises en place récemment par la CGE. Ces analyses ont pour but la comparaison des consommations, pour les mêmes abonnés, avant et après le changement du compteur. Il s'agit d'éléments importants pour un distributeur d'eau qui peut évaluer l'effet d'une stratégie de renouvellement mais aussi, en principe, valider les hypothèses métrologiques à la base de cette stratégie.

Les études sont encore en cours, mais les premiers résultats obtenus montrent pour les abonnés domestiques (en grande partie équipés de compteurs avec un DN de 15 mm) que le gain en volume mesuré est, en pourcentage, nettement plus important quand les consommations annuelles sont supérieures à 200 m^3/an.

Effectivement un retour aux données métrologiques confirme cette considération. La figure 6.18 montre que la proportion de "bons" compteurs parmi les

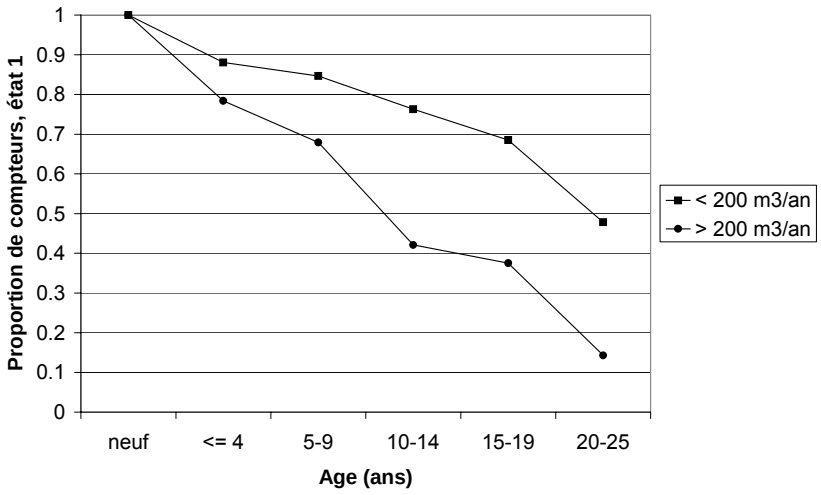


FIG. 6.18 – Proportion de "bons" compteurs dans les échantillons de la BMN en fonction de l'âge et de la consommation (DN 15 mm).

échantillons de même âge est fonction de la consommation moyenne (index/âge) et que les appareils enregistrant des consommations supérieures à 200 m^3/an se dégradent sensiblement plus vite. Les données sont relatives aux compteurs volumétriques de référence, dans le sens déjà précisé au début de ce chapitre.

Un résultat similaire concerne les compteurs de DN 20 mm, si on prend comme seuil pour la consommation la valeur de 500 m^3/an (figure 6.19). Pour les compteurs de diamètre supérieur le choix de cette valeur est plus compliqué, puisque les effectifs dans la base de données métrologiques deviennent de taille assez petite.

En définitive on prend en compte l'effet du volume total enregistré avec une variable discrète qui indique si la consommation annuelle moyenne dépasse ou non une valeur limite qui dépend du calibre du compteur selon le schéma suivant :

DN (mm)	Valeur limite de la consommation
15	200 m^3/an
20	500 m^3/an
30	2000 m^3/an
40	4000 m^3/an

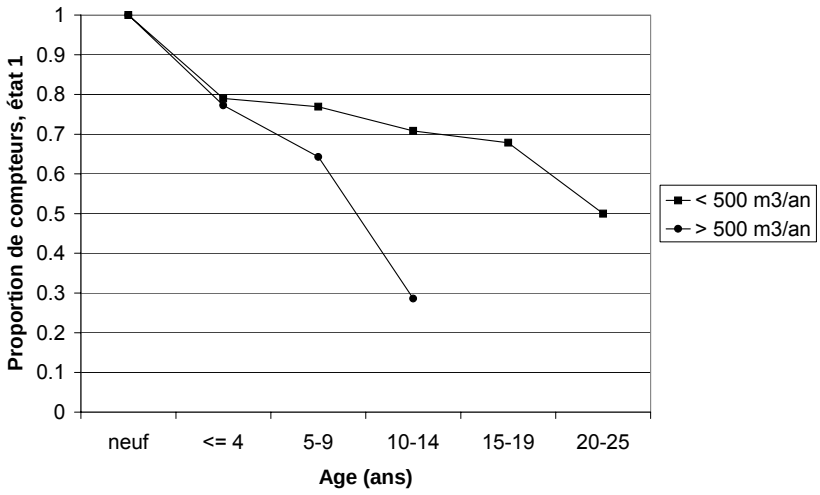


FIG. 6.19 – Proportion de "bons" compteurs dans les échantillons de la BMN en fonction de l'âge et de la consommation (DN 20 mm).

Ces valeurs de seuil se situent, selon les diamètres, autour des percentiles 0.8 à 0.9 des consommations annuelles pour les différents diamètres, comme le montrent les figures 6.20 et 6.21. Ces courbes de fréquence empirique ont été établies sur la base des dernières consommations annuelles d'environ 3.5 millions d'usagers.

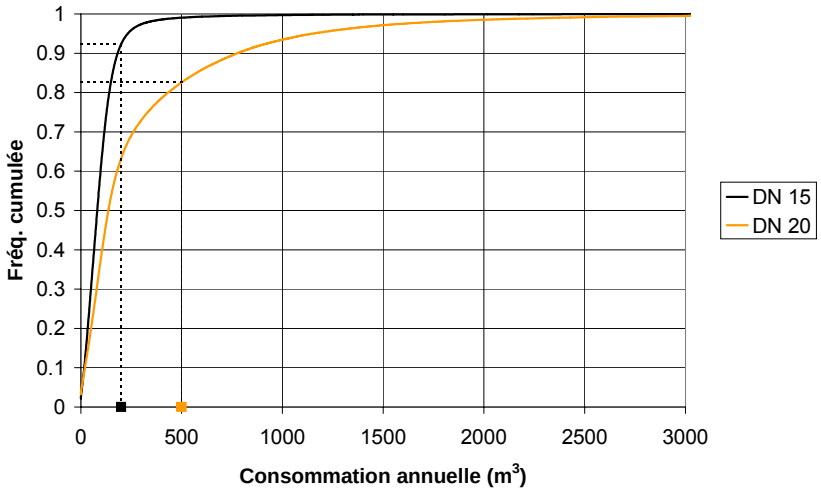


FIG. 6.20 – Fonctions de répartition empirique des consommations annuelles enregistrées par des compteurs de DN 15 et 20 mm.

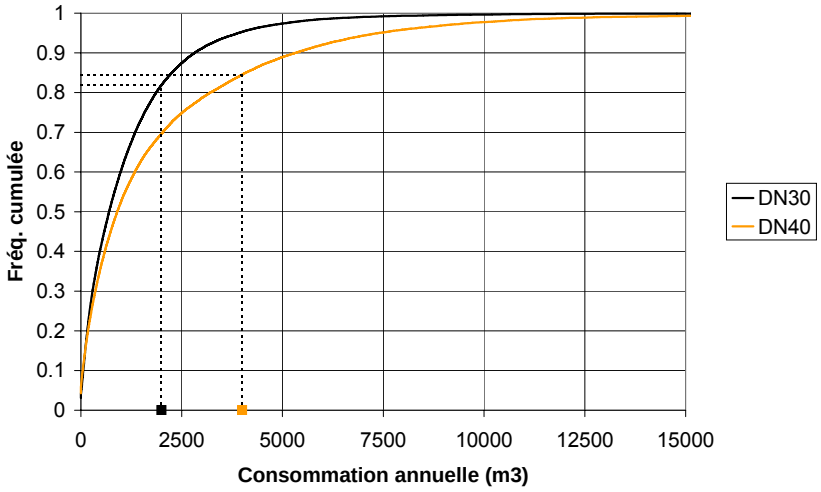


FIG. 6.21 – Fonctions de répartition empirique des consommations annuelles enregistrées par des compteurs de DN 30 et 40 mm.

Chapitre 7

De la théorie à l'aide à la décision

7.1 Utilisation pratique du modèle : établir des lois de vieillissement

Du point de vue du distributeur d'eau le but principal des études métrologiques est l'amélioration des hypothèses à la base des stratégies de gestion des parcs de compteurs. Dans cet esprit le modèle statistique développé fournit des éléments d'aide à la décision. Parmi les différentes réponses qu'on peut obtenir à l'aide du modèle proposé, la prédiction du rendement en fonction de l'âge des compteurs "survivants" (*loi de vieillissement*) est actuellement la plus importante du point de vue opérationnel. Les prochains paragraphes montreront un exemple de calculs de prévision, relatifs à un certain type de compteur volumétrique de DN 15 mm, en fonction de tous les facteurs explicatifs connus.

Dans le chapitre précédent nous avons montré que l'agressivité du site d'exploitation et la consommation annuelle sont des facteurs explicatifs de la dégradation des compteurs. Leur prise en compte peut se faire par la définition de 2 variables discrètes : la première exprimant une mesure globale de l'agressivité du contrat d'exploitation (1-peu agressif, 2-moyennement agressif, 3-très agressif), la seconde caractérisant l'importance de la consommation (*ordinaire* ou *élevée* selon que la valeur du rapport entre index et âge dépasse ou non une valeur de seuil).

Du point de vue de la modélisation statistique, l'introduction de ces nouvelles

variables peut être faite de plusieurs façons. La manière la plus simple est de décomposer l'échantillon selon les 6 différentes combinaisons possibles des valeurs des 2 variables. Le premier inconvénient est que, compte tenu surtout du fait que les consommations élevées sont relativement rares (les valeurs limites sont dépassées par 10 à 20% des compteurs uniquement), ce découpage des données métrologiques est excessif et donne lieu à des échantillons de taille trop petite. Le second inconvénient est que le mécanisme de dégradation partage des traits communs entre les 6 groupes qu'il faudrait respecter.

Ainsi, l'analyse préliminaire des données montre clairement que des consommations élevées accélèrent le vieillissement et, en particulier, (figures 6.18 et 6.19) favorisent le passage de l'état 1 de métrologie excellente aux états de fonctionnement dégradé. Cette réflexion est à la base de la technique de modélisation employée (comme il sera exposé dans la suite) qui prévoit l'introduction d'un paramètre "*perturbateur*" qui modifie la matrice de transition du modèle markovien quand la consommation est importante.

Concernant le nombre d'états à prendre en compte, nous ferons appel à un modèle à 3 états et donc nous utiliserons uniquement les données métrologiques. La raison de ce choix est la difficulté d'obtention des données de blocage. D'une part le calcul des taux de blocage devrait se faire sur la base de populations de tailles très différentes (les compteurs à consommation ordinaire sont de 4 à 9 fois plus nombreux), et surtout la technique actuelle d'obtention des informations de la base ICBC (figure 6.6) ne permet pas l'obtention de tous les index des compteurs bloqués. Comme il a été déjà souligné, la récupération des informations relatives aux compteurs déposés se fait à partir de la table "*Compteurs*" et dans cette table il n'y a pas un champ contenant le dernier index connu.

Des analyses exploratoires, menées sur la base d'un échantillon de compteurs bloqués dont on connaît l'index de dépose montrent que des consommations importantes augmentent la probabilité de blocage : la proportion d'appareils ayant enregistré des consommations élevées parmi les compteurs bloqués est significativement supérieure à celle observée parmi les compteurs en service. Toutefois, ces études n'ont pas encore permis d'établir la structure de dépendance des taux de blocage en fonction de la consommation et on préfère se fier à un modèle conceptuel plus simple qui fait abstraction du phénomène non observé.

7.2 Un cas d'étude : les données

Afin d'illustrer un exemple pratique d'application du modèle statistique de dégradation pour prédire le rendement des compteurs en service, nous montrons ici des calculs d'inférence et de prédiction relatifs à des appareils de type Volumag de DN 15 mm.

La base de données métrologiques fournit deux sortes d'informations : la composition (par état métrologique) des échantillons de compteurs de même âge et, pour les appareils dont on connaît la courbe métrologique complète, les rendements, calculés sur la base d'un histogramme type de consommation.

Environ 3 800 étalonnages sont disponibles pour estimer les paramètres du modèle markovien de dégradation. Ils sont répartis par état métrologique, groupe d'agressivité et niveau de consommation selon le tableau de la figure (7.1).

	Agressivité 1					
	Consommation ordinaire			Consommation élevée		
Age	y_1	y_2	y_3	y_1	y_2	y_3
0-4	30	4	1	2	1	0
5-9	300	41	12	12	0	0
10-14	84	24	8	3	5	0
15-19	110	38	10	6	6	2
20-25	65	40	31	1	4	2

	Agressivité 2					
	Consommation ordinaire			Consommation élevée		
Age	y_1	y_2	y_3	y_1	y_2	y_3
0-4	174	28	6	7	5	1
5-9	508	131	43	40	39	7
10-14	244	112	34	20	33	3
15-19	223	177	70	15	19	12
20-25	67	63	23	3	3	5

	Agressivité 3					
	Consommation ordinaire			Consommation élevée		
Age	y_1	y_2	y_3	y_1	y_2	y_3
0-4	32	9	4	1	0	0
5-9	125	50	17	4	12	4
10-14	121	155	43	11	24	7
15-19	41	63	83	1	8	7
20-25	14	22	28	4	4	3

FIG. 7.1 – Répartition des compteurs par état, âge, agressivité et consommation annuelle moyenne.

Concernant les rendements, on dispose de 1 271 signatures métrologiques complètes. Les figures 7.2, 7.3 et 7.4 montrent les histogrammes de fréquence empirique des observations regroupées par état métrologique et agressivité.

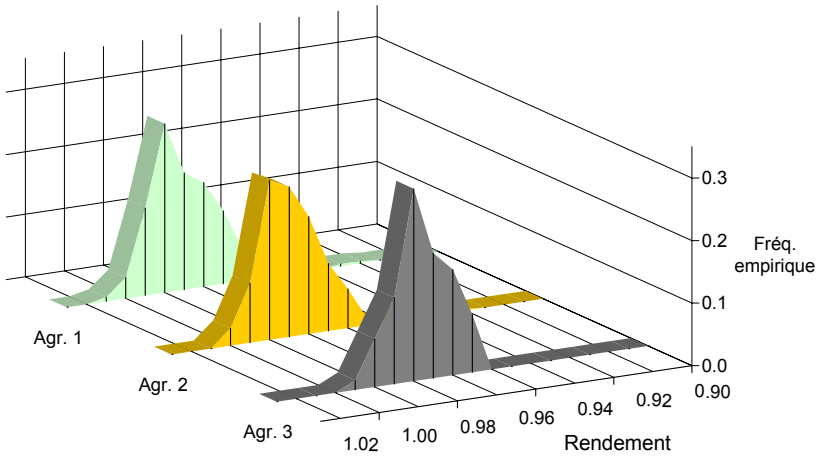


FIG. 7.2 – Histogrammes de fréquence empirique des rendements des compteurs appartenant à l'état 1.

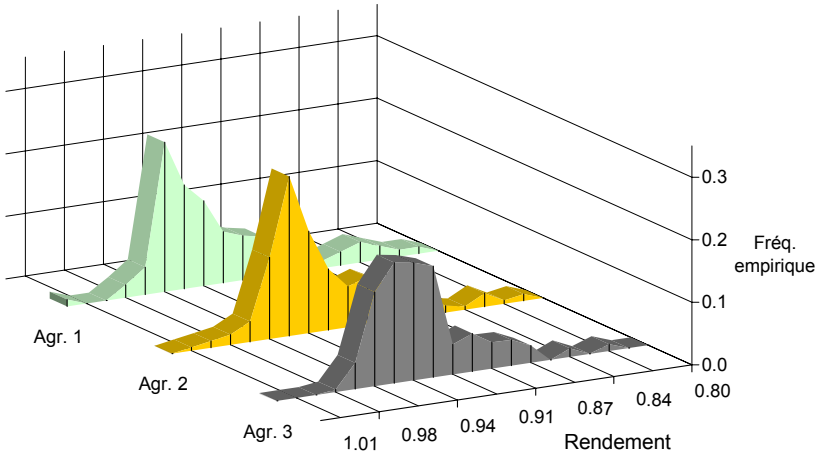


FIG. 7.3 – Histogrammes de fréquence empirique des rendements des compteurs appartenant à l'état 2.

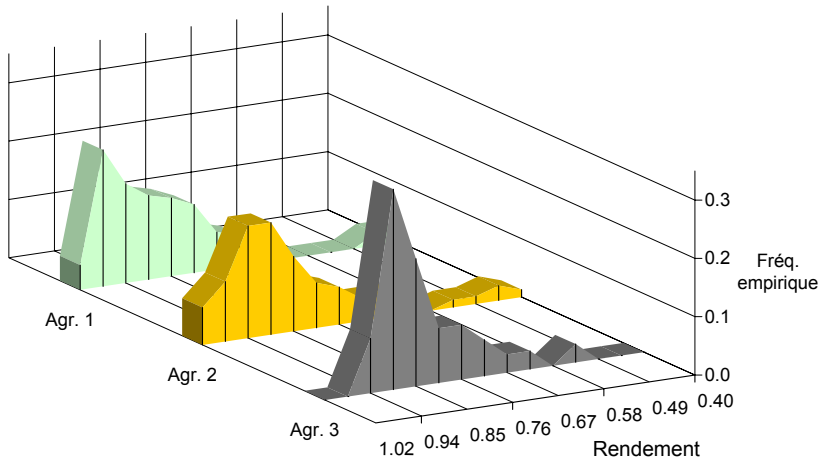


FIG. 7.4 – Histogrammes de fréquence empirique des rendements des compteurs appartenant à l'état 3.

7.3 Rappel de la méthode "indirecte" de prévision des rendements

Le travail de modélisation nécessaire pour la prévision des rendements en fonction de l'âge et des autres variables explicatives s'articule en trois phases différentes :

1. Dans un premier temps toutes les signatures métrologiques exploitables sont utilisées pour entreprendre des calculs d'inférence sur les paramètres du modèle markovien de dégradation. Au terme de cette phase, on dispose des matrices de transition qui permettent de reproduire le comportement d'un compteur en conditions réelles d'exploitation.
2. Ensuite on passe à l'observation des rendements. Les calculs d'inférence permettront d'établir leur distribution de probabilité selon l'agressivité du site et l'état métrologique.
3. Finalement le couplage du modèle dynamique de dégradation et des modèles d'observation des rendements à l'intérieur de chaque état métrologique permet de reconstruire, sur la base de la prévision de la composition par état d'un parc de compteurs d'âge donné, la loi prédictive des rendements en fonction des facteurs explicatifs pris en compte.

En pratique les calculs d'inférence et de prévision ont été menés séparément pour chacune des trois valeurs de la variable "agressivité". La seule modification dans la structure du modèle présenté dans le chapitre 5, consiste en l'introduction d'un paramètre supplémentaire pour prendre en compte la variable "consommation", comme il sera expliqué dans le paragraphe suivant.

7.4 La déformation de la matrice de transition. Un problème de choix : faut-il rajouter 1 ou 2 paramètres ?

La prise en compte d'une variable discrète à deux valeurs dans le modèle de dégradation décrit dans les chapitres 4 et 5 peut être faite de manière assez simple puisqu'on connaît déjà, par une analyse préliminaire des données, le rôle de ce facteur.

Dans le cas présent on souhaite introduire un paramètre qui favorise les transitions à partir de l'état 1 vers les états 2 et 3, ou de manière équivalente qui rende plus difficiles les transitions de l'état 1 à l'état 1 même, quand la consommation annuelle moyenne C dépasse la valeur limite C_{lim} (page 127). On rappelle que pour des petits compteurs (DN 15 mm) C_{lim} a été fixée à 200 m^3/an.

Du point de vue mathématique la modification de la probabilité de transition θ_{11} peut être faite en la multipliant pour un paramètre ξ, compris entre 0 et 1, qui réduit donc les chances de rester dans l'état de bonne métrologie d'un pas de temps à un autre. Après t pas de temps la probabilité d'appartenir à la catégorie des bons compteurs vaut donc :

$$[X(t)=1] = \begin{cases} \theta_{11}^t & \text{si } C < C_{lim} \\ (\xi \cdot \theta_{11})^t & \text{si } C > C_{lim} \end{cases} \tag{7.1}$$

Concernant les transitions de l'état 1 vers les états 2 et 3, une manière simple d'aborder le problème est de supposer que les probabilités relatives aux compteurs à consommation élevée s'obtiennent en multipliant les valeurs caractérisant les compteurs à consommation ordinaire par un unique coefficient de majoration $\varrho > 1$.

Puisque la somme des trois probabilités de transition de l'état 1 est égale à 1, la valeur de ϱ est immédiatement connue en fonction de ξ et de θ_{11} :

$$\xi \cdot \theta_{11} + \varrho \cdot (\theta_{12} + \theta_{13}) = 1 \Rightarrow \varrho = \frac{1 - \xi \cdot \theta_{11}}{1 - \theta_{11}} \tag{7.2}$$

Il suffit d'opérer la substitution :

$$\theta_{12} + \theta_{13} = 1 - \theta_{11} \tag{7.3}$$

pour obtenir l'expression (7.2).

Concernant les transitions à partir de l'état 2, on peut imaginer, dans un premier temps, qu'elles ne sont pas affectées par la consommation. Une autre façon d'aborder le problème est l'introduction d'une nouvelle variable ψ, comprise entre 0 et 1, qui multiplie la probabilité θ_{22} de rester dans la catégorie 2 quand la consommation est élevée.

En définitive deux modèles différents peuvent être imaginés. Dans le premier la prise en compte de la variable "consommation" se traduit par l'introduction d'une seule variable supplémentaire ξ, alors que dans le deuxième on introduit deux nouveaux paramètres ξ et ψ. Les matrices de transition dans les deux modèles sont :

Modèle 1

$$\begin{pmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ 0 & \theta_{22} & \theta_{23} \\ 0 & 0 & 1 \end{pmatrix} \text{ ou } \begin{pmatrix} \xi \cdot \theta_{11} & \varrho \cdot \theta_{12} & \varrho \cdot \theta_{13} \\ 0 & \theta_{22} & \theta_{23} \\ 0 & 0 & 1 \end{pmatrix} \tag{7.4}$$

si $C < C_{lim}$ si $C > C_{lim}$

Modèle 2

$$\begin{pmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ 0 & \theta_{22} & \theta_{23} \\ 0 & 0 & 1 \end{pmatrix} \text{ ou } \begin{pmatrix} \xi \cdot \theta_{11} & \varrho \cdot \theta_{12} & \varrho \cdot \theta_{13} \\ 0 & \psi \cdot \theta_{22} & \omega \cdot \theta_{23} \\ 0 & 0 & 1 \end{pmatrix} \tag{7.5}$$

si $C < C_{lim}$ si $C > C_{lim}$

avec $\omega = (1 - \psi\theta_{22}) / (1 - \theta_{22})$.

L'application de la méthode utilisée en statistique bayésienne de sélection entre modèles en compétition, basée sur le calcul du "*Facteur de Bayes*", ne montre aucun avantage visible en faveur de l'utilisation du modèle 2 par rapport au premier. Par conséquent le modèle 1, plus parcimonieux, sera préféré par l'ingénieur.

Le détail des calculs des Facteurs de Bayes, menés avec la méthode dite *de Raftery* sur la base d'échantillons des lois *a posteriori* des paramètres des deux modèles (Raftery, 1996) est montré dans l'annexe D.

7.5 L'observation des rendements

La variabilité des rendements R des compteurs observés est modélisée avec des lois de type *Bêta*, dont les paramètres dépendent de l'état métrologique et de l'agressivité. Les distributions de probabilité de la famille *Bêta* sont bien adaptées à la description de variables bornées et, en effet, les données montrent que les rendements des compteurs étalonnés ne dépassent jamais la valeur de 1.04. Cette observation s'explique par le fait que la dégradation des compteurs volumétriques se manifeste toujours avec un sous-comptage et reste valable en principe aussi pour les compteurs à jet unique. Si on introduit la variable auxiliaire Z, égale à $R/1.04$, on peut donc imaginer qu'elle est distribuée selon une loi *Bêta* standard (annexe B).

En pratique, pour chaque valeur de l'agressivité on détermine les lois *a posteriori* des 6 paramètres qui décrivent les distributions de probabilité des 3 variables Z_1, Z_2, Z_3, liées aux rendements des compteurs qui se trouvent dans les états 1, 2, 3 respectivement :

$$\begin{aligned} Z_1 &\sim \mathcal{B}(\alpha_1, \beta_1) \\ Z_2 &\sim \mathcal{B}(\alpha_2, \beta_2) \\ Z_3 &\sim \mathcal{B}(\alpha_3, \beta_3) \end{aligned} \tag{7.6}$$

Concernant les distributions *a priori* des paramètres α_i et β_i on a utilisé des lois *Gamma* dont les deux paramètres sont égaux à 10^{-3}. Il s'agit du choix classique non-informatif pour des variables positives.

Dans un travail précédent (Pasanisi et Parent, 2003) on avait supposé que les paramètres α_i et β_i étaient fonctions de l'âge des appareils. Dans le cas d'étude qu'on présente ici, un découpage ultérieur des observations selon la variable âge donnerait lieu à des échantillons de taille trop petite qui ne permettraient pas une estimation correcte des paramètres de dépendance de l'âge de α_i et β_i, et, pour cette raison, on a préféré garder l'agressivité comme seule variable explicative.

7.6 Retour sur le cas d'étude : résultats des calculs d'inférence

La figure 7.5 montre un tableau récapitulatif des résultats des calculs d'inférence, obtenus sur la base de la simulation de 3 chaînes indépendantes de 10 000 itérations chacune. Pour ces calculs on a fait appel au logiciel WinBUGS.

La convergence a été vérifiée grâce à la méthode de Brooks-Gelman que WinBUGS met automatiquement en place. Pour la simulation des lois *a posteriori* des paramètres on utilise les dernières 5 000 valeurs de chaque chaîne.

La figure 7.6 montre, une visualisation des lois *a posteriori* des variables ξ (pour les 3 valeurs de l'agressivité), obtenues comme histogrammes de fréquence empirique des échantillons extraits de ces lois avec la méthode MCMC.

7.7 Les lois prédictives des rendements

Les résultats des calculs d'inférence sont utilisés pour obtenir les distributions prédictives des rendements en fonction de l'âge. Ces lois sont obtenues par mélange, en proportions variables avec l'âge, des rendements des 3 classes :

$$\widetilde{R}(t) = R_1 \cdot Q_1(t) + R_2 \cdot Q_2(t) + R_3 \cdot Q_3(t) \tag{7.7}$$

Les R_i sont évalués à partir des équations (7.6), et le vecteur $\underline{Q}(t)$ de composantes $Q_1(t)$, $Q_2(t)$ et $Q_3(t)$ est une variable aléatoire multinomiale de paramètres $([\underline{X}(t)], 1)$:

$$\underline{Q}(t) \sim \mathcal{M}([\underline{X}(t)], 1) \Leftrightarrow \underline{q}(t) = \begin{cases} (1,0,0) \text{ avec prob. } [x_1(t)] \\ (0,1,0) \text{ avec prob. } [x_2(t)] \\ (0,0,1) \text{ avec prob. } [x_3(t)] \end{cases} \tag{7.8}$$

La structure graphique du modèle (figure 7.7) montre bien que les modèles d'observation des rendements et le modèle de dégradation, séparés dans la phase d'estimation des paramètres, se rassemblent en prédictif pour fournir la distribution des rendements d'un parc de compteurs en fonction de sa composition (en termes d'états métrologiques) qui dépend, à son tour, de l'âge par l'intermédiaire de θ et (éventuellement) ξ. Dans ce graphe l'indice "e" caractérise les grandeurs relatives aux consommations élevées.

Param.	Moyenne	Mediane	Ec. Type	Percentiles 2.50%	97.50%	
θ_{11}	0.8995	0.8997	0.0067	0.8859	0.9121	
θ_{12}	0.0859	0.0858	0.0082	0.0702	0.1021	
θ_{13}	0.0146	0.0142	0.0059	0.0041	0.0269	
θ_{22}	0.8671	0.8665	0.0489	0.7744	0.9640	
θ_{23}	0.1329	0.1335	0.0489	0.0360	0.2256	
ξ	0.9003	0.9028	0.0419	0.8115	0.9756	Agr. 1
α_1	258.00	257.00	21.10	219.40	301.40	
α_2	39.22	39.02	5.90	28.44	51.69	
α_3	5.27	5.19	1.16	3.21	7.74	
β_1	11.55	11.50	0.93	9.85	13.45	
β_2	4.24	4.22	0.61	3.14	5.52	
β_3	1.11	1.10	0.21	0.75	1.55	

Param.	Moyenne	Mediane	Ec. Type	Percentiles 2.50%	97.50%	
θ_{11}	0.8473	0.8474	0.0054	0.8366	0.8578	
θ_{12}	0.1251	0.1250	0.0065	0.1126	0.1381	
θ_{13}	0.0276	0.0276	0.0049	0.0179	0.0373	
θ_{22}	0.9279	0.9289	0.0288	0.8697	0.9835	
θ_{23}	0.0721	0.0711	0.0288	0.0165	0.1303	
ξ	0.8387	0.8390	0.0272	0.7845	0.8913	Agr. 2
α_1	300.10	299.50	21.05	260.70	343.50	
α_2	61.57	61.36	6.39	49.75	74.57	
α_3	6.08	6.03	1.03	4.24	8.24	
β_1	13.79	13.76	0.95	12.01	15.75	
β_2	6.28	6.26	0.63	5.13	7.56	
β_3	1.35	1.34	0.20	0.99	1.77	

Param.	Moyenne	Mediane	Ec. Type	Percentiles 2.50%	97.50%	
θ_{11}	0.7304	0.7305	0.0109	0.7087	0.7515	
θ_{12}	0.2362	0.2363	0.0147	0.207	0.2645	
θ_{13}	0.0333	0.0326	0.0112	0.01363	0.05699	
θ_{22}	0.7807	0.7794	0.0345	0.7172	0.8517	
θ_{23}	0.2193	0.2206	0.0345	0.1483	0.2828	
ξ	0.8616	0.8633	0.0536	0.7531	0.9619	Agr. 3
α_1	422.20	415.80	74.26	290.30	584.90	
α_2	84.07	83.15	14.25	58.51	114.60	
α_3	17.59	17.30	4.39	10.00	26.88	
β_1	23.51	23.14	4.10	16.17	32.52	
β_2	8.75	8.66	1.44	6.15	11.83	
β_3	4.30	4.23	1.03	2.52	6.48	

FIG. 7.5 – Tableau récapitulatif des résultats d'inférence.

