PRÉVISION DE L'INCIDENCE DES CANCERS DANS LE BAS-RHIN. APPROCHE BAYÉSIENNE. PRINCIPES ET APPLICATIONS

Daniel Eilstein

▶ **To cite this version:**

Daniel Eilstein. PRÉVISION DE L'INCIDENCE DES CANCERS DANS LE BAS-RHIN. APPROCHE BAYÉSIENNE. PRINCIPES ET APPLICATIONS. Santé publique et épidémiologie. Université Louis Pasteur Strasbourg I, 2001. Français. NNT : . tel-01595731

HAL Id: tel-01595731
https://hal.science/tel-01595731

Submitted on 26 Sep 2017

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

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

THÈSE DE DOCTORAT

DE L'UNIVERSITÉ LOUIS PASTEUR STRASBOURG I

FACULTÉ DE MÉDECINE DE STRASBOURG

ANNÉE : 2001

Spécialité : RECHERCHE CLINIQUE ET ÉPIDÉMIOLOGIQUE

présentée et soutenue publiquement par Daniel Eilstein

pour obtenir le grade de Docteur de l'Université Louis Pasteur Strasbourg I

le 18 octobre 2001 devant le jury composé de :

Professeur Simon Schraub (rapporteur interne et président)

Professeur Paul Schaffer

Professeur Jacques Estève (rapporteur externe)

Professeur Jean Faivre (rapporteur externe)

Professeur Élisabeth Quoix

Sujet de thèse :

PRÉVISION DE L'INCIDENCE DES CANCERS DANS LE BAS-RHIN. APPROCHE BAYÉSIENNE. PRINCIPES ET APPLICATIONS.

DIRECTEUR DE THÈSE : Paul Schaffer

THÈSE DE DOCTORAT

DE L'UNIVERSITÉ LOUIS PASTEUR STRASBOURG I

FACULTÉ DE MÉDECINE DE STRASBOURG

ANNÉE : 2001

Spécialité : RECHERCHE CLINIQUE ET ÉPIDÉMIOLOGIQUE

présentée et soutenue publiquement par Daniel Eilstein

pour obtenir le grade de Docteur de l'Université Louis Pasteur Strasbourg I

le 18 octobre 2001 devant le jury composé de :

Professeur Simon Schraub (rapporteur interne et président)

Professeur Paul Schaffer

Professeur Jacques Estève (rapporteur externe)

Professeur Jean Faivre (rapporteur externe)

Professeur Élisabeth Quoix

Sujet de thèse :

PRÉVISION DE L'INCIDENCE DES CANCERS DANS LE BAS-RHIN. APPROCHE BAYÉSIENNE. PRINCIPES ET APPLICATIONS.

DIRECTEUR DE THÈSE : Paul Schaffer

À Danièle,

À mes enfants : Cathy Anne, Joan, Lucien Niels

À mon père, à ma mère,

À ma sœur

À ma famille,

À mes amis.

À

Monsieur le Professeur Paul Schaffer pour l'accueil chaleureux qu'il nous a réservé dans son laboratoire et pour sa vision courageuse et engagée de la santé publique

Monsieur le Professeur Simon Schraub pour avoir bien voulu juger notre travail, pour ses qualités humaines et pour les marques d'amitié qu'il nous a toujours montrées

Monsieur le Professeur Jacques Estève pour les patients conseils qu'il nous a prodigués et pour la voie qu'il nous a ouverte dans l'un des domaines les plus passionnants des statistiques

Monsieur le Professeur Jean Faivre pour nous avoir montré que l'épidémiologie et la clinique sont indissociables dans la fabrication et la valorisation des concepts de santé publique

Madame le Professeur Élisabeth Quoix pour son enthousiasme et pour nous avoir fait l'amitié de juger notre travail

À

Monsieur le Professeur Michel Roos pour nous avoir révélé la pertinence de la complicité de l'intuition et du raisonnement dans la statistique qu'il a su replacer « dans le siècle »

Bernard Canguilhem qui, le tout premier, m'a accordé sa confiance d'enseignant avant de m'offrir son amitié

Guy Hédelin pour la rigueur de ses réponses à nos interrogations statistiques et pour l'acuité … de ses questions face au mêmes interrogations

Philippe Quénel qui nous a offert l'opportunité de notre premier travail d'épidémiologiste au service de la santé publique

Remerciements

À tout le personnel du laboratoire d'épidémiologie et de santé publique pour son accueil chaleureux et son aide logistique précieuse

À Saghir Bashir et Stephen W Dufy pour nous avoir aidé à concevoir la programmation de l'analyse.

PRÉVISION DE L'INCIDENCE DES CANCERS DANS LE BAS-RHIN.

APPROCHE BAYÉSIENNE. PRINCIPES ET APPLICATIONS.

« Il n'y a pas d'image vraie a priori. »[1]

[1] Ludwig Wittgenstein. Tractatus logico-philosophicus. Proposition 2.225. Traduction, préambule et notes de Gilles-Gaston Granger. Paris, Nrf, Gallimard 1993.

Liste des acronymes, sigles, symboles et notations utilisés

Acronymes

ADEMAS : Association pour le dépistage des maladies du sein
APC : âge-période-cohorte (modèle)
CIRC : Centre international pour la recherche sur le cancer
CNR : Comité national des registres
GAM : modèle additif généralisé
GLM : modèle linéaire généralisé
INSEE : Institut national de la statistique et des études économiques
MCMC : Monte Carlo par chaîne de Markov (méthodes de)

Notations

De façon générale :

Une lettre minuscule « non grasse » désigne un scalaire,
Une lettre minuscule « grasse » désigne un vecteur,
Une lettre majuscule « grasse » désigne une matrice,
Une lettre majuscule « non grasse » désigne une variable aléatoire,
Une lettre majuscule, grasse et italique désigne une variable aléatoire,

En particulier

Scalaires

$x, y, \mu, \eta, \ldots$

$x_i, y_i, \mu_i, \eta_i, \ldots$

Vecteurs

$\mathbf{x}, \mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\eta}, \ldots$

$\mathbf{x_i}, \mathbf{y_i}, \boldsymbol{\mu_i}, \boldsymbol{\eta_i}, \ldots$

a' désigne le vecteur transposé de **a**

Vecteurs colonnes

$\mathbf{y} = (y_1,y_2,\ldots,y_i,\ldots,y_n)'$

$\mu = (\mu_1,\mu_2,\ldots,\mu_i,\ldots,\mu_n)'$

$\eta = (\eta_1,\eta_2,\ldots,\eta_i,\ldots,\eta_n)'$

$\mathbf{x_1} = (x_{11},x_{21},\ldots,x_{i1},\ldots,x_{n1})'$

$\mathbf{x_2} = (x_{12},x_{22},\ldots,x_{i2},\ldots,x_{n2})'$

…

$\mathbf{x_j} = (x_{1j},x_{2j},\ldots,x_{ij},\ldots,x_{nj})'$

…

$\mathbf{x_p} = (x_{1p},x_{2p},\ldots,x_{ip},\ldots x_{np})'$

$\beta = (\beta_1,\beta_2,\ldots,\beta_j,\ldots,\beta_p)'$

Vecteurs lignes

$\mathbf{x_{[1]}} = (x_{11}, x_{12},\ldots, x_{1j},\ldots x_{1p})'$

$\mathbf{x_{[2]}} = (x_{21}, x_{22},\ldots, x_{2j},\ldots x_{2p})'$

…

$\mathbf{x_{[i]}} = (x_{i1}, x_{i2},\ldots, x_{ij},\ldots x_{ip})'$

…

$\mathbf{x_{[n]}} = (x_{n1}, x_{n2},\ldots, x_{nj},\ldots x_{np})'$

Matrices

X et (x_{ij}) désignent des matrices : $\mathbf{X} = (x_{ij})$ $((i,j) \in I \times J)$ avec I = [1,n], J = [1,p]

$$\mathbf{X} = \begin{bmatrix} x_{11} & \ldots & x_{1j} & \ldots & x_{1p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ x_{i1} & \ldots & x_{ij} & \ldots & x_{ip} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ x_{n1} & \ldots & x_{nj} & \ldots & x_{np} \end{bmatrix}$$

Variables aléatoires

Variables aléatoires individuelles

X, Y, Z, …

X_i, Y_i, Z_i, …

Vecteurs de variables aléatoires

$\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$

TABLE DES MATIÈRES

1. INTRODUCTION

« Une intelligence qui pour un instant donné connaîtrait toutes les forces dont la nature est animée et la situation respective des êtres qui la composent, si d'ailleurs elle était assez vaste pour soumettre ces données à l'analyse, embrasserait dans la même formule les mouvements des plus grands des corps de l'Univers et ceux du plus léger atome : rien ne serait incertain pour elle, et l'avenir comme le passé seraient présents à ses yeux.. »[2]

[2] Pierre Simon de Laplace. Essai philosophique sur les probabilités. Texte de la 5ème édition, 1825. Paris, Christian Bourgois, 1986.

Nos actions reposent presque toujours sur l'anticipation. Le mécanisme mental que l'on nomme prévision accompagne ainsi quotidiennement notre activité, observe notre passé, y recherche une certaine régularité et essaye de façon plus ou moins précise de (se) représenter un ou des événements futurs ou encore l'état futur d'un phénomène[3]. La prévision fait le postulat – ou le pari – de l'existence d'une certaine régularité dans le monde. La qualité de la prévision dépend cependant de la complexité des phénomènes et de la durée sur laquelle elle porte.

Prévision, prédiction, projection, prospective, autant de termes utilisés dans des acceptions proches et qu'il n'est pas toujours aisé de distinguer. La préférence sera accordée ici au terme de prévision mais sans exclure forcément les autres. En tout cas, à la base des prévisions établies dans ce travail, il y aura toujours la connaissance et l'analyse d'un passé. Ceci dit, pour couper cours à quelque ambiguïté que ce soit et afin d'exclure toute interprétation « exotique » assimilant, dans la confusion des sens, la prévision à la prédiction astrologique (qui, il est intéressant de le noter, a dorénavant doit de cité à l'Université !) ou à la divination. Paradoxalement d'ailleurs, la divination est plus proche de ce qui se fait ici car elle part du postulat que les dieux donnent aux hommes les clefs (cachées) à découvrir pour connaître leur avenir. Le seul divin pourtant, dans la prévision scientifique, est l'enfant issu du mariage de la mathématique et de la statistique : le modèle.

Et puisqu'il est question de modèle et de divin, donc de *déterminisme* et de prédiction ou plutôt de *prédictibilité*, ajoutons à la confusion en rappelant que les théories des systèmes dynamiques que l'on regroupe sous le nom de chaos déterministe (Glass, 1988 ; Bergé, 1988 ; Prigogine, 1994) ont invalidé le démon de Laplace, être à propos duquel, le mathématicien disait en 1776 : « L'état présent du système de la Nature est évidemment une suite de ce qu'il était au moment précédent, et si nous concevons une intelligence qui, pour un instant donné, embrasse tous les rapports des êtres de cet Univers, elle pourra déterminer pour un temps quelconque pris dans le passé ou dans l'avenir la position respective, les mouvements et, généralement, les affections de tous ces êtres. » (Laplace, 1776). Ce à quoi Henri Poincaré devait répondre en 1903 : « Si nous connaissions exactement les lois de la nature et la situation de l'Univers à l'instant initial, nous pourrions prédire exactement la situation de ce même Univers à un instant ultérieur. Mais, lors même que les lois naturelles n'auraient plus de secret pour nous, nous ne pourrions connaître la situation naturelle qu'approximativement. Si cela nous permet de prévoir la situation ultérieure avec la même approximation, c'est tout ce qu'il nous faut, nous disons que le phénomène a été prévu, qu'il est régi par des lois ; mais il n'en n'est pas toujours ainsi, il peut arriver que de petites différences dans les conditions initiales en engendrent de très grandes dans les phénomènes finaux ; une petite erreur sur les premières produirait une erreur énorme sur les derniers. La prédiction devient impossible et nous avons le phénomène fortuit. »

[3] Encyclopédie philosophique universelle. II Les notions philosophiques Tome 2. Presses universitaires de France, Paris, 1990

(Poincaré, 1903). Déterminisme et prédictibilité ne s'impliquent donc plus réciproquement (Ruelle, 1989).

En dépit des difficultés inhérentes au travail de prévision telles qu'elles ont été énoncées en préambule, l'augmentation de la fréquence de certains cancers et la part qu'ils prennent dans la mortalité générale incitent les pouvoirs publics à tenter d'anticiper la connaissance des paramètres futurs de ces pathologies et, entre autres, leur incidence. Les « décideurs » ont, en effet, pour mission de prévoir les infrastructures nécessaires à l'accueil et à la prise en charge des patients (nombre de lits, traitements, suivi des malades, mise en place d'éventuelles campagnes de dépistage, etc.). Cette nécessité montre, à l'évidence, l'intérêt de prévoir l'évolution des cancers en terme d'incidence (Hakulinen, 1991 ; Coleman, 1996). D'autre part, ces prévisions répondent à une demande licite d'information de la part du public et des professionnels de santé. Outre ces motifs que l'on pourrait qualifier de justifications *a priori*, il est un objectif *a postériori* non négligeable : les projections étant réalisées, il est utile de comparer la réalité à ce qui a été prévu et, s'il apparaît des différences, d'expliquer celles-ci par l'efficacité d'un traitement, d'une statégie de prévention, impact d'un dépistage, par exemple.

