

# REVUE DE STATISTIQUE APPLIQUÉE

ALBERTO PASANISI

ERIC PARENT

## **Modélisation bayésienne du vieillissement des compteurs d'eau par mélange de classes d'appareils de différents états de dégradation**

*Revue de statistique appliquée*, tome 52, n° 1 (2004), p. 39-65

[http://www.numdam.org/item?id=RSA\\_2004\\_\\_52\\_1\\_39\\_0](http://www.numdam.org/item?id=RSA_2004__52_1_39_0)

© Société française de statistique, 2004, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme  
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

## MODÉLISATION BAYÉSIENNE DU VIEILLISSEMENT DES COMPTEURS D'EAU PAR MÉLANGE DE CLASSES D'APPAREILS DE DIFFÉRENTS ÉTATS DE DÉGRADATION

Alberto PASANISI\*<sup>\*\*\*</sup>, Eric PARENT\*

\* *École Nationale du Génie Rural, des Eaux et des Forêts, Laboratoire GRESE,  
19, av. du Maine 75732 Paris Cedex 15*

\*\* *Compagnie Générale des Eaux, Direction Technique,  
18, Bd. 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.

**Keywords :** *Water meters, Degradation, Dynamic models, Bayesian inference, Graphical models, MCMC simulation.*

## 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<sup>3</sup> 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 littérature technique, p. ex. (*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 imaginer 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*) :

$[A]$  : 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  $[X = x]$ .

## 2. Comment se dégrade un compteur et quelles sont les conséquences

### 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.

$$e = \frac{v_{enr} - v_{réel}}{v_{réel}}. \quad (1)$$

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 ( $q$ ).

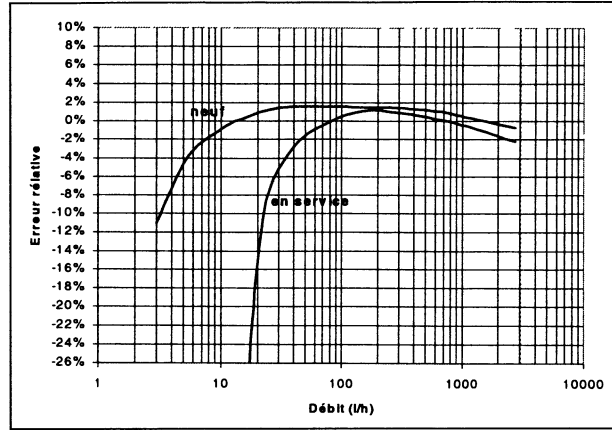


FIGURE 1

*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 :

$$r = \frac{v_{enr}}{v_{réel}} = 1 + \sum_i e_i x_i \quad (2)$$

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.

Les calculs développés dans la suite sont relatifs à des histogrammes «types» de consommation qu'on suppose complètement connus et invariants dans le temps.

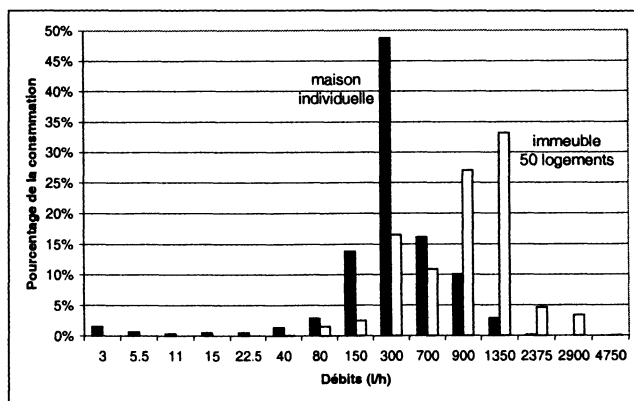


FIGURE 2  
*Histogrammes de consommation*

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).

## 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  $q_n = 1.5 \text{ m}^3/\text{h}$  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  $q_n < 10 \text{ m}^3/\text{h}$ ). 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 ( $q_n = 1.5 \text{ m}^3/\text{h}$ ) respectivement dans les états  $\mathcal{E}_1$ ,  $\mathcal{E}_2$ ,  $\mathcal{E}_3$ .

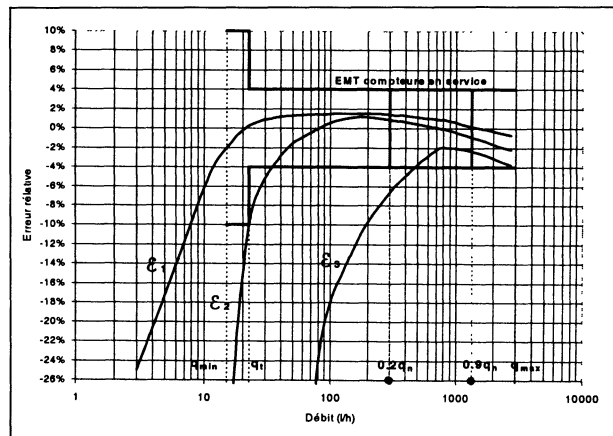


FIGURE 3  
« États métrologiques »

### 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 Veolia 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*,  $q_n = 1.5 \text{ m}^3/\text{h}$ , 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.