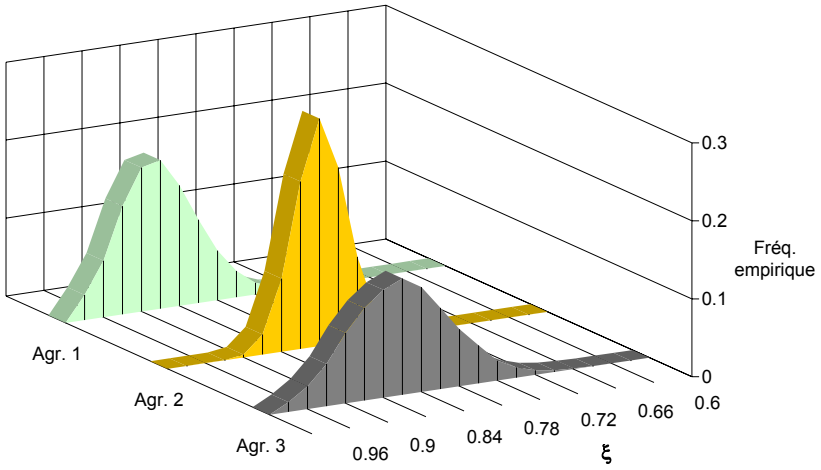


FIG. 7.6 – Lois *a posteriori* du paramètre ξ selon l'agressivité.

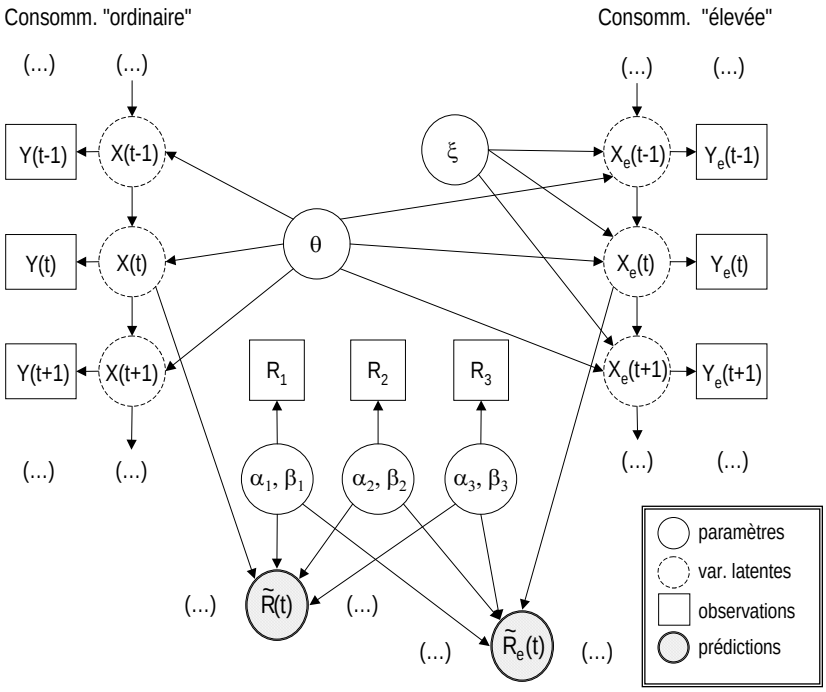


FIG. 7.7 – Représentation graphique du modèle de prévision des rendements.

Encore plus importante pour le distributeur est la prévision du rendement moyen d'un parc de compteurs d'âge donné, défini par la relation :

$$\widetilde{R}_{moy}(t) = R_1(t) \cdot [x_1(t)] + R_2(t) \cdot [x_2(t)] + R_3 \cdot [x_3(t)] \tag{7.9}$$

L'évolution de cette variable en fonction de l'âge est interprétée comme la fameuse "*loi de vieillissement*" du type de compteur examiné. Ces prévisions peuvent être utilisées directement pour les calculs technico-économiques à la base de la stratégie de gestion.

Les figures 7.8, 7.9 et 7.10 montrent les espérances *a posteriori* du rendement moyen R en fonction de l'âge et les longueurs relatives des intervalles de crédibilité à 95%. On remarque que l'incertitude sur la prévision augmente au fur et à mesure que l'espérance baisse. Plus un parc de compteurs est dégradé, plus son rendement est incertain.

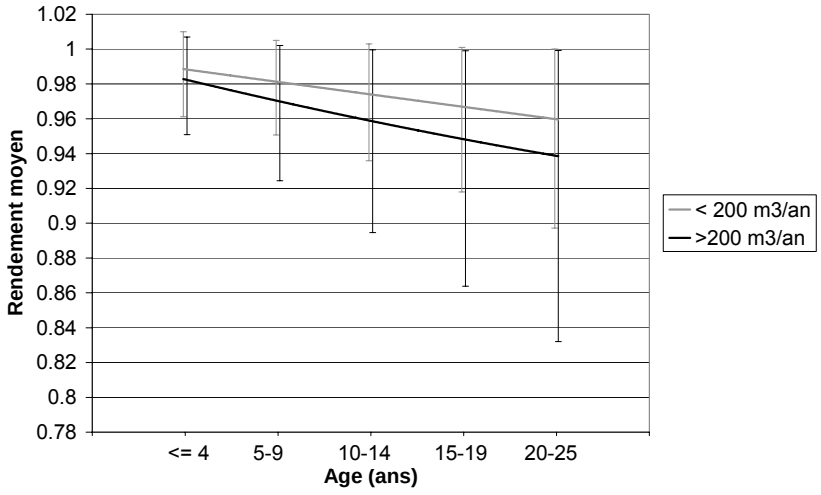


FIG. 7.8 – Espérances et intervalles de crédibilité du rendement moyen, pour des sites d'agressivité 1.

7.8 Discussion : l'importance de la prise en compte des nouveaux facteurs explicatifs

En pratique, l'espérance du rendement moyen est le paramètre technique employé en opérationnel pour définir les pertes en volume d'eau facturée.

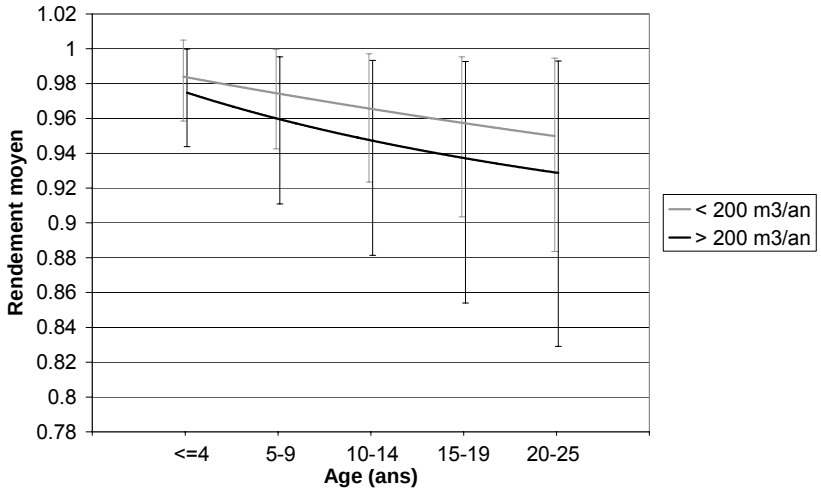


FIG. 7.9 – Espérances et intervalles de crédibilité du rendement moyen, pour des sites d'agressivité 2.

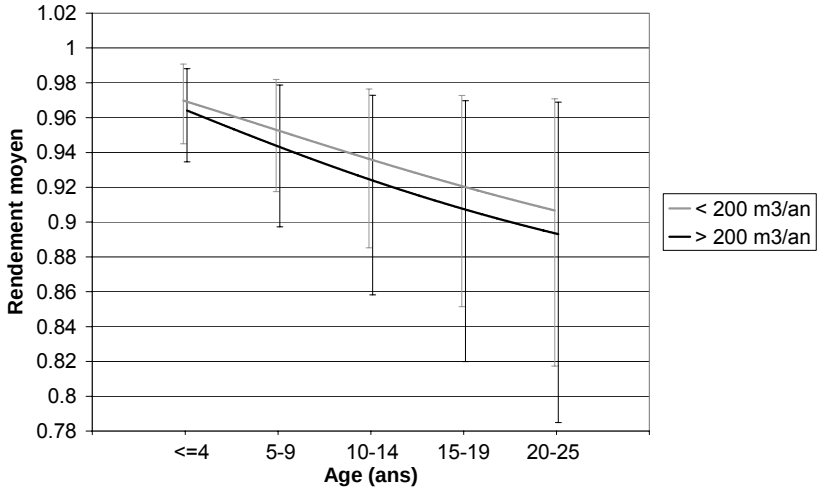


FIG. 7.10 – Espérances et intervalles de crédibilité du rendement moyen, pour des sites d'agressivité 3.

En imaginant un rendement unitaire pour les compteurs neufs, la perte annuelle de rendement moyen s'estime autour de 0.21%/an pour les compteurs exploités en sites peu agressifs avec consommations ordinaires mais peut arriver jusqu'à 0.58%/an dans le pire des cas (agressivité 3 et consommations supérieures à 200 m^3/an) :

Perte moyenne annuelle de rendement			
Conso annuelle	Agr. 1	Agr. 2	Agr. 3
< 200 m^3	0.21%	0.27%	0.50%
> 200 m^3	0.32%	0.40%	0.58%

Les différences sont sensibles et montrent l'intérêt de prendre en compte les deux variables explicatives supplémentaires.

Ces lois de vieillissement seront employées, dès l'année 2004, pour définir la stratégie de renouvellement des compteurs de la CGE sur la base du calcul du sous-comptage pour chaque appareil en service. Volontairement on ne communique pas ici les détails mais on peut facilement imaginer l'impact de ces nouvelles prévisions.

A titre d'exemple, si nous n'avions pas pris en compte consommation et agressivité, les calculs d'inférence et prévision menés sur le même jeu de données examinées dans ce chapitre nous auraient fait évaluer une perte annuelle moyenne de 0.31%/an. En comparant cette valeur avec les valeurs précédemment établies on peut calculer l'écart relatif dans la prévision de la perte en volume facturé :

"Erreur" relative de prévision			
Conso annuelle	Agr. 1	Agr. 2	Agr. 3
< 200 m^3	+50.1%	+15.6%	-37.6%
> 200 m^3	-4.3%	-22.6%	-46.9%

En pourcentage, on surestimerait considérablement (de l'ordre de la moitié) les volumes non facturés pour les conditions d'exploitations les plus communes (agressivité 1, consommation inférieure à 200 m^3/an) et on les sous-estimerait pour les consommations les plus élevées, c'est-à-dire dans les cas où les pertes sont les plus importantes.

La prise en compte de tous les facteurs explicatifs connus constitue un progrès important du point de vue de l'amélioration des stratégies de gestion. Les décideurs auront désormais à leur disposition des calculs plus adaptés à la réalité

du terrain et, de leur côté, les exploitants, chargés de l'exécution matérielle des programmes de renouvellement, seront plus confiants parce qu'ils sauront que les particularités locales, auxquelles ils sont particulièrement sensibles (comme tout homme de terrain) ont été considérées. Nous reviendrons sur ces arguments dans les conclusions de cette thèse.

Quatrième partie

Epilogue

Chapitre 8

Conclusions

8.1 Contributions à l'étude d'un problème complexe

Les résultats obtenus dans le cadre de cette thèse améliorent la connaissance du phénomène de dégradation des compteurs en conditions réelles d'exploitation.

Tout d'abord le modèle dynamique proposé est plus réaliste et plus adapté aux données expérimentales que les méthodologies couramment utilisées dans la pratique technique. Elles étaient brutalement basées sur la modélisation directe du rendement en fonction de l'âge avec des techniques de régression (éventuellement multiple). Nous avons montré que le modèle markovien de dégradation est capable de reproduire de manière satisfaisante la structure stochastique des données, et que le couplage avec un modèle d'observation des rendements permet facilement d'obtenir indirectement des distributions de probabilité *a posteriori* plus crédibles des rendements.

Cette technique de modélisation est aussi moins exigeante du point de vue des données, puisqu'on peut utiliser, pour les calculs d'inférence, des signatures métrologiques partiellement incomplètes, sans que les données perdent en signification. On peut établir l'état métrologique sur la base de quelques points de la courbe métrologique, alors que l'extrapolation d'un rendement sur la base d'une interpolation entre ces points est un exercice beaucoup plus périlleux.

Par conséquent, la nouvelle méthode de modélisation a permis de valoriser l'exploitation d'un grand nombre d'étalonnages, la plupart issus de l'ancienne base de la CEO (*Compagnie des Eaux et de l'Ozone*), qui autrement n'auraient

pas pu être utilisés. Au prix d'un modèle légèrement plus complexe et qui nécessite de méthodes d'inférence qui sortent de la pratique courante des ingénieurs, on a réalisé des économies importantes en termes d'étalonnages !

Un autre point fort du schéma proposé est la caractérisation des compteurs *sous-marins*. Leur présence a été souvent évoquée par les spécialistes mais jamais ils n'avaient été décrits ni leur occurrence quantifiée. L'identification des compteurs *sous-marins* avec les compteurs non-conformes selon la nouvelle réglementation est très importante du point de vue du distributeur. De cette manière, le modèle est conçu pour fournir la probabilité de non-conformité d'un compteur, en fonction des facteurs explicatifs qui résument ses conditions d'exploitation. Ces estimations peuvent être immédiatement utilisées pour prédire, en fonction de la composition d'un parc de compteurs, son taux de non conformité et pour mettre en place, si nécessaire, des stratégies de gestion pour le contenir en dessous de la valeur limite fixée (15% pour les 7 premières années d'application de la norme, puis 12.5% pour une seconde période de 7 ans et finalement 10%).

La technique d'estimation bayésienne mise en place permet d'évaluer globalement l'incertitude sur les estimations des paramètres du modèle, à travers leurs lois jointes de probabilités. Cette incertitude est prise en compte dans les prévisions, obtenues à l'aide du modèle, et peut être facilement exprimée sous forme d'écarts types ou d'intervalles de crédibilité. Du point de vue du décideur la quantification de l'incertitude sur une prévision a, en principe, la même importance que la prédiction même. En pratique, même si ce paramètre n'est pas intégré dans les règles pratiques de décision (souvent déterministes), la quantification de sa propre "*ignorance*" est toujours bénéfique pour le gestionnaire parce qu'elle le rend conscient des risques qu'il prend et de l'importance d'améliorer sa connaissance du problème

Enfin la contribution principale apportée par cette étude a été la possibilité de prendre en compte des facteurs explicatifs autres que l'âge.

Depuis toujours la qualité de l'eau a été retenue comme une variable explicative importante de la dégradation des compteurs mais aucune étude, à notre connaissance, n'avait encore permis de quantifier une relation de cause à effet entre une certaine typologie d'eau et la vitesse de la dégradation. La technique choisie, qui renonce à une compréhension approfondie du trop complexe mécanisme d'interaction entre compteurs, eau, réseau et habitudes locales de consommation pour définir un paramètre d'agressivité lié au site d'exploitation, s'est

avérée efficace. En vertu de la relation logique (et confirmée par les données) entre vitesse de dégradation et taux de blocage, on peut extrapoler l'agressivité pour un grand nombre de sites non représentés dans la base de données métrologiques.

En même temps, cette étude a permis une analyse détaillée des phénomènes de blocage et une quantification de l'occurrence de ces événements. Les résultats sont assez rassurants pour le distributeur : on peut conclure que les pannes des compteurs sont aujourd'hui rares et les arguments autrefois avancés concernant la fragilité des compteurs volumétriques ne sont plus valables aujourd'hui (du moins en France). En moyenne moins de 4 compteurs sur 1 000 installés se bloquent chaque année.

La prise en compte du volume total enregistré par l'intermédiaire de la consommation annuelle moyenne a aussi donné de bons résultats et sa caractérisation par une variable binaire (consommation ordinaire ou élevée) a permis une réelle exploitation de ce facteur explicatif.

Les résultats du chapitre 7 montrent l'importance de la prise en compte de ces variables supplémentaires dans l'estimation du rendement (et donc des pertes économiques). Des écarts relatifs parfois importants (de l'ordre de 50%) sont observés par rapport à une modélisation basée uniquement sur l'âge des compteurs.

L'intégration des nouvelles "*lois de vieillissement*", en fonction des variables identifiées par cette étude, dans les calculs technico-économiques est, pour le distributeur d'eau, une avancée considérable dans la gestion optimale des parcs de compteurs.

8.2 Retombées pour le distributeur d'eau

Dans le paragraphe précédent j'ai surtout insisté sur les résultats pratiques de ce travail de thèse. Evidemment, l'utilité de ces résultats est un point important pour le distributeur qui améliore sa connaissance du problème et, par conséquent, ses techniques de gestion.

Les nouvelles lois de vieillissement seront utilisées, à partir de l'année 2004, pour bâtir la politique de renouvellement. Les prévisions des taux de non-conformité ne sont pas encore intégrées dans les règles de gestion (il faut encore quelques années pour que la nouvelle réglementation soit opérationnelle) mais, de manière

informative, une première étude interne (confidentielle) a montré l'effet d'une politique de gestion donnée sur les taux de non-conformité régionaux. Cette notion commence donc à rentrer dans les préoccupations des gestionnaires, bien avant la mise en application de l'arrêté ministériel.

Au-delà de ces retombées importantes, il y en a une autre plus difficile à évaluer mais tout aussi significative.

Cette étude met en valeur, surtout aux yeux des exploitants, l'importance des données (métrologiques et ICBC). L'établissement de lois de vieillissement "*locales*" rapproche de quelque manière les hommes de terrain des "*prévisionnistes*" du siège et rend leurs études plus "*crédibles*". Les exploitants connaissent bien les particularités de chaque site et auront toujours du mal à faire confiance à des études "*nationales*" qui mélangent les résultats des compteurs de Rennes, Toulouse, Paris et Saint Laurent du Var !

Une des causes du succès de la campagne d'échantillonnage de l'année 2003 est le fait que les demandes d'étalonnages étaient ciblées sur un certain nombre de contrats choisis par les exploitants, désireux d'avoir des informations plus détaillées sur des sites qui les intéressent pour des raisons différentes. Les résultats de l'étude d'agressivité montrent que leur effort de contribution à l'alimentation de la base métrologique se traduit par des résultats de grande importance pratique.

D'autre part, ces résultats montrent aussi l'importance de bien consigner dans les bases de données des informations apparemment "accessoires" comme le motif de dépose. Evidemment, si cette information est localement négligée au moment de la dépose du compteur, au mieux le contrat sera exclu de l'analyse (trop de motifs inconnus) et au pire (motifs erronés) les conclusions en termes d'agressivité du site seront totalement fausses. Le fait de savoir qu'il s'agit d'éléments importants dans le cadre de la gestion des parcs et que se tromper de groupe d'agressivité peut se traduire par une stratégie non adaptée au site (avec des retombées économiques) motivera les exploitants à veiller de plus en plus sur l'alimentation de la base de données.

La richesse d'une base de données ne se mesure pas en Mega-octets ni même en Giga-octets mais dépend de la qualité de ses informations.

8.3 La méthodologie est transposable

La problématique de la dégradation des compteurs, à l'origine et au cœur de cette thèse, a inspiré des considérations méthodologiques de caractère absolument général. Il a été déjà mis en évidence dans le chapitre 4 que les modèles markoviens peuvent être utilisés en d'autres domaines d'application qui n'ont rien à voir avec les compteurs : en fiabilité, mais aussi en hydrologie, en météorologie, en bio-statistique et dans les sciences sociales.

Le problème de l'estimation des modèles markoviens en l'absence d'observations répétées sur les individus a été relativement peu exploré dans le passé. Dans le chapitre 5 nous avons montré que l'estimation bayésienne par méthode MCMC est une alternative plus que valable (et à notre avis plus simple) aux techniques classiques d'inférence par *Maximum de Vraisemblance* qui séparent le problème "*statistique*" de l'écriture du modèle du problème "*numérique*" de la maximisation de la vraisemblance. Pour cette raison, dans l'estimation classique des probabilités de transition il est nécessaire d'imposer le respect des conditions que la matrice trouvée soit effectivement une matrice stochastique (toute valeur comprise entre 0 et 1 et somme des valeurs par ligne égale à 1). Dans l'inférence bayésienne ce problème prend la forme de la construction de la probabilité *a priori* sur l'ensemble des paramètres (et alors il en est nécessairement de même pour la probabilité *a posteriori*).

En outre, la mise en place d'un algorithme d'inférence de type Metropolis-Hastings ne présente aucune difficulté particulière et permet d'obtenir les résultats en temps de calcul très raisonnables (de l'ordre de quelques minutes sur un PC).

Les techniques de modélisation utilisées peuvent être facilement transposées à d'autres domaines d'application et il s'agit là d'une particularité propre à la statistique. Par ailleurs l'idée du modèle dynamique de dégradation des compteurs a été grandement inspirée par le travail de Parent et Prevost (2003) et de Rivot (2003) sur la description du cycle de vie des saumons atlantiques en Bretagne.

8.4 Perspectives

Comme le premier paragraphe de ce chapitre aurait pu s'appeler "*Ce qu'on a fait*", de même la partie relative aux perspectives est plutôt celle de "*Ce qu'on n'a pas fait*".

Au début du travail de recherche le premier problème rencontré a été le manque de représentativité de la base de données métrologiques. Pour cette raison un important effort d'échantillonnage a été conduit en 2002 et 2003 où je me suis chargé personnellement de choisir les compteurs à étalonner parmi ceux destinés à la dépose, sur la base de listes détaillées (avec indication du numéro de série, du nom de l'abonné, de l'adresse etc.).

Si ce mode opératoire s'est révélé efficace, en revanche, les critères utilisés ont été assez simplistes alors qu'aujourd'hui la disponibilité d'un modèle statistique de dégradation pourrait inspirer la mise en œuvre de plans d'échantillonnage plus optimisés. Il s'agit d'un problème délicat de statistique décisionnelle qui demande des méthodes de calcul adaptées, comme par exemple l'algorithme de Muller (1999), mais aussi l'encodage d'une fonction d'utilité qui "*monétise*" l'intérêt du gestionnaire à réduire l'incertitude sur une ou plusieurs estimations, comme dans l'exemple de Parent et al. (1995) relatif au contrôle de la qualité par attributs.

La figure 8.1 montre une schématisation de la problématique bayésienne des plans d'expérience.

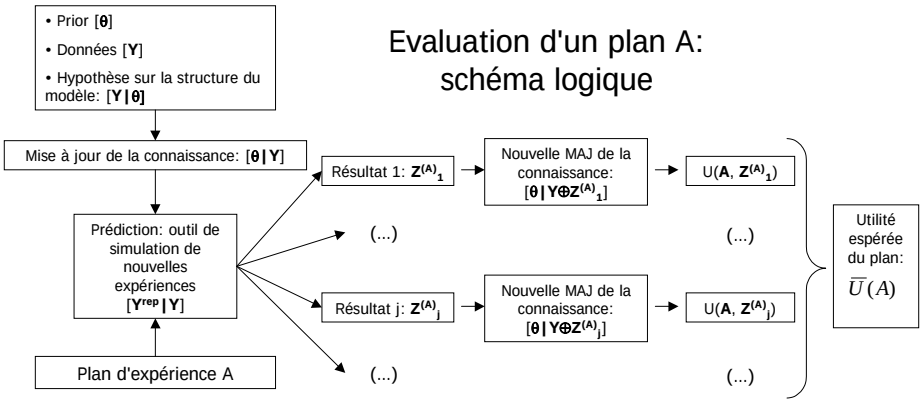


FIG. 8.1 – L'approche décisionnelle à la problématique des plans d'expérience.

L'*utilité espérée* d'un plan A se calcule en prenant en compte tous les possibles résultats Z des nouvelles expériences. L'interprétation bayésienne du choix d'un plan met au cœur du processus le modèle statistique, capable de simuler les

résultats (inconnus), mais aussi la fonction d'utilité qui transforme le problème mathématique du calcul de "l'*incertitude espérée*" en un problème décisionnel. Plus de détails sur ces arguments peuvent être trouvés dans (Tribus, 1972), (Lindley, 1997), (Joseph et Wolfson, 1997) et (Bernardo, 1997).

De même le modèle statistique pourrait, éventuellement, suggérer des règles de gestion des parcs de compteurs moins déterministes qui prennent en compte l'incertitude sur les prévisions des rendements. Ce type de travail a été exclu du cadre de cette thèse mais reste pour moi et pour le comité de suivi une perspective intéressante.

Enfin, la mise en application de la nouvelle réglementation pourrait aussi être source d'études intéressantes : le découpage du parc national en lots et la définition du nombre de compteurs à étalonner dans chaque lot est en principe un très "*beau*" problème décisionnel que le modèle de dégradation présenté dans cette thèse peut sans aucun doute aider à résoudre.

8.5 Ingénierie et statistique : bilan d'une thèse

La statistique est apparue comme une science appliquée. Le monde réel pose des problèmes qui à leur tour inspirent des créations purement intellectuelles : les modèles.

Ce passage du monde réel au monde des idées pour revenir enfin au monde réel avec les prévisions ou, dans le langage de l'ingénieur les "*résultats*", n'est pas sans danger, comme nous en a prévenu Platon, il y a 2 400 ans, dans son mythe de la caverne (La République, VIIème livre). La phase d'abstraction inductive du mécanisme qui produit les observations est difficile, ingrate et pas forcément valorisée aux yeux des praticiens qui restent évidemment liés à ce qu'ils observent.

La démarche intellectuelle consiste surtout en la conceptualisation d'un problème et la partie de calcul (difficile, fastidieuse, intéressante, passionnante selon les cas et les points de vue) n'est qu'une étape purement technique du processus. Quant aux "*résultats*" en interprétant encore librement les belles pages de Platon, ils ne sont que des manifestations d'un modèle, comme les ombres projetées sur un écran par des maquettes censées représenter la réalité. Si le modèle n'est pas une bonne schématisation du phénomène alors les résultats, même établis avec des techniques statistiques éprouvées et louées par des centaines de publications et avec des logiciels de calcul puissants et coûteux, ne seront d'aucune

utilité.

L'ingénieur et le statisticien, l'industriel et le chercheur, évoluent souvent dans des mondes différents. Ce dernier aura une naturelle tendance à se focaliser uniquement sur les aspects méthodologiques, sur les propriétés mathématiques d'un objet ou d'une classe de modèles. Etudiant, il m'est parfois arrivé d'avoir l'impression que les applications pratiques viennent dans une phase ultérieure, que les exemples sont choisis en fonction de leur "*adaptabilité*" à un modèle.

Pour sa part l'ingénieur est focalisé sur son problème pratique et recherche la mise en place d'outils d'analyse relativement simples avec des résultats d'interprétation immédiate. A cause aussi de sa formation déterministe, il aura du mal surtout à valoriser la partie de conceptualisation du modèle et il aura tendance à mettre en œuvre des recettes statistiques déjà prêtes et incorporées dans un logiciel "prestigieux", cher si possible et largement commércialisé, comme si c'était gage de qualité des résultats. D'autre part ses exigences temporelles ont une échelle bien différente de celles du théoricien : les unes se mesurent en semaines (rarement en mois) en compétition avec d'autres problèmes urgents, tandis que les autres se mesurent souvent en années consacrées à un seul et unique thème de recherche, en compagnie de son ordinateur chéri.

Je me suis permis cette description, un peu caricaturale (*absit iniuria verbis*), parce que, ingénieur de formation et désormais aussi statisticien, je comprends et respecte les deux points de vue. Tout au long de ce travail j'ai essayé d'être le trait d'union, souvent inconfortablement assis en équilibre, entre le monde académique des statisticiens, des modèles, des idées, des abstractions, des articles remplis de formules (souvent incompréhensibles pour le non-initiés), et le monde des ingénieurs d'une grande entreprise.De même j'ai essayé de concilier la difficulté d'un phénomène complexe et peu étudié avec la nécessité de produire des résultats utilisables dans la pratique courante et l'exigence de les présenter de manière claire et synthétique aux praticiens.

Le rôle unique, que seul mon statut d'ingénieur-thésard pouvait me donner, m'a permis de faire la synthèse entre les indications des experts du comptage et la théorie statistique.

L'enseignement que je tire de ce travail est l'importance de la conceptualisation, mais aussi du dialogue entre théoriciens et hommes de terrain. Si le problème est compliqué, il ne faut pas hésiter à prendre du recul et passer du temps à discuter avec les experts, à observer les données, à les décomposer, à

comprendre leur structure, à faire de la *statistique* (sans "s" !) et résister à la tentation de se limiter à faire *des statistiques*.