Paradoxalement, les travaux qui devraient contribuer à l'élaboration des connaissances nécessaires à l'anticipation des besoins, à la décision et aux modalités de la prise en charge sont très peu nombreux.
Aussi, l'objectif de ce travail est de réaliser une (la ?) prévision de l'incidence des cancers dans le Bas-Rhin, en tous cas des plus importants en termes d'impact sur la santé publique et de déterminer une méthode permettant de réaliser au mieux cet objectif. Les données d'incidences sont issues du registre des cancers du Bas-Rhin. De par sa nature, un registre présente les qualités d'exhaustivité, de fiabilité et de couverture géographique qui en font un outil idéal pour l'estimation des tendances et leur prévision (Coleman, 1996).

Différents moyens permettent de prédire l'évolution de l'incidence. Il est possible d'utiliser la connaissance actuelle des facteurs favorisants, de leur relation à la maladie et de leur évolution future (Aitio, 1990). Dans le cas du cancer du côlon, par exemple, l'hygiène de vie (habitudes alimentaires, sédentarité), les antécédents personnels et familiaux sont des facteurs de risque manifestes ; le cancer du sein est aussi relié à un ensemble de facteurs de risque, parmi lesquels l'absence de grossesse ou une première grossesse tardive, les défauts d'hygiène alimentaire (alimentation riche en graisse et/ou hypercalorique, consommation d'alcool) et certains facteurs héréditaires.
Ce qui est moins bien connu c'est la façon dont ceux-ci sont amenés à évoluer dans les prochaines années. D'autre part, il est plus que probable qu'un ensemble de facteurs psychologiques, culturels et ethniques modulent l'effet des autres facteurs rendant d'autant plus difficile l'élaboration d'un modèle exhaustif (Massé, 1995 ; Hertzman, 1996). Aussi est-il difficile d'utiliser cette méthode et la

préférence a été donnée ici à des méthodes de modélisation prenant en compte l'influence de l'âge (durée d'exposition aux facteurs de risque ou aux facteurs protecteurs), de la période (exposition simultanée de l'ensemble de la population à ces facteurs) et de la cohorte (exposition commune à des sujets nés au même moment). Quant à l'analyse statistique (i.e. le calcul des paramètres du modèle), plusieurs méthodes ont été testées pour retenir finalement une approche bayésienne utilisant un procédé de simulation itératif (échantillonnage de Gibbs).

Le plan adopté ici, propose, à la suite de ce chapitre de généralités, la revue des différentes méthodes de prévision (chapitre 2). Le chapitre 3 expose les différentes sources de données utilisées ainsi que les méthodes statistiques mises en œuvre dans l'analyse. Les cancers ayant fait plus particulièrement l'objet d'une prédiction ont été réunis dans le chapitre 4. Le chapitre 5 explore d'autres méthodes testées parallèlement à la méthode de référence. Les différents problèmes et difficultés, la mise en perspective des différentes méthodes utilisées et les conclusions sont rassemblés dans le chapitre 6. Enfin, et pour ne pas alourdir le texte, les notions statistiques ayant servi de base aux différentes analyses ont été placées en fin de document (chapitre 7).

2. REVUE DES DIFFÉRENTES MÉTHODES DE PRÉVISION

« Quant au troisième temps, qui découvre l'avenir – il signifie que l'événement, l'action ont une cohérence secrète excluant celle du moi, se retournant contre le moi qui leur est devenu égal, le projetant en mille morceaux comme si le gestateur du nouveau monde était emporté et dissipé par l'éclat de ce qu'il fait naître au multiple : ce à quoi le moi s'est égalisé, c'est l'inégal en soi. »[4]

[4] Gilles Deleuze. Différence et répétition. 8ème ed. Paris, Presses universitaires de France, 1996.

Le principe de la prévision repose sur une connaissance structurée du passé. De façon générale les méthodes adéquates procèdent en deux temps :

- Analyse du passé à la recherche d'une structure évolutive au sein des données ;
- Reproduction de cette structure dans le futur.

Les méthodes qui figurent dans ce chapitre ont, en général, accompli ces deux tâches. Certaines, cependant, ont pu s'arrêter à la première étape et réaliser une analyse de tendance. Mais il a été décidé de les citer tout de même car elles pourraient faire l'objet d'un travail complémentaire et opérer alors comme méthodes de prévision.

Les différentes méthodes de prévision envisagées ici sont déduites des techniques d'analyse des séries temporelles et concernent les séries chronologiques discrètes (appelées aussi séries temporelles discrètes) c'est à dire qu'elles sont adaptées à l'analyse de suites de valeurs indexées (par le temps, dans ce cas particulier).

La présentation de ces divers procédés a été réalisée en tentant de l'organiser selon une classification *raisonnée*. Deux ensembles généraux de techniques sont décrits selon qu'un terme aléatoire apparaît ou n'apparaît pas. Il s'agit, bien sur, d'une classification artificielle car il n'existe pas toujours de limite nette entre les différentes méthodes.

2.1. Méthodes non aléatoires

La série temporelle étudiée se présente sous la forme :

$$y_t, \text{ avec } t = 1, 2, \ldots, T.$$

2.1.1. Méthodes de prévision empiriques

Elles sont issues des méthodes empiriques utilisées en analyse des séries chronologiques (Gourieroux, 1983 ; Gourieroux, 1990 ; Coutrot B, 1990 ; Giraud, 1994 ; Bresson G, 1995). Ces techniques extraient des données un ensemble de paramètres dépendant du temps comme la tendance ou la saison (étape analytique) et *prolongent* ces fonctions du temps (étape de prévision).

2.1.1.1. Étape analytique

2.1.1.1.1. Analyse de la tendance

Plusieurs méthodes peuvent être utilisées qui sont bien décrites par Gourieroux (1990) :

- Le filtrage (ou désaisonnalisation) par moyennes mobiles,

Cette technique consiste à remplacer les termes d'une série temporelle y_t, t = 1, 2, …, T par une somme pondérée des valeurs de la série correspondant à des temps entourant celui de y_t.

Ainsi x_t est remplacé par la somme pondérée de m_1+m_2+1 valeurs de la série (dont y_t elle-même) :

$$y_t^* = \sum_{i=-m_1}^{m_2} \theta_i y_{t+i}$$

Si l'ordre de la série (m_1+m_2+1) est choisi correctement il est possible de faire disparaître le terme de saisonnalité.

Il est à noter que, si l'on conserve la même transformation pour toute la série chronologique, le nombre de valeurs (de points) dans la série filtrée est inférieur au nombre de valeurs dans la série initiale : les m_1 premier « points » et les m_2 derniers points de la série initiale n'ont pas de moyenne mobile.

- L'ajustement

La courbe de variation de la tendance est approchée par une fonction du temps (fonction polynomiale, exponentielle, logistique, de Gompertz) grâce à la méthode classique des moindres carrés. Ainsi, si f(t) est une fonction du temps approchée de la série y_t, la méthode des moindres carrés consiste à minimiser la somme des carrés des différences entre y_t et f(t) pour tous les instants t, ce qui revient à minimiser le carré de la distance G :

$$G = \sum_{t=1}^{T} [y_t - f(t)]^2$$

2.1.1.1.2. Détermination de la composante saisonnière

La détermination de la composante saisonnière passe par la mise en évidence de son existence. Cette étape est basée sur le calcul du coefficient d'autocorrélation de la série qui rend compte, pour chaque intervalle de temps, de la dépendance linéaire entre les mesures de la série. Le graphe représentant la variation de la valeur du coefficient d'autocorrélation en fonction du décalage est le corrélogramme.

2.1.1.2. Prédiction

- Moyenne mobile

Si l'analyse de la tendance s'est basée sur une moyenne mobile, il est possible d'imaginer un ensemble de techniques de prévision basées sur les valeurs obtenues par moyennes mobiles et les valeurs de la série initiale.

Si l'on suppose que la série initiale est :

$$y_t,\ t = 1, 2, \ldots, T$$

et que la série filtrée est

$$y^*_t,\ t = m_1+1, m_1+2, \ldots, T-m_2$$

La valeur prédite à T+1 peut être, par exemple, la moyenne arithmétique des $T - m_2 - m_1$ valeurs filtrées et des m_2 dernières valeurs de la série initiale soit, si $\hat{y}_T(1)$ désigne la valeur de y prédite à l'horizon T+1 :

$$\hat{y}_T(1) = \frac{y^*_{m_1+1} + y^*_{m_1+2} + \ldots + y^*_{T-m_2} + y_{T-m_2+1} + y_{T-m_2+2} + \ldots y_T}{T - m_1}$$

La prédiction est rarement de bonne qualité au delà de l'horizon T + 1.

Remarques :

1. Dans l'expression de la valeur prédite, il est possible de donner plus de poids aux dernières valeurs afin de tenir compte du fait que l'influence des observations sur la prédiction diminue avec le décalage.

2. Il est possible d'établir une prévision par moyenne mobile sans passer par le filtrage initial. En effet la valeur prédite à l'horizon peut-être calculée directement comme moyenne arithmétique de la série initiale.

- Ajustement

La prévision peut se calculer par application du modèle (établi par ajustement) au temps t + h. Il est même possible de rajouter un terme de saisonnalité, si celui-ci a été mis en évidence.

2.1.2. Les méthodes de lissage exponentiel

Le principe du lissage exponentiel est fondé sur l'hypothèse que l'influence des observations décroît exponentiellement au fur et à mesure que l'on s'éloigne de l'instant de prévision. Différents types de lissage exponentiel sont utilisables en prévision : le lissage exponentiel simple, le lissage exponentiel double et le lissage exponentiel généralisé.

Cette méthode réalise une prévision en un seul temps.

Le lissage exponentiel simple, par exemple, estime une prévision $\hat{y}_T(k)$ de la façon suivante :

$$\hat{y}_T(k) = (1-\beta)\sum_{j=0}^{T-1} \beta^j y_{T-j}$$

La constante de lissage β est comprise entre 0 et 1.

2.2. Méthodes avec modélisation

Les méthodes de modélisation diffèrent des précédentes car elles ajoutent une composante aléatoire à la composante déterministe du modèle.

2.2.1. Modèles de séries chronologiques

2.2.1.1. Processus ARMA

2.2.1.1.1. Définition

La notion de processus ARMA réunit celles de « processus autorégressif » et de « processus moyenne mobile » (Gourieroux, 1990). Dans les modèles ARMA, la valeur prise au temps t par la variable étudiée est une fonction linéaire de ses valeurs passées et des valeurs présentes ou passées d'un bruit blanc.

La forme générale d'un modèle ARMA (p, q) se présente de la façon suivante :

$$Y_t + \varphi_1 Y_{t-1} + \ldots + \varphi_p Y_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$$

ε_t est un bruit blanc.

Le processus ARMA (p, q) peut être représenté aussi par la symbolique suivante (équivalente de l'écriture précédente) :

$$\varphi_p(B) Y_t = \theta_q(B) \varepsilon_t$$

B est l'opérateur retard : $BY_t = Y_{t-1}$

$\varphi_p(B)$ est le polynôme $1 - \varphi_1 B - \varphi_2 B^2 - \ldots - \varphi_p B^p$

$\theta_q(B)$ est le polynôme $1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^p$

Le modèle ARIMA est un modèle ARMA auquel on applique un caractère de non stationnarité (le rajout d'un terme de tendance ou de saisonnalité par exemple). Le processus ARIMA s'écrit avec la symbolique suivante :

$$\varphi_p(B)Y_t \nabla^d Y_t = \theta_q(B)\varepsilon_t$$

Le terme $\nabla^d Y_t$ est un filtre différence d'ordre d : $\nabla^d = (1 - B)^d$

On suppose souvent que Y_t est un processus aléatoire normal stationnaire du deuxième ordre : les moments du 1er ordre ($E(Y_t)$) et du 2ème ordre ($E(Y_t^2)$) sont invariants par translation dans le temps[5].
Pour s et t, deux instants quelconques :

$$E(Y_t) = m,$$

$$Cov(Y_t, Y_s) = \gamma_{s-t}$$

avec γ_{s-t} ne dépendant que de (s-t).

Les modèles utilisés le plus fréquemment sont les modèles MA (stationnaire), AR (stationnaire sous certaines conditions), ARMA (stationnaire sous certaines conditions) et ARIMA (non stationnaire), SARIMA (non stationnaire).

2.2.1.1.2. Méthode de prévision de Box et Jenkins

La méthode de prévision de Box et Jenkins (Box, 1970 ; Brockwell, 1987 ; Gourieroux, 1990) est basée sur les processus ARIMA. Elle se présente sous forme d'un algorithme et comprend les phases suivantes (Figure 2.1.) :

- Identification du modèle ou plus précisément détermination des paramètres p, q et d.

Cette étape se base sur le corrélogramme et/ou le corrélogramme partiel. Des critères de sélection automatisée des paramètres ont été élaborés (Akaike, 1969 et 1977).

- Estimation des paramètres

Cette étape utilise le maximum de vraisemblance

[5] Pour plus de précision, voir le paragraphe 7.4.2.1. Processus aléatoire.

- Validité du modèle

Les résidus du modèle doivent se comporter comme un bruit blanc. Les test utilisés sont l'autocorrélation des résidus ou le test du portemanteau. Le modèle est alors accepté ou rejeté. Dans ce dernier cas il faut recommencer la procédure depuis le début. Si le modèle est accepté, il est possible de passer à l'étape suivante.