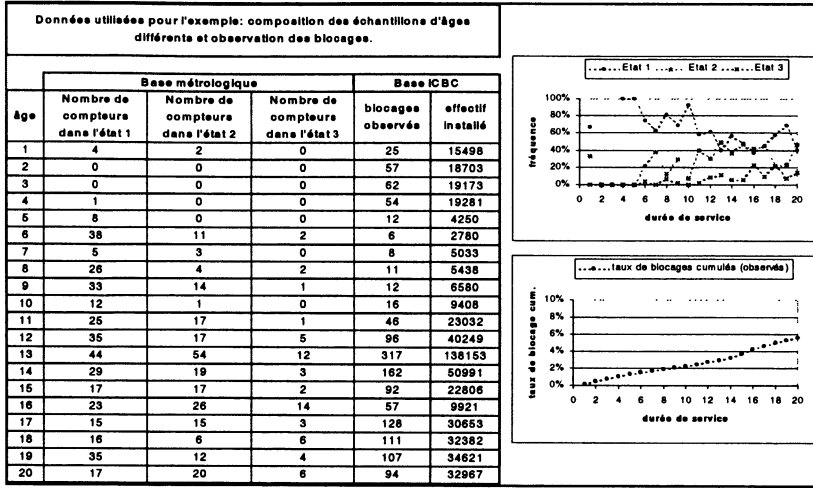


FIGURE 4  
Données utilisées pour l'exemple

### 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  $\theta_{ij}$  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 :

$$\mathbf{P}(t) = \mathbf{P}(t-1) \cdot \Theta \tag{3}$$

où  $\Theta$  est la matrice de transition.



À 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.

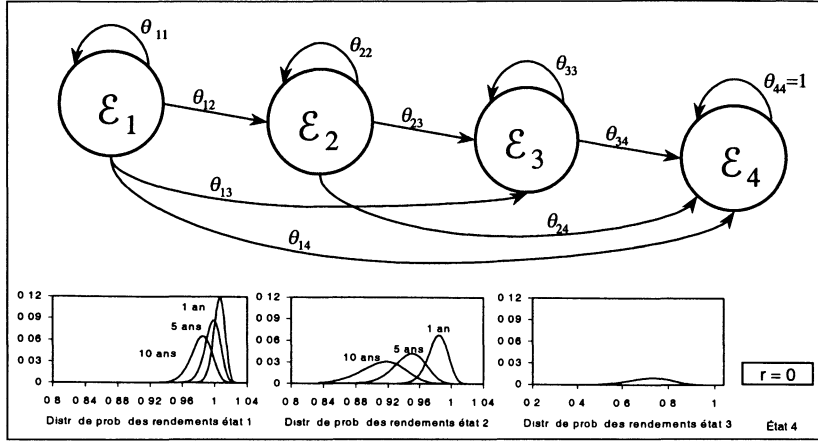


FIGURE 5

*Mécanisme markovien de dégradation*

### 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  $N(t)$  :

$$[\mathbf{n}(t)|\mathbf{P}(t)] = \frac{N(t)!}{n_1(t)!n_2(t)!n_3(t)!} W_1(t)^{n_1(t)} W_2(t)^{n_2(t)} W_3(t)^{n_3(t)} \quad (4)$$

où la probabilité conditionnelle d'être dans l'état  $\mathcal{E}_i$  (sachant que le compteur n'est pas bloqué) s'écrit :

$$W_i(t) = \frac{P_i(t)}{\sum_{j=1}^{j=3} P_j(t)} \quad (i = 1, 2, 3) \quad (5)$$

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  $\mathcal{E}_1$ .

### 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 :

$$p_b(t) = P_1(t-1) \cdot \theta_{14} + P_2(t-1) \cdot \theta_{24} + P_3(t-1) \cdot \theta_{34} = P_4(t) - P_4(t-1) \quad (6)$$

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 défectueux (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  $t$  :

$$[N_b(t) | P_4(t), P_4(t-1)] = \binom{N_C(t)}{N_b(t)} b(t)^{N_b(t)} (1 - b(t))^{(N_C(t) - N_b(t))} \quad (7)$$

La valeur 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  $\Theta$  est relié aux grandeurs observables par les équations d'observation (4) et (7) et par l'équation dynamique (3).