Chapitre 9

Références bibliographiques

- Abi-Zeid I., Bobée B. *Some reliability measures for non-stationary Markov chains.* Dans *Engineering Risk in Natural Resources management*, Ed. Duckstein L., Parent E., pp.149-158. Kluwer Academic Publishers, 1994.
- American Water Works Association, California Section Committee. *Determination of Economic Period for Water Meter Replacement.* Journal of American Water Works Association, vol. 58, n° 6, pp. 642-650, 1966.
- Anderson T.W., Goodman L.A. *Statistical Inference about Markov Chains.* Annals of Mathematical Statistics, vol. 28, n° 1, pp 89-110, 1957.
- Ang A.H.S., Tang W.H. *Probability Concepts in Engineering Planning and Design*, vol. 2. John Wiley & Sons, 1984.
- Association des Maires de France. *Guide de l'affermage du service de distribution d'eau potable.* 2001.
- Austin B., Devos F., Judet A. *Stratégie de comptage au Royaume-Uni.* Techniques Sciences Méthodes n° 7-8 pp. 46-53, 2000.
- Babillot P., Le Lourd Ph. *Y-a-t-il un marché de l'eau ?* La Houille Blanche, n° 2, pp. 39-54, 2000.
- Bayes T., *Essay Towards Solving a Problem in the Doctrine of Chances.* Philosophical Transactions of the Royal Society of London n° 53, pp. 370-418 et n° 54, pp. 296-325, 1763. Ré-imprimé dans Biometrika, vol. 45, pp. 293-315, 1958.
- Berchtold A. *Chaînes de Markov et modèles de transition. Applications aux sciences sociales.* Hermès, 1998.
- Berger J.O. *Statistical Decision Theory and Bayesian Analysis*, 2ème édition Springer-Verlag, 1985.
- Bernardo J.M. *Statistical inference as a decision problem : the choice of sample*

size. The Statistician, vol. 46, n° 2, pp. 145-149, 1997.

- Bernardo J.M., Smith A.F.M. *Bayesian Theory*. John Wiley & Sons, 1994.
- Bernier J., Parent E., Boreux J.J. *Statistiques pour l'Environnement. Traitement Bayésien des Incertitudes*. Tec & Doc, 2000.
- Berry D.A. *Statistics. A Bayesian perspective*. Duxbury Press, 1996.
- Berry D.A., Thor A., Cirrincione C., Edgerton S., Muss H., Marks J., Liu E., Wood W., Budman D., Perloff M., Peters W., Henderson I.C. *Scientific Inference and Predictions. Multiplicities and Convincing Stories : A Case Study in Breast Cancer Therapy* (avec discussion). Dans *Bayesian Statistics 5*, Ed. Bernardo J.M, Berger J.O., Dawid A.P., Smith A.F.M., pp. 45-67. Clarendon Press, 1996.
- Betta V., Cascetta F., Palombo A. *Cold potable water measurement by means of a combination meter*. Measurement, vol. 32, pp. 173-179, 2002.
- Bickenbach F., Bode E. *Markov or Not Markov, this should be a question*. Working paper n° 1086, Kiel Institute of World Economics, 2001.
- Boireau A., Mousty, P., Proucelle G., Mercier B. *Diagnostic de réseau de distribution de la ville de Perpignan suite à des phénomènes d'eaux rouges*. Techniques Sciences Méthodes, n° 11, pp. 37-41, 2000.
- Box G.E.P., Tiao G.C. *Bayesian inference in statistical analysis*. Addison-Wesley, 1973.
- Brooks S.P. *Markov Chain Monte Carlo Method and its application*. The Statistician, vol. 47, Part 1, pp. 69-100, 1998.
- Brooks S.P., Gelman A. *General Methods for Monitoring Convergence of Iterative Simulations*. Journal of Computational and Graphical Statistics, vol. 7, n° 4, pp. 434-455, 1998.
- Brooks S.P., Roberts G.O. *Assessing convergence of Markov chain Monte Carlo algorithms*. Statistics and Computing, vol. 8, pp. 319-335, 1998.
- Carlin J.B., Chib S. *Bayesian model choice via Markov Chain Monte Carlo methods*. Journal of the Royal Statistical Society, séries B, vol. 57 pp. 473-484, 1995.
- Chauveau D., Diebolt J. *An automated stopping rule for MCMC convergence assessment*. Computational Statistics, vol. 14, n° 3, pp. 449-465, 1999.
- Carlier M. *Machines hydrauliques*. Imprimerie Louis-Jean, 1968.
- Cascetta F., Sepe D. *Evaluation of long term performance of domestic turbine water meters*. Measurement & Control, vol. 30, pp. 233-237, 1997.
- Caswel H. *Matrix population models*. Sinauer Associates Inc., 1989.

- Chambolle T. *Le circuit de financement de l'eau potable et de l'eau usée en France*. Actes du colloque : *Coût et Prix de l'eau en Ville, Alimentation et Assainissement*. Paris, 6-8 décembre 1988, pp. 3-11. Presses de l'Ecole Nationale des Ponts et Chaussées, 1988.
- Chan K.S. *Asymptotic behaviour of the Gibbs sampler*. Journal of the American Statistical Association, vol. 88, pp. 320-326, 1993.
- Chapus R. *Droit Administratif Général*, Tome I, 7ème edition. Montchrestien, 1993.
- Comité Français d'Accréditation. *Guide technique pour un dossier d'accréditation débitmétrie liquide*. Document COFRAC n° 2023 (anciennement FRETAC n° 23), 1993.
- Compagnie Générale des Eaux (Direction Technique, Dép. QER). *Bilan Qualité Eau : Définition des modes de calculs des classes de qualité et définition du mode de présentation. Cahier des charges, version E5-2*, 2003.
- Congdon P. *Bayesian Statistical Modelling*. John Wiley & Sons, 2001.
- Coquillard P., Hill D.R.C. *Modélisation et simulation d'écosystèmes*. Masson, 1997.
- Costes A., Pia Y. *Les compteurs d'eau en France : la réglementation et son évolution*. Techniques Sciences Méthodes, n° 7-8, pp. 21-27, 2000.
- Cowell R.G. *Introduction to inference for bayesian networks*. Dans *Learning in Graphical Models*, Ed. Jordan M.I., pp 9-26. M.I.T. Press, 1999.
- Cowell R.G., Dawid, A.P., Lauritzen S.L., Spiegelhalter D.J. *Probabilistic Networks and Expert Systems*. Springer-Verlag, 1999.
- Cowles M.K., Carlin B.P. *Markov Chain Monte Carlo Convergence Diagnostics : A Comparative Review*. Journal of the American Statistical Association, vol. 91, pp. 883-904, 1996.
- Cullen A.C., Frey H.C. *Probabilistic Techniques in Exposure Assessment*. Plenum Press, 1999.
- De Finetti B. *La prévision : ses lois logiques, ses sources subjectives*. Annales de l'Institu Henri Poincaré, vol. 7, pp 86-133, 1937.
- De Finetti B. *Theory of probability, a critical introductory treatment*, vol. 1. John Wiley & Sons, 1974.
- Detoc S. *Stratégie d'économie d'eau dans l'habitat : l'exemple des villes de Bretagne*. Techniques Sciences Méthodes, n° 2, pp. 33-36, 2000.
- Dhillon B.S., Yang N. *Comparison of block diagram and Markov method : system*

reliability and mean time to failure results for constant and non-constant unit failure rates. Microelectronics Reliability, vol. 31, n° 3, pp. 505-509, 1997.

- Ditcham S.F. *Water demand management in developing countries*. Water Supply, vol. 15, n° 1, pp. 41-44, 1997.
- Duckstein L., Bogardi I. *Uncertainties in lake management*. Dans *Reliability in Water Resources Management*, Ed. McBean E.A., Hipel K.W., Unny T.E., pp. 253-279. Water Resources Publications 1979.
- Duroy S. *La distribution d'eau potable en France. Contribution à l'étude d'un service public local*. LGDJ, 1996.
- Erhard-Cassegrain A., Margat J. *Introduction à l'économie générale de l'eau*. Masson, 1983.
- Fontana S. *Le comptage de l'eau, une informatisation croissante*. Hydroplus, n° 62, pp. 49-59, 1996.
- Garthwaite P., O'Hagan, A. *Quantifying expert opinion in the water industry : an experimental study*. The Statistician, vol. 49, pp. 455-477, 2000.
- Gelfand A.E., Dey D.K. *Bayesian model choice : Asymptotics and exact calculations*. Journal of the Royal Statistical Society, séries B, vol. 56, n° 3, pp. 501-514, 1994.
- Gelfand A.E., Hills S.E. Racine-Poon A., Smith A.F.M., *Illustration of Bayesian inference in normal data models using Gibbs sampling*. Journal of the American Statistical Association, vol. 85, pp. 972-985, 1990.
- Gelfand A.E., Smith A.F.M. *Sampling-based approaches to calculating marginal densities*. Journal of American Statistical Association, vol. 85, pp. 398-409, 1990.
- Gelman A. *Inference and monitoring convergence*. Dans *Markov Chain Monte Carlo in Practice*, Ed. Gilks W.R., Richardson S., Spiegelhalter D.J., pp. 131-143. Chapman & Hall, 1996.
- Gelman A., Carlin J.B., Stern H.S., Rubin D.B. *Bayesian data analysis*. Chapman & Hall, 1995.
- Gelman A., Rubin D.B. *Inference from iterative simulation using multiple sequences* (with discussion). Statistical Science, vol. 7, pp. 457-511, 1992.
- Geman S., Geman D. *Stochastic Relaxation, Gibbs Distribution and the Bayesian Restoration of Images*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-6, n° 6, pp. 721-741, 1984.
- Gilks W.R. *Full conditional distributions*. Dans *Markov Chain Monte Carlo in Practice*, Ed. Gilks W.R., Richardson S., Spiegelhalter D.J., pp. 75-88. Chapman

& Hall, 1996.

- Gilks W.R., Richardson S., Spiegelhalter D.J. *Markov Chain Monte Carlo in Practice*. Chapman & Hall, 1996-a.
- Gilks W.R., Richardson S., Spiegelhalter D.J. *Introducing Markov Chain Monte Carlo*. Dans *Markov Chain Monte Carlo* in Practice, Ed. Gilks W.R., Richardson S., Spiegelhalter D.J., pp. 1-19. Chapman & Hall, 1996-b.
- Girard Ph. *Campagne "Histogrammes de consommation" (1994-1995). Analyse statistique et interpretation*. Rapport de stage, Anjou Recherche, 1995.
- Girard Ph., Parent E. *The deductive phase of statistical inference via predictive simulations : test, validation and control of a linear model with autocorrelated errors on a food industry case study*. Soumis au Journal of Statistical Planning and Inference, 2000.
- Good I.J. *The Estimation of Probabilities : An Essay on Modern Bayesian Methods*. M.I.T. Press, 1965.
- Goodman L.A. *A Further Note on Miller's "Finite Markov Processes in Psychology"*, Psychometrika, vol. 18, pp. 245-248, 1953.
- Grandjean Ph., Jannin B. *L'influence des gros consommateurs sur l'évolution des consommations d'eau à Paris*. Techniques Sciences Méthodes, n° 2, pp. 19-22, 2000.
- Grau P. *Expériences dans le comptage de l'eau et dans le renouvellement des compteurs*. Water Supply, vol. 3, n° 4, pp. 273-292, 1985.
- Guarino C.F. *Metering*. Journal of American Water Works Association, vol. 68, n° 9, pp. 17-27, 1976.
- Guédon Y. *Modélisation de séquences d'événements décrivant la mise en place d'éléments botaniques*. Dans *Modélisation et simulation de l'architecture des végétaux*, Ed. Bouchon J., de Reffye Ph., Barthélémy D., pp. 187-202. Editions INRA, 1997.
- Guerquin M., Grosjean M. *Prise en compte du comptage à l'international : exemple de Buenos Aires*. Techniques Sciences Méthodes, n° 7-8, pp. 59-61, 2000.
- Guihenneuc-Jouyaux C., Richardson S., Longini Jr. I.M. *Modeling markers of disease progression by a hidden Markov process : application to characterizing CD4 cell decline*. Biometrics, n° 56, pp. 733-741, 2000.
- Han C., Carlin B.P. *MCMC methods for computing Bayes factors : a comparative review*. Rapport de recherche n° 2000-001, Division of Biostatistics, University

of Minnesota, 2000.

- Hartigan J.A. *Bayes Theory*. Springer-Verlag, 1983
- Hastings W.K. *Monte Carlo Sampling Methods using Markov Chains and their applications*. Biometrika, vol. 57, pp. 97-109, 1970.
- Hazen A. *Meter Rates for Water Works*. John Wiley & Sons, 1918.
- Heckerman D. *A tutorial on learning with bayesian networks*. Dans *Learning in Graphical Models*, Ed. Jordan M.I., pp 301-354. M.I.T. Press, 1999.
- Howe J.W. *Flow Measurement*. Dans *Engineering Hydraulics*, Ed. Rouse H., pp 177-228. John Wiley & Sons, 1950.
- ISO 2859-0. *Règles d'échantillonnage pour les contrôles par attributs. Partie 0 : Introduction au système d'échantillonnage par attributs de l'ISO 2859*. Première édition, 1995.
- ISO 4064-1. *Mesurage de débit d'eau dans les conduites fermées – Compteurs d'eau potable froide. Partie 1 : Spécifications*. Deuxième édition, 1993.
- ISO 4064-3. *Mesurage de débit d'eau dans les conduites fermées – Compteurs d'eau potable froide. Partie 3 : Méthodes et matériels d'essai*. Deuxième édition, 1999.
- ISO/CEI 17025. *Prescriptions générales concernant la compétence des laboratoires d'étalonnages et d'essais*. Première édition, 1999.
- Jeffreys, H. *Theory of probabilities*, 3ème édition. Oxford University Press, 1961.
- Jimoh O.D., Webster P. *Stochastic modelling of daily rainfall in Nigeria : intra-annual variation of model parameters*. Journal of Hydrology n° 222, pp. 1-17, 1999.
- Joseph L., Wolfson D.B. *Interval-based versus decision theoretic criteria for the choice of sample size*. The Statistician, vol. 46, n° 2, pp. 151-153, 1997.
- Kaas R.E., Raftery A.E. *Bayes Factors*. Rapport Technique n° 254, Dép. de Statistiques de Université de Washington et Rapport Technique n° 571, Dép. de Statistiques de l'Université Carnegie-Mellon, 1994.
- Kadane J.B., Wolfson L.J. *Experiences in elicitation*. The Statistician, vol. 47 pp. 1-20, 1998.
- Krzysztofopowicz R. *Why should a forecaster and a decision maker use Bayes Theorem*. Water Resources Research, vol. 19, n° 2, pp 327-336, 1983.
- Lecoutre B., Poitevineau J. *Traitement statistique des données expérimentales : des pratiques traditionnelles aux pratiques bayésiennes*. CISIA-CERESTA, 1996.
- Legendre L., Legendre P. *Ecologie numérique*, vol. 2. Masson, 1979.

- Lee P.M. *Bayesian Statistics, an Introduction*. 2ème édition, Arnold, 1997.
- Lee T., Oliver J.L., Teniere-Buchot P.F., Travers L., Valiron F. *Economic and financial aspects*. Dans *Frontiers in Urban Water Management : Deadlock or Hope*, Ed. Maksimovic C., Tejada-Guibert J.A., pp. 313-343. I.W.A. Publishing, 2001.
- Lee T.C., Judge G.G., Zellner A. *Maximum Likelihood and Bayesian Estimation of Transition Probabilities*. Journal of the American Statistical Association, vol. 63, pp. 1162-1179, 1968.
- Lindley D.V. *Introduction to Probability and Statistics from a Bayesian Viewpoint*. University Press, 1965.
- Lindley D.V. *The choice of sample size*. The Statistician, vol. 46, n° 2, pp. 129-138, 1997.
- Lomborg B. *The Skeptical Environmentalist*. Cambridge University Press, 2001.
- Lund J.R. *Metering Utility Services : Evaluation and Maintenance*. Water Resources Research, vol. 24, n° 6, pp. 802-816, 1988.
- Machina M. *Generalised Expected Utility Analysis and the Nature of Observed Violations of the Independence Axiom*. Dans *Foundation of Utility and Risk with Applications*, Ed. Stigum B.P., Wenstop F., pp. 263-293. Reidel, 1983.
- Mackay, D.J.C. *Introduction to Monte Carlo methods*. Dans *Learning in Graphical Models*, Ed. Jordan M.I., pp. 175-204. M.I.T. Press, 1999.
- Madansky A. *Least Squares Estimation in Finite Markov Processes*. Psychometrika, vol. 24, pp. 509-520, 1959.
- Margat J. *Combien d'eau utilise-t-on ? Pour quoi faire ?*. La Houillle Blanche, n° 2, pp. 12-28, 2000.
- Marschak J. *Rational Behaviour, Uncertain Prospects and Measurable Utility*. Econometrica, vol. 55, pp. 111-141, 1950.
- Mayadatchevsky G. *Présence de "sable" et/ou d'eau rouge dans les circuits de distribution d'eau froide et chaude sanitaire*. Techniques Sciences Méthodes, n° 9, pp. 79-83, 2000.
- Mergensen K.L., Robert C.P., Guihenneuc-Jouyaux C. *MCMC Convergence Diagnostics : A "reviewww"*. Dans *Bayesian Statistics 6*, Ed. Bernardo J.M., Berger J.O., Dawid A.P., Smith A.F.M., pp. 415-440. Oxford University Press, 1999.
- Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H., Teller E. *Equation of State Calculations by Fast Computing Machine*. Journal of Chemical Physics, vol. 21, pp. 1087-1091, 1953.

- Miller, G.A. *Finite Markov Processes in Psychology.* Psychometrika, vol. 17, pp. 149-167, 1952.
- Muller P. *Simulation-based optimal design.* Dans *Bayesian Statistics 6*, Ed. Bernardo J.M., Berger J.O., Dawid A.P., Smith A.F.M., pp. 459-474. Oxford University Press, 1999.
- Neal R.M. *Probabilistic Inference Using Markov Chain Monte Carlo Methods.* Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto, 1993.
- Newman G.J., Noss R.R. *Domestic 5/8 Inch Meter Accuracy and Testing, Repair and Replacement Programs.* Proceedings 1982 American Water Works Association Annual Conference, Part 1, Paper n° 11-7, 1982.
- Newton M.A., Raftery A.E. *Approximate Bayesian Inference by the weighted likelihood bootstrap* (avec discussion). Journal of the Royal Statistical Society, séries B, vol. 56, pp. 3-48, 1994.
- O'Hagan A. *Eliciting expert beliefs in substantial practical applications.* The Statistician, vol. 47, pp. 21-35, 1998.
- Orr L.E., Enna V.A., Miller M.C. *Analysis of a water-meter replacement program.* Journal of American Water Works Association, vol. 69, n° 2, pp. 68-71, 1977.
- Parent E., Bernier J. *Une procédure bayésienne de sélection/validation différentielle pour déterminer le domaine d'attraction des valeurs extrêmes.* Accepté dans la Révue de Statistique appliquée, 2003.
- Parent E., Chaouche A., Girard Ph. *Sur l'apport des statistiques bayesiennes au contrôle de la qualité par attributs. Partie 1 : Contrôle simple.* Révue de Statistique appliquée, n° 4, pp. 5-18, 1995.
- Parent E., Lebdi F., Hurand P. *Stochastic modeling of a water resource system : analytical techniques versus synthetic approaches.* Dans *Water Resources engineering Risk Assessment*, Ed. Ganoulis J., pp. 418-434. Springer-Verlag 1991.
- Parent E., Prevost E. *Inférence bayésienne de la taille d'une population de saumons par utilisation de sources multiples d'informations.* Revue de Statistique Appliquée, n° 3, pp. 5-38, 2003.
- Pasanisi A. *Modélisation bayésienne de la dégradation métrologique des compteurs d'eau avec la prise en compte de sources multiples d'information.* Actes des XXXV Journées de Statistiques, Lyon 2-6 juin 2003, vol. 2, pp. 761-764, 2003.
- Pasanisi A., Parent E. *Modélisation bayésienne du vieillissement des compteurs*

d'eau par mélange de classes d'appareils de différents états de dégradation. Accepté dans la Revue de Statistique Appliquée, 2003. Texte intégral en Annexe E.

- Perreault L. *Analyse bayésienne rétrospective d'une rupture dans les séquences de variables aléatoires hydrologiques*. Thèse de doctorat, ENGREF, 2000.
- Press S.J., Tanur J.M. *The Subjectivity of Scientists and the Bayesian Approach*. John Wiley & Sons, 2001.
- Raftery A.E. *Hypothesis Testing and Model Selection*. Dans *Markov Chain Monte Carlo in Practice*, Ed. Gilks W.R., Richardson S., Spiegelhalter D.J., pp. 163-187. Chapman & Hall, 1996.
- Richard B., Huau M.C. *Gestion de la baisse de consommation de l'eau potable : l'expérience dans le PECO avec SAUR Neptun Gdansk*. Techniques Sciences Méthodes, n° 7-8, pp. 54-58, 2000.
- Rivot E. *Investigations Bayesiennes de la dynamique des populations de Saumon atlantique (Salmo salar L.). Des observations de terrain à la construction du modèle statistique pour apprendre et gérer*. Thèse de doctorat, ENSAR, 2003.
- Robert C.P. Casella G. *Monte Carlo Statistical Methods*. Springer-Verlag, 1999.
- Robert C.P. *L'Analyse Statistique Bayésienne*. Economica, 1992.
- Robert C.P. *Méthode de Simulation Monte Carlo par chaînes de Markov*. Economica, 1996.
- Roberts G.O. *Markov chain concepts related to sampling algorithms*. Dans *Markov Chain Monte Carlo in Practice*, Ed. Gilks W.R., Richardson S., Spiegelhalter D.J., pp. 45-57. Chapman & Hall, 1996.
- Roberts G.O., Polson N.G. *On the geometric convergence of the Gibbs sampler*. Journal of the Royal Statistical Society, série B, vol. 56, pp. 377-384, 1994.
- Roberts G.O. and Smith A.F.M. *Simple conditions for the convergence of the Gibbs sampler and Hastings-Metropolis algorithms*. Stochastic Processes and their Applications, n° 49 pp. 207-216, 1994.
- Savage L.J. *The Foundations of Statistics*, 2ème édition. Dover Publications, 1972.
- Shipman H. R. *Water Metering Practices*. Aqua, n° 2, pp. 2-12, 1978.
- Schervisch M.J., Carlin B.P. *On the convergence of successive substitution sampling*. Journal of Computational and Graphical Statistics, vol. 1, pp. 111-127, 1992.
- Sisco R.C. *The Case for Meter Replacement Programs*. Journal of American

Water Works Association, vol. 59, n° 11, pp. 1449-1455, 1967.

- Sivia D.S., *Data Analysis, a Bayesian Tutorial*. Oxford University Press, 1996.
- Spiegelhalter D.J., Best N.G., Gilks W.R., Inskip H. *Hepatitis B : A case study in MCMC*. Dans *Learning in Graphical Models*, Ed. Jordan M.I., pp. 575-598. M.I.T. Press, 1999 (une version précédente avec le même titre se trouve dans Gilks et al., 1996-a).
- Spiegelhalter DJ, Thomas A., Best N.G. *Computation on Bayesian Graphical Models* (avec discussion). Dans : *Bayesian Statistics 5*, Ed. Bernardo J.M, Berger J.O., Dawid A.P., Smith A.F.M., pp. 407-425. Clarendon Press, 1996-a.
- Spiegelhalter D.J., Thomas A., Best N.G. *WinBUGS Version 1.3 User Manual*. MRC Biostatistics Unit, 2000.
- Spiegelhalter D.J., Thomas A., Best N.G., Gilks W. *BUGS 0.5 Examples*, vol. 2. MRC Biostatistics Unit, 1996-b.
- Tierney L. *Markov chains for exploring posterior distributions* (avec discussion). Annals of Statistics, vol. 22, pp. 1701-1762, 1994.
- Tierney L. *Introduction to general state-space Markov chain theory*. Dans *Markov Chain Monte Carlo in Practice*, Ed. Gilks W.R., Richardson S., Spiegelhalter D.J., pp. 59-74. Chapman & Hall, 1996.
- Tort X., Valls M., Coll J., Asencio E. *Techniques of collecting data for a study of errors in measurements in water meters*. Aqua, n° 1, pp. 14-17, 1988.
- Tronskolanski A.T. *Théorie et pratiques des mesures hydrauliques*. Dunod, 1963.
- Tribus M. *Décisions rationnelles dans l'Incertain*. Masson, 1972.
- Van der Linden M.J. *Implementing a large-meter replacement program*. Journal of the American Water Works Association, vol. 90, n° 8, pp. 50-56, 1998.
- Varszegi C. *L'introduction des compteurs d'eau individuels en Hongrie*. Techniques Sciences Méthodes, n° 7-8, pp. 38-41, 2000.
- Vlontakis A. *La fiabilisation des opérations de comptage, relevé et facturation à Casablanca*. L'opération de recensement des compteurs. Techniques Sciences Méthodes, n° 7-8, pp. 62-65, 2000.
- Von Mises R., Geiringer H. *The Mathematical Theory of Probability and Statistics*. Academic Press, 1984.
- West M., Harrison J. *Bayesian Forecasting and Dynamic Models*, 2ème édition Springer-Verlag, 1997.
- Williams R.L. *Water meter maintenance*. Proceedings of 1976 AWWA Annual Conference, vol. 2, papier n° 2-3, pp. 1-6, 1976.