- Prévision

Il est possible de démontrer que si les données sont disponibles jusqu'à l'instant t (et donc jusqu'à la donnée y_t), alors (Coutrot, 1990 ; Gourieroux, 1990) :

$$y_{t+h} = \hat{y}_t(h) + e_t(h)$$

avec $\hat{y}_t(h) = \sum_{j=1}^{\infty} \alpha_j(h) y_{t+1-j}$,

$e_t(h) = \sum_{i=1}^{h-1} \beta_j \varepsilon_{t+h-i}$,

$\alpha_1(0) = \alpha_2(0) = \alpha_3(0) = \ldots = 0$,

et $\beta_0 = 1$

Les coefficients β_i, i = 1, 2, …, h-1, sont calculés par identification des coefficients correspondant à la même puissance dans l'égalité de polynômes

$$\varphi_p(B)(1 - B)^d(1 + \beta_1 B + \beta_2 B^2 + \ldots) = \theta_q(B)$$

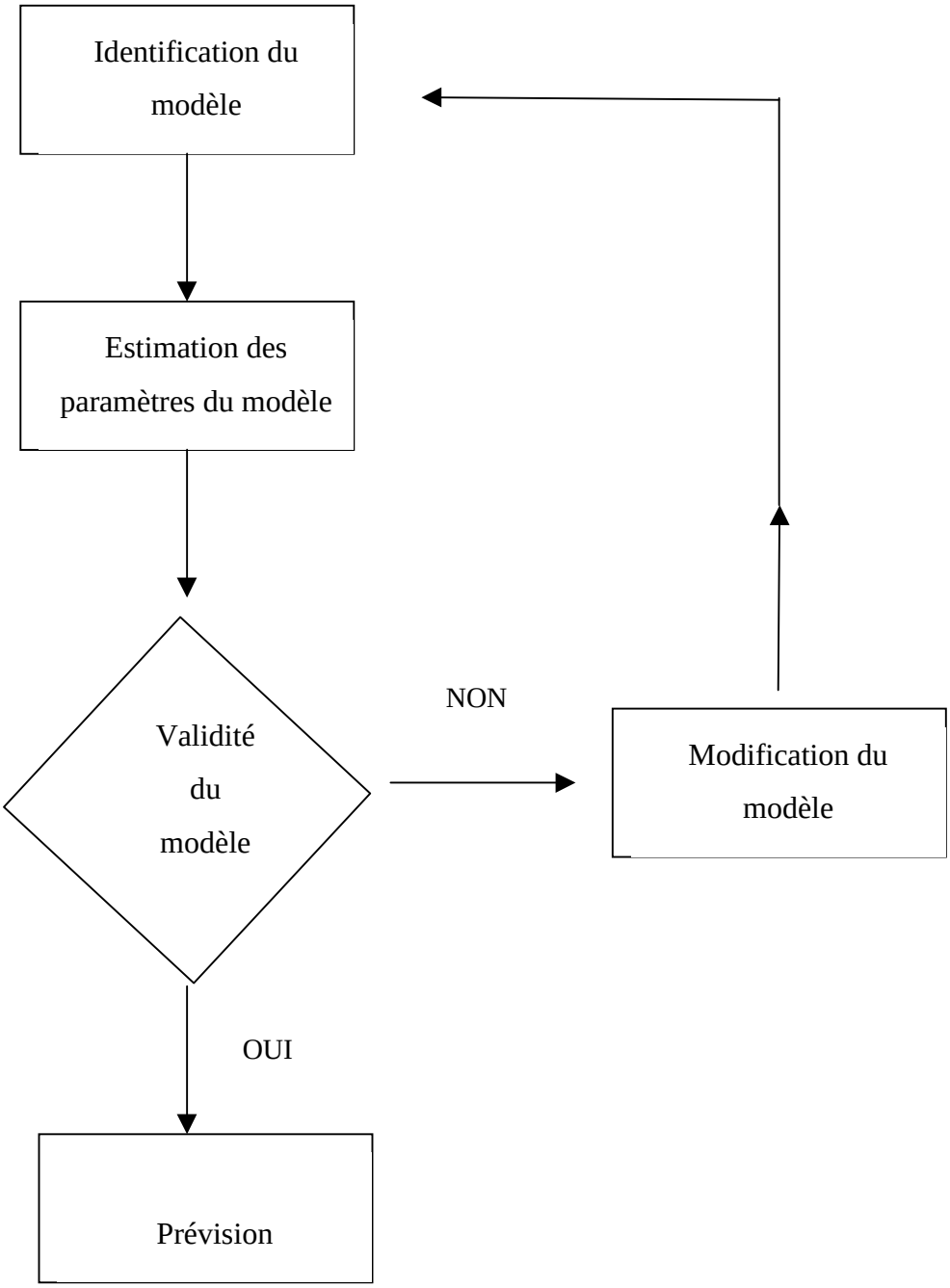
Les coefficients $\alpha_j(h)$ sont calculés grâce à la formule de récurrence :

$$\alpha_j(h) = \alpha_{j+1}(h-1) + \beta_{h-1}\alpha_j$$

Les coefficients α_j sont égaux à $\alpha_j(\mathbf{1})$ et sont calculés, comme les coefficients β_i par égalisation des coefficients de même rang (correspondant aux mêmes puissances) dans l'égalité de polynômes :

$$\theta_q(B)(1 - \alpha_1 B - \alpha_2 B^2 - \ldots) = \varphi_p(B)(1 - B)^d$$

Figure 2.1. Méthode de Box et Jenkins



2.2.1.2. Modèles non linéaires

Il peut arriver que les modèles linéaires ne soient pas adaptés à certaines données. Aussi, de nombreux modèles non linéaires ont ils été élaborés : processus de structure non linéaire (où Y_t est une fonction non linéaire des Y_{t-k}), modèles autorégressifs conditionnellement hétéroscédastiques ou modèles ARCH (qui permettent de prendre en considération des variances conditionnelles dépendant du temps), etc. (Droesbeke, 1994 ; Gourieroux, 1992 ; Guégan, 1994 ; Bresson , 1995).
Certains modèles non linéaires peuvent engendrer des systèmes chaotiques (modèles déterministes chaotiques) pouvant expliquer une partie de la variabilité qui ne serait alors plus laissée entièrement au hasard seul (May 1976 ; Sugihara, 1990 ; Guégan, 1994).

2.2.2. Modèles explicatifs

Ces modèles tentent de résumer l'incidence présente par un ensemble de facteurs. Puis ils déduisent l'évolution future de l'incidence à partir de la connaissance de celle des facteurs.

2.2.2.1. Principe

Le modèle le plus naturel et le plus opérationnel est le modèle linéaire généralisé (GLM).

2.2.2.1.1. Le modèle linéaire généralisé

Le modèle linéaire généralisé (Nelder, 1972 ; McCullagh, 1989, Fahrmeir, 1996 ; Lindsey, 1997) est sans doute l'outil le plus général, le plus utile et, par conséquent, le plus utilisé de la panoplie dévolue à la modélisation. La présentation détaillée du GLM est exposée au chapitre 7.

Avant de définir ce modèle, il convient de préciser quelques notations et définitions :

Soit y une grandeur à expliquer, y_i, i de 1 à n, les valeurs prises par la grandeur y lors de n mesures.
Soient x_1, x_2, …, x_j, …, x_p, p grandeurs explicatives. La variable générique x_j prend n valeurs x_{ij}, i de 1 à n.

Les n mesures sont présentées sous forme d'un tableau, comme suit (Tableau 2.1.).

Tableau 2.1. Variable expliquée et variables explicatives. Notations

n° observation	**y**	$\mathbf{x_1}$	$\mathbf{x_2}$	…	$\mathbf{x_j}$	…	$\mathbf{x_p}$
1	y_1	x_{11}	x_{12}	…	x_{1j}	…	x_{1p}
2	y_2	x_{21}	x_{22}	…	x_{2j}	…	x_{2p}
…	…	…	…	…	…	…	…
i	y_i	x_{i1}	x_{i2}	…	x_{ij}	…	x_{ip}
…	…	…	…	…	…	…	…
n	y_n	x_{n1}	x_{n2}	…	x_{nj}	…	x_{np}

Y_i est une variable aléatoire correspondant à la valeur mesurée y_i,

L_{exp} est une loi quelconque de la famille exponentielle,

μ_i est l'espérance de Y_i,

g() désigne une fonction,

η_i est appelé prédicteur,

β_j est un des paramètres à déterminer,

x_{ij} sont les valeurs prises par le facteur $\mathbf{x_j}$.

Le GLM s'écrit alors :

$$Y_i \sim L_{exp} \quad \text{et} \quad \mu_i = E[Y_i]$$
$$\eta_i = g(\mu_i)$$
$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$$

Ou encore, en utilisant la notation vectorielle et matricielle :

$$\boldsymbol{Y} \sim L_{exp} \quad \text{et} \quad \boldsymbol{\mu} = E[\boldsymbol{Y}]$$
$$\boldsymbol{\eta} = g(\boldsymbol{\mu})$$
$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

Y est le vecteur des variables aléatoires,

μ est le vecteur des espérances,

η est le vecteur des prédicteurs,

β est le vecteur des paramètres à déterminer,

X est la matrice des cofacteurs

De plus l'écriture générale de la densité de probabilité d'une loi exponentielle se présente de la façon suivante :

$$f_Y(y|\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)$$

2.2.2.1.2. Hypothèses particulières à l'analyse de l'incidence

Les hypothèses propres à l'analyse de l'incidence ont été largement exposées par Esteve et *al.* (Estève, 1993).
Les notions principales sont les suivantes :

1° L_{exp} est une loi de Poisson (Breslow, 1984). En effet, le numérateur de l'incidence (nombre de cas) est un nombre entier et, de plus, il est petit, relativement à l'effectif de la population à risque.
2° g est la fonction ln

Si y_i est le nombre de cas incidents au cours de la période i, Y_i la variable aléatoire correspondante, m_i le nombre de personnes-années (ou mois, …) et λ_i l'incidence, les deux définitions précédentes deviennent :

$$Y_i \sim P(m_i\lambda_i) \qquad \text{et} \qquad m_i\lambda_i = E[Y_i]$$
$$\eta_i = \ln(\lambda_i)$$
$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$$

Avec :

$$P(Y_i = y_i) = e^{-\lambda_i m_i} \frac{(\lambda_i m_i)^{k_i}}{y_i!}$$

Il est à noter qu'une petite variation a été introduite par rapport à la définition de base : c'est λ_i et non μ_i (soit $m_i\lambda_i$) qui est modélisé par l'intermédiaire de η_i. Mais ceci ne change pas grand chose à la structure et à la signification du modèle.

Il est encore possible d'écrire le modèle en utilisant la notation vectorielle et matricielle :

$$\boldsymbol{Y} \sim P(\mathbf{M}\boldsymbol{\lambda}) \quad \text{et} \quad \mathbf{M}\boldsymbol{\lambda} = E[Y]$$
$$\boldsymbol{\eta} = \ln(\boldsymbol{\lambda})$$
$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

Avec

$$\mathbf{M} = \begin{bmatrix} m_1 & & & \\ & . & & 0 \\ & & m_i & \\ & 0 & & . \\ & & & m_n \end{bmatrix}$$

2.2.2.2. Différentes applications du modèle linéaire généralisé

La prévision est établie, là aussi, à partir de la connaissance de l'incidence passée (il s'agit d'approcher au mieux les propriétés évolutives / attributs de l'incidence passée et de projeter celles-ci). D'où, de manière globale, il existe 3 façons d'aborder ce problème.

2.2.2.2.1. Prévision réalisée en *prolongeant* l'incidence

2.2.2.2.1.1. Principe

La variation temporelle de l'incidence est modélisée, c'est à dire que la courbe représentative de la variation temporelle de l'incidence est approchée au mieux par une fonction du temps (linéaire, polynomiale, exponentielle, etc.).
Donc, ici, l'attribut est le temps lui-même. L'avantage de ce modèle réside dans sa simplicité
mais ses inconvénients sont issus de cette même simplicité ! En effet, l'écriture de la dépendance au temps est relativement rigide, aussi s'adapte t'elle mal aux variations des données ce qui rend la prédiction peu fiable pour le long terme.

En fait, cette méthode est très semblable à la méthode empirique de type ajustement. Cependant, l'approche est plus formalisée. De plus, l'écriture introduit une partie aléatoire.

Les composantes des vecteurs **y**, $\boldsymbol{\lambda}$ et $\boldsymbol{\eta}$ sont indexées par le temps (les événements peuvent être comptabilisés à des temps également espacés ou non, t_1, t_2,…, t_i,…, t_n).

$$\mathbf{y} = \left(y_{t_1}, y_{t_2}, ..., y_{t_i}, ..., y_{t_n}\right)'$$

$$\boldsymbol{\lambda} = \left(\lambda_{t_1}, \lambda_{t_2}, ..., \lambda_{t_i}, ..., \lambda_{t_n}\right)'$$

$$\mathbf{M} = \begin{bmatrix} m_{t_1} & & & & \\ & . & & 0 & \\ & & m_{t_i} & & \\ & 0 & & . & \\ & & & & m_{t_n} \end{bmatrix}$$

$$\boldsymbol{\eta} = \left(\eta_{t_1}, \eta_{t_2}, ..., \eta_{t_i}, ..., \eta_{t_n}\right)'$$

$$\mathbf{X} = \begin{bmatrix} f_1(t_1) & ... & f_j(t_1) & ... & f_p(t_1) \\ ... & ... & ... & ... & ... \\ f_1(t_i) & ... & f_j(t_i) & ... & f_p(t_i) \\ ... & ... & ... & ... & ... \\ f_1(t_n) & ... & f_j(t_n) & ... & f_p(t_n) \end{bmatrix}$$

Les expressions f_j, j de 1 à p, représentent des fonctions du temps.

Remarque : la suite de variables aléatoires $(Y_1, Y_2, \ldots, Y_i, \ldots, Y_n)$ forme un processus.