### 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  $\alpha$  et  $\beta$ .

$$[z_i | \alpha_i, \beta_i] = \frac{1}{B(\alpha_i, \beta_i)} z_i^{(\alpha_i-1)} (1-z_i)^{(\beta_i-1)} \quad (i = 1, 2, 3) \quad (8)$$

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 :

$$\ln(\alpha_i(t)) = \alpha_i^0 - \alpha_i^1 \cdot t \quad (9)$$

$$\ln(\beta_i(t)) = \beta_i^0 - \beta_i^1 \cdot t \quad (i = 1, 2) \quad (10)$$

où  $\alpha_i^0, \alpha_i^1, \beta_i^0, \beta_i^1 > 0$ . Dans notre cas particulier (voir annexe), les équations (9) et (10) donnent lieu à des distributions de probabilité unimodales à espérance décroissante et à variance croissante en fonction du temps  $t$  (figure 5).

### 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 déposés) d'âge donné  $t$ . C'est un mélange, de proportions variables avec l'âge, des rendements des 3 classes :

$$\tilde{r}(t) = r_1(t) \cdot \gamma_1(t) + r_2(t) \cdot \gamma_2(t) + r_3 \cdot \gamma_3(t). \quad (11)$$

Les  $r_i$  sont évalués à partir des équations (8), et  $\gamma(t)$  est la réalisation d'un tirage multinomial de paramètres  $(\mathbf{W}(t), 1)$  :

$$\gamma(t) = \{\gamma_1(t), \gamma_2(t), \gamma_3(t)\} = \begin{cases} \{1, 0, 0\} & \text{avec probabilité } W_1(t) \\ \{0, 1, 0\} & \text{avec probabilité } W_2(t) \\ \{0, 0, 1\} & \text{avec probabilité } W_3(t) \end{cases} \quad (12)$$

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 :

$$\tilde{r}_{moy}(t) = r_1(t) \cdot W_1(t) + r_2(t) \cdot W_2(t) + r_3 \cdot W_3(t) \quad (13)$$

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.

#### 4. Inférence

##### 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).

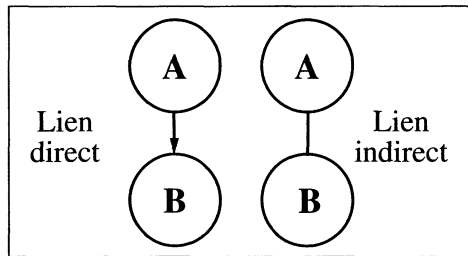


FIGURE 6  
*Types de liens entre 2 variables*

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 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  $\xi_j$  est conditionnellement indépendante de ses non descendants, étant donné ses parents :

$$[\xi_j | nd(\xi_j), pa(\xi_j)] = [\xi_j | pa(\xi_j)]. \quad (14)$$

On montre (*Cowell et al., 1999*) que l’équation (14) est valide si et seulement si la loi de probabilité conjointe de toutes les variables se factorise de la manière suivante :

$$[\xi_1, \xi_2, \dots, \xi_d] = \prod_{i=1}^d [\xi_i | pa(\xi_i)]. \quad (15)$$

L’équation (15) simplifie énormément 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  $\Theta$  dans lequel les probabilités d’occurrence des 4 états métrologiques jouent un rôle de variables latentes. Les *rejets* 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.

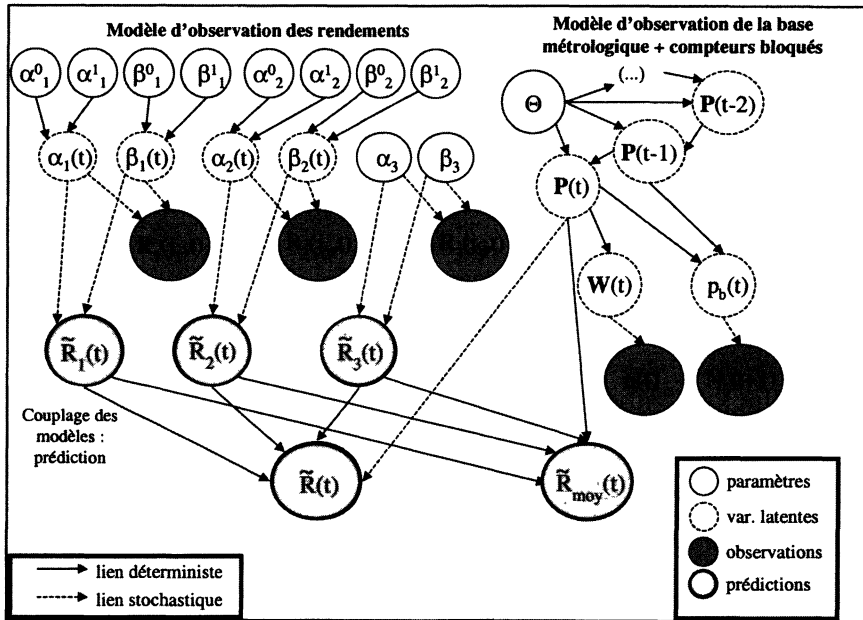


FIGURE 7  
 Modèle statistique de vieillissement sous forme de DAG

4.2. Inférence bayésienne par les méthodes MCMC

L'objectif de l'inférence bayésienne est l'obtention de la loi *a posteriori*  $[\xi|y]$  des paramètres  $\xi$ , du modèle, que le statisticien bayésien interprète comme la « mise à jour » d'une loi *a priori*  $[\xi]$  (exprimant la connaissance préliminaire sur ces grandeurs) suite à l'observation des données disponibles  $y$ .