Cinquième partie

Annexes

Annexe A

Résultats de l'étude d'agressivité

A.1 Vitesses de dégradation métrologique par contrat

Code Contrat	Libellé Contrat	Estimation λ	ec. Type	Interv. Crédibilité à 95%		
K2110	Roche Sur Yon (Ville de La)	0.0557	0.02217	0.0204	0.1055	
x0001	Avranches (villle Eau)	0.0687	0.02385	0.0298	0.1229	
X0596	Blois (Ville de)	0.0848	0.02119	0.0485	0.1310	
x0599	Région des Essards (S.I.A.E.P. de la)	0.0917	0.02102	0.0551	0.1370	
x0076	Communauté de Communes du Sud Estuaire	0.0933	0.03689	0.0352	0.1790	
x0006	Granville (ville Eau)	0.0966	0.03287	0.0426	0.1714	
S8020	Barbizon	0.0971	0.02906	0.0489	0.1616	
x0524	SAINT MALO (Ville) (eau)	0.0988	0.02209	0.0603	0.1456	
x0011	Vierzon (Ville de)	0.0999	0.01833	0.0671	0.1396	
I7340	Blagnac-Eau Potable	0.1002	0.03449	0.0439	0.1772	
G4700	Auchy les Mines (eau)	0.1055	0.02215	0.0663	0.1541	
B3110	Montbeliard (com.Agglom.) (eau)	0.1073	0.02290	0.0666	0.1568	
K8240	Ainay le Chateau (Commune de)	0.1120	0.03250	0.0576	0.1838	
x0243		0.1175	0.04018	0.0525	0.2077	
x0110	Champagnole (eau)	0.1192	0.03054	0.0669	0.1858	**Groupe 1**
x0074	Vals de Sévres (S.I.A.E.P. des)	0.1220	0.03159	0.0683	0.1907	
V7300	Région du Plessis Trévise (siaep De)	0.1240	0.03862	0.0602	0.2124	
x0539	BREST (Communauté Urbaine) (eau)	0.1258	0.01701	0.0954	0.1601	
x0030	C.U.N (Siaep Reze)	0.1277	0.01899	0.0932	0.1675	
s8240	Melun - Dammarie	0.1281	0.04639	0.0543	0.2322	
x0115	Moissac Aep	0.1345	0.04626	0.0607	0.2380	
D4020	BINIC (Commune) (eau)	0.1346	0.04328	0.0636	0.2307	
b1120	Lyon C.U.	0.1364	0.02047	0.0990	0.1799	
j3510		0.1366	0.04023	0.0701	0.2239	
x0482	Frouard (eau)	0.1376	0.02935	0.0868	0.2007	
I1100	COBAS	0.1404	0.02668	0.0926	0.1981	
D1900	RENNES (Ville) (eau)	0.1430	0.02801	0.0931	0.2034	
x0582		0.1437	0.04455	0.0701	0.2427	
x0108	Haut Bocage Siaep (eau)	0.1446	0.04962	0.0633	0.2568	
G6260	Lens C.A.L.L. (eau)	0.1479	0.03325	0.0899	0.2196	
K3418	Cognac (S I A A de)	0.1485	0.02538	0.1031	0.2024	
x0016	Bergerac Aep (eau)	0.1488	0.02898	0.0980	0.2100	
x0179		0.1504	0.03887	0.0839	0.2357	
K1110	Mayenne (Ville de)	0.1524	0.03228	0.0958	0.2213	
v6550	Lagny (siaep de la Région De)	0.1533	0.04248	0.0818	0.2459	
D3150	ST JACUT LES PINS (Synd) (eau)	0.1547	0.03668	0.0914	0.2343	
x0128	Joinville (eau)	0.1555	0.03854	0.0898	0.2400	
F5350	Villedieu Ouest (siaep Eau)	0.1570	0.03731	0.0935	0.2374	
x0014	COMACH	0.1572	0.02716	0.1086	0.2141	
x0004	Chatellerault (Ville de)	0.1708	0.02094	0.1322	0.2143	
K5310	Puiseaux (Commune de)	0.1750	0.03935	0.1061	0.2612	
j2600		0.1780	0.03267	0.1199	0.2485	
x0044	Ville d' Agen - Eau	0.1807	0.01993	0.1433	0.2221	**Groupe 2**
I4000	Ville de Toulouse-Eau Potable	0.1808	0.01511	0.1526	0.2110	
H2110	CC de l'Agglomération de Forbach (eau)	0.1808	0.04094	0.1110	0.2674	
x0255		0.1821	0.02626	0.1346	0.2367	
x0107		0.1838	0.05216	0.0960	0.2999	
i6010	Villeneuve-Sur-Lot - Eau	0.1851	0.03922	0.1157	0.2714	
K7570	Ancenis (Siaep de la Région d')	0.1852	0.03668	0.1209	0.2641	
h1160	SIE Florange et Sérémange Erzange	0.1873	0.05792	0.0922	0.3169	
x0644	Val Saint Martin (SIVOM du)	0.1889	0.02858	0.1372	0.2494	
x0087		0.1964	0.01008	0.1773	0.2164	
F4830	Caen (ville Eau)	0.2008	0.02438	0.1562	0.2517	
X0267	Durtal (S.I.A.E.P Région de)	0.2033	0.02781	0.1526	0.2612	
x0022	Verdun (eau)	0.2065	0.02511	0.1608	0.2584	
x0068	Angervilliers (siaep d')	0.2090	0.02960	0.1550	0.2721	
x0080	Su du Sud d'Agen - Eau	0.2142	0.05056	0.1272	0.3246	
x0157	Hurepoix (sie de la Région Du)	0.2171	0.03995	0.1465	0.3019	
x0261	Marbache (eau)	0.2240	0.04701	0.1427	0.3234	
G7300		0.2342	0.02754	0.1839	0.2912	
H4110	Syndicat d'Eau St Louis - Huningue et Environs	0.2419	0.06571	0.1325	0.3853	
x0005	Briey (eau)	0.2480	0.05197	0.1575	0.3584	
K4110	Argenton Sur Creuse (Ville d')	0.2495	0.05865	0.1496	0.3771	
x0939		0.2552	0.06168	0.1482	0.3885	
x0062	Nérondes (Siaep de)	0.2571	0.05045	0.1696	0.3641	
i7730	Lagarde - Aep	0.2846	0.03567	0.2194	0.3593	
x0042		0.2883	0.04414	0.2083	0.3836	**Groupe 3**
X0579		0.3129	0.07500	0.1847	0.4773	
V1250	Meulan	0.3135	0.07195	0.1899	0.4691	
x0082		0.3741	0.09052	0.2215	0.5720	
x0627	Maron (eau)	0.3857	0.06470	0.2687	0.5241	
x0210	Rosières Aux Salines (eau)	0.3858	0.06102	0.2767	0.5174	
x0008	Oissel (ville Eau)	0.4042	0.04796	0.3160	0.5057	
x0385		0.4135	0.09537	0.2503	0.6248	
F4420	Plateau Ouest Lisieux (si Eau)	0.4392	0.08053	0.2960	0.6151	
YA010		0.5321	0.03338	0.4684	0.5998	
x0035		0.5364	0.05797	0.4324	0.6545	
x0020	Sainte Adresse (ville Eau)	0.5426	0.08092	0.3999	0.7159	

A.2 Attribution "probabiliste" du groupe d'agressivité sur la base des taux de blocage

Code contrat	Libellé contrat	Groupe agr.	Proba gr 1	Proba gr 2	Proba gr 3
B2180	Tarare (Eau)	1	1	0	0
B3210	Pays Beaunois S.I (eau)	1	1	0	0
B3220	Beaune Ville (eau)	1	1	0	0
B4310	Pouilly Ss Charlieu S.I. (eau)	1	0.9975	0.0025	0
B4320	Sornin (le) S.I. (eau)	1	0.9975	0.0025	0
B4520	Cournon d'Auvergne (eau)	1	1	0	0
B5370	Rumilly (eau)	1	1	0	0
B6110	Valence (eau)	1	1	0	0
B6162	Belfortaine Agglomération (eau)	1	1	0	0
B6410	Romans / Mours Saint-Eusèbe (e)	1	1	0	0
B6930	Saint Peray Sivom (eau)	1	1	0	0
B7110	C.U.C.M. (eau)	1	1	0	0
B7120	C.U.C.M. (eau Brute)	1	1	0	0
B7250	Autun (eau)	1	1	0	0
D1370	SUD DE RENNES (Syndicat) (eau)	1	0.9135	0.086	0.0005
D1380	CHANTEPIE - VERN (Syndicat) (eau)	1	0.9995	0.0005	0
D1430	NORD DE RENNES (Syndicat) (eau)	1	1	0	0
D2050	FORET DU THEIL (la) (Synd) (eau)	1	0.9935	0.0065	0
D2070	CHATEAUBOURG (Synd) (eau)	3	0	0.1025	0.8975
D2080	MONTAUBAN-ST MEEN (Synd) (eau)	2	0.118	0.882	0
D2210	COGLAIS (Synd) (eau)	1	0.548	0.163	0.289
D2530	VITRE (Commune) (eau)	1	0.8845	0.1155	0
D3090	REDON (Commune) (eau)	2	0.0025	0.9955	0.002
D3120	MUZILLAC (Synd) (eau)	1	0.6505	0.3495	0
D4110	GOELO (Synd) (eau)	3	0	0.026	0.974
D4150	TRAOUIERO (Synd) (eau)	3	0	0.002	0.998
D4210	MORLAIX -ST MARTIN DES CHAMPS (SIVOM)(eau)	1	1	0	0
D4260	LANMEUR (Synd) (eau)	3	0	0	1
D5250	GOYEN (le) (Synd) (eau)	1	1	0	0
E1130	Plateau du Gatinais (sivom Du)	3	0.004	0.068	0.928
E3070	Bonnières (sme de la Région De)	1	0.5525	0.445	0.0025
E3300	Mantes Yvelines (c. d'Agg.)	1	0.992	0.008	0
E3460	Jouars Ponchartrain Maurepas (siaep De)	1	1	0	0
E4610	Provins	1	0.986	0.014	0
E4970	Montereau Fault Yonne	1	1	0	0
F1110	Banlieue Sud Rouen (si Eau)	1	1	0	0
F2010	Vexin Normand (si Eau)	2	0.2075	0.7775	0.015
F2210	Poses (si Eau)	1	0.999	0.001	0
F2390	CASE- LOUVIERS (eau)	1	1	0	0
F4410	Lisieux (ville Eau)	1	1	0	0
F4810	Mondeville (siaep Eau)	1	0.9995	0.0005	0
F5010	Cherbourg (cu Eau)	1	1	0	0
F5040	Val de Saire (saep Eau)	1	0.875	0.125	0
F5090	Saint-Lô (district Eau)	1	1	0	0
F5680	Messei (si Eau)	1	0.713	0.286	0.001
G1270	Arras C.U. (eau)	1	1	0	0
G3070	Chauny (eau)	3	0	0	1
G3110	Abbeville (eau)	3	0.0105	0.074	0.9155
G3960	U S E S A (eau)	1	0.9995	0.0005	0
G4190	SABALFA S.I. (eau)	1	1	0	0
G4410	Isbergues S.I (eau)	1	0.9975	0.002	0.0005
G4610	S.A.C.R.A. (eau)	1	1	0	0
G6480	Harnes C.A.L.L. (eau)	1	0.9995	0.0005	0
G7220	Lievin S.I. CALL (eau)	1	0.826	0.174	0
G7350	Avion C.A.L.L. (eau)	1	0.7585	0.2415	0
G8140	Henin Beaumont C.A.H.C. (eau)	1	1	0	0
G8510	Carvin C.A.H.C. (eau)	1	0.9975	0.0025	0
G9120	Boulonnais (C. A. - Eau)	1	1	0	0
H2150	S.I. Winborn (eau)	3	0	0	1
H2210	Sier Sarralbe (eau)	2	0.2365	0.7635	0
H5080	Sedan (eau)	1	0.9115	0.0885	0
H6190	Saint-Dizier (eau)	1	0.995	0.005	0
I4610	La Gartempe S.I - Eau	1	0.991	0.009	0
I5510	Commune de Cestas - Eau	1	0.9965	0.0035	0
I6020	Marmande - Eau	1	1	0	0
I7030	Ville de Muret-Eau Potable	1	1	0	0
I7160	Syndicat d'A.E.P. de Grisolles	1	1	0	0
I7570	Centre & Nord	1	1	0	0
I7600	S.I. d'A.E.P. Tarn et Girou	1	1	0	0

Code contrat	libellé contrat	Groupe agr.	Proba gr 1	Proba gr 2	Proba gr 3
I8000	Ville d'Auch - Eau Potable	1	1	0	0
I8600	Synd d'A.E.P. Canton Sud Tarbes	1	1	0	0
J6310	Bourg Saint Andeol S I (eau)	1	1	0	0
K1210	Château Gontier (SIGEA de l'Agglomération de)	1	0.975	0.024	0.001
K1370	Martinière (S.I.A.E.P de la)	2	0.328	0.6715	0.0005
K1410	Coutures (SI.A.E.P de)	1	1	0	0
K1510	Champtoceaux (S.I de)	2	0.0705	0.905	0.0245
K1560	eaux de la Loire (S.I)	2	0.056	0.9435	0.0005
K2250	Fontenay le Comte (Ville de)	3	0	0	1
K2710	Val de Loire (Syndicat Mixte du)	1	1	0	0
K3110	Romorantin Lanthenay (Ville de)	1	0.9225	0.0775	0
K3370	Saintes (Ville de)	1	1	0	0
K3480	Champniers (Syndicat Intercommunal de)	1	0.9965	0.0015	0.002
K5110	Olivet (Ville d')	3	0.003	0.455	0.542
K7060	MEE (S.I.A.E.P du Pays de La)	2	0	0.5065	0.4935
K7310	Saint-Gildas des Bois (S.I. de)	2	0.0135	0.976	0.0105
K7780	Guemene Penfao (S.I)	1	0.871	0.129	0
K8110	Saint Amand Orval (S.I.V.M de)	1	0.6645	0.3355	0
K8350	Levet (Syndicat intercommunal de)	1	0.9945	0.0055	0
K8410	Marche & Boischaut (S.I de La)	3	0	0.3595	0.6405
S8130	Fontainebleau - Avon (C.C.)	1	0.925	0.075	0
T3410	SIVOM Casinca (eau)	1	0.787	0.2035	0.0095
T3610	Porto Vecchio (eau)	2	0.1175	0.6985	0.184
U0082	Vallée de la Risle (siae Eau)	3	0	0.001	0.999
U0085	Saint Pierre des Corps (Ville de)	1	1	0	0
U0096	Joue les Tours (Ville de)	1	1	0	0
U0120	Rouesse Fontaine (S.I.A.E.P de)	1	1	0	0
U0179	Bourgueil (S.I.A.E.P de la Région de)	1	1	0	0
U0250	Dieppe (ville Eau)	1	0.93	0.066	0.004
V1370	Mureaux (les)	3	0	0.0055	0.9945
V2000	Marne la Vallée (san De)	2	0.3275	0.672	0.0005
V3580	Bellefontaine (sie De)	1	0.9985	0.0015	0
V3690	Dammartin En Goële (c.C.)	1	0.884	0.115	0.001
V4000	Tremblay (siaep De)	1	1	0	0
V5000	Cergy Pontoise (san De)	1	0.9995	0	0.0005
V7200	Ozoir la Ferrière	2	0.092	0.8405	0.0675
W1480	Yvetot (commune Eau)	1	0.9935	0.005	0.0015
W1520	Austreberthe (siaep Eau)	1	1	0	0
W1540	Boos (siaep Eau)	1	1	0	0
W1590	Malaunay - Montville (si Eau)	1	0.999	0.001	0
W1850	CODAH (Pour Montivilliers synd eau)	1	0.93	0.0695	0.0005
W4690	Piennes Se (eau)	1	0.9975	0.0025	0
W9220	Aniche - Auberchicourt Si	1	0.956	0.044	0
W9230	Sin Le Noble	1	1	0	0
W9460	Région de Machy Si	1	0.9885	0.0115	0
X0007	Argentan (ville Eau)	1	0.9915	0.0005	0.008
X0032	PLOEMEUR (Commune) (eau)	1	0.5195	0.4795	0.001
X0069	BEAUFORT (Syndicat) (eau)	3	0.4125	0	0.5875
X0073	Forêt de Rambouillet (siaep De)	3	0	0	1
X0079	Sarlat - Eau	1	1	0	0
X0116	PONT SCORFF (Synd) (eau)	1	1	0	0
X0383	Sivom Est d'Agen Aep (eau)	2	0.2935	0.3925	0.314
X0927	LANDERNEAU (SIVU) (eau)	1	1	0	0
X1049	Tulle Aep (eau)	1	0.9205	0.078	0.0015
YB100	Rambouillet	1	0.6505	0.201	0.1485
YF500	Touquet Paris Plage (le)	1	1	0	0
YG010	SENLIS (Ville de)	1	1	0	0
YG030	Beauvais (Ville de)	1	1	0	0
YG040	S.I.E.A.B.	1	0.999	0.001	0
YH600	Cambrai	1	1	0	0
ZB110	Trouville-Deauville District-E	1	1	0	0
ZG011	Communauté Communes Epernay (eau)	1	0.7935	0.2065	0
ZH001	Metz (eau)	1	0.897	0.096	0.007
ZH014	Verny Se (eau)	1	1	0	0
ZK201	Dadou - Eau	1	1	0	0
ZL920	Montivilliers (ville Eau)	1	1	0	0
ZT130	Amboise (Ville d')	1	0.9965	0.0035	0
ZW950	Mâcon (Eau - distribution)	1	1	0	0

Annexe B

Quelques lois de probabilités d'usage courant

B.1 La loi binomiale

La loi de probabilité la plus simple qu'on puisse imaginer est la loi dite de *Bernoulli*. Elle est définie sur l'ensemble discret {0,1} et elle affecte la valeur $p \in [0, 1]$ au nombre 1 et $1 - p$ au nombre 0. Cette loi est normalement associée au modèle d'urne : si on considère une urne remplie de boules de deux couleurs (blanches et noires, par exemple) aléatoirement mélangées, telle que le rapport entre le nombre de boules blanches et et le nombre total de boules vaille p, alors la variable aléatoire (v.a.) qui associe à un tirage dans l'urne, le chiffre 1 si la boule extraite est blanche et 0 si elle est noire, est une v.a. de *Bernoulli* de paramètre p.

On peut facilement vérifier que l'espérance et la variance de cette variable valent p et $p(1 - p)$ respectivement.

Imaginons maintenant de réaliser m tirages avec la précaution de remettre dans l'urne la boule extraite après chaque tirage, de manière que la proportion entre boules blanches et noires soit toujours constante, et de remélanger aléatoirement l'urne à chaque fois. Le nombre de fois qu'on a extrait une boule blanche (appelons le x) est alors la somme de m v.a. de *Bernoulli* de même paramètre p et indépendantes. On exprime cette circonstance en disant que les m v.a. sont indépendantes et identiquement distribuées (i.i.d.).

La loi *binomiale* de paramètres p et m est celle suivie par la somme de m

v.a. de *Bernoulli* i.i.d.

C'est une loi discrète et son support est l'ensemble $\{0, 1, \ldots\ m\}$. Elle s'écrit :

$$[x|p, m] = \binom{m}{x} p^x (1-p)^{m-x} \tag{B.1}$$

où $\binom{m}{x} = \dfrac{m!}{x!(m-x)!}$.

On exprime le fait que x est la réalisation d'une v.a. binomiale X de paramètres p et m avec la notation :

$$X \sim \mathcal{B}in\,(p, m) \tag{B.2}$$

L'espérance et la variance s'obtiennent facilement à partir de l'espérance et de la variance de la loi de *Bernoulli* :

$$\begin{aligned} \mathbb{E}(X) &= mp \\ \mathbb{V}(X) &= mp(1-p) \end{aligned} \tag{B.3}$$

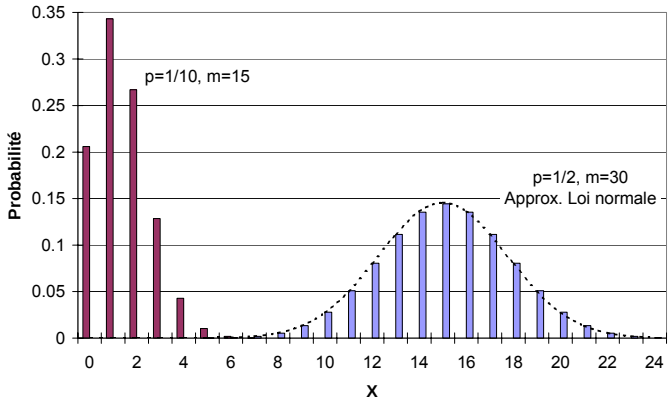


FIG. B.1 – Exemples de lois binomiales.

Pour des valeurs de m très grandes (en pratique supérieures ou égales à 30) la loi binomiale peut être approximée par une loi normale (figure B.1) de même espérance et même variance. Cette approximation trouve sa justification théorique dans le thérorème de la *limite centrale* (cf. paragraphe suivant).

B.2 La loi normale

La loi *normale*, dite aussi *gaussienne* ou (parfois) de *Laplace-Gauss* est une loi continue définie sur l'ensemble des nombres réels. Son expression, paramétrée par μ et σ^2 qui s'écrit :

$$\left[x|\mu,\sigma^2\right] = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{B.4}$$

donne lieu à la célèbre courbe en cloche de Gauss (figure B.2). On exprime le fait que X soit une v.a. normale avec la notation :

$$X \sim \mathcal{N}\left(\mu,\sigma^2\right) \tag{B.5}$$

L'espérance et la variance d'une loi normale sont μ et σ^2 respectivement. En particulier pour μ=0 et σ^2=1 on parle de loi normale *centrée-réduite* ou *standard* (figure B.2)

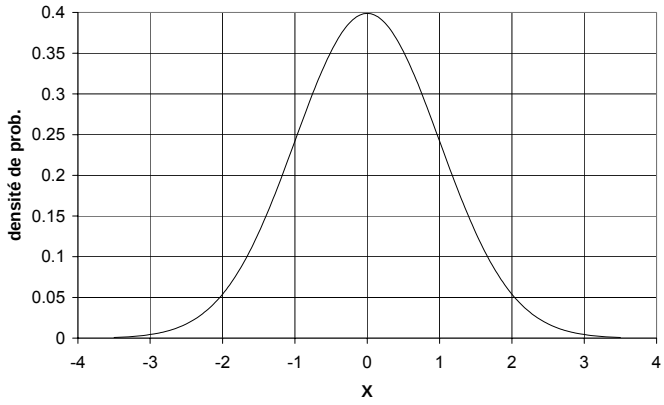


FIG. B.2 – Loi normale centrée réduite.

La loi normale a un rôle de premier plan en statistique en vertu du théorème de la *limite centrale* qui affirme que si $(X_1, X_2, ..X_n, ...)$ est une suite de v.a réelles (non nécessairement normales) i.i.d. d'espérance μ et variance σ^2, alors pour $n \to +\infty$ la v.a. :

$$\frac{\sqrt{n}}{\sigma}\left(\sum_{i=1}^{n}\frac{X_i}{n} - \mu\right)$$

tend, en loi, vers la loi normale centrée-réduite $\mathcal{N}(0,1)$.

En vertu du théorème de la limite centrale (Bernier et al. 2000) la loi normale peut être utilisée pour décrire un phénomène dont la variabilité résulte de la combinaison d'un grand nombre de causes dont les effets s'additionnent mais restent individuellement petits par rapport à leur somme, comme dans le cas des erreurs de mesure. La loi normale est ainsi utilisée pour décrire la variabilité naturelle. Dans un livre fameux de Science-Fiction (Jurassic Park de M. Chrichton), la preuve définitive que des animaux théoriquement incapables de se reproduire arrivent à procréer naturellement est que la taille des jeunes individus est distribuée selon la courbe en cloche de Gauss.

D'autres propriétés mathématiques de la loi normale (symétrie, additivité) et des échantillons gaussiens (théorème de Cochran), qui simplifient énormément les calculs dans le cadre des techniques statistiques le plus couramment utilisées (modèles linéaires, analyse de la variance), ont contribué à consolider le statut prédominant de la loi normale en statistique.

Si $X \sim \mathcal{N}(\mu, \sigma^2)$ alors l'intervalle $\mu \pm 2\sigma$ contient 95.45% de toutes les valeurs possibles de X, ce qui justifie la pratique courante d'utiliser l'intervalle dit "du 2σ" pour décrire l'essentiel de la variabilité d'un phénomène aléatoire (supposé normal).

B.3 La loi *Gamma*

La densité de probabilité d'une loi *Gamma*, définie sur l'intervalle $]0,+\infty[$ est exprimée par la relation :

$$[x|\alpha,\beta] = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) \tag{B.6}$$

où $\Gamma(\alpha)$ est la fonction *Gamma* où intégrale d'Euler de deuxième espèce :

$$\Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1} \exp(-t)\, dt \tag{B.7}$$

On écrit alors, selon la notation usuelle :

$$X \sim \mathcal{G}(\alpha,\beta) \tag{B.8}$$

La loi *Gamma* est paramétrée par les deux réels positifs α et β et son espérance et variance valent respectivement :

$$\begin{aligned} \mathbb{E}(X) &= \frac{\alpha}{\beta} \\ \mathbb{V}(X) &= \frac{\alpha}{\beta^2} \end{aligned} \tag{B.9}$$

Cette loi est habituellement utilisée pour décrire des v.a. strictement positives et notamment en statistique bayésienne comme loi *a priori* du paramètre d'une loi de Poisson ou de l'inverse de la variance d'une loi normale.

En fonction des valeurs des paramètres α e β cette loi peut prendre des formes très différentes (figure B.3).

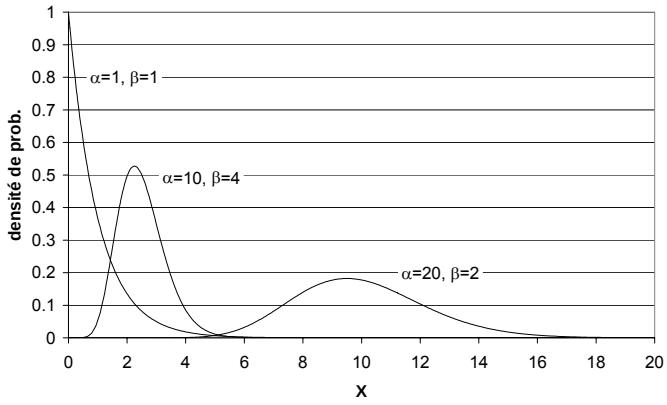


FIG. B.3 – Exemples de lois Gamma.

Certains logiciels de programmation codifient différemment l'expression (B.6) et notamment pour MatLab la loi *Gamma* est paramétrée par α et $\theta = 1/\beta$.

La loi *exponentielle* de paramètre λ et la loi *du* χ^2 de paramètre n (à n "*degrés de liberté*") peuvent être regardées comme des cas particuliers de lois *Gamma* de paramètres (1,λ) et (n/2,2) respectivement.

B.4 La loi *Bêta* (standard)

Une v.a. X, définie dans l'intervalle $[0, 1]$, distribuée selon une loi *Bêta* (*standard*), a une densité de probabilité exprimée par la relation :

$$[x|\alpha, \beta] = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{B.10}$$

où $B(\alpha, \beta)$ est la fonction *Bêta* ou intégrale d'Euler de première espèce :

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{B.11}$$

L'appellation *standard* distingue la loi *Bêta* ainsi définie de sa généralisation à des variables comprises dans n'importe quel intervalle réel $[a, b]$. Quand il n'y a pas de risque de confusion, on omet le terme "standard" et on parlera de loi *Bêta* tout court.

On exprime le fait que X suit une loi *Bêta* avec la notation :

$$X \sim \mathcal{B}(\alpha, \beta) \tag{B.12}$$

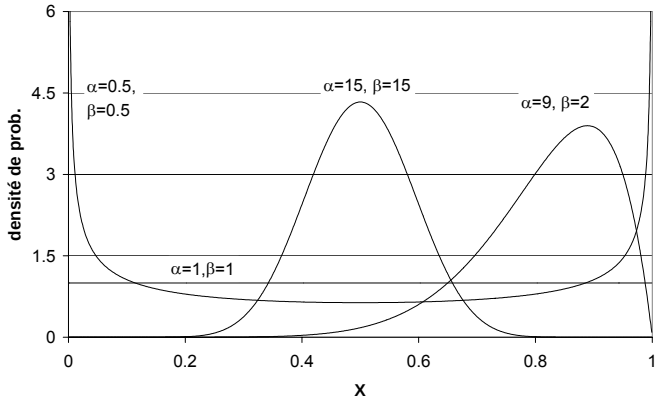


FIG. B.4 – Exemples de lois Bêta.

La moyenne et l'espérance d'une v.a. de type *Bêta* valent respectivement :

$$\begin{aligned} \mathbb{E}(X) &= \frac{\alpha}{\alpha+\beta} \\ \mathbb{V}(X) &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \end{aligned} \tag{B.13}$$

Les lois de la famille *Bêta* sont utilisées pour décrire des variables bornées et notamment en statistique bayésienne comme lois *a priori* conjuguées du paramètre p des lois binomiales (cf. chapitre 5). La figure B.4 montre comment, en fonction des valeurs de α et de β, la loi *Bêta* peut assumer des configurations assez variées, ce qui rend cette famille de lois très commode pour décrire des v.a. très différentes entre elles.

B.5 La loi Multinomiale

La loi *multinomiale* s'obtient en généralisant le modèle d'urne décrit précédemment, dans le cas où le nombre de résultats aléatoires possibles est supérieur à 2 (par exemple on peut imaginer une urne avec des boules de 3 ou 4 couleurs différentes).

Si on associe le nombre 1 au résultat 1, le nombre 2 au résultat 2 etc. la v.a. ainsi construite, qui généralise la v.a. de *Bernoulli*, est appelée *catégorielle*.

Imaginons de répéter m fois une expérience qui peut résulter en k valeurs possibles. Le phénomène est régi par le vecteur de $\mathbb{R}^k$:

$$\underline{p} = (p_1, p_2, ...p_k) \tag{B.14}$$

dont les composantes expriment les probabilités de réalisation de chacun des k événements. Les p_i vérifient les conditions :

$$0 \leq p_i \leq 1 \qquad \sum_{i=1}^{k} p_i = 1 \tag{B.15}$$

Le résultat des m répétitions peut être exprimé sous forme d'un vecteur $\underline{x}$=$(x_1, x_2, ...x_k)$, dont la composante *ième* est le nombre de fois que l'expérience a donné le résultat i. Evidemment on a :

$$0 \leq x_i \leq m \qquad \sum_{i=1}^{k} x_i = m \tag{B.16}$$

Si les expériences sont répétées dans les mêmes conditions, de manière que tous les tirages puissent être décrits par des v.a. catégorielles i.i.d. de paramètre $\underline{p}$, alors le vecteur $\underline{x}$ est la réalisation d'une variable $\underline{X}$ dite *multinomiale* de paramètres $\underline{p}$ et m :

$$\underline{X} \sim \mathcal{M}\left(\underline{p}, m\right) \tag{B.17}$$

La loi de probabilité d'une v.a. multinomiale est donnée par :

$$\left[\underline{x}|\underline{p}, m\right] = \frac{m!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} p_i^{x_i} \tag{B.18}$$

Les espérances et variances des X_i sont exprimées par les relations :

$$\begin{aligned} \mathbb{E}(X_i) &= mp_i \\ \mathbb{V}(X_i) &= mp_i(1-p_i) \end{aligned} \tag{B.19}$$

B.6 La loi de *Dirichlet*

La loi de *Dirichlet* de taille $k \in \mathbb{N}$ et de paramètre $\underline{\alpha}=(\alpha_1, \alpha_2, ..\alpha_k) \in]0, +\infty[^k$ est la loi du vecteur :

$$\left(\frac{Z_1}{Z_1 + ...Z_k}, \frac{Z_2}{Z_1 + ...Z_k}, ..., \frac{Z_k}{Z_1 + ...Z_k} \right) \tag{B.20}$$

$Z_1, Z_2..., Z_k$ étant des variables aléatoires i.i.d. distribuées selon des lois exponentielles de paramètres $\alpha_1, \alpha_2, ..\alpha_k$ respectivement.

Cette loi est portée par le simplèxe $\mathcal{S}(k, 1)$:

$$\mathcal{S}(k,1) = \left\{ \underline{x} \in]0, +\infty[^k, x_i \geq 0, \sum_{i=1}^{k} x_i = 1 \right\} \tag{B.21}$$

et la densité de probabilité de $(x_1, x_2, ..x_{k-1})$ est exprimée par :

$$\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdot \Gamma(\alpha_2) \cdot ... \cdot \Gamma(\alpha_k)} \left(\prod_{i=1}^{k-1} x_i^{\alpha_i - 1} \right) \left(1 - \sum_{i=1}^{k-1} x_i \right)^{\alpha_k - 1} \tag{B.22}$$

où $\alpha_0 = \sum_{i=1}^{k} \alpha_i$. On écrit alors :

$$\underline{X} \sim \mathcal{D}(\underline{\alpha}) \tag{B.23}$$

Les espérances et variances d'une loi de *Dirichlet* de paramètre $\underline{\alpha}$ sont :

$$\begin{aligned} \mathbb{E}(X_i) &= \frac{\alpha_i}{\alpha_0} \\ \mathbb{V}(X_i) &= \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)} \end{aligned} \tag{B.24}$$

Les lois de *Dirichlet* sont normalement utilisées en statistique bayésienne comme lois *a priori* conjuguées du paramètre $\underline{p}$ d'une loi multinomiale. Avec ce choix, la loi *a posteriori* issue de la formule de Bayes, proportionnelle au

produit entre une densité de probabilité de *Dirichlet* (B.22) et une vraisemblance multinomiale :

$$\prod_{i=1}^{k} p_i^{\alpha_i - 1} \prod_{i=1}^{k} p_i^{x_i} \tag{B.25}$$

est encore une *Dirichlet* de paramètre $\underline{\alpha} + \underline{x}$, en parfaite analogie avec le cas *Bêta-binomiale* (cf. chapitre 5).