Pour simplifier la notation, l'indexation par les naturels sera substituée à l'indexation par le temps :

$$\mathbf{y} = (y_1, y_2, \ldots, y_i, \ldots, y_n)'$$

$$\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_i, \ldots, \lambda_n)'$$

$$\mathbf{M} = \begin{bmatrix} m_1 & & & \\ & . & 0 & \\ & & m_i & \\ & 0 & & . \\ & & & m_n \end{bmatrix}$$

$$\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_i, \ldots, \eta_n)'$$

$$\mathbf{X} = \begin{bmatrix} f_1(t_1) & \ldots & f_j(t_1) & \ldots & f_p(t_1) \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ f_1(t_i) & \ldots & f_j(t_i) & \ldots & f_p(t_i) \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ f_1(t_n) & \ldots & f_j(t_n) & \ldots & f_p(t_n) \end{bmatrix}$$

La fonction f est une fonction quelconque du temps (linéaire, quadratique, etc.).

$$\beta = (\beta_1, \beta_1, \ldots, \beta_j, \ldots, \beta_p)'$$

Le GLM prend la forme suivante :

$$Y_i \sim P(m_i \lambda_i) \qquad \text{et} \qquad m_i \lambda_i = E[Y_i]$$

$$\eta_i = \ln(\lambda_i)$$

$$\eta_i = \sum_{j=1}^{p} \beta_j f_j(t_i)$$

Un cas simple peut être caractérisé, par exemple, par : $f_1 = x$ et $f_2 = x^2$

Alors :

$$\lambda = e^{\beta_1 t + \beta_2 t^2}$$

2.2.2.2.1.2. Exemple

Hakulinen et Dyba (Hakulinen, 1994), Dyba, Hakulinen et Päivärinta (Dyba, 1997), puis Dyba et Hakulinen (Dyba, 2000) comparent différentes types de prédiction de l'incidence basées sur des interpolations. Les modèles testés sont simples.

Si λ_t et λ_{it} sont respectivement l'incidence standardisée selon l'âge, au temps t et l'incidence spécifique pour l'âge i, au temps t, les modèles sont les suivants :

- Lorsque le cancer étudié présente une incidence croissante : $E(\lambda_t) = \alpha + \beta t$,
- Lorsque le cancer étudié présente une incidence décroissante : $\ln[E(\lambda_t)] = \alpha + \beta t$,
- Pour réaliser des prévisions plus spécifiques de la tranche d'âge, les deux modèles précédents peuvent être remplacés par : $E(\lambda_{it}) = \alpha_i + \beta_i t$, $\ln[E(\lambda_{it})] = \alpha_i + \beta_i t$ et $\ln[E(\lambda_{it})] = \alpha_i + \beta t$

Les auteurs font d'abord l'hypothèse que les distributions des incidences spécifiques sont normales et indiquent le calcul de l'intervalle de confiance. Puis, afin d'estimer conjointement les intervalles de confiance des nombres de cas incidents et des taux d'incidences spécifiques, il font l'hypothèse que les nombres de cas spécifiques au temps t sont indépendants et suivent une loi de Poisson.
Une réflexion générale et exhaustive au sujet des variances et intervalles de confiance a été menée par Hakinen *et al*. (1994).

Wiklund *et al.* avaient utilisé des modèles approchés des modèles précédents pour estimer la tendance future de l'incidence standardisée pour prévoir la mortalité par cancer en fonction de différentes interventions. La prédiction est basée sur les modèles non spécifiques (Wiklund, 1992).

2.2.2.2.2. En tenant compte des facteurs externes influençant l'incidence

2.2.2.2.2.1. Principe

Les méthodes de prédiction de l'incidence dépendent de la connaissance que l'on a des facteurs qui l'influencent (Hakulinen, 1991).
Cette approche est idéale mais nécessite la connaissance des facteurs influençant l'incidence (Aitio, 1990) ainsi que celle de leur évolution temporelle. Ceci est rarement, sinon jamais, réalisé. Tout au plus est-il possible d'imaginer certains scénarios.

Le GLM prend, dans ce cas, la forme générale vue plus haut mais la matrice X représente les valeurs prises au cours du temps par les variables explicatives.

2.2.2.2.2.2. Exemples

Kessler (1991), par exemple, a construit un modèle basé sur une somme de facteurs, témoins d'une tendance à long terme et d'un facteur représentant l'influence du dépistage :

$$\ln [E(n_{say})] = \ln(N_{ay}) + \beta_s + \beta_{as} * \text{UNITS}_y$$

n_{say} est la variable nombre de cas de cancer pour le stade s, l'âge a et la période y : n_{say} est supposé suivre une loi de Poisson,
$E(n_{say})$ représente le nombre de cas attendu pour le stade s, l'âge a et la période y,
N_{ay} représente le nombre de cas de base pour l'âge a et la période y (donc en l'absence de dépistage),
β_s est un terme représentant le nombre de cas de base pour le stade s ; c'est ce terme qui servira à la prévision à long terme,
UNITS_y est le nombre d'appareils radiographiques destinés aux mammographies, opérationnels durant la période ; ce nombre est un *proxy* du nombre de mammographies pratiquées durant la période y,
β_{as} est le coefficient (pour l'âge a et le stade s) de UNITS_y.
Le graphe représentant les variations de UNITS en fonction de la période est une courbe en S, pouvant être modélisée par une fonction logistique :

$$\text{UNITS}_y = \frac{\max \text{UNITS}}{1 + e^{\alpha + \beta y}}$$

Hakulinen (1991) a réalisé une revue des méthodes de prévision de l'incidence basées sur la projection de la tendance des facteurs de risque. Hakulinen et Pukkala avaient réalisé dix ans plus tôt une prévision de la tendance du cancer du poumon en imaginant un ensemble de scénarios relatifs à la consommation de tabac (proportion de sujets arrêtant de fumer dans différentes catégories d'âge) (Hakulinen, 1981). Il avaient conçu un modèle tenant compte du risque lié au nombre de cigarettes quotidien, à la durée du tabagisme, au délai écoulé depuis l'arrêt du tabagisme. Le modèle est représenté par le rapport entre l'incidence des sujets ayant arrêté de fumer et les sujets n'ayant jamais fumé au même âge :

$$S_{iklm} = 1 + \frac{s_k - 1}{s_{20} - 1}(b_m - 1)x_l$$

Avec s_k, le risque relatif lié au nombre k de cigarettes fumées par jour,

b_m, le risque relatif lié au nombre m d'année de tabagisme,
w_l, le risque relatif lié au nombre d'années écoulées l depuis l'arrêt du tabac (normalement, $w_l < 1$).
Le résultat de cette étude est exprimé grâce à un ensemble de courbes représentant l'incidence standardisée prévues jusqu'en 2050.

2.2.2.2.3. Modèle âge-période-cohorte

2.2.2.2.3.1. Principe général

Cette approche occupe une place intermédiaire en regard des deux méthodes précédentes : elle n'impose pas la connaissance des facteurs explicatifs extrinsèques et ne présente pas la rigidité des modèles strictement temporels. Ici, d'une certaine façon, les facteurs en jeu sont « intrinsèques » (Estève, 1993). Le facteur « âge » est une variable physiologique ; il représente la durée de l'exposition à des facteurs de risque. Le facteur « période » correspond à la variation de l'exposition de l'ensemble de la population à des facteurs de risques ou protecteurs au cours du temps. La « cohorte de naissance » témoigne d'une exposition touchant des sujets à un moment clef de leur vie, en particulier lors de la petite enfance ou même à la période embryonnaire.

Lorsque les différents paramètres ont été calculés par l'analyse du modèle, la prévision s'établit sur la base d'une projection des effets « période » et « cohorte », tout en maintenant, constant, l'effet « âge » si les âges maximal et minimal de la population étudiée sont considérés comme invariants.

Pour chacun des sujets inclus dans l'étude d'une pathologie, les paramètres connus sont, en général, l'âge au diagnostic (ou au décès) et la date de diagnostic (ou du décès). Les valeurs de ces paramètres sont groupées, respectivement, par tranches d'âge et par périodes. Les amplitudes de ces intervalles peuvent être fixes ou variables selon le cas. Généralement, elles sont fixes et de même valeur pour les deux variables, de l'ordre de quelques années (le plus souvent cinq ans). Les intervalles ainsi déterminés sont indicés et forment un repère à deux dimensions (âge et période). Tout événement (décès, cas incident) occupe une place précise dans le plan ainsi formé et l'ensemble des nombres d'événements correspondant aux âges et aux périodes est réuni en un tableau appelé tableau âge-période (Tableau 2.2).

Par la suite il sera convenu que l'étude porte sur I tranches d'âge et J périodes. L'indice i est compris entre 1 et I, l'indice j est compris entre 1 et J.

Tableau 2.2. Tableau âge-période

âge \ période	**période 1**	**période 2**	**…**	**période j**	**…**	**période J**
âge 1	n_{11}	n_{12}	…	n_{1j}	…	n_{1J}
âge 2	n_{21}	n_{22}	…	n_{2j}	…	n_{2J}
…	…	…	…	…	…	…
âge i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{iJ}
…	…	…	…	…	…	…
âge I	n_{I1}	n_{I2}	…	n_{Ij}	…	n_{IJ}

Les nij représentent le nombre de cas correspondant à l'âge i et à la période j.

Les cohortes de naissance (Tableau 2.3.) se disposent de la façon suivante (la première cohorte ou cohorte n° 1 est la plus ancienne et se trouve dans la case inférieure gauche du tableau, la dernière cohorte ou cohorte n° I+J-1 est la plus récente et se trouve dans la case supérieure droite) :

Tableau 2.3. Tableau âge-période et numéro de la cohorte de naissance correspondante

âge \ période	**période 1**	**période 2**	**…**	**période j**	**…**	**période J**
âge 1	I	I+1	…	I+j -1	…	I+J-1
âge 2	I-1	I	…	I+j -2	…	I+J-2
…	…	…	…	…	…	…
âge i	I+1-i	I+2-i	…	I+j-i	…	I+J-i
…	…	…	…	…	…	…
âge I	1	2	…	j	…	J

Il existe, bien sur, une relation entre les niveaux (ou numéros ou indices) de ces trois facteurs.

- Si I est le nombre total de classes d'âge, J le nombre total de périodes et K le nombre total de cohortes, alors :

$$K=I+J-1$$

- Si une case correspond à la classe d'âge i et à la période j, le numéro de la cohorte correspondante est :

$$k = j + I - i$$

Dans la « case » (i, j) ou (i, j, k) puisque k est défini dès que i et j le sont, les « valeurs » respectives des effets « âge », « période » et « cohorte » sont notés, respectivement, a_i, p_j et c_k. Pour alléger l'écriture et lorsqu'il n'y aura pas d'ambiguïté, ces effets serons notés, respectivement, a, p et c.

La relation générale, dans la case (i, j, k), entre le nombre de cas incidents (y_{ijk}) ou, plus précisément, la variable aléatoire correspondante (Y_{ijk}), le nombre de personnes-années (m_{ijk}) et le taux d'incidence spécifique (λ_{ijk}) est :

$$E[Y_{ijk}] = \lambda_{ijk}\, m_{ijk}$$
$$Y_{ijk} \sim P(\lambda_{ijk}\, m_{ijk})$$

D'où, le modèle devient :

$$Y_{ijk} \sim P\left(m_{ijk}\lambda_{ijk}\right) \quad \text{et} \quad m_{ijk}\,\lambda_{ijk} = E\left[Y_{ijk}\right]$$
$$\eta_{ijk} = \ln(\lambda_{ijk})$$
$$\eta_{ijk} = a_i + p_j + c_k$$

2.2.2.2.3.2. Différentes approches et différents modèles

Le modèle âge-période-cohorte est un canevas relativement général dans lequel se reconnaissent un ensemble de modèles utilisés en épidémiologie. Les différences entre les divers modèles viennent de la nature des facteurs a, p et c. Ces derniers peuvent être des variables quantitatives ou qualitatives.

(a) Variables quantitatives

(α) Méthode du CIRC

Cette approche a été utilisée par le Centre international pour la recherche sur le cancer (CIRC) pour modéliser les tendances de l'incidence et de la mortalité (Coleman, 1993).

Soient i, j et k, (i, j, k) $\in N^3$, les niveaux respectifs de l'âge de la période et de la cohorte. Le modèle s'écrit :

$$\ln(\lambda_{ijk}) = a(i) + p(j) + c(k)$$

Les lettres a, p et c représentent des polynômes en i, j et k, respectivement .
La modélisation détermine le degré de chaque polynôme. Les modalités relatives au choix du meilleur modèle sont largement explicité au chapitre « Méthode » du travail de Coleman et al. (1993) et sont rappelées dans le chapitre 5.2. du présent travail (Méthode du CIRC).

Remarque : en raison du problème d'identification des effets âge période et cohorte lié à la dépendance des niveaux des paramètres (k = j + I – i), les modèles de Coleman *et al.* (1993) ne comportent pas de terme de degré 0 (donc de constante) dans les polynômes p et c ni de terme de degré 1 (donc de terme linéaire) dans le polynôme c.
Le modèle s'écrit :

$$Y_{ijk} \sim P(m_{ijk}\lambda_{ijk}) \quad \text{et} \quad m_{ijk}\lambda_{ijk} = E[Y_{ijk}]$$
$$\eta_{ijk} = \ln(\lambda_{ijk})$$
$$\eta_{ijk} = a(i) + p(j) + c(k)$$

Lorsque la modélisation est achevée (par le choix du meilleur modèle et le calcul des paramètres a, p et c), la prévision est réalisée par projection des effets « période » et « cohorte », tout en maintenant constant l'effet « âge ». Il est à noter que, lorsque le modèle contient des termes de degré supérieur à 1 - ce qui le cas, le plus souvent - la prévision devient hasardeuse au delà du pas 1, en raison de l' « explosion » des fonctions respectives de la période et de la cohorte.

Ahsan (1995) a utilisé la même méthode pour l'étude de l'incidence des tumeurs cérébrales malignes en se limitant à un modèle âge-période :

$$\ln\lambda = \alpha + \beta_1 i + \beta_2 j$$

Avec i, la classe d'âge et j, la période.