La relation entre ces deux lois de probabilité est exprimée par la formule de Bayes qui, dans le cas de variables continues s'écrit :

$$[\xi|y] = \frac{[\xi][y|\xi]}{\int_{\Xi} [\xi][y|\xi] d\xi} \tag{16}$$

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 ( $\xi_1, \xi_2, \dots$ ) 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  $[\xi]$  (avec  $\xi \in \Xi \subseteq \mathbb{R}^d$ ) connue à une constante près.

L'algorithme de *Metropolis-Hastings* construit une chaîne de Markov apériodique et irréductible dans l'espace  $\Xi$ , ayant comme distribution stationnaire  $[\xi]$ . Alors, si l'on produit de nombreuses simulations de la chaîne, à partir d'une configuration initiale  $\xi_0$ , les « dernières valeurs obtenues » sont distribuées selon la loi stationnaire  $[\xi]$  (indépendante de  $\xi_0$ ) 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  $\xi = \{\xi_1, \xi_2, \dots, \xi_d\}$  conditionnellement à toutes les autres, appelées *conditionnelles complètes*. Pour démarrer la simulation on donne des valeurs initiales aux variables  $\xi^{(0)} = \{\xi_1^{(0)}, \xi_2^{(0)}, \dots, \xi_d^{(0)}\}$  et ensuite on procède itérativement de la manière suivante pour générer  $\xi^{(i)}$ , composante par composante :

$$\left. \begin{array}{l} \text{ITERATION } i \\ \text{Tirage de } \xi_1^{(i)} \text{ de } [\xi_1 | \xi_2^{(i-1)}, \dots, \xi_d^{(i-1)}] \\ \text{Tirage de } \xi_2^{(i)} \text{ de } [\xi_2 | \xi_1^{(i)}, \dots, \xi_d^{(i-1)}] \\ \dots \\ \text{Tirage de } \xi_d^{(i)} \text{ de } [\xi_d | \xi_1^{(i)}, \xi_2^{(i)}, \dots, \xi_{d-1}^{(i)}] \end{array} \right\} \quad (17)$$

La chaîne de Markov ainsi construite a une distribution stationnaire et cette distribution est la loi visée  $[\xi]$ . 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 répliques 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*.

### 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 (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 (14) chaque variable  $\xi_j$  (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 :

$$[\xi_j | \{\xi_i, i \neq j\}] = [\xi_j | bl(\xi_j)], \tag{18}$$

où :

$$bl(\xi_j) = pa(\xi_j) \cup ch(\xi_j) \cup \{\xi_i : ch(\xi_j) \cap ch(\xi_i) \neq \emptyset\}. \tag{19}$$

Dans la pratique des DAG, les conditionnelles complètes apparaissant dans les formules (17) se simplifient considérablement puisque seule la couverture markovienne est utilisée dans le conditionnement.

La *couverture markovienne* de  $\xi_j$ , 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  $\xi_j$  dans ce nouveau graphe.

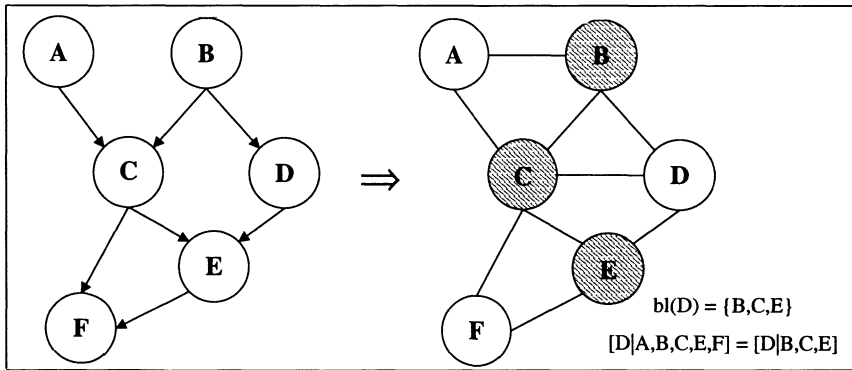


FIGURE 8  
Moralisation d'un DAG

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  $E$ .