Ces lois peuvent être interprétées comme des généralisation de lois *Bêta* à des v.a. multidimensionnelles définies dans l'espace $[0,1]^k$ et pour cette raison sont parfois appelées "*Bêta multidimensionnelles*".

Annexe C

Méthodes de simulation MCMC

C.1 Généralités

Le but des méthodes dites MCMC (*Markov Chain Monte Carlo*) est d'obtenir des tirages aléatoires d'une loi de probabilité $f(x)$, éventuellement multidimensionnelle, dont l'expression est connue à une constante près c :

$$f(x) = c \cdot \varphi(x) \tag{C.1}$$

Ce problème se présente typiquement dans les calculs d'inférence bayésienne, où l'expression de la loi *a posteriori* des paramètres :

$$[\theta|y] = \frac{[\theta] \cdot [y|\theta]}{\int_\Omega [y|\theta] \cdot [\theta]\, d\theta} \tag{C.2}$$

est en générale inconnue, à cause de la difficulté du calcul de l'intégrale à dénominateur. Cette intégrale peut être considérée comme une constante puisqu'elle ne dépend pas des paramètres θ. En revanche, l'expression du numérateur, fonction de θ, est connue, puisqu'il s'agit du produit entre la loi *a priori* de θ et de la vraisemblance du modèle $[y|\theta]$.

Revenons au problème général consistant à réaliser des tirages selon la loi de densité $f(x)$, avec $x \in \Omega$, sous-espace de $\mathbb{R}^m$. On suppose que l'expression de $\varphi(x) = f(x)/c$ est connue.

Un algorithme MCMC est une procédure itérative qui construit à partir d'une valeur initiale x_0, une chaîne de Markov dans l'espace Ω :

$$x_0, x_1, ..., x_i, ...$$

dont la distribution stationnaire est la loi visée $f(x)$.

Alors, une fois vérifié que la chaîne s'est "*suffisamment*" approchée de sa loi limite (c'est-à-dire que les valeurs obtenues sont indépendantes de la valeur initiale x_0) tout x_i peut être considéré comme extrait de la distribution de probabilité $f(x)$.

Dans la suite on décrira les deux principaux algorithmes MCMC (qui ont été utilisés dans le cadre de cette thèse) et on abordera aussi le problème du contrôle de la convergence.

C.2 L'algorithme de Metropolis-Hastings

La procédure itérative de Metropolis-Hastings génère, à partir d'une valeur x_i, la valeur suivante x_{i+1} sur la base d'un algorithme en deux temps :

1. D'abord on choisit une valeur candidate x_* tirée aléatoirement d'une distribution de probabilité $j(y|x_i)$, éventuellement dépendante de x_i. Cette loi est dite *instrumentale*, mais est aussi appelée "*fonction de saut*" parce qu'elle permet à la chaîne de "bouger" dans Ω à partir d'un point donné.
2. En suite le candidat est accepté avec une probabilité $\pi(x_i, x_*)$ définie par :

$$\pi(x_i, x_*) = \min\left(1, \frac{\varphi(x_*) \cdot j(x_i|x_*)}{\varphi(x_i) \cdot j(x_*|x_i)}\right) \tag{C.3}$$

"Accepter" le candidat signifie le choisir comme valeur suivante de la chaîne : $x_{i+1} = x_*$. *A contrario*, si le candidat est refusé, alors la chaîne ne bouge pas de x_i : $x_{i+1} = x_i$.

En pratique si le rapport de la formule C.3 est supérieur à 1 on accepte le candidat. S'il est inférieur à 1 on tire une valeur u d'une loi uniforme $\mathcal{U}(0,1)$ et on définit :

$$\begin{aligned} x_{i+1} &= x_* \quad \text{si } u \leq \pi(x_i, x_*) \\ x_{i+1} &= x_i \quad \text{si } u > \pi(x_i, x_*) \end{aligned} \tag{C.4}$$

Il est relativement simple (Gilks et al., 1996b) de vérifier que la distribution stationnaire de la chaîne ainsi construite est la loi visée $f(x)$. Nous allons présenter cette preuve.

Tout d'abord on remarque que, puisque :

$$\frac{\varphi(x_*)}{\varphi(x_i)} = \frac{f(x_*)}{f(x_i)} \tag{C.5}$$

la probabilité d'acceptation peut être écrite de la manière suivante :

$$\pi(x_i, x_*) = \min\left(1, \frac{f(x_*) \cdot j(x_i|x_*)}{f(x_i) \cdot j(x_*|x_i)}\right) \tag{C.6}$$

Le noyau de transition de la chaîne, exprimant la probabilité $[x_{i+1}|x_i]$ de passage de x_i à x_{i+1}, est défini en fonction de la probabilité d'acceptation (équation (C.6)) par la formule :

$$\begin{aligned}[x_{i+1}|x_i] &= j\left(x_{i+1}|x_i\right) \cdot \pi(x_i, x_{i+1}) + \\ &\quad + \mathbf{1}_{x_{i+1}=x_i}\left(1 - \int_\Omega j\left(y|x_i\right) \cdot \pi(x_i, y)dy\right)\end{aligned} \tag{C.7}$$

Dans l'expression (C.7) le premier terme est relatif à la probabilité d'accepter le candidat proposé par la loi instrumentale, et le deuxième au refus de tous les candidats possibles (pour cette raison on effectue une intégration sur l'espace Ω).

De la formule (C.6) on peut déduire que :

$$f(x_i) \cdot j(x_{i+1}|x_i) \cdot \pi(x_i, x_{i+1}) = f(x_{i+1}) \cdot j(x_i|x_{i+1}) \cdot \pi(x_{i+1}, x_i) \tag{C.8}$$

et la multiplication de $f(x_i)$ et de $[x_{i+1}|x_i]$, compte tenu de C.8 nous fait conclure que :

$$f(x_i) \cdot [x_{i+1}|x_i] = f(x_{i+1}) \cdot [x_i|x_{i+1}] \tag{C.9}$$

Par intégration de la formule (C.9) sur la variable x_i on arrive finalement à déduire que :

$$\int_\Omega f(x_i) \cdot [x_{i+1}|x_i]\, dx_i = f(x_{i+1}) \tag{C.10}$$

La formule (C.10) nous montre que la loi stationnaire de la chaîne est $f(x)$. En fait, dans cette expression, l'intégrale à gauche du signe d'égalité est la loi marginale de x_{i+1}, dans l'hypothèse que x_i est issu de la loi $f(x)$:

$$\int_{\Omega} [x_{i+1}, x_i]\, dx_i$$

Donc la formule (C.10) montre que si x_i est issu de la loi $f(x)$ il en sera ainsi pour x_{i+1} et pour toutes les valeurs successives $x_{i+2}, x_{i+3}, ...$

Ces considérations ne constituent pas une démonstration complète de la validité de l'algorithme de Metropolis-Hastings. Il resterait à démonter la convergence asymptotique de la chaîne vers la loi stationnaire. Des détails ultérieurs peuvent être trouvés dans (Robert, 1996 ; Roberts, 1996 ; Tierney 1996).

L'importance de l'algorithme de Metropolis-Hastings réside dans sa généralité. Sous conditions de régularité peu restrictives, par exemple la positivité du noyau de transition, la convergence de la chaîne vers la distribution visée $f(x)$ est assurée indépendamment du choix de la loi instrumentale.

Cependant, même si la convergence théorique n'est pas mise en cause, en pratique, le choix des lois instrumentales est un moment décisif dans la mise en place de la méthode. Des lois instrumentales trop dispersées qui génèrent des candidats souvent refusés peuvent laisser la chaîne longtemps "bloquée" sur une valeur et demander un nombre d'itérations pratiquement incompatible avec les exigences du modélisateur. Inversement, avec des fonctions de saut peu dispersées la chaîne peut bouger trop lentement. L'effet est le même : trop d'itérations sont nécessaires pour atteindre la convergence.

D'autres algorithmes MCMC peuvent être considérés comme des cas particuliers de l'algorithme de Metropolis-Hastings. Parmi eux l'*Independence Sampler* (Tierney, 1994), avec des fonctions de saut indépendantes de la dernière valeur retenue, l'algorithme de *Metropolis* (Metropolis et al. 1953) avec des fonctions de saut symétriques, et l'algorithme de *Gibbs* (Geman et Geman, 1984 ; Gelfand et Smith., 1990), objet du paragraphe suivant.

C.3 L'algorithme de Gibbs

La méthode dite de l'*échantillonneur de Gibbs* permet d'obtenir des tirages de lois multidimensionnelles $f(\underline{x}) = \left[x_{(1)}, x_{(2)}, \ldots, x_{(m)}\right]$ à partir des lois de chaque composante $x_{(j)}$ du vecteur $\underline{x}$, conditionnellement à toutes les autres :

$$\left[x_{(j)} | x_{(1)}, .., x_{(j-1)}, x_{(j+1)}, ...x_{(m)}\right] \tag{C.11}$$

Pour mettre en œuvre l'algorithme, il faut réaliser, à chaque itération i, m tirages successifs dans les susdites lois conditionnelles selon le schéma suivant :

$$\left|\begin{array}{l} \text{Itération } i \\ \text{1) Tirage de } x_{(1)_i} \text{ de } \left[x_{(1)}|x_{(2)_{i-1}}, x_{(3)_{i-1}}..., x_{(m)_{i-1}}\right] \\ \text{2) Tirage de } x_{(2)_i} \text{ de } \left[x_{(2)}|x_{(1)_i}, x_{(3)_{i-1}}..., x_{(m)_{i-1}}\right] \\ ... \\ \text{m) Tirage de } x_{(m)_i} \text{ de } \left[x_{(m)}|x_{(1)_i}, x_{(2)_i}..., x_{(m-1)_i}\right] \end{array}\right. \tag{C.12}$$

On montre (cf., par exemple, Brooks, 1998 ou Gelman et al., 1995) que cette procédure itérative construit une chaîne de Markov dont le noyau de transition peut être vu comme un cas particulier du noyau de Metropolis-Hastings dont la formule (C.7) représente l'expression générale. Cette chaîne admet comme distribution stationnaire la loi visée $f(x)$.

La convergence est assurée sous des conditions de régularité très peu contraignantes. Tierney (1996) propose une liste d'articles qui abordent ce problème : (Schervish et Carlin, 1992 ; Chan, 1993 ; Roberts et Smith, 1994 ; Roberts et Polson, 1994).

Cet algorithme est très utilisé en statistique bayésienne parce qu'il permet de décomposer le problème du tirage dans une loi *a posteriori* définie dans un espace de dimension m, en m sous-tirages de variables unidimensionnelles (et donc en m sous-problèmes plus simples). L'écriture pratique de ces lois se fait sur la base des relations d'indépendance conditionnelle mises en évidence par la structure graphique du modèle (comme il a été montré, par exemple, dans le chapitre 5).

L'implémentation pratique de cette méthode est particulièrement simple quand il est possible d'obtenir l'expression analytique des lois conditionnelles. Dans le cas contraire, il faut réaliser des pas de Metropolis-Hastings pour obtenir des tirages dans les lois inconnues selon la technique hybride décrite, entre autres, par Brooks (1998) et Gilks (1996).

C.3.1 Le logiciel WinBUGS

Comme l'indique son nom (acronyme de "*Bayesian Inference Using Gibbs Sampler*"), ce programme, développé par l'Unité de Bio-Statistique du MRC (*Medical Research Coucil*) de Cambridge (Royaume Uni), permet de mener les calculs d'inférence bayésienne sur des modèles complexes, avec l'algorithme de

Gibbs. Un des avantages majeurs est la possibilité de définir le modèle statistique directement sous forme de DAG : la conversion du graphe en code de calcul est faite automatiquement par le logiciel (Spiegelhalter et al., 1996).

Sa facilité d'utilisation (Spiegelhalter et al., 2000) et sa souplesse ont contribué à son énorme succès dans la communauté bayésienne. La licence d'utilisation est gratuite et le logiciel peut être téléchargé à partir du site http ://www.mrc-bsu.cam.ac.uk/bugs/.

De nombreux exemples d'application de WinBUGS ont été publiés (une liste non exhaustive est disponible sur le site web du projet). Parmi eux un célèbre cas d'étude dû à Spiegelhalter et al. (1999), concernant l'immunisation d'une population africaine au virus de l'hépatite B, peut être considéré comme un véritable didacticiel : toutes les étapes de la modélisation, de la conceptualisation du problème à l'écriture du code de calcul, puis l'analyse de la convergence, y sont détaillées. Plus récemment, Congdon (2001) fournit plus de 200 exemples d'implémentation pratique sous WinBUGS des principaux modèles utilisés en statistique appliquée.

Si WinBUGS peut être efficacement employé, même pour l'estimation de modèles à structure complexe (notamment hiérarchique), en revanche son utilisation pour les calculs d'inférence de modèles avec un certain nombre de paramètres peut demander des temps de calcul importants quand la structure du modèle ne permet pas l'obtention des expressions analytiques des lois conditionnelles, comme dans le cas du modèle de dégradation des compteurs décrit dans le chapitre 5. C'est pour cette raison qu'on a construit également une procédure de Metropolis-Hastings *ad hoc* sous MatLab.

A titre d'exemple, on conclut cette description avec le code utilisé, sous WinBUGS, pour l'estimation du modèle hiérarchique introduit dans le chapitre 6 pour prendre en compte l'effet du site sur la dégradation des compteurs :

```
# MODELE HIERARCHIQUE DE DEGRADATION
# EN FONCTION DU CONTRAT D'EXPLOITATION
model;
{
# HYPERPRIORS
alpha~dgamma(0.001,0.001)
beta~dgamma(0.001,0.001)
for(j in 1:78) {
```

```
#TIRAGE DES LAMBDA[j] DU PRIOR COMMUN
lambda[j]~dgamma(alpha,beta)
for(i in 1:5) {
# EQUATIONS D'ETAT
P[i,j]<-exp((-lambda[j])*t[i])
# EQUATIONS D'OBSERVATION
y[i,j]~dbin(P[i,j],n[i,j])
}}}
```

C.4 Contrôle de la convergence

On a déjà anticipé que la convergence des méthodes MCMC (Metropolis-Hastings et Gibbs) est, dans la quasi-totalité des cas, assurée. Le problème majeur dans les applications pratiques est de déterminer, sur la base des valeurs calculées avec les algorithmes proposés, quand on peut considérer que "la convergence a été atteinte", ou en d'autres termes, à partir de quelle itération on peut considérer, avec une qualité d'approximation suffisante pour les besoins de l'étude, que les valeurs sont issues de la loi visée $f(x)$.

Des règles empiriques pour calculer, en première approximation, le nombre d'itérations nécessaires sont fournies, entre autres, par Gelman et al. (1995) et Mackay (1999).

Les méthodes rigoureuses de contrôle de la convergence sont de deux types différents. Une première famille de diagnostics est basée sur la mise en place d'une chaîne unique, alors que d'autres méthodes prévoient la réalisation de plusieurs chaînes indépendantes, à partir de points initiaux très dispersés dans l'espace Ω. L'esprit de ces diagnostics est très simple : pour chaque itération on vérifie si les dernières valeurs obtenues de chaque chaîne constituent des échantillons de la même loi. Si c'est le cas, puisque asymptotiquement toutes les chaînes convergent vers $f(x)$, alors on peut conclure que la convergence a été atteinte. Ces techniques *multi-chaînes* sont aujourd'hui particulièrement répandues, et généralement elles sont réputées plus efficaces que les méthodes *mono-chaîne* (Gelman, 1996 ; Brooks et Roberts, 1998 ; Chauveau et Diebolt, 1998).

Une liste des nombreux diagnostics proposés dans les dernières années est donnée par Cowles et Carlin (1996) ou, plus récemment, par Mergensen et al. (1999). On se focalise ici sur une technique très courante initialement proposée

par Gelman et Rubin (1992).

Cette méthode se base sur la comparaison entre la variance *inter-chaînes* (entre les valeurs de chaînes différentes) et la variance *intra-chaînes* (entre valeurs d'une même chaîne). Dans un premier temps, les chaînes subissent l'influence des points initiaux (très dispersés) et la variance *inter-chaînes* est sensiblement supérieure à la variance *intra-chaînes*. Quand le rapport entre ces deux grandeurs s'approche de la valeur asymptotique de 1, alors on peut considérer que toutes les chaînes sont arrivées à convergence.

Si on note $x_{i,j}$ la *ième* valeur de la chaîne j, la méthode de Gelman-Rubin prévoit le calcul des variabilités *intra* et *inter-chaînes* (B et W respectivement) sur la base des variances empiriques des dernières n valeurs obtenues des m chaînes (à partir de l'itération $k > n$), selon la procédure classique de l'analyse de la variance :

$$B = \frac{n}{m-1} \sum_{i=k-n}^{k} (x_{\cdot,j} - x_{\cdot,\cdot}) \tag{C.13}$$

$$W = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{n-1} \sum_{i=k-n}^{k} (x_{i,j} - x_{\cdot,j}) \tag{C.14}$$

où :

$$x_{\cdot,j} = \frac{1}{n} \sum_{i=k-n}^{k} x_{i,j} \qquad x_{\cdot,\cdot} = \frac{1}{m} \sum_{j=1}^{m} x_{\cdot,j} \tag{C.15}$$

La quantité :

$$\widehat{\sigma}_{+}^{2} = \frac{n-1}{n} W + \frac{1}{n} B \tag{C.16}$$

peut être interprétée comme un estimateur de la variance de la loi visée $f(x)$. Gelman et Rubin montrent que $\widehat{\sigma}_{+}^{2}$ surestime systématiquement la variance de la loi $f(x)$ tant que les chaînes n'ont pas atteint la convergence. Le contrôle de la convergence s'appuie sur la statistique $\widehat{R}_{GR}$, définie par la relation :

$$\sqrt{\widehat{R}_{GR}} = \sqrt{\frac{\widehat{\sigma}_{+}^{2}}{W}} \tag{C.17}$$

Cette statistique tend vers 1 pour $n \to +\infty$. En pratique on considère que les chaînes sont arrivées à convergence si $\sqrt{\widehat{R}_{GR}} < 1.2$.

La méthode de Brooks et Gelman (1998) est une variante plus générale de celle de Gelman et Rubin, basée sur les longueurs des intervalles de crédibilité (de niveau fixé 1-α) des dernières n valeurs de chacune des m chaînes.

Si on note par δ_j la longueur de l'intervalle de crédibilité des n valeurs :

$$x_{k-n,j}, x_{k-(n-1),j}, ...x_{k,j}$$

et Δ la longueur de l'intervalle de crédibilité des $n \cdot m$ valeurs mélangées :

$$\begin{aligned}
&x_{k-n,1}, x_{k-(n-1),1}, ...x_{k,1}\\
&x_{k-n,2}, x_{k-(n-1),2}, ...x_{k,2}\\
&...\\
&x_{k-n,m}, x_{k-(n-1),m}, ...x_{k,m}
\end{aligned}$$

on définit la statistique de Brooks-Gelman comme :

$$\widehat{R}_{BG} = \frac{\Delta}{\overline{\delta}} \tag{C.18}$$

où $\overline{\delta}$ est la moyenne empirique des δ_j :

$$\overline{\delta} = \frac{1}{m}\sum_{j=1}^{m}\delta_j \tag{C.19}$$

En pratique, les intervalles de crédibilité sont calculés comme différences entre les percentiles empiriques d'ordre $\alpha/2$ et $1-\alpha/2$ et la valeur limite de $\widehat{R}_{BG}$, en dessous de laquelle on peut figurer que la convergence a été atteinte est de 1.2. Pour les calculs réalisés dans le cadre de cette thèse on a choisi 1-$\alpha = 80\%$ (valeur utilisée aussi par le logiciel WinBUGS).

En définitive, les longueurs des intervalles de crédibilité de niveau 1-α étant fonctions des statistiques d'ordres des échantillons $\{x_{ij}\}$ on peut considérer la méthode de Brooks-Gelman comme une variante *non-paramétrique* du diagnostic de Gelman-Rubin.

Si la variable x est multidimensionnelle, dans la pratique courante, on calcule indépendamment la statistique de convergence pour chaque composante du vecteur $\underline{x}$. On considère que les échantillons de $\underline{x}$ sont issus de la même loi, si chacune de ses composantes vérifie *marginalement* la condition : $\widehat{R}_{BG} < 1.2$.

Annexe D

Eléments de sélection bayésienne de modèle

D.1 Généralités : les facteurs de Bayes

Soit $\mathfrak{M} = \{M_1, M_2, ...M_k\}$ un ensemble fini de modèles (paramétriques) possibles pour expliquer les données y, paramétrés par $\theta_1, \theta_2, ..., \theta_k$ respectivement. En statistique bayésienne le problème de la sélection de modèles en compétition se fait sur la base de la comparaison entre les probabilités *a posteriori* de chacun d'entre eux : le meilleur modèle parmi les k proposés, est celui ayant la probabilité *a posteriori* la plus élevée.

La probabilité *a posteriori* du modèle M_j s'écrit en vertu de la formule de Bayes :

$$[M_j|y] = \frac{[M_j] \cdot [y|M_j]}{\sum_{M_j \in \mathfrak{M}} [M_j] \cdot [y|M_j]} \tag{D.1}$$

Dans cette formule $[M_j]$ est une probabilité *a priori* de la validité de M_j, respectant la propriété :

$$\sum_{M_j \in \mathfrak{M}} [M_j] = 1 \tag{D.2}$$

et le terme $[y|M_j]$:

$$[y|M_j] = \int_{\Omega_j} [y|\theta_j] \cdot [\theta_j] \, d\theta_j \tag{D.3}$$

est la *loi prédictive a priori* de y sous le modèle M_j. Ce terme, comme le font observer Parent et Bernier (2003), peut être aussi interprété comme une *vraisemblance moyenne a priori*.

Revenons, juste un instant, sur le problème d'estimation des paramètres θ_j du modèle M_j (chapitre 5). La loi *a posteriori* $[\theta_j|y]$ exprimée par la formule de Bayes :

$$[\theta_j|y] = \frac{[\theta_j] \cdot [y|\theta_j]}{\int_{\Omega_j} \left[y|\theta'_j\right] \cdot \left[\theta'_j\right] d\theta'_j} \tag{D.4}$$

nous montre que la *vraisemblance moyenne a priori* de la formule (D.3) est le dénominateur de la formule de Bayes "*inférentielle*" (D.4).

Considérons maintenant deux modèles en compétition M_1 et M_2. Le rapport entre les probabilités *a posteriori* $[M_2|y]$ et $[M_1|y]$:

$$\frac{[M_2|y]}{[M_1|y]} = \frac{[M_2]}{[M_1]} \frac{[y|M_2]}{[y|M_1]} \tag{D.5}$$

peut être interprété comme un pari (*a posteriori*) à faveur du modèle M_2, par rapport à M_1. Dans cette formule le rapport :

$$B_{M_2M_1} = \frac{[y|M_2]}{[y|M_1]} = \frac{\int_{\Omega_2} [y|\theta_2] \cdot [\theta_2] \, d\theta_2}{\int_{\Omega_1} [y|\theta_1] \cdot [\theta_1] \, d\theta_1} \tag{D.6}$$

qui modifie le pari *a priori* $[M_2] / [M_1]$ est dit *facteur de Bayes* du modèle M_2 relativement à M_1.

Les facteurs de Bayes sont à la base de la méthode bayésienne de sélection de modèle : ils quantifient la validité d'un modèle relativement à un autre, sur la base de l'observation des données. Des rapports proches de 1 traduisent le fait que les deux modèles sont substantiellement équivalents alors que des valeurs significativement supérieures à 1 montrent que le modèle au numérateur est préférable.

Kaas et Raftery (1994) suggèrent le barème suivant des facteurs de Bayes, basé sur une échelle de valeurs de $2 \ln (B_{M_2M_1})$:

Interprétation de $B_{M_2M_1}$ (Kaas et Raftery, 1994)		
$B_{M_2M_1}$	$2\ln(B_{M_2M_1})$	Evidence a faveur de M_2
1 à 3	0 à 2	Aucune
3 à 12	2 à 5	Positive
12 à 150	5 à 10	Forte
>150	>10	Très forte

Ce barème indicatif peut être valable pour un premier jugement sur la comparaison de deux modèles mais il est loin d'être général. D'autres interprétations des facteurs de Bayes, légèrement différentes, ont été proposées par Jeffreys (1961) et Congdon (2001).

Une intéressante interprétation "*décisionnelle*" des facteurs de Bayes est donnée par Parent et Bernier (2003). Le choix du modèle M_i a des répercussions qu'on peut quantifier avec le coût C_{ij} de la décision d'utiliser M_i pour représenter un certain phénomène, alors que le "*vrai*" modèle (représentatif de l'état de la nature) est M_j.

Pour le cas de 2 modèles en compétition, en imaginant (logiquement) que $C_{11} = C_{22} = 0$, le problème de décision est résumé par la table suivante :

	"Vrai" modèle	
Décision	M_1	M_2
d_1 : Choisir M_1	0	C_{12}
d_2 : Choisir M_2	C_{21}	0

Les coûts espérés *a posteriori* des deux décision sont alors :

$$\begin{aligned} W_y(d_1) &= [M_1|y]\cdot 0 + [M_2|y]\cdot C_{12} = [M_2|y]\cdot C_{12} \\ W_y(d_2) &= [M_1|y]\cdot C_{21} + [M_2|y]\cdot 0 = [M_1|y]\cdot C_{21} \end{aligned} \tag{D.7}$$

et retenir la décision qui mène au coût minimal nous fait conclure qu'il faut choisir le modèle M_2 si :

$$\frac{W_y(d_2)}{W_y(d_1)} = \frac{[M_1|y]}{[M_2|y]}\frac{C_{21}}{C_{12}} = \frac{1}{B_{M_2M_1}}\frac{[M_1]}{[M_2]}\frac{C_{21}}{C_{12}} \leq 1 \tag{D.8}$$

autrement dit, si :

$$B_{M_2M_1} \geq \frac{[M_1]}{[M_2]}\frac{C_{21}}{C_{12}} \tag{D.9}$$

D.2 Le calcul des facteurs de Bayes

Le calcul des facteurs de Bayes se base sur l'évaluation du dénominateur de la formule (D.4). Par simplicité on réécrira cette intégrale de la façon suivante (en supprimant l'indice j relative au modèle M_j) :

$$[y] = \int_{\Omega} [y|\theta] \cdot [\theta] \, d\theta \tag{D.10}$$

Parmi les différentes méthodes proposées, on distingue les techniques analytiques approximées et les techniques de simulation par méthodes MCMC.

Dans la première famille de méthode (Kaas et Raftery, 1994) on trouve notamment des techniques basées sur l'approximation (dite de Laplace) de la loi *a posteriori* par une loi multinormale dont la moyenne est le mode *a posteriori* (pour plus de détails cf. (Gelfand et Dey, 1994)) et une autre qui prévoit l'approximation (dite de Schwartz) du logarithme de $B_{M_1 M_2}$ par la quantité :

$$S = \ln\left(\left[y|\widehat{\theta}_1\right]\right) - \ln\left(\left[y|\widehat{\theta}_2\right]\right) - \frac{1}{2}(d_1 - d_2)\ln(n) \tag{D.11}$$

où $\widehat{\theta}_1$ et $\widehat{\theta}_2$ sont les estimateurs par Maximum de Vraisemblance, d_1 et d_2 les dimensions de θ_1 et θ_2 et n la taille de l'échantillon.

Ici on s'intéresse surtout aux techniques d'estimation par simulation. La méthode de base consiste en un calcul approximé de l'intégrale (D.10) par méthode Monte Carlo. En principe le calcul pourrait être mené sur la base de tirages aléatoires dans la loi *a priori* $[\theta]$. Alors après un nombre important de tirages :

$$\theta^{(1)}, \theta^{(2)}, ..., \theta^{(m)}$$

l'intégrale (D.10) est approximée par la quantité :

$$[y] \simeq \frac{1}{m} \sum_{i=1}^{m} \left[y|\theta^{(m)}\right], \tag{D.12}$$

moyenne des vraisemblances des valeurs échantillonnées. L'inconvénient de cette méthode, qu'on peut qualifier de *naïve* (Parent et Bernier, 2003), est qu'en général la loi *a priori* est beaucoup plus plate que la vraisemblance et, par conséquent, à la plupart des valeurs tirées de cette loi correspondent des valeurs faibles et non significatives de la vraisemblance.

Une procédure nettement plus efficace est celle dite de *Raftery*, qui prévoit des tirages de θ dans la loi *a posteriori* $[\theta|y]$ (Newton et Raftery, 1994).

La technique se fonde directement sur la formule de Bayes (D.4) d'où on peut immédiatement déduire que :

$$\frac{[\theta]}{[y]} = \frac{[\theta|y]}{[y|\theta]} \tag{D.13}$$

L'intégration de la formule (D.13) sur l'espace Ω donne :

$$\int_\Omega \frac{[\theta]}{[y]} d\theta = \int_\Omega \frac{[\theta|y]}{[y|\theta]} d\theta \tag{D.14}$$

Dans cette dernière formule le terme à gauche du signe d'égalité est égal à :

$$\frac{1}{[y]} \int_\Omega [\theta]\, d\theta = \frac{1}{[y]} \tag{D.15}$$

puisque $[y]$ ne dépend pas de θ et l'intégrale sur Ω de la loi *a priori* $[\theta]$ est égale à 1.