(β) Méthode de Price

Price (1997) adopte un modèle basé sur une courbe de fonction logistique :

$$\lambda = \frac{\beta_1}{1 + \beta_2 e^{-\beta_3 j}}$$

λ est l'incidence standardisée ou une incidence spécifique, j est l'année de diagnostic, β_1, β_2 et β_3 sont des coefficients à déterminer.

(b) Variables qualitatives

Ici a, b et c sont des variables factorielles donc des variables à plusieurs niveaux, correspondant aux classes d'âge, aux périodes et aux cohortes.

Comme précédemment, il est supposé qu'il existe I classes d'âge, J périodes et K cohortes. Les valeurs prises par les trois facteurs sont respectivement :

$a_1, a_2, \ldots, a_I$ pour la classe d'âge,
$p_1, p_2, \ldots, p_J$ pour l'année,
$a_1, a_2, \ldots, a_K$ pour la cohorte.

Le modèle s'écrit :

$$\ln(\lambda_{ijk}) = a_i + p_j + c_k$$

Le triplet (i,j,k) appartient à N^3 et a, p, c sont les valeurs des effets des facteurs à déterminer.
Le modèle complet s'écrit de la façon suivante :

$$Y_{ijk} \sim P(m_{ijk}\lambda_{ijk}) \quad \text{et} \quad m_{ijk}\lambda_{ijk} = E[Y_{ijk}]$$
$$\eta_{ijk} = \ln(\lambda_{ijk})$$
$$\eta_{ijk} = a_i + p_j + c_k$$

Le modèle âge-période-cohorte à variables qualitative a fait l'objet de nombreuses approches. Celles-ci se distinguent par le mode de résolution du problème d'identification des trois effets.

(α) Méthode de Decarli et La Vecchia

Cette méthode (Decarli, 1987) est basée sur l'approche d'Osmond et Gardner (Osmond, 1982 a). Cette dernière a été proposée dans le cadre d'une analyse de tendance (mortalité) dans le but – comme il a été rappelé ci-dessus – de traiter le problème récurrent d'identification des effets âge, période et cohorte au sein du modèle basé sur ces trois facteurs. Cette technique a été largement utilisée depuis car elle propose une approche séduisante de la résolution du problème tant dans le domaine du cancer (Osmond, 1982 b ; Negri, 1990 ; McNally, 1997 ; Evstifeeva, 1997) que dans d'autres disciplines

(O'Callaghan, 2000) : elle construit trois modèles incluant deux des trois composantes en fixant respectivement l'effet âge, l'effet cohorte et l'effet période puis calcule leur adéquation aux données (déviance) et choisit le modèle final comme barycentre des trois modèles, pondérés par l'inverse de leur déviance. De plus un programme a été élaboré sous GLIM (Decarli, 1987)[6].
Cette méthode est critiquable comme moyen de résolution du problème d'identification mais dans le cas d'une prédiction l'indétermination de la part prise par chacun des trois facteurs n'a pas d'effet car ce qui est calculé ici est l'effet total et non pas les effets partiels.

(β) Méthode de Holford

Holford (1983 et 1991), Vioque (1993), Zheng (1995) et Dyba (1997) traitent le problème de l'identification en modélisant l'incidence avec des termes linéaires et des termes de courbure représentant les effets âge, période et cohorte. Les termes de courbure sont, en fait, des polynômes du second degré au moins, orthogonaux. Cette méthode permet ainsi d'estimer les trois effets sur la tendance. La prévision de l'incidence à partir de ce modèle est également possible (Dyba, 1997).

(γ) Méthode de Clayton et Schifflers

Clayton et Schifflers (Clayton, 1987 a et b) testent successivement différents modèles dont la complexité augmente : « âge seul » puis « âge et dérive » puis « âge et période », « âge et cohorte », « âge, période et cohorte ».

Remarque : la dérive est l'effet linéaire de la période ou / et de la cohorte.

Clayton et Schifflers tentent de lever le problème de l'identification de trois façons :

- En rajoutant un terme de dérive au paramètre de la période $p^*_j = p_j + \delta(j-1)$ et δ est calculé de telle façon que l'effet dérive disparaisse.
- En créant les différences premières des coefficients : $p_2 - p_1$, $p_3 - p_2$, …, $p_j - p_{j-1}$, … .
- En créant des différences secondes $(p_i - p_{i-1}) - (p_{i-1} - p_{i-2})$ qui sont le reflet de la courbure.

D'autres travaux ont repris ce principe pour analyser les tendances de l'incidence et de la mortalité du cancer du sein (Ewertz, 1988 ; Bornefalk, 1995) ou la mortalité due à la maladie de Hodgkin (Heuer, 1994) ou même, en dehors du champ du cancer, la mortalité par maladie cérébro-vasculaire (Maheswaran, 1997).

[6] GLIM : Generalized Linear Interactive Modelling, Royal Statistical Society (Aitkin, 1989 ; Francis, 1994).

(δ) Méthode de Breslow et Clayton et approche bayésienne

Ce procédé utilisé par Breslow et Clayton (Breslow, 1993) puis par Berzuini et Clayton associé à une approche bayésienne (Berzuini, 1994) ainsi que par Bashir et Estève (Bashir, 2001) a été choisi comme méthode de référence pour les analyses et les prévisions réalisées dans ce travail, aussi le principe et les modes d'application de cette méthode feront l'objet du chapitre 3.2. (Analyse).

(ε) Modèles non-linéaires

Lee (1995) introduit des termes non linéaires qui représentent des interactions entre les trois effets. Par exemple :

$$\lambda ij = (a_i + c_k \delta_i)\, p_j$$

a_i est l'effet « âge », p_j est l'effet « période », c_k est l'effet « cohorte » et δ_i est un terme d'interaction de l' « âge » avec la « cohorte ».

Andreasen (1994) étudie l'incidence et la mortalité relatives au cancer du sein à l'aide d'un modèle comprenant des termes dépendant de l'âge, de la période, de la région (différentes régions du Danemark) ainsi que des termes d'interaction entre ces trois facteurs.

(η) Modèles avec contraintes

Dubrow (1993 et 1994), Aragonés (1997) imposent des contraintes aux coefficients du modèle : ils imposent à l'effet « période » une pente nulle, partant du principe que l'effet cohorte est plus important.

Lee et Lin (Lee, 1996) ont proposé une contrainte autorégressive du premier ordre sur la cohorte qui rappelle les contraintes imposées au modèle de Breslow et Clayton (Breslow, 1993) et de Berzuini et Clayton (Berzuini, 1994), modèle qui a été choisi pour l'analyse présente :

$$\ln(\lambda_{ijk}) = \mu + a_i + p_j + c_k$$

Avec $c_k = \varphi c_{k-1} + \delta_k$

(θ) Autres modèles âge-période-cohorte avec composantes qualitatives

Robertson et Boyle (Robertson, 1986 et 1998 a), par exemple, ont testé un modèle âge-période cohorte avec un découpage de chaque case « âge-période(-cohorte) » en deux demi cases, chacune correspondant à une cohorte de classe d'âge différente.
Le modèle se présente de façon classique :

$$\ln(\lambda_{ijk}) = \mu + a_i + p_j + c_k$$

Avec i =1, 2, …, I ; j = 1, 2, …, J mais avec k = j – i + I pour la cohorte la plus âgée et k = j – i + I + 1 pour la cohorte la plus jeune.

Cette méthode permet de rompre la relation linéaire entre les trois facteurs.

(c) Variables quantitatives et variables qualitatives

Il est possible, bien sur de combiner les approches quantitative et qualitative. L'âge peut ainsi être considéré comme une variable qualitative, la cohorte et la période comme des variables quantitatives.

2.3. Autres méthodes

D'autres méthodes intéressantes ont été utilisées. Elles ont recours à des modèles mathématiques et statistiques quelque peu différents comme les modèles stochastiques à compartiments (Cherruault, 1977), les processus de diffusion multidimensionnels ou les modèles à états flous (Manton, 1993 a). Une application basée sur un modèle réunissant cohorte et compartiments a été réalisée par Manton *et al.* (Manton, 1993 b).

3. MATÉRIEL ET MÉTHODES

« Tu trouves tel homme chez lequel la faculté de conjecturer et de deviner est tellement forte et juste, que presque tout ce que, dans son imagination, il croit être, est (réellement) tel qu'il se l'est imaginé, ou l'est (du moins) en partie. Les causes en sont nombreuses, (et cela arrive) par un enchaînement de nombreuses circonstances, antérieures, postérieures et présentes ; mais par la force de cette (faculté de) divination, l'esprit parcourt toutes ces prémisses et en tire les conclusions en si peu de temps qu'on dirait que c'est l'affaire d'un instant. C'est par cette faculté que certains hommes avertissent de choses graves qui doivent arriver. »[7]

[7] Moïse Maïmonide. Le guide des égarés. Traduction de Salomon Munk. Paris, Verdier 1979.

Dans ce chapitre, la description des données disponibles sera envisagée de façon générale. Les informations concernant les données relatives aux localisations qui ont fait l'objet d'une prévision pourront être consultées au chapitre 4. (Exemples). En ce qui concerne la méthode, seule l'analyse statistique sera exposée ici. Le chapitre 4. rendra compte des méthodes utilisées plus particulièrement dans chacun des exemples.

3.1. Matériel

3.1.1. Population

La population est constituée de femmes ou d'hommes dont les âges sont compris, selon les cancers, entre 20 ou 25 ans et 88, 89 ou 94 ans. En deçà de cette tranche d'âge, le nombre de cas incidents et l'incidence sont, pour les cancers étudiés, en général, très faibles. Au delà, la connaissance de l'incidence est moins fiable.
Les effectifs de la population bas-rhinoise ont été obtenus auprès de l'Institut national de la statistique et des études économiques (INSEE). L'INSEE a estimé les populations par année à partir des résultats de trois recensements (1975, 1982 et 1990) et fait une projection pour estimer les populations au delà de 1990 à l'aide du modèle PRUDENT (méthode d'interpolation dépendant de la cohorte de naissance).

Les effectifs des populations étudiées, année par année, 5 ans par 5 ans ou 4 ans par 4 ans ainsi que les effectifs des populations de référence (population mondiale et européenne) sont présentés en annexe 1.

3.1.2. Données d'incidence

Le nombre de cas incidents de cancer, par âge et par année, est fourni par le registre des tumeurs du Bas-Rhin (Schaffer, 1981). Ce dernier est conforme à la définition que donne le Comité national des registres (CNR) : « Un registre de morbidité est une structure épidémiologique qui réalise l'enregistrement continu et exhaustif des cas d'une pathologie donnée dans une région géographique donnée et qui, à partir de cet enregistrement, effectue, seule ou en collaboration avec d'autres équipes,

des études visant à améliorer les connaissances concernant cette pathologie. » (Comité national des registres, 1989).

Le registre des tumeurs du Bas-Rhin a été créé en 1975 et couvre une population de 1,06 millions d'habitants. Il procède à un enregistrement actif des cas (auprès des laboratoires d'anatomie pathologique et de l'ensemble des services hospitaliers publics et privés) ce qui lui permet de disposer de données exhaustives (MacLennan, 1978 ; Powell, 1991 ; Skeet, 1991 ; Ducimetière, 1992 ; Jensen, 1996).

Le registre des cancers est un outil de mesure précis de l'incidence (Comité national des registres, 2000). L'intérêt du registre comme outil de mesure en vue de la prévision est incontestable (Coleman, 1996).

3.2. Analyse

3.2.1. Principe

La méthode utilisée pour la prévision des incidences se base sur l'analyse bayésienne d'un modèle âge-période-cohorte :

$$Y_{ijk} \sim P(m_{ijk}\lambda_{ijk}) \quad \text{et} \quad m_{ijk}\lambda_{ijk} = E[Y_{ijk}]$$
$$\eta_{ijk} = \ln(\lambda_{ijk})$$
$$\eta_{ijk} = a_i + p_j + c_k$$

Le principe de la prévision est, comme il a été dit en introduction, d'estimer un ensemble de valeurs futures (ici des incidences spécifiques) à partir de la connaissance d'un ensemble de valeurs passées (ici, des incidences spécifiques, également).

Le calcul met, ainsi, en jeu :

- Un ensemble de données : ce sont les valeurs passées (connues) des variables à prédire ainsi que des valeurs connues, prises par des covariables, tant dans le passé que dans le futur ;
- Un ensemble de valeurs futures (inconnues) « prises » par les variables à prédire ainsi que des paramètres représentant les effets des covariables.

Dans le cas présent :

- Les valeurs passées des variables à prédire sont les nombres de cas incidents spécifiques (selon la classe d'âge et la période) passés, extraits de la base du registre ;
- Les covariables sont les effectifs de population correspondant aux sous-populations définies par la tranche d'âge et la période ainsi que les niveaux des trois paramètres « âge », « période » et « cohorte » pour les périodes passées et futures ;
- Les valeurs futures sont les nombres de cas incidents spécifiques (selon la classe d'âge et la période) futurs à prédire ;
- Les paramètres inconnus sont les effets des covariables « âge », « période » et « cohorte ». Il font lien entre ces covariables et le nombre de cas incidents.

Soit y la valeur sur laquelle doit porter la prédiction. Les données disponibles sont représentées par y_1, y_2, …y_n mesurées respectivement aux temps t_1, t_2, …t_n (passé). Les variables aléatoires correspondantes sont Y_1, Y_2, … Y_n. Le paramètre de la distribution de probabilité des Y_i est Θ.
La prédiction doit être établie pour les temps (futurs) t_{n+1}, t_{n+2}, …t_{n+p}. Les valeurs et les variables aléatoires correspondantes sont, respectivement, y_{n+1}, y_{n+2}, …y_{n+p} et Y_{n+1}, Y_{n+2}, …Y_{n+p}.
Le modèle bayésien exprime la distribution de probabilité prédictive *a posteriori* des Y_{n+1}, Y_{n+2}, …Y_{n+p} comme suit (Mouchart, 1998) :

$$g\left(y_{n+1}, y_{n+2},..., y_{n+p} \middle| y_1, y_2,..., y_n\right) = \int f\left(\theta \middle| y_1, y_2,..., y_n\right) g\left(y_{n+1}, y_{n+2},..., y_{n+p} \middle| y_1, y_2,..., y_n, \theta\right) d\theta$$

$f(\theta|y_1, y_2, \ldots y_n)$ est la distribution de probabilité a posteriori de Θ.