## 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 Council) 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 :

$$\begin{aligned} [\theta_{11}, \theta_{12}, \theta_{13}, \theta_{14}] &= \frac{\Gamma(\sum_{j=1}^4 \alpha_{1j})}{\prod_{j=1}^4 \Gamma(\alpha_{1j})} \cdot \theta_{11}^{\alpha_{11}-1} \cdot \theta_{12}^{\alpha_{12}-1} \cdot \theta_{13}^{\alpha_{13}-1} \cdot \theta_{14}^{\alpha_{14}-1} \quad (20) \\ [\theta_{22}, \theta_{23}, \theta_{24}] &= \frac{\Gamma(\sum_{j=2}^4 \alpha_{2j})}{\prod_{j=2}^4 \Gamma(\alpha_{2j})} \cdot \theta_{22}^{\alpha_{22}-1} \cdot \theta_{23}^{\alpha_{23}-1} \cdot \theta_{24}^{\alpha_{24}-1} \\ [\theta_{33}, \theta_{34}] &= \frac{\Gamma(\sum_{j=3}^4 \alpha_{3j})}{\prod_{j=3}^4 \Gamma(\alpha_{3j})} \cdot \theta_{33}^{\alpha_{33}-1} \cdot \theta_{34}^{\alpha_{34}-1} \end{aligned}$$

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  $\alpha_{ij}$ . Un autre choix, plus classique, est la loi non informative de *Jeffreys* dans laquelle  $\alpha_{ij} = 1/2$ .

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  $\theta_{ij}$  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

TABLEAU 1

Variable	Moyenne	Percentiles		
		2.5 %	50 %	97.5 %
$\theta_{11}$	0.951	0.946	0.951	0.956
$\theta_{12}$	0.043	0.038	0.043	0.049
$\theta_{13}$	0.001	0.000	0.001	0.003
$\theta_{14}$	0.004	0.004	0.004	0.005
$\theta_{22}$	0.965	0.956	0.966	0.975
$\theta_{23}$	0.033	0.024	0.033	0.043
$\theta_{24}$	0.001	0.000	0.001	0.003
$\theta_{33}$	0.972	0.963	0.972	0.980
$\theta_{34}$	0.028	0.020	0.028	0.037

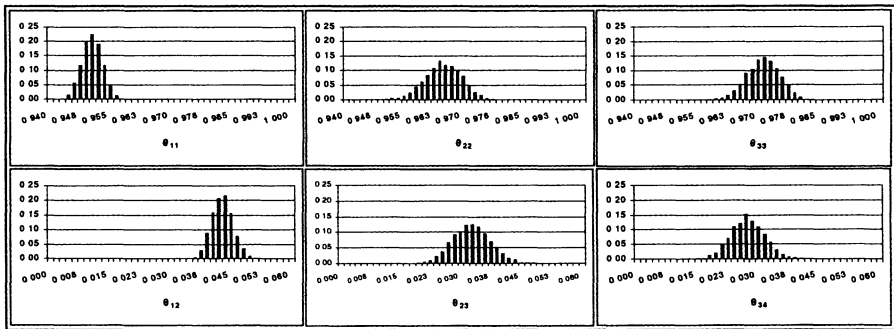


FIGURE 9  
 Estimation par méthode MCMC des distributions a posteriori des  $\theta_{ij}$

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).

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.

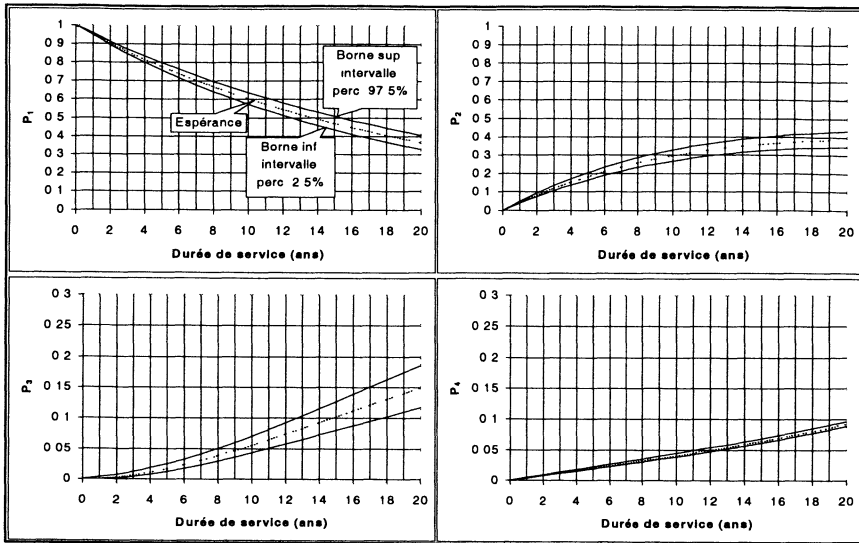


FIGURE 10

*Intervalles de crédibilité des probabilités d'appartenance aux 4 états*

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 (11).