Par conséquent :

$$\frac{1}{[y]} = \int_\Omega \frac{1}{[y|\theta]} [\theta|y]\, d\theta \tag{D.16}$$

et alors un estimateur de $1/[y]$ peut être obtenu par intégration Monte Carlo d'inverses de vraisemblances évaluées en correspondance de tirages de θ dans sa loi *a posteriori* $[\theta|y]$:

$$\frac{1}{[y]} \simeq \frac{1}{m} \sum_{i=1}^{m} \frac{1}{\left[y|\theta^{(i)}\right]} \tag{D.17}$$

d'où la *formule de Raftery* :

$$[y] \simeq \left(\frac{1}{m} \sum_{i=1}^{m} \frac{1}{\left[y|\theta^{(i)}\right]} \right)^{-1} \tag{D.18}$$

La mise en œuvre de cette méthode est assez simple. Il suffit, dans les calculs d'inférence par méthode MCMC de *garder de coté*, à chaque itération i, la valeur de la vraisemblance $\left[y|\theta^{(i)}\right]$ (dont le calcul est d'ailleurs nécéssaire pour le déroulement de l'algorithme). Une fois la convergence atteinte, ces valeurs peuvent être employées pour le calcul exprimé par la formule (D.18).

Pour cette raison, malgré un risque d'instabilité (Kaas et Raftery, 1994), cette méthode est très répandue et il s'agit notamment de la technique qu'on a utilisée pour les calculs pratiques dans le cadre de cette thèse.

D'autres méthodes par simulation MCMC sont disponibles. Un aperçu des différentes techniques est donné, par exemple, dans (Raftery, 1996) et (Han et Carlin, 2000).

Parmi elles on signale la procédure proposée par Carlin et Chib (1995) de sélection simultanée entre les k modèles en compétition :

$$M_j \qquad j \in \{1, ..k\}$$

avec des tirages dans l'espace produit :

$$\{1, ..k\} \times \Omega_1 \times ... \times \Omega_k \tag{D.19}$$

réalisés avec un algorithme de Gibbs.

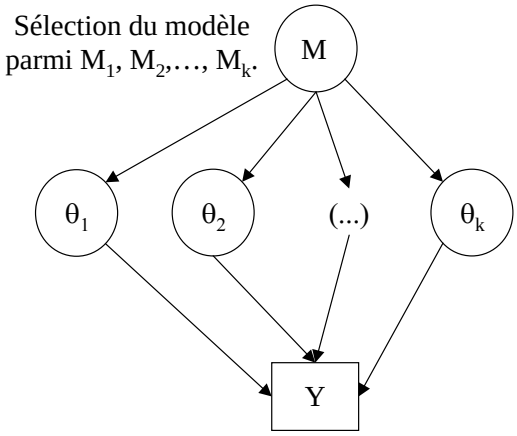


FIG. D.1 – Représentation sous forme de DAG de la procédure de sélection de modèle de Carlin et Chib (1995).

Cette méthode qui permet d'obtenir directement des tirages dans la loi *a posteriori* de j (soit les probabilités *a posteriori* $[M_j|y]$) est dite aussi du "*Super-Modèle*". En effet, tout se passe comme si aux modèles en compétition on superposait un mécanisme de sélection qui, à chaque itération, choisissait le *bon* modèle par un tirage dans un modèle d'urne à k dimensions, les paramètres de ce *super-modèle* étant les probabilités *a posteriori* des k alternatives (figure D.1).

La mise en œuvre de cette méthode demande beaucoup d'attention. En particulier quand on effectue les tirages des θ_i des lois conditionnelles :

$$[\theta_i|\theta_{i'\neq i}, j, y] \begin{cases} \propto [y|\theta_i, j=i]\,[\theta_i|j=i] & \text{si } j=i \\ = [\theta_i|j\neq i] & \text{si } j\neq i \end{cases} \tag{D.20}$$

il est important que la valeur échantillonnée de θ_i soit *plausible* même quand on imagine que le modèle M_i est faux (cas $j \neq i$).

Pour cette raison Carlin et Chib recommandent de choisir les lois $[\theta_i|j \neq i]$ (qu'ils appellent "*pseudo-priors*") comme approximations des lois *a posteriori* $[\theta_i|y]$ obtenues en réalisant séparément les calculs d'inférence pour chacun des k modèles. Des exemples d'implémentation pratique de cette méthode sous WinBUGS sont fournis par Spiegelhalter et al. (1996-b) et Congdon (2001).

D.3 Le problème de choix de modèle du chapitre 7

Les deux modèles en compétition sont, dans ce problème (page 136) :

1. Modèle M_1, paramétré par la matrice de transition (3×3) $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ 0 & \theta_{22} & \theta_{23} \\ 0 & 0 & 1 \end{pmatrix}$$

   et un coefficient ξ qui modifie la premiere ligne de $\boldsymbol{\theta}$ (en rendant plus difficiles les transitions de l'état 1 à l'état 1 lui même) quand les consommations dépassent une valeur limite.

2. Modèle M_2 où l'on introduit aussi un paramètre supplémentaire ψ qui modifie, pour des consommations élevées, la deuxième ligne de la matrice $\boldsymbol{\theta}$ (en réduisant ainsi la probabilité de rester dans l'état 2).

Notons, par simplicité, $\eta_1 = \{\boldsymbol{\theta}, \xi\}$ et $\eta_2 = \{\boldsymbol{\theta}, \xi, \psi\}$ l'ensemble des paramètres de M_1 et de M_2 respectivement.

Dans les deux schémas proposés, la vraisemblance est le produit de deux termes multinomiaux, un relatif aux consommations ordinaires et un autre relatif aux consommations élevées (indice e) :

$$[y|\eta_i] = \prod_{t=0}^{t_{obs}} \frac{n(t)!}{\prod_{j=1}^{3} y_j(t)!} \prod_{j=1}^{3} [x_j(t)]^{y_j(t)} \cdot \frac{n_e(t)!}{\prod_{j=1}^{3} y_{j,e}(t)!} \prod_{j=1}^{3} [x_{j,e}(t)]^{y_{j,e}(t)} \qquad i \in \{1,2\} \tag{D.21}$$

où l'indice i designe le modèle (M_1 ou M_2). Dans cette expression, formellement valable pour les deux modèles, les probabilités des 3 états $[x_j(t)]$ et $[x_{j,e}(t)]$ sont fonction de η_1 sous M_1 et de η_2 sous M_2.

On observe aussi que dans la formule (D.21) seul les termes dépendant des $[x_j(t)]^{y_j(t)}$ et des $[x_{j,e}(t)]^{y_{j,e}(t)}$ sont fonctions des paramètres des deux modèles et le terme :

$$K = \prod_{t=0}^{t_{obs}} \frac{n(t)!}{\prod_{j=1}^{3} y_j(t)!} \frac{n_e(t)!}{\prod_{j=1}^{3} y_{j,e}(t)!} \tag{D.22}$$

peut être regardé comme une simple constante multiplicative (invariante quand on passe de M_1 à M_2).

Puisque dans la formule (D.6) les deux vraisemblances moyennes *a priori* sont proportionnelles à K, on en déduit que dans les calculs pratiques du facteur de Bayes on peut se débarrasser de ce produit, un peu fastidieux à cause des termes factoriels, et travailler avec les quantités (qu'on pourrait qualifier de "*pseudo-log-vraisemblances*") :

$$\mathcal{L}_i(y, \eta_i) = \sum_{t=0}^{t_{obs}} \left(\sum_{j=1}^{3} y_j(t) \cdot \ln\left([x_j(t)]\right) + \sum_{j=1}^{3} y_{j,e}(t) \cdot \ln\left([x_{j,e}(t)]\right) \right) \tag{D.23}$$

plus simples à manipuler, liées aux vraisemblances $[y|\eta_i]$ par les relations :

$$[y|\eta_i] = K \cdot \exp\left(\mathcal{L}_i(y, \eta_i)\right) \tag{D.24}$$

Pour chacune des 3 valeurs de la variable "agressivité" on a estimé les lois *a posteriori* des paramètres $[\eta_1|y]$ et $[\eta_2|y]$ dans les 2 modèles M_1 et M_2 avec le logiciel WinBUGS. Pour prévenir les risques d'instabilité dans le calcul du facteur de Bayes on a calculé les *vraisemblances moyennes a priori* avec la formule de *Raftery* (D.18) sur la base d'un grand nombre (20 000) de valeurs de $\mathcal{L}_1(y, \eta_1)$ et $\mathcal{L}_2(y, \eta_2)$.

Le calcul des facteurs de Bayes pour les 3 cas examinés donne :

	Agress. 1	Agress. 2	Agress. 3
$B_{M_2 M_1}$	0.297	0.501	0.248

Comme déjà anticipé dans le chapitre 7, ces résultats, qui montrent une légère évidence en faveur du modèle M_1, justifient l'utilisation de ce dernier, plus parcimonieux, par rapport à M_2 qui a un paramètre de plus.

Annexe E

Article accepté par la Revue de Statistique Appliquée

Modélisation bayésienne du vieillissement des compteurs d'eau par mélange de classes d'appareils de différents états de dégradation

Alberto Pasanisi*, Eric Parent***

** Ecole Nationale du Génie Rural, des Eaux et des Forêts, Lab. GRESE.*
19, avenue du Maine. 75732 Paris Cedex 15.
*** Compagnie Générale des Eaux, Direction Technique.*
18, boulevard Malesherbes. 75008 Paris.

Résumé

Les compteurs d'eau, en vieillissant, fournissent une mesure de plus en plus imprécise de la consommation d'eau. Cette dégradation se traduit généralement par un sous-comptage. Ce phénomène est source de problèmes pour les distributeurs d'eau qui ont mis en place des stratégies de gestion des parcs compteurs ayant comme objectif la réduction des pertes économiques (représentées par les volumes d'eau non facturés) et le respect d'une politique de comptage équitable entre les différents usagers. Toute stratégie nécessite préalablement la compréhension du mécanisme de dégradation et la quantification du sous-comptage. Dans cet article le vieillissement des compteurs est décrit à travers un modèle dynamique à états discrets, représentant chacun une certaine qualité métrologique. Ce modèle, couplé avec l'observation des erreurs de mesure à l'intérieur de chaque état, permet l'estimation notamment du taux de compteurs défaillants et de l'évolution de la précision de la mesure en fonction de la durée de service du

dispositif. L'estimation des paramètres du modèle et la prédiction des valeurs des grandeurs d'intérêt pratique, ont été réalisées dans un cadre bayésien, avec l'utilisation de techniques de simulation MCMC. Les résultats montrent un bon comportement général des compteurs examinés dans l'exemple proposé mais aussi une incertitude sensible sur l'estimation de certaines grandeurs.

Mots clés : Compteurs d'eau, dégradation, modèles dynamiques, inférence bayésienne, modèles graphiques, simulation MCMC.

Abstract

Water meters give a more and more inaccurate measure of water consumption, when getting older. Such a degradation generally gives rise to an underestimation of consumption. That originates several problems to water distribution companies who have developed management strategies in order to attempt two different goals : the reduction of financial losses (caused by unaccounted-for water) and equity between customers. Every strategy needs, as pre-requisite, the understanding of the degradation process and the evaluation of the loss of accuracy. In this article the ageing of meters is described by a dynamic discrete state model, every state of which characterises a given measurement quality. This model, together with the observation of measurement errors within each state, allows to evaluate the ill-behaved meters rate and the accuracy, as a function of the operating age of the device. Model parameters estimation and prediction of practically interesting quantities have been made by MCMC simulation techniques. Results show that meters examined hereby have globally a good behaviour, but a noticeable uncertainty still remains on estimates and predictions of several parameters.

Key-words : Water meters, degradation, dynamic models, Bayesian inference, graphical models, MCMC simulation.

E.1 Introduction

La dégradation métrologique des compteurs est source de problèmes pour les distributeurs d'eau. Les stratégies de gestion du parc-compteurs, mises en place dans les dernières années, poursuivent deux objectifs majeurs : améliorer le rendement global afin de réduire les pertes économiques du distributeur d'eau représentées par les volumes non facturés et, en même temps, assurer un comptage (et donc une facturation) équitable entre les consommateurs.

Pour avoir une idée des enjeux économiques on peut chiffrer, par exemple, le manque à gagner de la *Générale des Eaux* (environ 2.1 milliards de m^3 d'eau

distribués par an et 6 millions de compteurs) : par rapport à une situation idéale de comptage parfait, il s'établit autour de 45 millions d'Euros par an.

Le cadre de la gestion d'un parc de compteurs d'eau est largement décrit dans la litterature technique, par exemple (*Newman et Noss, 1982*), (*Grau, 1985*), mais rares sont les exemples où les critères de gestion sont formulés explicitement en termes mathématiques. Des procédures d'optimisation ont été développées d'abord dans un cadre déterministe (*Noss et al., 1987*), puis dans un cadre stochastique (*Lund, 1988*).

Toute procédure de gestion nécessite préalablement la compréhension du mécanisme de dégradation de la qualité métrologique des compteurs en fonction de plusieurs facteurs explicatifs, et notamment de leur durée de service, afin de pouvoir estimer les volumes d'eau non facturés. Ces estimations représentent le point de départ pour l'évaluation des opportunités technico-économiques de remplacement. En général, les conditions d'exploitation sont variables et les types de compteurs installés sont multiples. Pour les très gros consommateurs, compte tenu de leur faible nombre et des enjeux économiques, on peut ima-giner un suivi au cas par cas (*Van der Linden, 1998*), alors que pour la grande majorité des usagers on a recours à l'approche statistique.

Cet article présente un modèle statistique de vieillissement qui s'inscrit dans le cadre de la gestion optimale du parc compteurs de la *Générale des Eaux*.

Le papier s'articule en quatre parties.

La première partie donne des définitions de base sur les compteurs d'eau, nécessaires pour comprendre la problématique de l'étude, et une description des données à disposition.

La deuxième partie du papier représente la dégradation des compteurs à travers un modèle markovien à états discrets. Ce modèle, associé avec l'observation des déposes des compteurs défaillants et des distributions des rendements à l'intérieur d'un même état, permet de donner des indications directement exploitables pour la gestion des parcs.

Le modèle se met naturellement sous forme de *DAG* (*Directed Acyclic Graph*) et cette représentation facilite l'estimation bayésienne des paramètres notamment grâce à l'algorithme de Gibbs (*Spiegelhalter et al., 1996*). Dans la troisième partie on présente les résultats d'inférence et de calcul prédictif obtenus à l'aide du logiciel *WinBUGS* (*Spiegelhalter et al., 1999*). Ces résultats montrent une incertitude sensible sur les estimations de certaines grandeurs et notamment sur

le taux de compteurs à métrologie insatisfaisante. On note aussi le bon comportement général des appareils examinés, puisque, après 20 ans de service, de l'ordre de 85% des compteurs présentent encore une métrologie correcte.

La dernière partie est consacrée aux perspectives du travail, notamment la possibilité d'introduire des covariables explicatives et la validation du modèle.

Dans la suite on utilisera la notation *entre crochets* pour les lois de probabilité (*Gelfand et Smith, 1990*) :

$$[A] : \text{ Probabilité de l'événement aléatoire } A$$

Si X est une variable aléatoire à valeurs $x \in \Omega \subseteq \mathbb{R}^d$, alors on utilisera la notation simplifiée $[x]$ pour représenter la probabilité $[X = x]$.

E.2 Comment se dégrade un compteur et quelles sont les conséquences

E.2.1 Définitions de base

Comme tous les instruments de mesure, les compteurs d'eau sont susceptibles d'erreur : sur un branchement, le volume enregistré (v_{enr}), et donc facturé, est généralement différent du volume effectivement consommé ($v_{réel}$).

Une description détaillée des principes de fonctionnement des différents types de compteurs est donnée dans (*Troskolanski, 1963*) et (*Carlier, 1968*).

L'étude statistique de ce papier, se limitera exclusivement aux données des compteurs dits *volumétriques* (largement majoritaires en France) caractérisés par la présence d'un organe de mesure qui se déplace sous l'effet de la poussée hydrodynamique, refoulant un volume déterminé d'eau à chaque tour.

L'erreur de mesure d'un compteur (e) est obtenue en divisant la différence entre le volume enregistré (v_{enr}) et le volume réellement écoulé ($v_{réel}$) par ce dernier.

$$e = \frac{v_{enr} - v_{réel}}{v_{réel}}. \tag{E.1}$$

La *courbe métrologique* (dite aussi "signature" métrologique) est la représentation graphique (figure 1) de la relation entre l'erreur relative de mesure (e) et le débit circulant (q) .

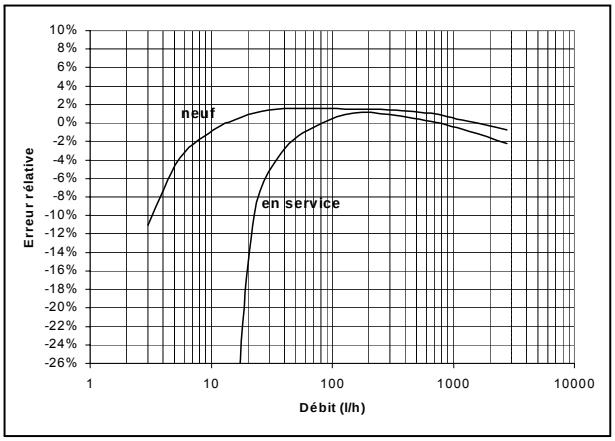


FIG. E.1 – Courbes métrologiques d'un compteur neuf et en service.

Quand le débit est faible, le transfert d'énergie mécanique entre le courant et les organes de mesure est plus difficile et donc les erreurs de mesure sont négatives et très importantes en valeur absolue, pouvant atteindre la totalité (-100%). À mesure que le débit augmente, l'erreur diminue en valeur absolue jusqu'à atteindre la "zone haute" de la courbe avec des valeurs très faibles et parfois positives. Cette plage de fonctionnement apparait nettement sur la figure 1.

Le vieillissement d'un compteur entraîne une dégradation de la qualité de la mesure. Dans la quasi-totalité des cas, la largeur de la plage de fonctionnement diminue et donc un sous-comptage se manifeste (figure 1).

Les courbes métrologiques sont obtenues expérimentalement dans des laboratoires spécialisés, en faisant circuler dans le compteur un volume connu d'eau à débit constant. La courbe est alors construite point par point avec les résultats des essais aux différents débits d'étalonnage.

Le rendement d'un compteur (r) est le rapport entre le volume enregistré et le volume consommé :

$$r = \frac{v_{enr}}{v_{réel}} = 1 + \sum_i e_i x_i \tag{E.2}$$

où e_i et x_i sont les erreurs de mesure et les proportions de la consommation ayant lieu au débit q_i. Pour le calculer, il faut superposer, le long de l'axe horizontal des débits, la courbe métrologique du compteur avec la répartition de la consommation sur les différents débits (*histogramme de consommation*).

qui dépend des modalités d'utilisation de l'eau. Par exemple, dans la figure 2, le premier histogramme de consommation a été enregistré dans une maison individuelle tandis que le second caractérise un immeuble de 50 appartements.

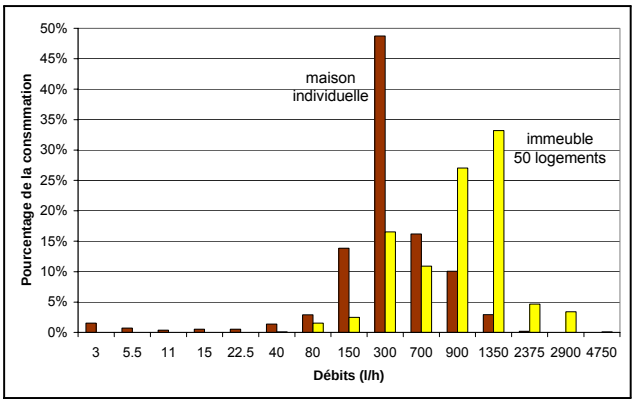


FIG. E.2 – Histogrammes de consommation.

Les calculs développés dans la suite sont relatifs à des histogrammes "types" de consommation qu'on suppose complètement connus et invariants dans le temps.

Le rendement est le paramètre le plus important pour les gestionnaires parce qu'il permet le calcul direct de la perte économique par eau non facturée. Dans les rares études disponibles, l'évolution du rendement en fonction de l'âge est soit reportée sous forme de tableau (*AWWA, 1966*), (*Grau, 1985*), soit est empiriquement considérée comme une fonction linéaire, (*Newman et Noss, 1982*).

E.2.2 Classification des compteurs selon leur qualité de mesure

L'approche suivie dans cette étude s'appuie sur un modèle de vieillissement qui prévoit le passage par 4 états métrologiques, de qualité décroissante $\mathcal{E}_1$, $\mathcal{E}_2$, $\mathcal{E}_3$, $\mathcal{E}_4$.

La définition des quatre états est inspirée par la réglementation actuelle et ses développements (*Costes et Pia, 2000*).

Le décret n. 76-130 du 29/01/1976 (qui reprend la directive *CEE* n. 75/33) fixe les erreurs maximales tolérées (*EMT*) en fonction du débit pour les compteurs neufs :

$\pm 5\%$ entre un *débit minimal* q_{min} et un débit dit *de transition* q_t

$\pm 2\%$ entre le *débit de transition* et le *débit maximal* de fonctionnement q_{max}.

Ces valeurs sont doublées pour les compteurs en service. Les débits q_{min} et q_t sont définis en fonction du débit nominal (q_n), caractéristique du compteur, et de sa classe métrologique (A, B ou C en ordre croissant de précision). La figure 3 montre les "canaux de tolérance" pour des compteurs de débit nominal $q_n = 1.5$ m^3/h de classe C (les compteurs domestiques les plus utilisés en France).

La définition des états métrologiques part de la constatation que les compteurs dont les erreurs de mesure à tous les débits d'étalonnage sont inférieures aux *EMT* ont aussi des rendements très proches de 1. On réserve l'état $\mathcal{E}_1$ aux compteurs qui respectent la conformité aux *EMT* sur toute la gamme de débits. L'état $\mathcal{E}_2$ est caractéristique des compteurs ayant une métrologie imparfaite mais encore acceptable. On a décidé de l'assigner aux compteurs dont la courbe métrologique sort en quelques points des canaux de tolérance mais qui respectent les *EMT* dans la gamme des débits entre $0.2q_n$ et $0.9q_n$. La raison de ce choix est que le projet d'arrêté concernant la vérification périodique des compteurs en service (*Costes et Pia*, 2000) prévoit deux essais d'exactitude : le premier à un débit compris entre $0.8q_n$ et q_n et le deuxième à un débit compris entre $0.1q_n$ et $0.3q_n$ (pour q_n <10 m^3/h). Les compteurs qui se trouvent dans l'état $\mathcal{E}_3$ présentent donc des rendements très mauvais et sont aussi potentiellement non-conformes. Enfin dans l'état $\mathcal{E}_4$ on a mis les compteurs bloqués à tout débit qui n'enregistrent pas.

La figure 3 montre les courbes métrologiques de trois compteurs domestiques ($q_n = 1.5$ m^3/h) respectivement dans les états $\mathcal{E}_1$, $\mathcal{E}_2$, $\mathcal{E}_3$.

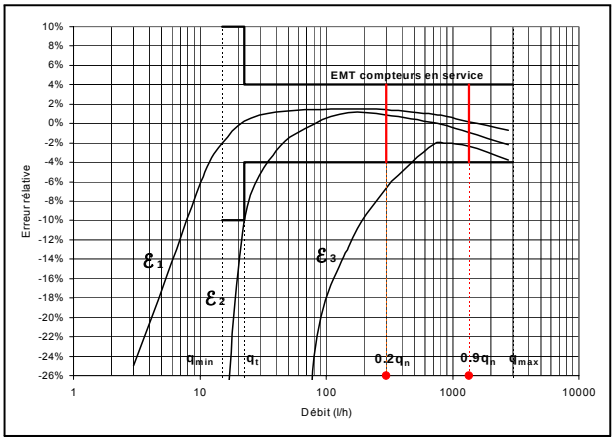


FIG. E.3 – "Etats métrologiques"

E.2.3 Les données

Deux sources différentes de données sont disponibles pour l'étude.

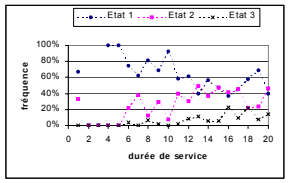
D'une part on utilise les résultats d'étalonnage de compteurs en service, réalisés dans les laboratoires d'essai du Groupe Vivendi Water (*base métrologique*) et d'autre part on se sert des informations issues d'un outil interne de gestion des parcs de compteurs (*ICBC*) pour analyser les phénomènes de blocages. Le recours aux deux bases de données est nécessaire puisque l'état absorbant $\mathcal{E}_4$ est pratiquement inobservé dans la *base métrologique*.

L'exemple détaillé dans la suite de cet article concerne 682 essais de compteurs volumétriques de classe C, type *Volumag,* $q_n = 1.5\ m^3/h$, d'âges compris entre 6 et 20 ans. À partir de la courbe métrologique de chaque compteur, on peut déterminer son état et son rendement. Les individus sont en suite regroupés et dénombrés par état et par âge pour obtenir les $n_i(t)$, soit le nombre de compteurs qui se trouvent dans l'état $\mathcal{E}_i$ à l'âge t. Leur rendements sont utilisés pour estimer les paramètres des lois de probabilités des rendements à l'intérieur de chaque état.

La figure 4 décrit les données dont on s'est servi pour les calculs d'inférence.

Données utilisées pour l'exemple: composition des échantillons d'âges différents et observation des blocages.

	Base métrologique			Base ICBC	
âge	Nombre de compteurs dans l'état 1	Nombre de compteurs dans l'état 2	Nombre de compteurs dans l'état 3	blocages observés	effectif installé
1	4	2	0	25	15498
2	0	0	0	57	18703
3	0	0	0	62	19173
4	1	0	0	54	19281
5	8	0	0	12	4250
6	38	11	2	6	2780
7	5	3	0	8	5033
8	26	4	2	11	5438
9	33	14	1	12	6580
10	12	1	0	16	9408
11	25	17	1	46	23032
12	35	17	5	96	40249
13	44	54	12	317	138153
14	29	19	3	162	50991
15	17	17	2	92	22806
16	23	26	14	57	9921
17	15	15	3	128	30653
18	16	6	6	111	32382
19	35	12	4	107	34621
20	17	20	6	94	32967



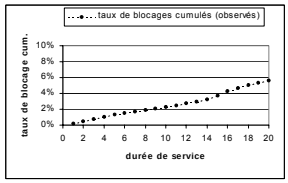


FIG. E.4 – Données utilisées pour l'exemple.

E.3 Le modèle de dégradation

Un mécanisme markovien homogène de dégradation a été imaginé sur la base de considérations techniques (figure 5). En effet des tests d'endurance réalisés

par des fabricants sur banc d'essai montrent que le "vieillissement naturel" des compteurs est très lent et ne provoque de changement d'état, à partir de $\mathcal{E}_1$, qu'après des périodes extrêmement longues. Les changements d'état sont donc essentiellement liés à des accidents de parcours.

En outre, il est réaliste de faire l'hypothèse que la dégradation est irréversible et donc la chaîne a pour état absorbant $\mathcal{E}_4$. Le modèle est à temps discret, et l'unité temporelle est une année. Les éléments de la matrice de transition θ_{ij} sont les probabilités de passage de l'état $\mathcal{E}_i$ (âge *t-1*) à l'état $\mathcal{E}_j$ (âge t).

Le vecteur ligne $\mathbf{P}(t)=\{P_1(t), P_2(t), P_3(t), P_4(t)\}$ des probabilités des quatre états, à l'âge t, est exprimé en fonction de $\mathbf{P}(t-1)$ par l'équation :

$$\mathbf{P}(t) = \mathbf{P}(t-1)\cdot\mathbf{\Theta} \tag{E.3}$$

où $\mathbf{\Theta}$ est la matrice de transition.

A l'intérieur de chaque état, le rendement suit une loi de probabilité caractéristique de l'état, dont les paramètres sont éventuellement fonction de l'âge. Le couplage entre le modèle de dégradation et les lois de probabilité des rendements permet de modéliser l'évolution de la métrologie d'un ensemble de compteurs par mélange (en proportions variables avec l'âge) d'individus appartenant aux différents états.

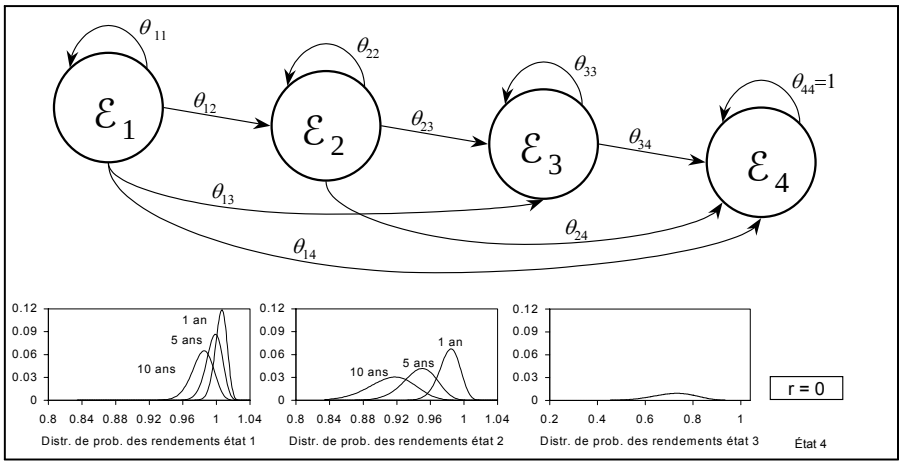


FIG. E.5 – Mécanisme markovien de dégradation.