Si, de plus, il existe des covariables Z_1, Z_2, …Z_n, Z_{n+1}, Z_{n+2}, …, Z_{n+p} dont les réalisations respectives z_1, z_2, …z_n, z_{n+1}, z_{n+2}, …, z_{n+p}, sont connues, la formule précédente devient :

$$g\left(y_{n+1},..., y_{n+p} \middle| y_1,..., y_n, z_1,...z_{n+p}\right) = \int f\left(\theta \middle| y_1,..., y_n, z_1,...z_{n+p}\right) g\left(y_{n+1},..., y_{n+p} \middle| y_1,..., y_n, z_1,...z_{n+p}, \theta\right) d\theta$$

Cette écriture peut être résumée par l'expression suivante :

$$g\left(y_F \middle| y_P, z\right) = \int f\left(\theta \middle| y_P, z\right) g\left(y_F \middle| y_P, z, \theta\right) d\theta$$

Avec $y_P = y_1, y_2, \ldots y_n$ les valeurs passées de y, $y_F = y_{n+1}, y_{n+2}, \ldots y_{n+p}$ les valeurs futures de y, $z = z_1, z_2, \ldots z_n, z_{n+1}, z_{n+2}, \ldots, z_{n+p}$ les covariables.

Les bases statistiques de ces calculs sont rassemblées dans le Chapitre 7.3. (Rappels de statistique bayésienne)

En résumé, une probabilité jointe est donnée pour l'ensemble des variables et il faut trouver la probabilité marginale de chacune des variables, ce qui nécessite une intégration compliquée ; il est alors possible d'avoir recours à l'échantillonnage de Gibbs qui produit des échantillons pour chaque variable, tirés de la densité marginale de cette variable sans calculer l'intégrale marginale elle-même mais en calculant les densités de probabilité conditionnelle d'une variable par rapport aux autres ; ces densités conditionnelles sont en effet plus facile à calculer.

Le modèle impose, de plus, des contraintes entre les paramètres successifs (les effets) des trois covariables « âge », « période » et « cohorte » : ce sont des relations autorégressives déduites des réflexions et des modèles de Breslow et Clayton (Breslow, 1993) et de Berzuini et Clayton (Berzuini, 1994 ; Bashir, 2001).

Ainsi, pour l'âge, la relation de dépendance générales entre les effets successifs est :

$$\alpha_i \sim \frac{4\alpha_{i-1} + 4\alpha_{i+1} - \alpha_{i-2} - \alpha_{i+2}}{6}$$

Pour la période :

$$\beta_j \sim 2\beta_{j-1} - \beta_{j-2} ;$$

Pour la cohorte :

$$\gamma_k \sim 2\gamma_{k-1} - \gamma_{k-2}.$$

Ces relations incluent des termes généraux et ne peuvent être appliquées comme telles aux paramètres extrêmes ; pour ces derniers, les relations particulières sont détaillées au Chapitre 3.2.2.1.2. (Principe et hypothèses de la modélisation à la base du programme).

Les relations autorégressives ne sont pas imposées au hasard : elles sont déduites d'une appréhension intuitive des relations de dépendance entre les niveaux des effets des trois paramètres de base (Breslow, 1993 ; Berzuini, 1994). Il est relativement aisé de concevoir que les effets de deux tranches d'âge, de deux périodes ou de deux cohortes de naissance successives soient relativement proches si l'on exclue *a priori* la possibilité d'accidents.
Il apparaît que la relation la plus naturelle entre termes successifs est que le risque au temps t est la moyenne géométrique des risques au temps t-1 et t+1. Ce qui s'écrit, du point de vue mathématique :

$$\lambda_t^2 = \lambda_{t-1}\, \lambda_{t+1}$$

Cette expression, transposée aux risques correspondant à t-1, t-2 :

$$\lambda_{t-1}{}^2 = \lambda_{t-2}\,\lambda_t$$

Et, en prenant le logarithme de chacun des termes de l'égalité, cette relation peut être considérée comme une extrapolation à t, des valeurs prises par le risque aux temps t-1 et t-2 :

$$\ln\lambda_t = 2\ln\lambda_{t-1} - \ln\lambda_{t-2}$$

Si la relation est considérée d'un point de vue probabiliste, elle devient :

$$\ln\lambda_t = 2\ln\lambda_{t-1} - \ln\lambda_{t-2} + \varepsilon_t$$

Avec $\varepsilon_t \sim N(0,\sigma^2)$

Cette relation, applicable comme telle à la période (équivalente du temps), peut être transposée à la classe d'âge et à la cohorte de naissance :

$$\ln\lambda_n = 2\ln\lambda_{n-1} - \ln\lambda_{n-2} + \varepsilon_n$$

Avec $\varepsilon_n \sim N(0,\sigma^2)$

Ainsi, pour la période, si l'on désigne $\ln\lambda_n$ par β_j, la relation précédente devient :

$$\beta_j = 2\beta_{j-1} - \beta_{j-2} + \varepsilon_t$$

De même, pour la cohorte, si l'on désigne $\ln\lambda_n$ par γ_k, il vient :

$$\gamma_k = 2\gamma_{k-1} - \gamma_{k-2} + \varepsilon_t$$

Ces résultats (qui sont équivalents à une égalité entre différences premières des effets) sont conformes aux relations présentées plus haut pour la période et la cohorte.

En ce qui concerne l'âge, à une égalité entre différence premières des effets, il a été préféré une égalité entre différences secondes ce qui peut se traduire par :

$$(\alpha_{i+1} - \alpha_i) - (\alpha_i - \alpha_{i-1}) = \frac{(\alpha_i - \alpha_{i-1}) - (\alpha_{i-1} - \alpha_{i-2}) + (\alpha_{i+2} - \alpha_{i+1}) - (\alpha_{i+1} - \alpha_i)}{2}$$

En développant les deux termes de cette égalité, on obtient :

$$\alpha_i = \frac{4\alpha_{i-1} + 4\alpha_{i+1} - \alpha_{i-2} - \alpha_{i+2}}{6}$$

L'utilisation de relations autorégressives a un autre intérêt. Un modèle âge-période-cohorte se fonde sur un tableau disposé selon l'âge et la période (Tableau 2.1.). Les cohortes extrêmes (Tableau 2.2.) sont représentées par un nombre faible de cases et, par conséquent, disposent d'un nombre réduit de données. L'information les concernant étant pauvre, l'estimation de l'effet de ces cohortes risque de souffrir d'une grande variabilité. L'instauration de relations autorégressives a pour effet de stabiliser

ces cohortes et de permettre d'estimer leur effet de façon plus précise (Breslow, 1993 ; Berzuini, 1994).

3.2.2. Calcul (programme)

3.2.2.1. Principe du calcul

Le principe du calcul est fondé sur l'échantillonnage (ou algorithme) de Gibbs qui est un cas particulier de méthode de Monte Carlo par chaînes de Markov. Le mode opératoire de cette technique est exposé de façon détaillée au chapitre 7.4. L'échantillonnage de Gibbs est utilisé en raison des difficultés à calculer correctement certaines intégrales nécessaires à l'inférence bayésienne (voir 7.4.1. Problématique).

Les travaux faisant figurer l'échantillonnage de Gibbs au nombre de leurs outils sont pléthore. En outre, en dehors de l'épidémiologie, les domaines tirant partie de cette méthode sont nombreux et diversifiés : statistique bayésienne, analyse d'images, géographie, intelligence artificielle, etc. La littérature évoquant cette méthode en la citant, l'expliquant, la transformant, l'adaptant est, par conséquent, abondante. Mais Gibbs, qui est-il ? Il est possible de passer en revue des centaines d'articles et autant de sites Internet citant tantôt la méthode d'échantillonnage de Gibbs, tantôt l'existence d'un certain Josiah Willard Gibbs, mathématicien et physicien sans trouver trace d'un lien quelconque entre l'algorithme et la vie d'un homme qui serait à l'origine d'un outil si précieux !
Après de multiples recherches, la relation entre le cerveau et l'outil a été enfin reconstituée et matérialisée à la lumière de la lecture d'un article d'essence multidisciplinaire (philosophe, statisticiens et informaticien) paru il y a quelques années (Glymour, 1996). Le résumé de la vie et de l'œuvre de J Willard Gibbs fait l'objet de l'annexe 3.

3.2.2.1. Mise en œuvre

3.2.2.1.1. Logiciel

Les calculs sont effectués sur le logiciel BUGS (Bayesian inference Using Gibbs Sampling). Développé par David Spiegegelhalter, Andrew Thomas, Nicky Best et Wally Gilks au sein de l'unité

MRC de biostatistique de Cambridge en 1995, ce programme se base sur un échantillonnage de Gibbs (Spiegelhalter, 1996 a ; Spiegelhalter, 1996 b ; Spiegelhalter, 1996 c).
Toutes les grandeurs (paramètres, données manquantes et données connues) sont considérées comme des variables aléatoires. Le modèle prend en compte la distribution jointe de toutes ces variables, élabore une distribution *a posteriori* des paramètres et des données cachées et calcule les densités marginales des paramètres recherchés grâce à l'algorithme d'échantillonnage de Gibbs.

Les coordonnées de l'unité MRC de biostatistique sont les suivantes :
MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR.
Tel 44-1223-330300
Fax 44-1223-330388.
e-mail : bugs@mrc-bsu.cam.ac.uk
http://www.mrc-bsu.cam.ac.uk/bugs

Il est possible de télécharger les programmes BUGS (pour DOS et Windows) sur le site de l'unité.
Les manuels relatifs au programme sont également téléchargeables ; il s'agit d'un manuel didactique et de deux manuels d'exemples :
Spiegelhalter D, Thomas A, Best N, Gilks W. BUGS 0.5 Bayesian inference using Gibbs sampling Manual (version ii). Août 1996.
Spiegelhalter D, Thomas A, Best N, Gilks W. BUGS 0.5 Examples Volume 1 (version i). Août 1996.
Spiegelhalter D, Thomas A, Best N, Gilks W. BUGS 0.5 Examples Volume 2 (version ii). Août 1996.

3.2.2.1.2. Principe et hypothèses de la modélisation à la base du programme

Le programme utilisé ici est inspiré de l'exemple cité dans le volume exemples numéro 2 : « Ice: non-parametric smoothing in an age-cohort model » (Spiegelhalter, 1996 d).

Notations : n est le nombre de sujets, I est le nombre de classes d'âge, J est le nombre de périodes, K est le nombre de cohortes, E() désigne l'espérance, τ désigne la précision c'est-à-dire l'inverse de la variance ($\tau = \frac{1}{\sigma^2}$).

La variable « nombre de cas » suit une loi de Poisson :
cas ~ $P(\mu_n)$

Le principe et les hypothèses servant de base à la modélisation sont les suivants :

Écriture du modèle :

$\ln\mu_n = \ln popn_n + \alpha_{age} + \beta_{period} + \gamma_{cohort}$

Distribution conditionnelle des paramètres relatifs à l'âge :

$\alpha_i|\alpha_{i'}, i' \neq i \sim N(E(\alpha_i),\tau_a)$

Corrélation des paramètres relatifs à l'âge :

$E(\alpha_1) = 2\alpha_2 - \alpha_3$

$E(\alpha_2) = (2\alpha_1 + 4\alpha_3 - \alpha_4) / 5$

$E(\alpha_i) = (4\alpha_{i-1} + 4\alpha_{i+1} - \alpha_{i-2} - \alpha_{i+2}) / 6$, pour $3 \leq i \leq I-2$

$E(\alpha_{I-1}) = (2\alpha_I + 4\alpha_{I-2} - \alpha_{I-3}) / 5$

$E(\alpha_I) = 2\alpha_{I-1} - \alpha_{I-2}$

Distribution conditionnelle des paramètres relatifs à la période :

$\beta_j|\beta_{j'}, j' \neq j \sim N(E(\beta_j),\tau_p)$

Corrélation des paramètres relatifs à la période :

$E(\beta_1) = 0$

$E(\beta_2) = 0$

$E(\beta_j) = 2\beta_{j-1} - \beta_{j-2}$, pour $3 \leq i \leq J$

Distribution conditionnelle des paramètres relatifs à la cohorte :

$\gamma_k|\gamma_{k'}, k' \neq k \sim N(E(\gamma_k),\tau_c)$

Corrélation des paramètres relatifs à l'âge :

$E(\gamma_1) = 0$

$E(\gamma_2) = 0$

$E(\gamma_k) = 2\gamma_{k-1} - \gamma_{k-2}$, pour $3 \leq i \leq K$

Les densités de distribution des précisions, τ_a,τ_p et τ_c sont des lois gamma.

L'écriture du modèle dans BUGS est contenue dans un fichier particulier et est détaillée dans l'Annexe 2.

3.2.2.1.3. Modèles graphiques directionnels

Les modèles graphiques (Figure 3.1.) aident à représenter les relations de dépendance entre les constantes, les variables et les paramètres du modèle bayésien (Clayton, 1991 ; Jordan, 1997 ; Albert, 1998). Ils sont construits en premier et servent de canevas à l'écriture du modèle dans BUGS.

La figure 3.1. représente graphiquement le programme utilisé pour effectuer les prévisions d'incidence.