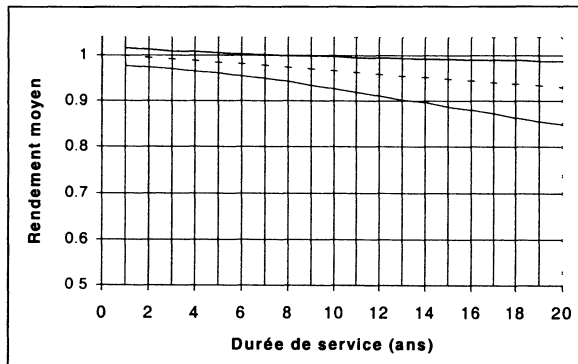


FIGURE 11

*Intervalles de crédibilité à 95% du rendement moyen d'un parc de compteurs*

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

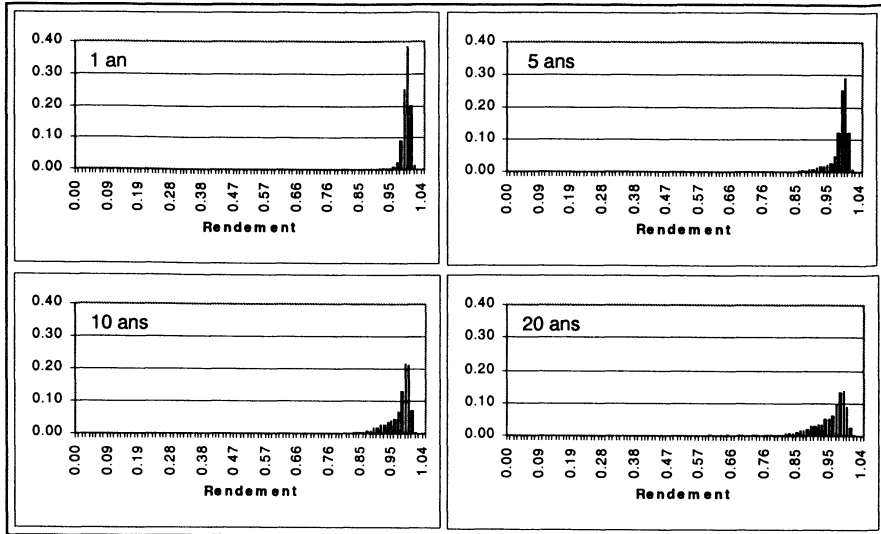


FIGURE 12

*Distributions prédictives du rendement d'un parc de compteurs*

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.

### 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  $\delta_i$  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  $\Delta$  de l'intervalle de crédibilité respectif (de niveau  $1 - \alpha$ ).

Le diagnostic de convergence se base sur la statistique (dite de *Brooks-Gelman*) :

$$\widehat{R}_{BG} = \frac{\Delta}{\frac{1}{k} \sum_{i=1}^{i=k} \delta_i}. \quad (21)$$

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*.

## 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  $p_{ob}$ .

La figure 13 montre les espérances *a posteriori* (obtenues par simulation *MCMC*) des probabilités d'appartenance aux 4 états météorologiques 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 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.

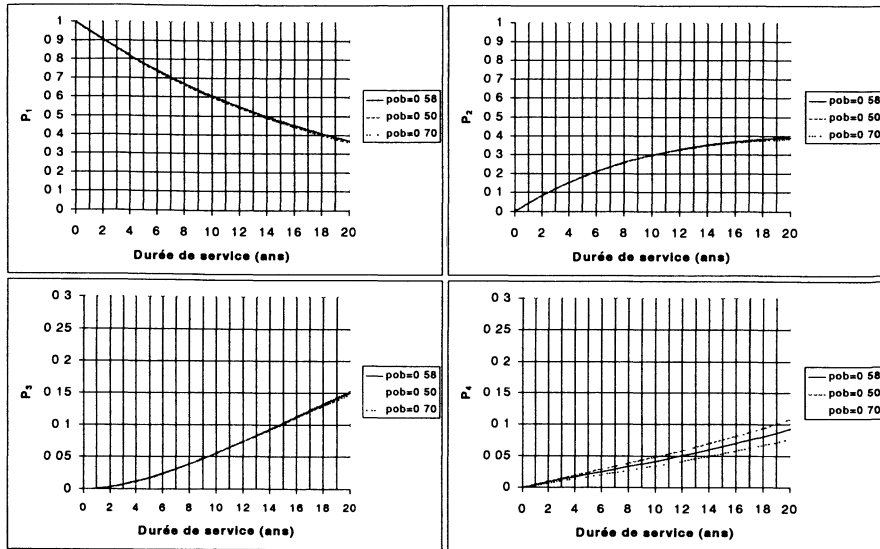


FIGURE 13  
 Estimateurs des probabilités d'appartenance aux 4 états  
 avec différentes valeurs de  $p_{ob}$

### 7. Discussion

#### 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  $[\xi|y]$ . On obtient ainsi des lois *prédictives a posteriori*  $[\tilde{y}|y]$  des «nouvelles» observations  $\tilde{y}$  sachant les «anciennes» données  $y$  :

$$[\tilde{y}|y] = \int_{\Xi} [\tilde{y}, \xi|y] d\xi = \int_{\Xi} [\tilde{y}|\xi][\xi|y] d\xi. \tag{22}$$

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 (p. ex.) 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.

La formule (22) est également utilisable dans un contexte de validation où l'on sépare les données en un échantillon d'apprentissage ( $y$ ) et un échantillon de test ( $\tilde{y}$ ).