E.3.1 Observation des rapports entre les différents états dans la base métrologique

Dans la base métrologique, on trouve dans un état donné des compteurs d'âge divers. L'essai d'un compteur est une mesure *destructive*, parce que l'appareil testé n'est pas remis en service. Il faut donc utiliser la composition des échantillons d'individus du même âge, pour estimer indirectement les paramètres du modèle. Le problème majeur est que dans les données métrologiques on ne rencontre que très peu de compteurs dans l'état $\mathcal{E}_4$. Cela est dû au fait que les compteurs bloqués sont repérés facilement sur le terrain par les releveurs et remplacés aussitôt. Il y a donc une absence de représentativité de l'échantillon pour l'état $\mathcal{E}_4$. Par conséquent, le nombre d'individus dans la base métrologique ne permet aucune estimation sérieuse de $P_4(t)$.

Pour ce qui concerne les états $\mathcal{E}_1$, $\mathcal{E}_2$, $\mathcal{E}_3$, si on note $n_i(t)$, le nombre de compteurs qui se trouvent dans l'état $\mathcal{E}_i$ à l'âge t, et $N(t) = n_1(t) + n_2(t) + n_3(t)$ la taille de l'échantillon des compteurs non bloqués d'âge t, le vecteur $\mathbf{n}(t)=\{n_1(t), n_2(t), n_3(t)\}$ suit une loi multinomiale de paramètres $\mathbf{W}(t)$ et $N(t)$:

$$[\mathbf{n}(t)|\mathbf{P}(t)] = \frac{N(t)!}{n_1(t)!n_2(t)!n_3(t)!} W_1(t)^{n_1(t)} W_2(t)^{n_2(t)} W_3(t)^{n_3(t)} \qquad \text{(E.4)}$$

où la probabilité conditionnelle d'être dans l'état $\mathcal{E}_i$ (sachant que le compteur n'est pas bloqué) s'écrit :

$$W_i(t) = \frac{P_i(t)}{\sum_{j=1}^{j=3} P_j(t)} \quad (i = 1, 2, 3) \qquad \text{(E.5)}$$

Concernant l'état initial des compteurs, on imagine que pour $t = 0$ tous les appareils se trouvent dans l'état $\mathcal{E}_1$, c'est-à-dire : $\mathbf{P}(0) = \{1, 0, 0, 0\}$. En effet à la sortie d'usine les compteurs doivent nécessairement respecter les *EMT* des compteurs neufs en correspondance des débits q_{min}, q_t et q_{max} et le respect de ces conditions implique aussi le respect des critères d'appartenance à l'état $\mathcal{E}_1$.

E.3.2 Observation des déposes de compteurs bloqués

Pour estimer l'occurrence de l'état $\mathcal{E}_4$, on utilise les informations contenues dans la base *ICBC*. En fait on peut imaginer, en première approche qu'un compteur qui se bloque dans la période (*t-1, t*) est détecté dans l'année t et remplacé avec probabilité p_r dans l'année t même.

La probabilité de blocage $p_b(t)$ dans la période (*t-1, t*) se calcule en fonction des probabilités d'état et des probabilités de transitions :

$$p_b(t) = P_1(t-1)\cdot\theta_{14} + P_2(t-1)\cdot\theta_{24} + P_3(t-1)\cdot\theta_{34} = P_4(t) - P_4(t-1) \quad \text{(E.6)}$$

Dans l'état actuel de la base ICBC, on n'est pas capable d'observer toutes les déposes, et en outre leur causes ne sont pas toujours correctement renseignées. Une partie des remplacements a lieu à titre "préventif" : le compteur fonctionne mais le gestionnaire pense qu'il existe un intérêt technico-économique à le remplacer. D'autres compteurs, par contre, sont remplacés parce que détectés défaillants (bloqués, fuyards, illisibles, gelés) ou pour des raisons administratives (résiliation du contrat).

Dans notre étude, seuls les compteurs déposés pour cause de blocage nous intéressent, les autres causes n'étant aucunement liées à la dégradation métrologique. On note p_{ob} la probabilité qu'une dépose de compteur soit observée et correctement renseignée dans la base *ICBC*.

Le nombre $N_b(t)$ de compteurs d'âge t, déposés pour blocage et renseignés dans la base *ICBC* suit alors une loi binomiale de paramètres $b(t) = p_b(t)\cdot p_{ob}\cdot p_r$ et $N_C(t)$, ce dernier étant le nombre de compteurs installés d'âge t :

$$[N_b(t)|P_4(t), P_4(t-1)] = \binom{N_C(t)}{N_b(t)} b(t)^{N_b(t)}(1-b(t))^{(N_C(t)-N_b(t))} \quad \text{(E.7)}$$

La valeurs de p_r utilisée dans les calculs ci-après est égale à 1. On suppose donc que la probabilité qu'un compteur bloqué ne soit pas remplacé aussitôt est négligeable.

Concernant p_{ob}, on a choisi une valeur de 0.58, sur la base de la comparaison entre le nombre d'observations recensées correctement dans la base *ICBC*, relatives à l'année 2001 et le nombre de compteurs neufs posés par les exploitants la même année. On pourra, par la suite, étudier la robustesse des résultats vis à vis de la valeur choisie pour p_{ob} et p_r.

Le modèle, paramétré par les coefficients de la matrice $\mathbf{\Theta}$ est relié aux grandeurs observables par les équations d'observation (4) et (7) et par l'équation dynamique (3).

E.3.3 Observation des rendements des compteurs en fonction de l'âge et de l'état

Pour décrire la variabilité du rendement r on a supposé que, pour chaque classe, ce dernier suit une loi de type Bêta. Ces lois sont adaptées au cas où la variable aléatoire est bornée et unimodale. En effet, pour les compteurs volumétriques, la dégradation est toujours synonyme de sous-comptage et donc, il est impossible d'avoir des compteurs qui, au long de leur vie, ont un rendement qui augmente avec l'âge. Grâce aux histogrammes de consommations et aux *EMT*, on sait que la valeur théorique maximale qu'un compteur de ce type peut atteindre est 1.04.

La variable aléatoire $z = r/1.04$ est donc comprise entre 0 et 1. On fait l'hypothèse qu'elle est distribuée selon une loi Bêta standard de paramètres α et β.

$$[z_i|\alpha_i, \beta_i] = \frac{1}{B(\alpha_i, \beta_i)} z_i^{(\alpha_i - 1)} (1 - z_i)^{(\beta_i - 1)} \qquad (i = 1, 2, 3) \tag{E.8}$$

Les paramètres de cette loi sont fonction de l'âge pour les états $\mathcal{E}_1$ et $\mathcal{E}_2$, alors que pour l'état $\mathcal{E}_3$ ils sont constants. Cette dernière hypothèse traduit le comportement erratique des compteurs défaillants dont le rendement ne semble pas dépendre de l'âge (un jeune défaillant peut être bien pire qu'un vieux). Evidemment le problème ne se pose pas pour les compteurs de l'état $\mathcal{E}_4$ qui ont tous un rendement égal à 0.

La dégradation du rendement pour les états $\mathcal{E}_1$ et $\mathcal{E}_2$ est décrite par les équations :

$$\ln(\alpha_i(t)) = \alpha_i^0 - \alpha_i^1 \cdot t \tag{E.9}$$

$$\ln(\beta_i(t)) = \beta_i^0 - \beta_i^1 \cdot t \qquad (i = 1, 2) \tag{E.10}$$

où $\alpha_i^0, \alpha_i^1, \beta_i^0, \beta_i^1 > 0$. Dans notre cas particulier (voir annexe), les équations (E.9) et (E.10) donnent lieu à des distribution de probabilité unimodales à espérance décroissante et à variance croissante en fonction du temps t (figure 5).

E.3.4 Assemblage des différents modèles dans un cadre prédictif

Les résultats de l'inférence sont utilisés dans un cadre prédictif pour obtenir la distribution prédictive des rendements des compteurs "survivants" (non dé-

posés) d'âge donné t. C'est un mélange, de proportions variables avec l'âge, des rendements des 3 classes :

$$\widetilde{r}(t) = r_1(t) \cdot \gamma_1(t) + r_2(t) \cdot \gamma_2(t) + r_3 \cdot \gamma_3(t). \tag{E.11}$$

Les r_i sont évalués à partir des équations (E.8), et $\boldsymbol{\gamma}(t)$ est la réalisation d'un tirage multinomial de paramètres ($\mathbf{W}(t)$, 1) :

$$\boldsymbol{\gamma}(t) = \{\gamma_1(t), \gamma_2(t), \gamma_3(t)\} = \begin{array}{l} \{1,0,0\} \text{ avec probabilité } W_1(t) \\ \{0,1,0\} \text{ avec probabilité } W_2(t) \\ \{0,0,1\} \text{ avec probabilité } W_3(t) \end{array} . \tag{E.12}$$

Une autre grandeur d'intérêt pour le distributeur est le "*rendement moyen*" du compteur, qui exprime la dégradation moyenne d'un ensemble de compteur, défini par la relation :

$$\widetilde{r}_{moy}(t) = r_1(t) \cdot W_1(t) + r_2(t) \cdot W_2(t) + r_3 \cdot W_3(t) \tag{E.13}$$

L'évolution de cette variable en fonction de l'âge est aussi dite "*loi de vieillissement*" d'un parc de compteurs exempt d'appareils bloqués.

E.4 Inférence

E.4.1 Représentation du modèle statistique sous forme de *DAG*

Le modèle statistique, décrit dans le paragraphe précédent peut être représenté sous forme de graphe. Dans cette schématisation, de plus en plus répandue dans la communauté bayésienne, les variables sont représentées avec des nœuds éventuellement liés entre eux par des connexions exprimant des relations de nature déterministe ou probabiliste. L'absence de liens entre deux variables représente leur indépendance conditionnelle.

Les liens peuvent être de deux types : directs et indirects (figure 6).

Un lien direct entre deux variables A et B, représenté par une flèche directe de A à B, exprime le fait que B dépend de A (on dit aussi que A est un *parent* de B), alors qu'un lien indirect représente plutôt une corrélation entre les deux variables. Par exemple (*Cowell et al., 1999*) si A est la variable binaire "*Fumeur ?*" (Oui ou

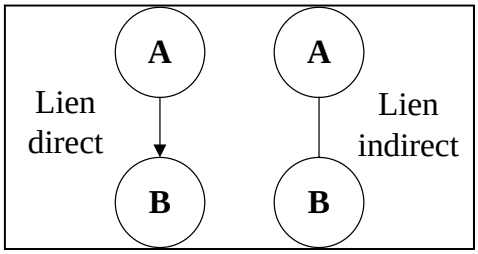


FIG. E.6 – Types de liens entre 2 variables.

Non) et B est la variable "*Stress ?*" (p.ex. mesuré par la tension artérielle), un lien direct entre A et B traduit l'assertion que fumer a un effet sur la tension artérielle. Vice versa, un lien indirect exprime une association entre les deux variables, non explicitée dans le modèle : par exemple un gène inconnu qui prédispose, en même temps, à l'habitude de fumer et au stress.

Un *DAG* (*Directed Acyclic Graph*) est un graphe d'influence où tous les liens sont directs et qui ne contient pas de boucles. De cette manière, la hiérarchie des parents et des descendants est définie sans ambiguïté. Les relations d'indépendance conditionnelle représentées avec les *DAG* sont exprimées par la "*Local Directed Markov Property*" : toute variable ξ_j est conditionnellement indépendante de ses non descendants, étant donné ses parents :

$$[\xi_j|nd(\xi_j), pa(\xi_j)] = [\xi_j|pa(\xi_j)]. \tag{E.14}$$

On montre (*Cowell et al., 1999*) que l'équation (15) est valide si et seulement si la loi de probabilité conjointe de toutes les variables se factorise de la manière suivante :

$$[\xi_1, \xi_2, ..., \xi_d] = \prod_{i=1}^{d} [\xi_i|pa(\xi_i)]. \tag{E.15}$$

L'équation (E.15) simplifie énormement les calculs d'inférence dans un cadre bayésien. La figure 7 montre la représentation du modèle statistique de vieillissement des compteurs sous forme de *DAG*.

Cette schématisation montre que le modèle statistique proposé est l'assemblage de deux blocs différents :

1. Un modèle dynamique de population de paramètre $\mathbf{\Theta}$ dans lequel les probabilités d'occurrence des 4 états métrologiques jouent un rôle de variables

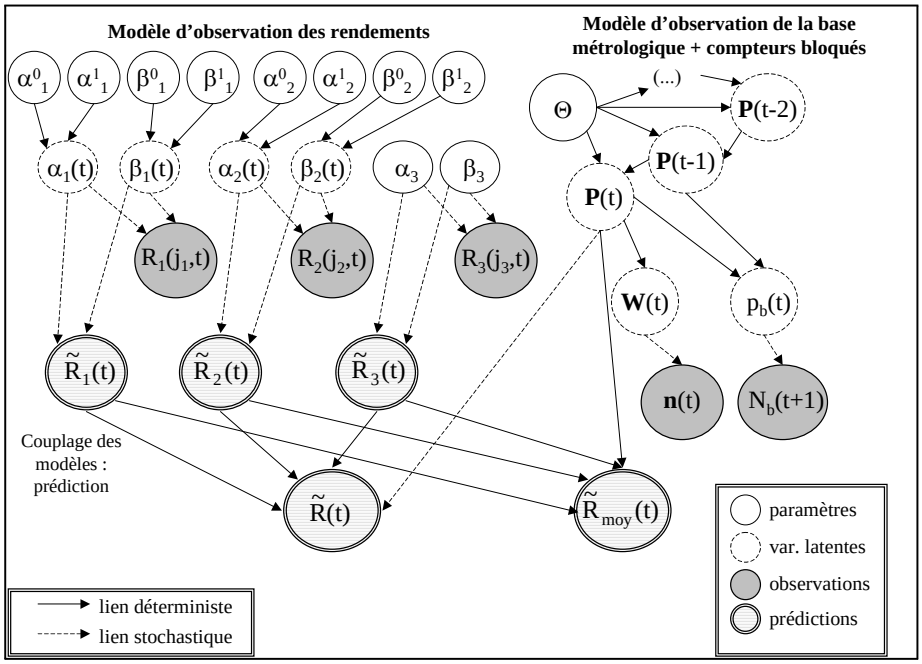


FIG. E.7 – Modèle statistique de vieillissement sous forme de DAG.

latentes. Les *rejetons* de ce modèle sont les grandeurs observables (de deux natures différentes) : la répartition des trois premiers états dans la base métrologique et les blocages observés sur le terrain.

2. Trois modèles d'observation des rendements, à l'intérieur de chacun des 3 états métrologiques de fonctionnement. La présence d'un niveau supplémentaire dans le modèle d'observation concernant les états $\mathcal{E}_1$ et $\mathcal{E}_2$, par rapport à celui relatif à l'état $\mathcal{E}_3$, montre bien la différence de comportement des 3 groupes de compteurs : dans le dernier groupe il n'y a pas de dépendance du rendement de l'âge.

L'analyse du *DAG* met en évidence que l'inférence de chacun de ces 4 modèles est menée séparément et que le couplage a lieu uniquement en phase prédictive pour reconstruire le rendement d'un parc de compteurs en fonction de la proportion de chaque groupe à l'intérieur du parc et du comportement métrologique des différents groupes.

E.4.2 Inférence bayésienne par les méthodes MCMC

L'objectif de l'inférence bayésienne est l'obtention de la *loi a posteriori* $[\boldsymbol{\xi}|\mathbf{y}]$des paramètres $\boldsymbol{\xi}$ du modèle, que le statisticien bayésien interprète comme la "mise à jour" d'une loi *a priori* $[\boldsymbol{\xi}]$ (exprimant la connaissance préliminaire sur ces grandeurs) suite à l'observation des données disponibles $\mathbf{y}$.

La relation entre ces deux lois de probabilité est exprimée par la formule de Bayes qui, dans le cas de variables continues s'écrit :

$$[\boldsymbol{\xi}|\mathbf{y}] = \frac{[\boldsymbol{\xi}][\mathbf{y}|\boldsymbol{\xi}]}{\int_{\Xi}[\boldsymbol{\xi}][\mathbf{y}|\boldsymbol{\xi}]d\boldsymbol{\xi}}. \tag{E.16}$$

Le passage exprimé par la formule de Bayes, simple sur le plan formel et conceptuel, est très difficile sur le plan opérationnel dans la plupart des cas, du fait de la difficulté de calcul de l'intégrale au dénominateur. D'autre part, dans les applications, on est plus intéressé à obtenir des estimateurs et des intervalles de crédibilité *a posteriori* pour chaque paramètre $(\xi_1, \xi_{2,\ldots})$ que l'expression mathématique de leur loi jointe de probabilité. Les algorithmes de simulation *MCMC* (*Monte Carlo Markov Chains*) sont aujourd'hui des méthodes efficaces d'inférence bayésienne.

D'abord utilisées dans le domaine physique (*Metropolis et al., 1953*), ces méthodes ont été adaptées à la simulation statistique (*Hastings, 1970*) pour l'obtention de tirages aléatoires dans une loi de probabilité $[\boldsymbol{\xi}]$ (avec $\boldsymbol{\xi} \in \Xi \subseteq \mathbb{R}^d$) connue à une constante près.

L'algorithme de *Metropolis-Hastings* construit une chaîne de Markov apériodique et irréductible dans l'espace Ξ, ayant comme distribution stationnaire $[\boldsymbol{\xi}]$. Alors, si l'on produit de nombreuses simulations de la chaîne, à partir d'une configuration initiale $\boldsymbol{\xi}_0$, les "dernières valeurs obtenues" sont distribuées selon la loi stationnaire $[\boldsymbol{\xi}]$ (indépendante de $\boldsymbol{\xi}_0$) et peuvent être utilisées pour en évaluer les principales caractéristiques statistiques (moyenne, médiane, percentiles). Ces méthodes ont connu dans les dernières années un tel succès dans la communauté bayésienne, que leur application est proposée et décrite dans tous les textes récents de statistique bayésienne : (*Robert, 1992*), (*Bernardo et Smith, 1994*), (*Gelman et al., 1995*), (*Lee, 1997*).

Pour résoudre le problème traité dans cet article on a utilisé un algorithme de *Metropolis-Hastings* avec une chaîne apériodique et irréductible particulière : *l'échantillonneur de Gibbs*, décrit pour la première fois par *Geman et Geman*

(*1984*). Pour appliquer cette méthode, il est nécessaire de connaître les lois de chacune des composantes du vecteur $\boldsymbol{\xi} = \{\xi_1, \xi_2, ..., \xi_d\}$ conditionnellement à toutes les autres, appelées *conditionnelles complètes*. Pour démarrer la simulation on donne des valeurs initiales aux variables $\boldsymbol{\xi}^{(0)} = \{\xi_1^{(0)}, \xi_2^{(0)}, ..., \xi_d^{(0)}\}$ et ensuite on procède itérativement de la manière suivante pour générer $\boldsymbol{\xi}^{(i)}$, composante par composante :

$$\left|\begin{array}{l} ITERATION\ i \\ \text{Tirage de } \xi_1^{(i)} \text{ de } [\xi_1|\xi_2^{(i-1)}, ..., \xi_d^{(i-1)}] \\ \text{Tirage de } \xi_2^{(i)} \text{ de } [\xi_2|\xi_1^{(i)}, ..., \xi_d^{(i-1)}] \\ ... \\ \text{Tirage de } \xi_d^{(i)} \text{ de } [\xi_d|\xi_1^{(i)}, \xi_2^{(i)}, ..., \xi_{d-1}^{(i)}] \end{array}\right. \tag{E.17}$$

La chaîne de Markov ainsi construite a une distribution stationnaire et cette distribution est la loi visée $[\boldsymbol{\xi}]$. La démonstration de la validité de l'algorithme de Gibbs et les méthodes pratiques d'implémentation se trouvent dans beaucoup de textes et monographies entièrement consacrés aux méthodes *MCMC* parmi lesquels (*Neal, 1993*), (*Gilks et al., 1996*), (*Robert, 1996*), (*Brooks, 1998*).

Dans la pratique courante on itère plusieurs fois l'algorithme de Gibbs et, on considère que, si on exclut les premiers vecteurs obtenus, les réplications restantes sont issues de la distribution de probabilité visée et sont alors utilisées pour déduire de façon empirique la moyenne et les percentiles des lois *a posteriori*.

E.4.3 Modélisation graphique et inférence bayésienne

L'avantage d'une représentation graphique du modèle pour l'inférence est que les relations d'indépendance statistique entre les variables sont explicites et la propriété (E.14) fait apparaître très simplement l'écriture des expressions des probabilités conditionnelles que l'algorithme de Gibbs utilise successivement à chaque itération. En fait, selon l'équation (E.14) chaque variable ξ_j (y compris les variables latentes) est conditionnellement dépendante uniquement de sa *couverture markovienne* (*Markov blanket*) $bl(\xi_j)$, ensemble formé de ses parents $pa(\xi_j)$, de ses descendants immédiats $ch(\xi_j)$ et des co-parents de ses descendants. Formellement :

$$[\xi_j|\{\xi_{i,i\neq j}\}] = [\xi_j|bl(\xi_j)], \tag{E.18}$$

où :

$$bl(\xi_j) = pa(\xi_j) \cup ch(\xi_j) \cup \{\xi_i : ch(\xi_j) \cap ch(\xi_i) \neq \emptyset\}. \tag{E.19}$$

Dans la pratique des *DAG*, les conditionnelles complètes apparaissant dans les formules (E.17) se simplifient considérablement puisque seule la couverture markovienne est utilisée dans le conditionnement.

La *couverture markovienne* de ξ_j, peut être déterminée à partir d'une simple transformation du graphe (*moralisation*), qui prévoit le rajout de liens fictifs entre les variables qui ont les mêmes descendants directs et effectue la transformation des liens directs en liens indirects. On peut facilement voir que $bl(\xi_j)$ est l'ensemble des voisins de ξ_j dans ce nouveau graphe.

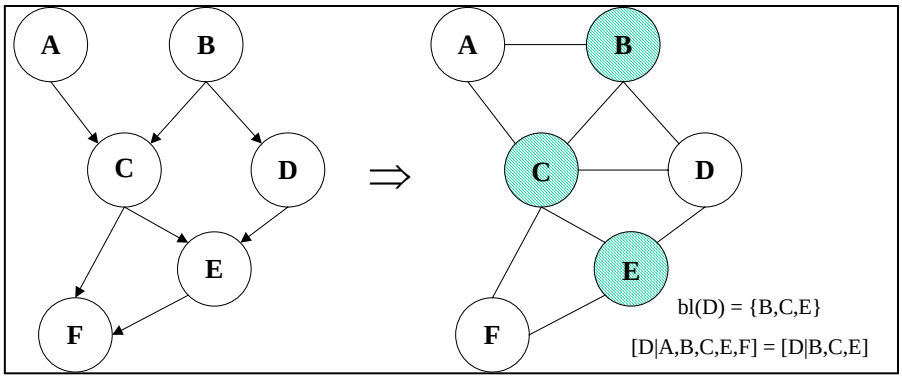


FIG. E.8 – Moralisation d'un *DAG*

La figure 8 montre un exemple de *moralisation* d'un graphe. Le nœud D est séparé des nœuds A et F par sa couverture markovienne (B, C, E). Par conséquent, sachant (B, C, E), D et (A, F) sont conditionnellement indépendants et la loi conditionnelle complète de D (sachant tous les autres nœuds) n'est fonction que de B, C et E.

E.5 Estimation du modèle de vieillissement des compteurs avec l'algorithme de Gibbs

Les résultats de la dernière partie de cet article, ont été obtenus à l'aide du logiciel *WinBUGS* (*Spiegelhalter et al., 1996, 1999*). Comme l'indique son nom (*BUGS = Bayesian Inference Using Gibbs Sampler*), ce programme, né d'un projet de l'Unité de Bio-Statistique du *MRC* (Medical Research Coucil) de

Cambridge (Royaume Uni), permet de mener les calculs d'inférence bayésienne sur des modèles complexes, à travers l'algorithme de Gibbs. Un des avantages majeurs est la possibilité de définir le modèle statistique directement sous forme de *DAG* : la conversion du graphe en code de calcul est faite automatiquement par le logiciel.

Sa relative facilité d'utilisation, sa souplesse, et le fait que jusqu'à présent sa licence d'utilisation est gratuite (téléchargeable du site http ://www.mrc-bsu.cam.ac.uk/bugs/) ont contribué à son succès dans la communauté bayésienne. Des nombreux exemples d'application de *WinBUGS* ont été publiés (une liste d'articles est disponible sur le site web du projet). *Congdon* (*2001*) fournit plus de 200 exemples d'implémentation pratique sous *WinBUGS* des principaux modèles utilisés en statistique appliquée.

Les distributions *a priori* $[\theta_{ij}]$ des probabilités de transition utilisées pour cet exemple sont *non informatives*. En particulier la loi *a priori* jointe est le produit de trois *distributions de Dirichlet* indépendantes, définies dans les espaces des probabilités de transition à partir des états $\mathcal{E}_1$, $\mathcal{E}_2$, $\mathcal{E}_3$ respectivement :

$$
\begin{aligned}
[\theta_{11,}\theta_{12,}\theta_{13,}\theta_{14}] &= \frac{\Gamma(\Sigma_{j=1}^{j=4}\alpha_{1j})}{\Pi_{j=1}^{j=4}\Gamma(\alpha_{1j})} \cdot \theta_{11}^{\alpha_{11}-1} \cdot \theta_{12}^{\alpha_{12}-1} \cdot \theta_{13}^{\alpha_{13}-1} \cdot \theta_{14}^{\alpha_{14}-1} \qquad \text{(E.20)} \\
[\theta_{22,}\theta_{23,}\theta_{24}] &= \frac{\Gamma(\Sigma_{j=2}^{j=4}\alpha_{2j})}{\Pi_{j=2}^{j=4}\Gamma(\alpha_{2j})} \cdot \theta_{22}^{\alpha_{22}-1} \cdot \theta_{23}^{\alpha_{23}-1} \cdot \theta_{24}^{\alpha_{24}-1} \\
[\theta_{33,}\theta_{34}] &= \frac{\Gamma(\Sigma_{j=3}^{j=4}\alpha_{3j})}{\Pi_{j=3}^{j=4}\Gamma(\alpha_{3j})} \cdot \theta_{33}^{\alpha_{33}-1} \cdot \theta_{34}^{\alpha_{34}-1}
\end{aligned}
$$

Les *distributions de Dirichlet*, couramment utilisées en statistique bayésienne comme lois *a priori* conjuguées des distributions multinomiales (*Good, 1965*), (*Gelman et al. 1995*), sont des généralisations des lois Bêta dans le cas de variables multidimensionnelles. Une loi *a priori* non informative (uniforme) est obtenue en affectant la valeur 1 à tous les paramètres α_{ij}. Un autre choix,

plus classique, est la loi non informative de *Jeffreys* dans laquelle $\alpha_{ij} = 1/2$.

		Percentiles		
Variable	Moyenne	2.5%	50%	97.5%
θ_{11}	0.951	0.946	0.951	0.956
θ_{12}	0.043	0.038	0.043	0.049
θ_{13}	0.001	0.000	0.001	0.003
θ_{14}	0.004	0.004	0.004	0.005
θ_{22}	0.965	0.956	0.966	0.975
θ_{23}	0.033	0.024	0.033	0.043
θ_{24}	0.001	0.000	0.001	0.003
θ_{33}	0.972	0.963	0.972	0.980
θ_{34}	0.028	0.020	0.028	0.037

(Tab. 1)

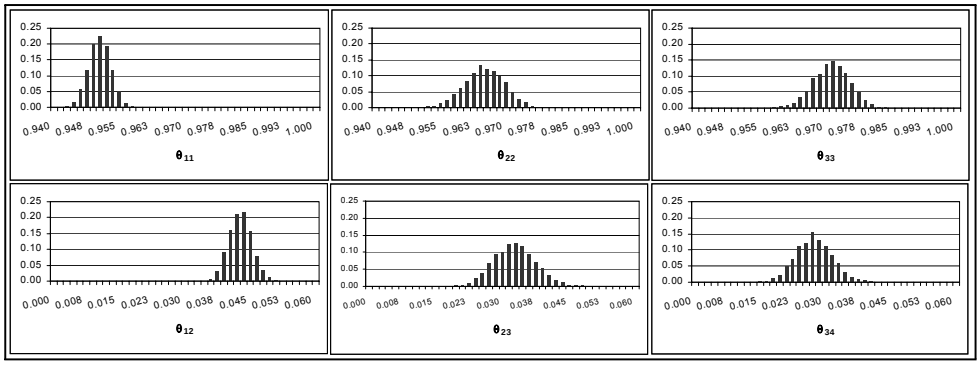


FIG. E.9 – Estimation par *méthode MCMC* des distributions *a posteriori* des θ_{ij}.

Le tableau 1 montre les résultats de l'inférence bayésienne concernant les probabilités de transitions et notamment les moyennes et les percentiles des lois *a posteriori*. Ces valeurs sont obtenues grâce à l'analyse des derniers 5000 résultats de la simulation *MCMC* (sur un total de 6000 itérations de l'algorithme de Gibbs). La représentation des mêmes résultats sous forme de diagrammes en bâtons est fournie en figure 9 pour les transitions θ_{ij} les plus intéressantes.

L'intervalle compris entre le percentile 0.025 et le percentile 0.975 constitue *l'intervalle de crédibilité a posteriori* à 95% : autrement dit, sous forme de jugement probabiliste direct que permet l'approche bayésienne, "il y a 95% de chance que les paramètres soient compris entre les bornes de ces intervalles". L'interprétation différente de l'*intervalle de crédibilité* (bayésien) et de l'*intervalle de*

confiance (fréquentiste) constitue une des différences majeures entre ces deux approches (*Lecoutre et Poitevineau, 1996*), (*Bernier et al., 2000*).