Les constantes, les variables et les paramètres sont représentés par des nœuds et sont reliés par des flèches. La flèche est orientée de la grandeur qui influence (nœud parent) vers la grandeur qui est influencée (nœud enfant).

Les nœuds sont de trois types :
Les rectangles (à bordure simple ou double) représentent les constantes (nœuds de base qui n'ont pas de parents) ;
Les cercles représentent des variables disposant d'une loi de distribution. Ce sont des nœuds stochastiques et peuvent correspondre à des grandeurs d'intérêt ou des paramètres à caractériser ;
Le troisième type de nœud regroupe les nœuds déterminés qui correspondent à des fonctions déterministes d'autres nœuds.

Les flèches sont de deux types :
Les flèches en trait plein représentent des relations aléatoires ;
Les flèches en tirets représentent des relations déterministes.

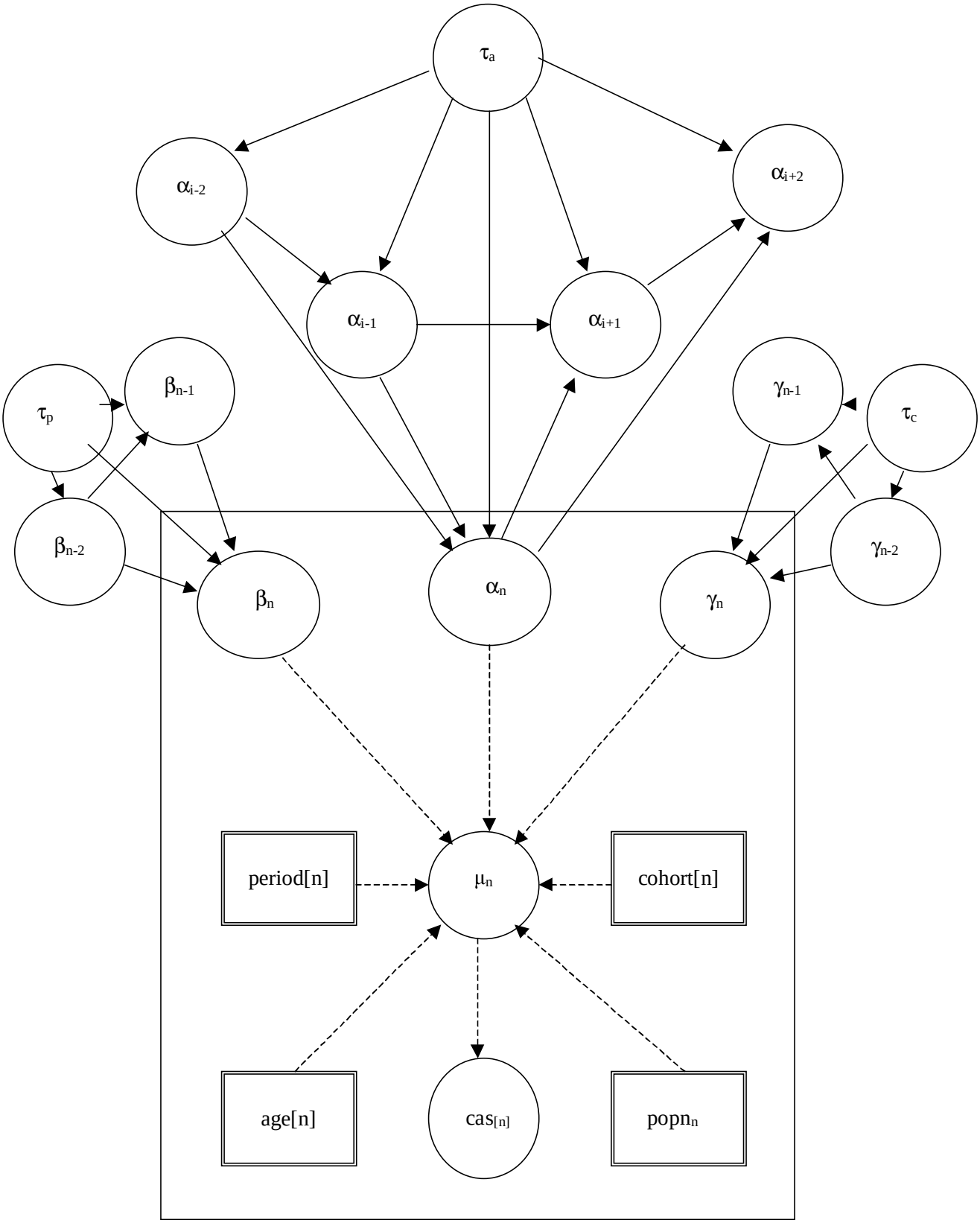
3.2.2.1.4. Fichier de valeurs initiales

La construction des échantillons conformément à la méthode de Gibbs nécessite de fixer un ensemble de valeurs servant de base à la première itération. Ces valeurs sont les paramètres de précision et les valeurs initiales des paramètres des effets âge, période et cohorte. Ces valeurs sont rassemblées dans un fichier.

L'écriture du fichier de valeurs initiales est détaillée dans l'Annexe 2.

Figure 3.1. Modèle graphique directionnel

(Légende voir paragraphe 2.2.2.1.3. Modèles graphiques directionnels)



3.2.2.1.5. Fichiers de données

Un fichier de données renferme les données de population avec les classes d'âge, les périodes et les cohortes.
Le deuxième fichier contient les nombres de cas incidents.

Un exemple de fichiers de données est présenté en Annexe 2.

3.2.2.1.6. Fichier de commande

Le fichier de commande donne les instructions destinées au logiciel : la compilation du programme, un premier cycle d'itérations dont le nombre peut être choisi puis un deuxième cycle d'itérations dont le nombre peut également être choisi et qui servira à calculer les paramètres et statistiques. Le fait d'effectuer un premier cycle d'itérations – ce cycle est appelé « tour de chauffe » – et de l'éliminer de la base de calcul des statistiques a pour effet de supprimer ou tout au moins de réduire la sensibilité aux conditions initiales.

L'écriture du fichier de commande est détaillée dans l'Annexe 2.

3.2.2.1.7. Fichier de résultats

Il contient le rappel du programme dans BUGS et la sortie des résultats du calcul.

L'écriture du fichier de commande est détaillée dans l'Annexe 2.

4. EXEMPLES

« Seize ans avant, en l'an 1888, quand Bloom avait l'âge actuel de Stephen, celui-ci avait 6 ans. Seize ans après, en 1920, quand Stephen aurait l'âge actuel de Bloom, celui-ci aurait 54 ans. En 1936, quand Bloom aurait 70 ans et Stephen 54, leur âge, initialement dans le rapport de 10 à 0, serait comme 17 1/2 à 13 1/2, la proportion augmentant et la différence diminuant selon que de futures années arbitraires seraient ajoutées, car si la proportion qui existait en 1883 avait continué immuablement, en concevant que ce fût possible, jusqu'à l'actuel 1904 quand Stephen avait 22 ans, Bloom aurait 374 ans, et en 1920 quand Stephen aurait 38 ans, comme Bloom avait actuellement, Bloom aurait 646 ans ; d'autre part en 1952, quand Stephen aurait atteint l'âge maximum postdiluvien de 70 ans, Bloom, ayant vécu 1190 ans étant né en l'année 714, aurait dépassé de 221 ans l'âge maximum antédiluvien, celui de Mathusalem, 969 ans, tandis que, si Stephen continuait à vivre jusqu'à ce qu'il eût atteint cet âge en l'année 3072 après J.-C., Bloom aurait été obligé d'avoir vécu 83300 ans, ayant été obligé d'être né en l'année 81 396 avant J.-C.
Quels événements pouvaient anéantir ces calculs ? »[8]

[8] James Joyce. Ulysse. Traduction d'Auguste Morel revue par Valéry Larbaud, Stuart Gilbert et l'auteur. Paris, Gallimard, 1957.

La prévision a été appliquée à l'incidence de quatre cancers importants en termes de santé publique (importance du nombre de cas incidents, mortalité liée au cancer, notion de dépistage). Ce sont les cancer du sein invasif, du col de l'utérus *in situ* et invasif, colorectal et pulmonaire. Les deux derniers ont été étudiés pour les deux sexes.

4.1. Cancer du sein

4.1.1. Introduction

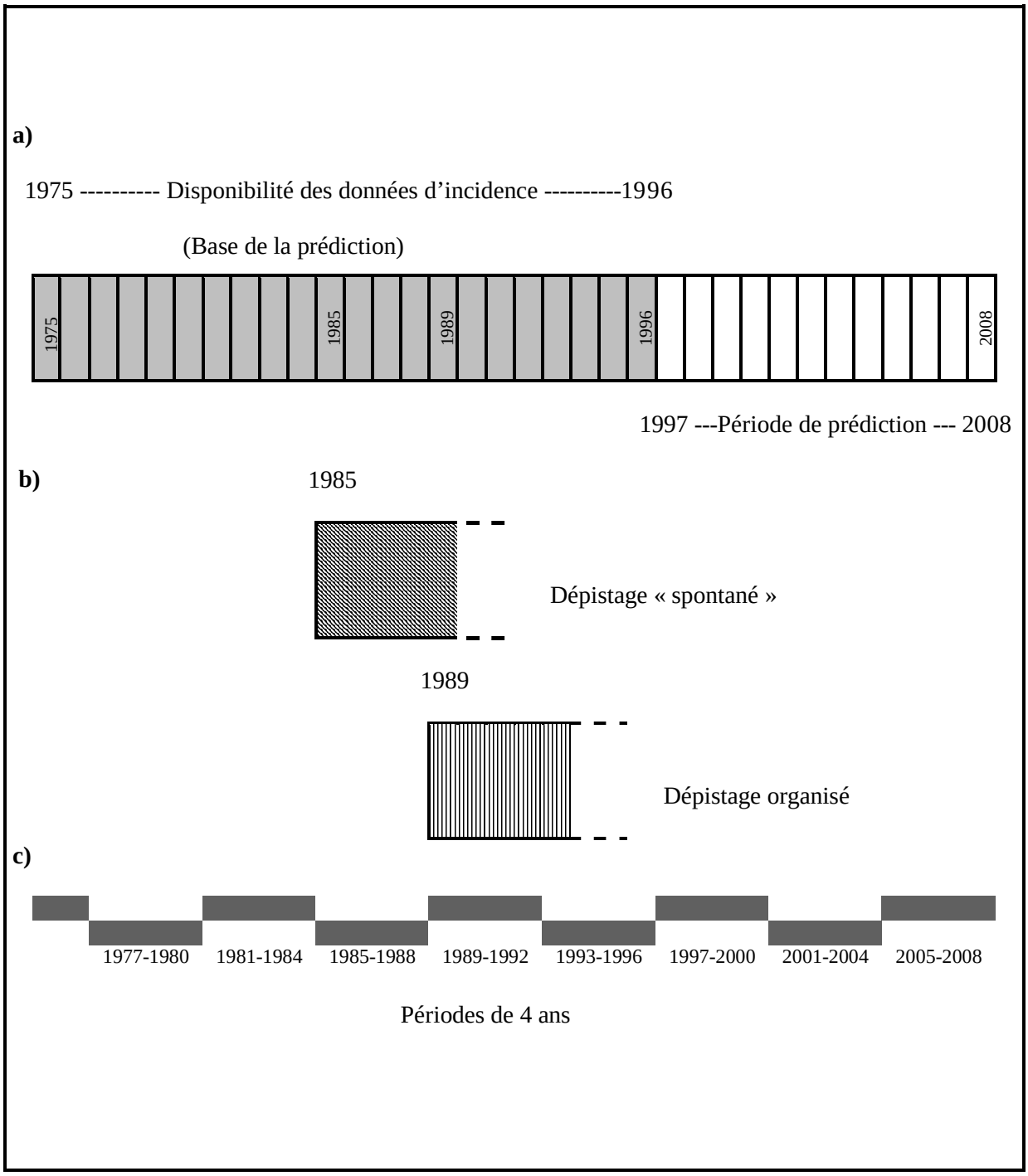
Le cancer du sein est le plus fréquent des cancers féminins. Il est la cause d'un grand nombre de décès (Coleman, 1993 ; Parkin, 1997). En France, en 1995, les nouveaux cas représentaient 32 % de l'ensemble des tumeurs malignes de la femme et cette maladie était à l'origine de 19 % des décès féminins par cancer (Réseau Francim, 1998). Dans le Bas-Rhin (France), l'incidence et la mortalité sont légèrement supérieures à celles de l'ensemble de la France : leurs valeurs, standardisées selon la population européenne, étaient, en 1995, respectivement de 106,6 et 29,3 pour 10^5 dans le Bas-Rhin et de 103,0 et 28,2 pour 10^5 en France (Réseau Francim, 1998).

Entre 1975 et 1990, dans le Bas-Rhin, l'incidence du cancer du sein, tous âges et tous types de tumeurs confondus, a augmenté de 32 % (1,9 % par an) (Zerr-Fuhrmann, 1992). En France, l'incidence a augmenté de 60 % entre 1975 et 1995 (2,4 % par an) ; cette évolution est retrouvée un peu partout dans le monde (Coleman, 1993 ; Ewertz, 1988 ; Andreasen, 1994 ; Holford, 1991 ; Kesley, 1993 ; Garfinkel, 1994, a ; Sondik, 1995 ; Tominaga, 1995 ; Wakai, 1995).

Les projections, réalisées en différentes régions du globe, sur la base d'un modèle A-P-C, prévoient toutes une augmentation de l'incidence du cancer du sein (Dyba, 1997 ; Engeland, 1993 ; Vioque, 1993 ; Hristova, 1997).