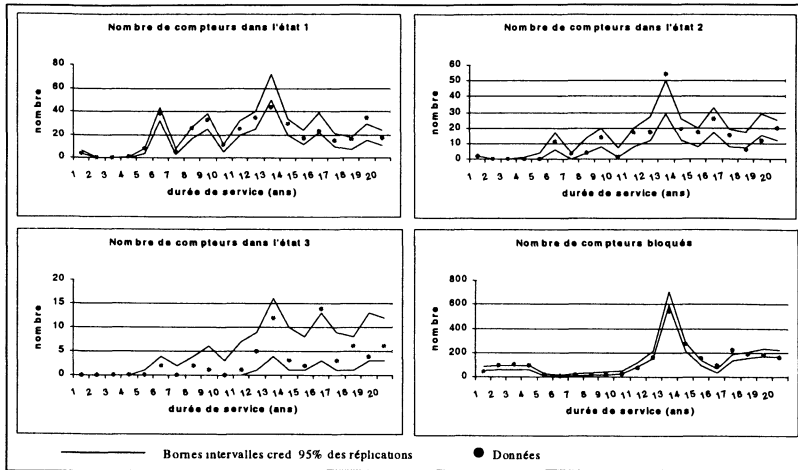


FIGURE 14

Tubes prédictifs à 95 % des répliations des observables et vraies données

Nous ne nous sommes pas livrés à cette opération, faute d'avoir un nombre suffisant de données.

### 7.2. Le poids des hypothèses

Les paramètres fondamentaux du modèle markovien sont les probabilités de transition  $\theta_{ij}$  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  $\theta_{ij}$  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 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  $\mathcal{E}_i \rightarrow \mathcal{E}_j$  ( $i \leq j$ ) 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

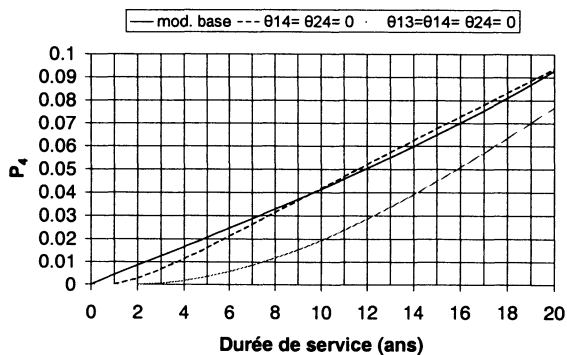


FIGURE 15

*Sensibilité de l'occurrence de l'état absorbant aux hypothèses sur les transitions possibles*

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 base ICBC.

### 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 littérature 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 appartiennent 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 donner 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<sup>3</sup>/an), le même impact qu'une année de vieillissement du compteur.

## 8. Référence

AMERICAN WATER WORKS ASSOCIATION (1966), 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.

BERNARDO J.M., SMITH A.F.M. (1994), *Bayesian Theory*. Wiley & Sons.

- BERNIER J., PARENT E., BOREUX J.J. (2000), *Statistiques pour l'Environnement. Traitement Bayésien des Incertitudes*. Tec & Doc.
- BROOKS S.P. (1998), *Markov Chain Monte Carlo Method and its application*. The Statistician, n. 47, Part 1, pp. 69-100.
- BROOKS S.P., GELMAN A. (1998), *General Methods for Monitoring Convergence of Iterative Simulations*. Journal of Computational and Graphical Statistics, Vol. 7, n. 4, pp. 434-455.
- CARLIER M. (1968), *Machines Hydrauliques*. Imprim. Louis-Jean.
- CONGDON P. (2001), *Bayesian Statistical Modelling*. Wiley & Sons.
- COSTES A., PIA Y. (2000), *Les compteurs d'eau en France : la Réglementation et son évolution*. Techniques, Sciences et Méthodes, n. 7, pp. 21-27.
- COWELL R.G., DAWID A.P., LAURITZEN S.L., SPIEGELHALTER D.J. (1999), *Probabilistic Networks and Expert Systems*. Springer-Verlag.
- COWLES M.K., CARLIN B.P. (1996), *Markov Chain Monte Carlo Convergence Diagnostics : A Comparative Review*. Journal of the American Statistical Association, Vol. 91, n. 434, pp. 883-904.
- GELFAND A.E., SMITH A.F.M. (1990), *Sampling-based approaches to calculating marginal densities*. Journal of American Statistical Association, n. 85, pp.398-409.
- GELMAN A., RUBIN D.B. (1992), *Inference from Iterative Simulation Using Multiple Sequences (with discussion)*. Statistical Science, n. 7, pp. 457-511.
- GELMAN A., CARLIN J.B., STERN H.S., RUBIN D.B. (1995), *Bayesian data analysis*. Chapman & Hall.
- GEMAN S., GEMAN D. (1984), *Stochastic Relaxation, Gibbs Distribution and the Bayesian Restoration of Images*. IEEE Transactions on Pattern Analysis and Machine Intelligence, n. 6, pp. 721-741.
- GILKS W.R., RICHARDSON S., SPIEGELHALTER D.J. (1996), *Markov Chain Monte Carlo in Practice*. Chapman & Hall.
- GOOD I.J. (1965), *The Estimation of Probabilities : An Essay on Modern Bayesian Methods*. M.I.T. Press.
- GRAU P. (1985), *Expériences dans le comptage de l'eau et dans le renouvellement des compteurs*. Water Supply, Vol. 3, n. 4 pp. 273-292.
- HASTINGS W.K. (1970), *Monte Carlo Sampling Methods using Markov Chains and their applications*. Biometrika, n. 57, pp. 97-109.
- LECOUTRE B., POITEVINEAU J. (1996), *Traitement statistique des données expérimentales : des pratiques traditionnelles aux pratiques bayésiennes*. CISIA-CERESTA.
- LEE P.M. (1997), *Bayesian Statistics : An Introduction*, 2nd Edition. Oxford University Press.
- LUND J.R. (1988), *Metering Utility Services : Evaluation and Maintenance*. Water Resources Research, Vol. 24, n. 6, pp 802-816.