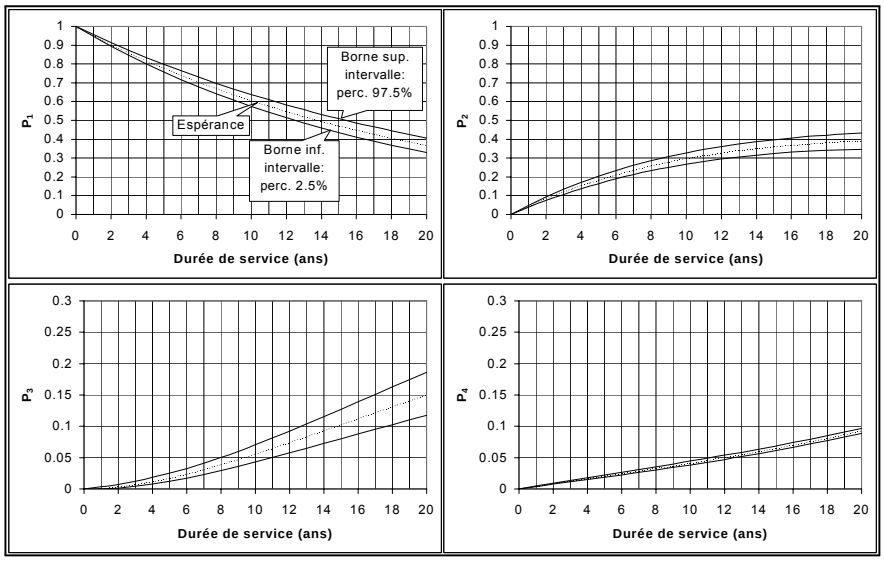


FIG. E.10 – Intervalles de crédibilité des probabilités d'appartenance aux 4 états.

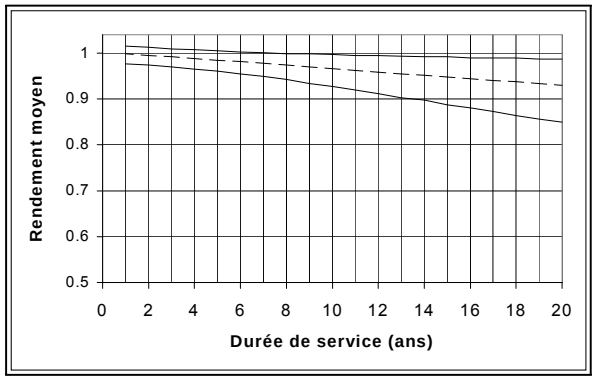


FIG. E.11 – Intervalles de crédibilité à 95% du rendement moyen d'un parc de compteurs.

La figure 10 trace les intervalles de crédibilité prédictifs des probabilités d'occurrence des 4 états en fonction de l'âge de l'appareil. La grandeur $P_3(t)$ est particulièrement intéressante parce qu'elle constitue une estimation de la proportion de compteurs potentiellement non conformes au sein d'une certaine population. L'horizon temporel choisi est de 20 ans de service.

La figure 11 donne l'évolution du rendement moyen pour ce type de compteur, alors que la figure 12 montre la distribution prédictive des rendements pour des compteurs survivants d'âge 1, 5, 10, 20 ans, obtenue par simulation de *Monte Carlo* à l'aide de l'équation (E.11).

Sur la figure 12, on observe assez bien l'effet du mélange des différentes sous-populations : la distribution très pointue caractéristique des jeunes compteurs se transforme, au fil des années, en une distribution beaucoup plus plate avec une queue de plus en plus étalée vers des valeurs très faibles de rendements.

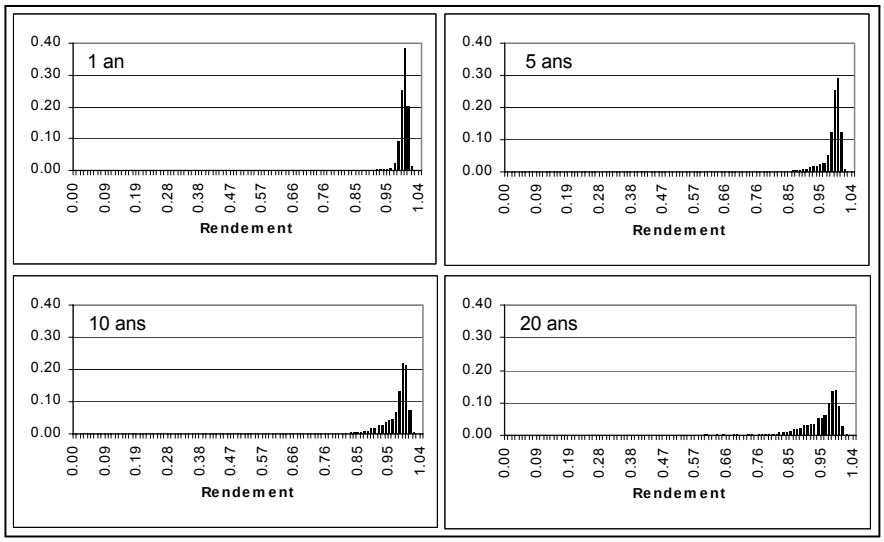


FIG. E.12 – Distributions prédictives du rendement d'un parc de compteurs.

E.5.1 Contrôle de convergence de l'algorithme *MCMC*

Le contrôle de convergence et l'estimation de la vitesse de convergence font l'objet de recherches actives chez les statisticiens bayésiens. *Cowles et Carlin (1996)* et, plus récemment, *Mergensen et al. (1999*) dressent une liste de techniques statistiques pour décider si la convergence de la chaîne *MCMC* a été atteinte.

Une procédure particulièrement répandue chez les praticiens est celle proposée par *Gelman et Rubin* (*1992*) qui consiste à lancer plusieurs chaînes en parallèle, avec des valeurs initiales très dispersées, et arrêter l'algorithme lorsque le rapport entre la variance inter-chaînes et la variance intra-chaînes est suffisamment proche de 1.

En effet si les valeurs initiales des chaînes sont différentes, la variance inter-chaînes reste significativement supérieure à la variance intra-chaînes tant que les chaînes n'ont pas atteint leur distribution stationnaire.

Dans *WinBUGS* l'analyse de convergence est menée à l'aide d'une variante de la méthode de *Gelman-Rubin* introduite par *Brooks et Gelman* (*1998*). À chaque itération, pour chacune des k chaînes, on considère les m dernières valeurs simulées.

Sur la base de ces valeurs, on calcule empiriquement les longueurs δ_i des k intervalles de crédibilité (de chaque suite) à un certain niveau fixé $(1-\alpha)$. Ensuite on considère l'ensemble des km valeurs extraites et on en détermine, de même, la longueur Δ de l'intervalle de crédibilité respectif (de niveau $1-\alpha$).

Le diagnostique de convergence se base sur la statistique (dite de *Brooks-Gelman*) :

$$\widehat{R}_{BG} = \frac{\Delta}{\frac{1}{k}\sum_{i=1}^{i=k}\delta_i}. \tag{E.21}$$

Dans l'équation le numérateur représente une mesure de la variabilité inter-chaîne et le dénominateur (longueur moyenne des intervalles de crédibilité issus de chacune des k chaînes) une mesure de la variabilité intra-chaîne.

WinBUGS calcule automatiquement la statistique $\widehat{R}_{BG}$ avec $m = 50$ et $1-\alpha = 80\%$.

Dans l'exemple proposé dans cet article, l'analyse menée sur la base des valeurs de 6 chaînes parallèles pour chaque paramètre, montre qu'après 1000 itérations la valeur de $\widehat{R}_{BG}$ est pratiquement égale à 1 pour toutes les variables. Par conséquent, toutes les valeurs, à partir de la 1000ème itérations sont utilisables pour simuler les lois *a posteriori*.

E.6 Analyse de sensibilité à la valeur de la probabilité d'observation des blocages p_{ob}

L'utilisation des données en provenance de la base de facturation peut engendrer des biais dus au fait qu'une partie des opérations de dépose se trouve dans la base mais leur cause n'est pas renseignée. Dans un premier temps on a supposé que la répartition des motifs de déposes parmi la population inconnue est la même que celle parmi la population connue, mais on peut imaginer que

cette hypothèse ne corresponde pas à la réalité (par exemple on a des raisons de croire qu'une bonne partie des déposes non renseignées est due à des raisons administratives). Cette réflexion a motivé une analyse de sensibilité des résultats à la valeur de p_{ob}.

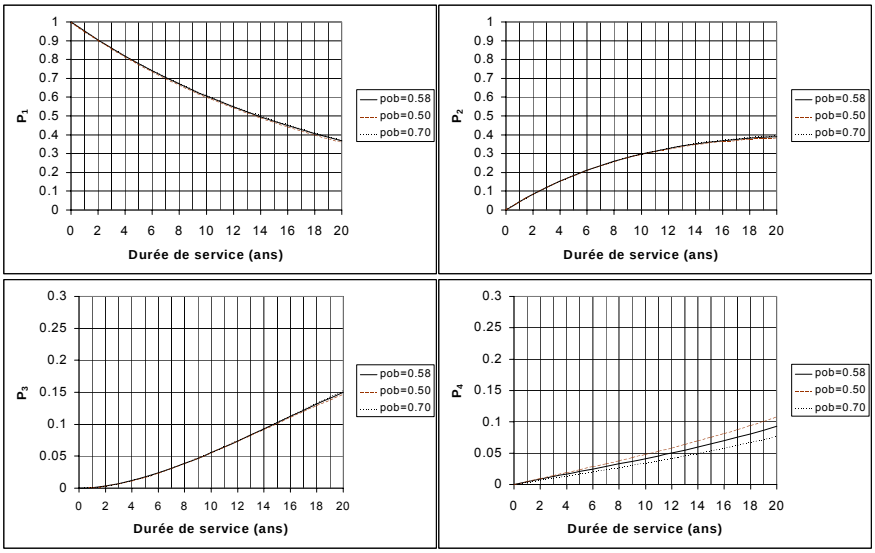


FIG. E.13 – Estimateurs des probabilités d'appartenance aux 4 états avec différentes valeurs de p_{ob}.

La figure 13 montre les espérances *a posteriori* (obtenues par simulation *MCMC*) des probabilités d'appartenance aux 4 états métrologiques pour 3 valeurs différentes de $p_{ob} \cdot p_r$. Par rapport à la valeur de référence (0.58), on s'est positionné dans deux cadres un peu extrêmes : une situation *optimiste*, où l'on imagine que les blocages observés représentent 70% des cas réels et une situation *pessimiste* où les données observées ne sont que la moitié des occurrences qui ont lieu effectivement.

Les courbes montrent que, dans ce cas particulier, les estimations des probabilités $P_1(t)$, $P_2(t)$, $P_3(t)$ ne semblent pas être trop sensibles aux valeurs de $p_{ob} \cdot p_r$, alors que, évidemment, la probabilité $P_4(t)$, c'est-à-dire le taux de blocage cumulé, est lui très affectée par la valeur de ce paramètre. On remarque aussi qu'à des taux de blocage supérieurs, correspondent des appareils qui sont globalement de qualité inférieure.

E.7 Discussion

E.7.1 Un retour aux données montre la bonne adéquation du modèle aux observations

Une simple méthode pour vérifier l'adéquation du modèle aux observations se fonde sur la simulation de nouvelles données, à l'aide du modèle même, sur la base des lois *a posteriori* de ses paramètres $[\boldsymbol{\xi}|\mathbf{y}]$. On obtient ainsi des lois *prédictives a posteriori* $[\widetilde{\mathbf{y}}|\mathbf{y}]$ des "nouvelles" observations $\widetilde{\mathbf{y}}$ sachant les "anciennes" données $\mathbf{y}$:

$$[\widetilde{\mathbf{y}}|\mathbf{y}] = \int_{\Xi} [\widetilde{\mathbf{y}}, \boldsymbol{\xi}|\mathbf{y}] d\boldsymbol{\xi} = \int_{\Xi} [\widetilde{\mathbf{y}}|\boldsymbol{\xi}]][\boldsymbol{\xi}|\mathbf{y}] d\boldsymbol{\xi}. \tag{E.22}$$

Un premier jugement sur l'adéquation du modèle consiste à vérifier que les *vraies* données sont contenues dans les intervalles de crédibilité des prédictions, obtenus (par exemple) avec des tirages de *Monte Carlo*.

La figure 14 montre que les "tubes" prédictifs à 95% du nombre de compteurs appartenant à chacun des 4 états contiennent pratiquement toujours les observations (nombre de compteurs d'âge donné appartenant à un état donné et nombre de blocages) et que donc le modèle est bien calé sur les données utilisées pour son estimation.

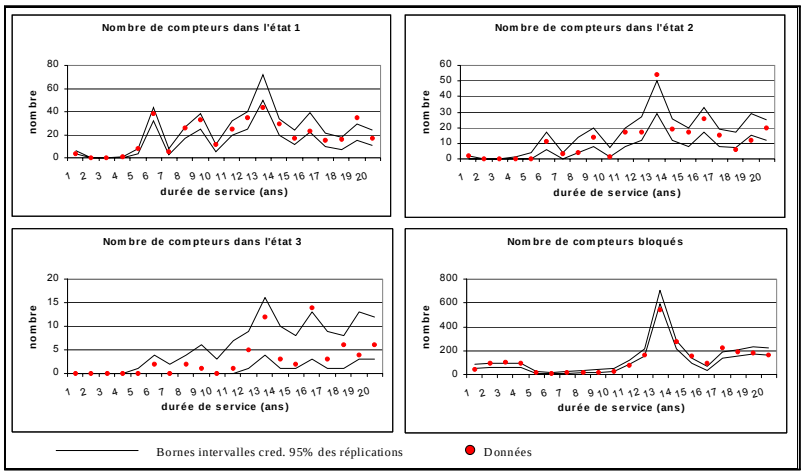


FIG. E.14 – *Tubes prédictifs* à 95% des réplications des observables et *vraies* données.

La formule (E.22) est également utilisable dans un contexte de validation où

l'on sépare les données en un échantillon d'apprentissage ($\mathbf{y}$) et un échantillon de test ($\widetilde{\mathbf{y}}$). Nous ne nous sommes pas livrés à cette opération, faute d'avoir un nombre suffisant de données.

E.7.2 Le poids des hypothèses

Les paramètres fondamentaux du modèle markovien sont les probabilités de transition θ_{ij} d'un état $\mathcal{E}_i$ à un état $\mathcal{E}_j$. Comme la collecte des données est destructive on n'observe jamais réellement la transition d'un compteur entre ces deux états et par conséquent l'inférence sur les θ_{ij} se fait par l'intermédiaire des probabilités d'occurrence de chaque état en fonction de l'âge : $P_i(t)$ avec $P_2(0) = P_3(0) = P_4(0) = 0$ et $P_1(0) = 1$. Par hypothèse tous les compteurs installés commencent leur vie dans l'état $\mathcal{E}_1$. Cette assertion est justifiée par le fait que le test de vérification primitive des compteurs, selon la réglementation en vigueur, oblige les dispositifs à respecter les *EMT* des compteurs neufs (décrites au le paragraphe 2.2) aux débits q_{min}, q_t, q_{max}. Compte tenu du fait que les *EMT* des compteurs en service, qui ont servi à la définition de l'état $\mathcal{E}_1$, sont le double (en valeur absolue) des *EMT* "à neuf", on a des bonnes raisons de croire qu'à la pose, les appareils répondent aux conditions prévues pour l'état $\mathcal{E}_1$

Dans l'exemple proposé, on a imaginé que toutes les transitions $\mathcal{E}_i \longrightarrow \mathcal{E}_j$ ($i \leq j$) sont possibles. Une autre hypothèse réaliste est d'imposer $\theta_{14} = \theta_{24} = 0$ (c'est-à-dire qu'un compteur, avant de se bloquer passe nécessairement dans un état de métrologie très incorrect), et on peut d'ailleurs aussi imaginer encore un autre modèle en imposant $\theta_{13} = \theta_{14} = \theta_{24} = 0$ (le compteur transite toujours par l'état $\mathcal{E}_{j-1}$ avant d'atteindre l'état $\mathcal{E}_j$). Dans la figure 15 on montre un estimateur de la probabilité d'occurrence de l'état absorbant $\mathcal{E}_4$ en fonction de l'âge (espérance *a posteriori*) calculé selon les 3 modèles proposés. On observe que dans le cas où l'on impose le passage à travers tous les états, le blocage est atteint nettement plus lentement.

Cette sensibilité marquée des résultats aux hypothèses (on remarque aussi que quand on limite les transitions, l'état $\mathcal{E}_3$ semble être sous-estimé) nous pousse à des réflexions nouvelles sur la structure du modèle et la qualité des données. En effet le mécanisme binomial d'observation des blocages pourrait provoquer des phénomènes de *surcalage* (*overfitting*) du modèle sur les données *ICBC* par rapport aux données de répartition des compteurs dans la base métrologique, les proportions étant calculées sur des échantillons de taille plus importante dans la

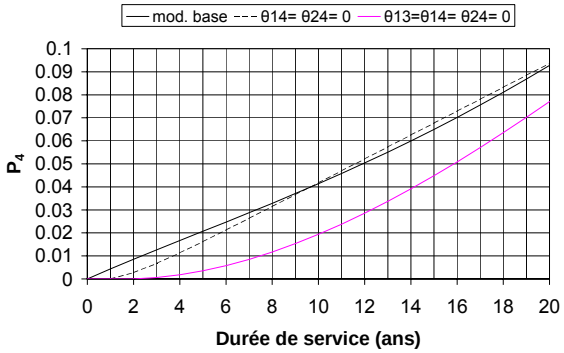


FIG. E.15 – Sensibilité de l'occurrence de l'état absorbant aux hypothèses sur les transitions possibles.

base *ICBC.*

E.7.3 Conclusions et perspectives

Les résultats montrent un bon comportement des compteurs examinés. Après 20 ans de service, entre 30% et 40% des dispositifs présentent une métrologie assez proche de celle d'un compteur neuf et moins de 20% se trouvent dans un état incompatible avec les exigences réglementaires.

La perte de rendement moyenne s'établit autour de 0.3 à 0.4% par an. Ce résultat est en accord avec l'opinion des experts du comptage et recoupe les résultats de la litterature technique quant à la meilleure qualité des compteurs volumétriques par rapport aux compteurs de vitesse, qui peuvent afficher des pertes de rendements 2 ou 3 fois supérieures.

D'autre part, on remarque aussi qu'une incertitude assez importante persiste sur l'estimation des paramètres les plus importants du point de vue opérationnel, c'est-à-dire le rendement moyen et la probabilité P_3 d'appartenir à la classe des défectueux.

Cette incertitude est due, d'une part au comportement naturellement stochastique des compteurs, et d'autre part au faible nombre de compteurs appartenant à l'état $\mathcal{E}_3$ qui ne permet pas une estimation très précise de leur proportion. Sur les 682 compteurs objet de l'étude, il y en a que 61 qui appartient à l'état $\mathcal{E}_3$. Ainsi, paradoxalement, la bonne qualité des dispositifs examinés rend difficile l'estimation du taux de défaillance.

Une autre explication de cette incertitude est le fait que l'âge n'est pas le seul

facteur explicatif de la dégradation. Le vieillissement d'un compteur en réelles conditions d'exploitation peut être accéléré par les propriétés physico-chimiques de l'eau circulante (pH, dureté, présence de fer, concentration des particules en suspension) et/ou de l'état du réseau de distribution (passage accidentel de particules solides dans le réseau). L'effet de chacun de ces facteurs est difficile à comprendre parce que la prise en compte d'un seul d'entre eux n'est pas, en général, suffisante pour expliquer la dégradation des compteurs. Si on peut imaginer que ces facteurs explicatifs ont une cohérence géographique (la qualité de l'eau dépend du site d'exploitation), malheureusement, la seule indication de la ville de provenance ne permet pas une réelle prise en compte : on devrait développer un modèle *ad hoc* pour chaque ville ! Une solution au problème pourrait être représentée par l'exploitation des résultats des expertises des compteurs (une partie des compteurs étalonnés est, en fait, démontée et expertisée au laboratoire d'essai). L'observation de l'état d'usure d'appareils de provenance de villes différentes, éventuellement couplée avec l'opinion des experts, pourrait permettre un renseignement de cette variable.

Enfin, la validation du modèle pourrait être menée à partir de l'analyse des séries temporelles des volumes d'eau facturés sur un branchement. Puisque la consommation enregistrée dépend du rendement, quand sur un point de desserte un vieux compteur est remplacé par un compteur neuf, on devrait observer une rupture dans la série des consommations due au fait que la consommation réelle n'est pas affectée par l'âge du compteur. Le changement de compteur devrait logiquement provoquer une augmentation de la consommation ou réduire la baisse annuelle. L'importance de cette rupture devrait être liée à la différence de rendement entre le compteur remplacé et le compteur remplaçant (pratiquement égal à 1 pour un appareil neuf).

Des études préliminaires montrent que le problème de modélisation est plutôt compliqué puisque la consommation réelle est une variable latente qui évolue dans le temps selon un modèle de dynamique inconnue.

À cette évolution *naturelle* se superposent le vieillissement des compteurs, une éventuelle composante *psychologique* (qui pousse l'abonné à réduire sa consommation au moment de la pose du nouveau compteur) et un bruit blanc qui a un écart type très important par rapport aux tendances recherchées. Pour en doner un ordre de grandeur, on peut dire que deux bains en plus ou en moins pris au long d'une année ont, sur une consommation domestique (environ 90 à

100 m^3/an), le même impact qu'une année de vieillissement du compteur.

E.8 Bibliographie

- American Water Works Association, California Section Committee, *Determination of Economic Period for Water Meter Replacement.* Journal of the American Water Works Association, Vol. 58, n. 6, pp. 642-650, 1966.

- Bernardo J.M., Smith A.F.M., *Bayesian Theory.* Wiley & Sons, 1994.

- Bernier J., Parent E., Boreux J.J., *Statistiques pour l'Environnement. Traitement Bayésien des Incertitudes.* Tec & Doc, 2000.

- Brooks S.P., *Markov Chain Monte Carlo Method and its application.* The Statistician, n. 47, Part 1, pp. 69-100, 1998.

- Brooks S.P., Gelman A., *General Methods for Monitoring Convergence of Iterative Simulations.* Journal of Computational and Graphical Statistics, Vol. 7, n. 4, pp. 434-455, 1998.

- Carlier M., *Machines Hydrauliques.* Imprim. Louis-Jean, 1968.

- Congdon P., *Bayesian Statistical Modelling.* Wiley & Sons, 2001.

- Costes A., Pia Y., *Les compteurs d'eau en France : la Réglementation et son évolution.* Techniques, Sciences et Méthodes, n. 7, pp. 21-27, 2000.

- Cowell R.G., Dawid, A.P., Lauritzen S.L., Spiegelhalter D.J., *Probabilistic Networks and Expert Systems.* Springer-Verlag, 1999.

- Cowles M.K., Carlin B.P., *Markov Chain Monte Carlo Convergence Diagnostics : A Comparative Review.* Journal of the American Statistical Association, Vol. 91, n. 434, pp. 883-904, 1996.

- Gelfand A.E., Smith A.F.M.. *Sampling-based approaches to calculating marginal densities.* Journal of American Statistical Association, n. 85, pp.398-409, 1990.

- Gelman A., Rubin D.B., *Inference from Iterative Simulation Using Multiple Sequences (with discussion).* Statistical Science, n. 7, pp. 457-511, 1992.

- Gelman A., Carlin J.B., Stern H.S., Rubin D.B., *Bayesian data analysis.* Chapman & Hall, 1995.

- Geman S., Geman D., *Stochastic Relaxation, Gibbs Distribution and the Bayesian Restoration of Images.* IEEE Transactions on Pattern Analysis and Machine Intelligence, n. 6, pp. 721-741, 1984.

- Gilks W.R., Richardson S., Spiegelhalter D.J., *Markov Chain Monte Carlo in Practice.* Chapman & Hall, 1996.

- Good I.,J. *The Estimation of Probabilities : An Essay on Modern Bayesian Methods*. M.I.T. Press, 1965.

- Grau P., *Expériences dans le comptage de l'eau et dans le renouvellement des compteurs*. Water Supply, Vol. 3, n. 4 pp. 273-292, 1985.

- Hastings W.K., *Monte Carlo Sampling Methods using Markov Chains and their applications*. Biometrika, n. 57, pp. 97-109, 1970.

- Lecoutre B., Poitevineau J., *Traitement statistique des données expérimentales : des pratiques traditionnelles aux pratiques bayésiennes*. CISIA-CERESTA, 1996.

- Lee P.M., *Bayesian Statistics : An Introduction*, 2nd Edition. Oxford University Press, 1997.

- Lund J.R., *Metering Utility Services : Evaluation and Maintenance*. Water Resources Research, Vol. 24, n. 6, pp 802-816, 1988.

- Mergensen K.L., Robert C.P., Guihenneuc-Jouyaux C., *MCMC Convergence Diagnostics : A "reviewww"*. In *Bayesian Statistics 6*, pp. 415-440. Ed. Bernardo J.M., Berger J.O., Dawid A.P. et Smith A.F.M. Oxford University Press, 1999.

- Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H., Teller E., *Equation of State Calculations by Fast Computing Machine*. Journal of Chemical Physics, n. 21, pp. 1087-1091, 1953.

- Neal R.M., *Probabilistic Inference Using Markov Chain Monte Carlo Methods*, Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto, 1993.

- Newman G.J., Noss R.R., *Domestic 5/8 Inch Meters Accuracy and Testing, Repair and Replacement Programs*. Proceeding 1982 American Water Works Association Annual Conference, Part 1, Paper n. 11-7, 1982.

- Noss R.R., Newman G.J., Male J.W., *Optimal Testing Frequency for Domestic Water Meters*. Journal of Water Resources Planning and Management, Vol 113, n. 1, pp. 1-14, 1987.

- Robert C.P., *L'Analyse Statistique Bayesienne*. Economica, 1992.

- Robert C.P., *Méthode de Simulation Monte Carlo par chaînes de Markov*. Economica, 1996.

- Spiegelhalter D.J., Thomas A., Best N.G., *Computation on Bayesian Graphical Models*. In *Bayesian Statistics 5*, pp. 407-425. Ed. Bernardo J.M., Berger J.O., Dawid A.P., Smith A.F.M. Oxford University Press, 1996.

- Spiegelhalter D.J., Thomas A., Best N.G., *WinBUGS Version 1.2 User Manual*. MRC Biostatistics Unit, 1999.

- Tronskolanski A.T., *Théorie et Pratiques des Mesures Hydrauliques*. Dunod, 1963.

- Van der Linden M.J., *Implementing a large-meter replacement program*. Journal of the American Water Works Association, Vol. 90, n.8, pp. 50-56, 1998.

Annexe

L'espérance E et la variance V d'une variable aléatoire qui suit une loi Bêta, de paramètres α et β, s'écrivent :

$$E = \frac{\alpha}{\alpha+\beta} \tag{A1}$$

$$V = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{A2}$$

Dans les formules (A1) et (A2), en accord avec les formules (10) et (11), E et V sont fonctions de la durée de service t.

Pour étudier la croissance ou décroissance de la fonction $E(t)$ il suffit d'étudier le signe de sa dérivée $E'(t)$. L'expression de $E'(t)$ s'écrit facilement sur la base de la règle de dérivation du rapport de fonctions et du fait que :

$$\alpha'(t) = -\alpha_1 \cdot e^{\alpha_0 - \alpha_1 \cdot t} = -\alpha_1\alpha \qquad \beta'(t) = -\beta_1\beta. \tag{A3}$$

On a donc :

$$E'(t) = \frac{\alpha'(\alpha+\beta) - \alpha(\alpha'+\beta')}{(\alpha+\beta)^2} = \frac{\alpha\beta}{(\alpha+\beta)^2}(\beta_1 - \alpha_1). \tag{A4}$$

$E'(t)$ est alors négative, pour $t > 0$, si et seulement si $\alpha_1 > \beta_1$. Dans le cas des distributions des rendements des compteurs d'eau, puisque la moyenne est proche de 1 on a généralement $\alpha \gg \beta$ et aussi $\alpha_1 > \beta_1$. Les formules (10) et (11) donnent lieu à des distributions de probabilité dont l'espérance est une fonction décroissante de l'âge t.

L'étude des propriétés de la fonction $V(t)$ est plus compliquée. Cependant il est équivalent d'étudier les propriétés de la fonction $v(t) = \ln(V(t))$ qui s'écrit :

$$v(t) = \ln(\alpha) + \ln(\beta) - 2\ln(\alpha+\beta) - \ln(\alpha+\beta+1). \tag{A5}$$

La dérivée de cette fonction, compte tenu des formules (10) et (11) est :

$$v'(t) = -(\alpha_1+\beta_1)+\left(\frac{2}{\alpha+\beta}+\frac{1}{\alpha+\beta+1}\right)(\alpha_1\alpha+\beta_1\beta). \qquad \text{(A6)}$$

L'interprétation de la formule (A6) n'est pas immédiate. Par contre, pour les valeurs des paramètres évalués dans notre cas particulier, on vérifie $v\prime(t) > 0$ et donc que la variance est fonction croissante de l'âge.

D'autre part, dans le cas où $\alpha \gg \beta$ et $\alpha \gg 1$ on peut obtenir une expression approximée de $v(t)$ en utilisant l'approximation $\ln(\alpha+\beta) \simeq \ln(\alpha+\beta+1) \simeq \ln(\alpha)$:

$$v(t) \simeq \ln(\beta) - 2\ln(\alpha) = \beta_0 - 2\alpha_0 + (2\alpha_1 - \beta_1)\cdot t \qquad \text{(A7)}$$

qui est évidemment croissante si $\alpha_1 > \beta_1$.