- MERGENSEN K.L., ROBERT C.P., GUIHENNEUC-JOUYAUX C. (1999), *MCMC Convergence Diagnostics: A "review"*. 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.
- METROPOLIS N., ROSENBLUTH A.W., ROSENBLUTH M.N., TELLER A.H., TELLER E. (1953), *Equation of State Calculations by Fast Computing Machine*. Journal of Chemical Physics, n. 21, pp. 1087-1091.
- NEAL R.M. (1993), *Probabilistic Inference Using Markov Chain Monte Carlo Methods*, Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto.
- NEWMAN G.J., NOSS R.R. (1982), *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.
- NOSS R.R., NEWMAN G.J., MALE J.W. (1987), *Optimal Testing Frequency for Domestic Water Meters*. Journal of Water Resources Planning and Management, Vol 113, n. 1, pp. 1-14.
- ROBERT C.P. (1992), *L'Analyse Statistique Bayésienne*. Economica.
- ROBERT C.P. (1996), *Méthode de Simulation Monte Carlo par chaînes de Markov*. Economica.
- SPIEGELHALTER D.J., THOMAS A., BEST N.G. (1996), *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,.
- SPIEGELHALTER D.J., THOMAS A., BEST N.G. (1999), *WinBUGS Version 1.2 User Manual*. MRC Biostatistics Unit.
- TRONSKOLANSKI A.T. (1963), *Théorie et Pratiques des Mesures Hydrauliques*. Dunod.
- VAN DER LINDEN M.J. (1998), *Implementing a large-meter replacement program*. Journal of the American Water Works Association, Vol. 90, n.8, pp. 50-56.

### Annexe

L'espérance  $E$  et la variance  $V$  d'une variable aléatoire qui suit une loi Bêta, de paramètres  $\alpha$  et  $\beta$ , s'écrivent :

$$E = \frac{\alpha}{\alpha + \beta} \quad (A1)$$

$$V = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \quad (A2)$$

Dans les formules (A1) et (A2), en accord avec les formules (9) et (10),  $E$  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 :

$$\alpha'(t) = -\alpha_1 \cdot e^{\alpha_0 - \alpha_1 \cdot t} = -\alpha_1 \alpha \quad \beta'(t) = -\beta_1 \beta. \quad (A3)$$

On a donc :

$$E'(t) = \frac{\alpha'(\alpha + \beta) - \alpha(\alpha' + \beta')}{(\alpha + \beta)^2} = \frac{\alpha\beta}{(\alpha + \beta)^2} (\beta_1 - \alpha_1). \quad (A4)$$

$E'(t)$  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 (9) et (10) 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 :

$$v(t) = \ln(\alpha) + \ln(\beta) - 2 \ln(\alpha + \beta) - \ln(\alpha + \beta + 1). \quad (A5)$$

La dérivée de cette fonction, compte tenu des formules (9) et (10) est :

$$v'(t) = -(\alpha_1 + \beta_1) + \left( \frac{2}{\alpha + \beta} + \frac{1}{\alpha + \beta + 1} \right) (\alpha_1 \alpha + \beta_1 \beta). \quad (A6)$$

L'interprétation de la formule (A6) n'est pas immédiate. Par contre, pour les valeurs des paramètres évalués dans notre cas particulier, on vérifie  $v'(t) > 0$  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  $\ln(\alpha + \beta) \simeq \ln(\alpha + \beta + 1) \simeq \ln(\alpha)$  :

$$v(t) \simeq \ln(\beta) - 2 \ln(\alpha) = \beta_0 - 2\alpha_0 + (2\alpha_1 - \beta_1) \cdot t \quad (A7)$$

qui est évidemment croissante si  $\alpha_1 > \beta_1$ .