En 1989, une campagne de dépistage a été mise en place dans le Bas-Rhin (ADEMAS) (Figure 4.1.). Depuis cette date, une mammographie est proposée aux femmes âgées de 50 à 64 ans, tous les deux ans. En 1985, cependant, un dépistage spontané, dit « dépistage sauvage », avait déjà eu lieu. Or, l'existence d'un dépistage a une influence sur l'incidence : celle-ci croît brusquement au moment du démarrage du dépistage (en raison de la révélation anticipée de nouveaux cas) puis, selon la participation, l'incidence revient à la tendance initiale, ou reste élevée (Feuer, 1992 ; Wun, 1995). Cette perturbation peut fausser la projection en augmentant artificiellement la base de la prédiction.

Figure 4.1. Disponibilité des données. a) Période d'étude : la période de disponibilité des données est ombrée, la période de projection apparaît en blanc ; b) Dépistage « spontané » (1985) et campagne de dépistage organisé (1989 et suivantes) ; c) Périodes groupées par tranches de 4 ans.



L'objectif de cette étude est de décrire, pour le Bas-Rhin, la tendance de l'incidence du cancer du sein chez la femme entre 1975 et 1994, de repérer les perturbations éventuelles liées aux dépistages puis d'établir une projection de cette tendance jusqu'en 2008.

4.1.2. Matériels et méthodes

4.1.2.1. Données

La période de disponibilité des données d'incidence était de 22 ans : de 1975 à 1996.
La population étudiée était constituée des femmes de 25 à 89 ans.

4.1.2.2. Analyse

L'**analyse descriptive** s'est attachée à examiner, séparément, l'évolution de l'incidence des tumeurs in situ et des tumeurs invasives en fonction du temps (périodes de 1975 à 1996). Les incidences, standardisées selon la population européenne, ont été décrites pour la population de 25 à 89 ans et pour les trois tranches d'âge 25-49 ans, 50-64 ans (population dépistée) et 65-89 ans. Les taux d'incidence spécifiques, par tranches d'âge de 4 ans (voir ci-dessous), ont été également examinés. Enfin, les taux d'incidence ont été calculés pour les différentes cohortes de naissance. L'observation des courbes d'incidence devait permettre de repérer les perturbations liées aux dépistages, pouvant fausser les calculs de prédiction, et de sélectionner le type de tumeur (*in situ* et/ou invasive) pouvant faire l'objet d'une projection.

La **prévision** s'est basée sur un modèle A-P-C appliqué à un tableau âge-période-cohorte (nombre de cas incidents et nombre de personnes-années) découpé selon des tranches d'âge de quatre ans et des périodes de quatre années. Ce découpage inhabituel a été choisi afin d'optimiser l'information apportée par les données, compte tenu des périodes disponibles et des périodes de dépistage (Figure 4.1). Il a aussi permis de réaliser une analyse de sensibilité et d'apprécier la variation de la prédiction lorsqu'on modifiait artificiellement l'incidence au moment des dépistages.
Ainsi, l'analyse a pris en compte les classes d'âge de 25-28 ans, 29-32, …, 85-88 ans ainsi que les périodes 1975-1976, 1977-1980, 1981-1984, …, 2005-2008. La base de prédiction est constituée par

les périodes comprises entre 1975 et 1996 (Figure 4.1.). La première période prise en compte est de deux ans (1975 et 1976) pour que le découpage fasse apparaître les périodes 1985-1988 (dépistage « sauvage ») et 1989-1992 (début du dépistage organisé).

Afin de vérifier que la prédiction n'avait pas été faussée par les dépistages (augmentation brusque de l'incidence au début du dépistage), il a été décidé de réaliser plusieurs **analyses de sensibilité**.

1°) Prédiction à l'horizon 2004, basée sur les données enregistrées de 1975 à 1988 (données non influencées par le dépistage) ;

2°) Prédiction à l'horizon 2004, basée sur les données enregistrées de 1975 à 1992 (données incluant le début du dépistage) ;

3°) Prédiction à l'horizon 2008, basée sur les données enregistrées de 1975 à 1996 (donc comme le calcul initial) après avoir remplacé, le cas échéant (i.e. quand elles paraissaient anormalement élevées), les incidences enregistrées au moment du dépistage par une interpolation (linéaire) des valeurs d'incidence antérieures et postérieures à celui-ci. Cette modification devait concerner les cohortes touchées par le dépistage ;

4°) Deux autres analyses ont été réalisées en incluant une variable « dépistage » dans le modèle. Celle-ci est la proportion de femmes des classes d'âge concernées (49 à 52 ans, 53 à 56 ans, 57 à 60 ans, 61 à 64 ans et 65 à 68 ans) ayant bénéficié d'une *première* mammographie (rapport du nombre de femmes dépistées – il s'agit d'une première vague – au nombre total de femmes de la classe d'âge considérée) (Rostgaard, 2000). Les valeurs des variables « dépistage » durant les périodes 1989-1992 et 1993-1996 sont calculées à partir des données (Tableau 4.1.) et sont donc les mêmes pour les deux analyses. Par contre, pour les périodes de prévisions (1997-2000, 2001-2004 et 2005-2008), le nombre de femmes dépistées n'est pas connu et a du être estimé. Deux scénarios ont alors été envisagés.
La première analyse a réalisé une extrapolation. Ainsi, pour une cohorte donnée (Tableau 4.2.), un premier groupe de femme est dépisté (cases notées 1) puis une deuxième groupe (cases notées 2), etc. Les tableaux 4.1. et 4.3. montrent que, pour les cases grisées notées 1 (premier « contact » d'une cohorte avec le dépistage), le nombre de femmes dépistées (femmes de 49 à 64 ans) est compris entre 8 500 et 10 300. Aussi, le nombre de femmes dépistées dans les cases 1 non grisées (périodes 1997-2000, 2001-2004 et 2005-2008) est choisi comme moyenne des nombres de dépistages dans les cases 1 grisées (Tableau 4.3.). Pour l'ensemble des cohortes (Tableau 4.1 et 4.3.), les cases 2 grises affichent un nombre de dépistages compris entre 2 000 et 3 000. Aussi, le nombre de femmes dépistées dans les cases 2 non grisées (périodes 1997-2000, 2001-2004 et 2005-2008) est choisi comme moyenne des nombres de dépistages dans les cases 2 grisées (Tableau 4.3.).

Tableau 4.1. Cancer du sein dans le Bas-Rhin. Nombre de femmes ayant bénéficié d'une première vague de dépistage. Le dépistage a commencé en 1989 et concerne par principe les femmes de 50 à 64 ans. Ce nombre est connu pour les périodes 1989-1992 et 1993-1996 et doit être extrapolé pour les périodes suivantes.

Ag\An	1975-1976	1977-1980	1981-1984	1985-1988	1989-1992	1993-1996	1997-2000	2001-2004	2005-2008
25-28	0	0	0	0	0	0	0	0	0
29-32	0	0	0	0	0	0	0	0	0
33-36	0	0	0	0	0	0	0	0	0
37-40	0	0	0	0	0	0	0	0	0
41-44	0	0	0	0	0	0	0	0	0
45-48	0	0	0	0	0	0	0	0	0
49-52	0	0	0	0	10336	8653	?	?	?
53-56	0	0	0	0	10787	3041	?	?	?
57-60	0	0	0	0	10412	2254	?	?	?
61-64	0	0	0	0	9475	2176	?	?	?
65-68	0	0	0	0	1345	335	?	?	?
69-72	0	0	0	0	0	0	0	0	0
73-76	0	0	0	0	0	0	0	0	0
77-80	0	0	0	0	0	0	0	0	0
81-84	0	0	0	0	0	0	0	0	0
85-88	0	0	0	0	0	0	0	0	0

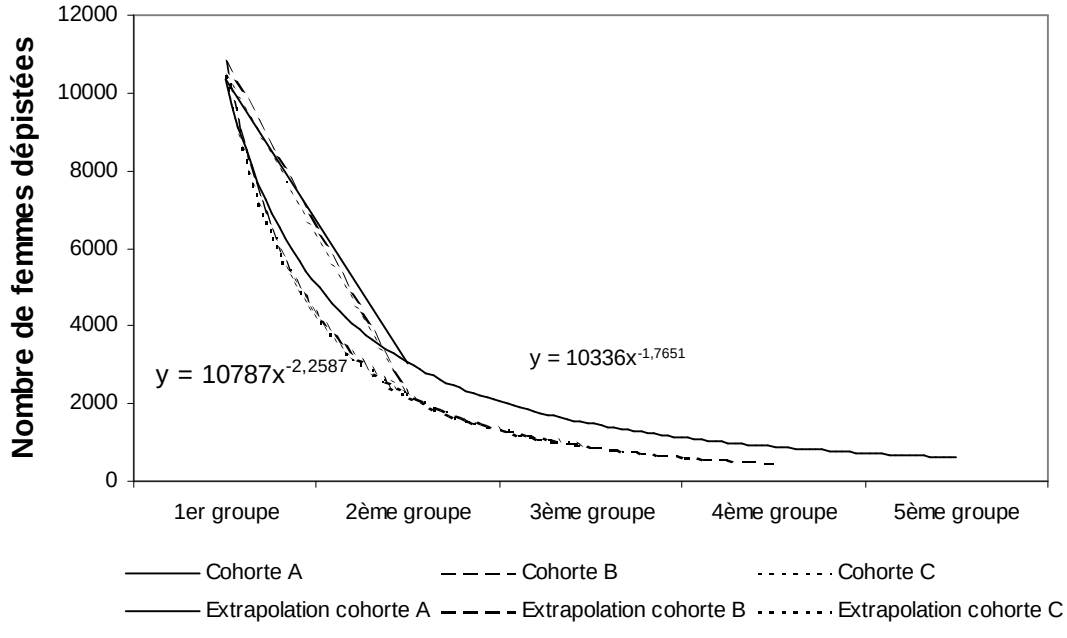
Tableau 4.2. Premières vagues de dépistage du cancer du sein dans le Bas-Rhin. Pour une cohorte donnée, un premier groupe de femme est dépisté (cases notées 1) puis une deuxième groupe (cases notées 2), etc. Les cases grisées correspondent aux données. Dans les cases blanches numérotées, les nombres correspondants ne sont pas connus et doivent être extrapolés.

Ag\An	1975-1976	1977-1980	1981-1984	1985-1988	1989-1992	1993-1996	1997-2000	2001-2004	2005-2008
25-28									
29-32									
33-36									
37-40									
41-44									
45-48									
49-52					1	1	1	1	1
53-56					1	2	2	2	2
57-60					1	2	3	3	3
61-64					1	2	3	4	4
65-68					…	…	3	4	5
69-72									
73-76									
77-80									
81-84									
85-88									

Tableau 4.3. Premières vagues de dépistage du cancer du sein dans le Bas-Rhin. Dans les cases blanches numérotées, les nombres de femmes dépistées ont été extrapolés selon la méthode explicitée dans le texte (voir § 4.1.2.2.4.). Dans les autres cases , il s'agit de valeurs extraites des données.

Ag\An	1975-1976	1977-1980	1981-1984	1985-1988	1989-1992	1993-1996	1997-2000	2001-2004	2005-2008
25-28	0	0	0	0	0	0	0	0	0
29-32	0	0	0	0	0	0	0	0	0
33-36	0	0	0	0	0	0	0	0	0
37-40	0	0	0	0	0	0	0	0	0
41-44	0	0	0	0	0	0	0	0	0
45-48	0	0	0	0	0	0	0	0	0
49-52	0	0	0	0	10336	8653	9495	9495	9495
53-56	0	0	0	0	10787	3041	2490	2490	2490
57-60	0	0	0	0	10412	2254	1100	1100	1100
61-64	0	0	0	0	9475	2176	1100	600	600
65-68	0	0	0	0	1345	335	1100	600	400
69-72	0	0	0	0	0	0	0	0	0
73-76	0	0	0	0	0	0	0	0	0
77-80	0	0	0	0	0	0	0	0	0
81-84	0	0	0	0	0	0	0	0	0
85-88	0	0	0	0	0	0	0	0	0

Figure 4.2. Extrapolation par fonction puissance du nombre de femmes dépistées en fonction du groupe de mise en présence de la première phase du dépistage pour chacune des cohortes concernées.



Une difficulté se présente pour les cases numérotées 3, 4 et 5 (Tableau 4.2.) car les données ne sont pas disponibles pour ces 3ème, 4ème et 5ème mises en présence de la première vague du dépistage. Une extrapolation a été faite à partir des deux premières mises en présence en faisant l'hypothèse que la variation du nombre de dépistages se faisait comme une fonction puissance du numéro du groupe de femmes dépistées (Figure 4.1. et tableau 4.3.). Les valeurs de la variable dépistage (proportion de femmes dépistées) figurent dans le tableau 4.4.
La deuxième analyse est partie d'une hypothèse extrême : l'interruption du dépistage en 1997 (Tableaux 4.5. et 4.6.). Cette analyse devrait donner une limite inférieure à l'effet du dépistage au cours des périodes de prédiction.

5°) Enfin, pour éliminer des incertitudes éventuelles dues à des erreurs d'enregistrement des cas incidents au moment du début du fonctionnement du registre, le calcul a été réitéré en enlevant de la base de prédiction les années 1975 et 1976.

Tableau 4.4. Cancer du sein dans le Bas-Rhin. Valeurs de la variable de dépistage (proportion de femmes dépistées). Dans les cases encadrées, il s'agit de valeurs déduites de l'extrapolation selon la méthode explicitée dans le texte (voir § 4.1.2.2. 4.). Dans les autres cases , il s'agit de valeurs extraites des données.

Ag\An	1975-1976	1977-1980	1981-1984	1985-1988	1989-1992	1993-1996	1997-2000	2001-2004	2005-2008
25-28	0	0	0	0	0	0	0	0	0
29-32	0	0	0	0	0	0	0	0	0
33-36	0	0	0	0	0	0	0	0	0
37-40	0	0	0	0	0	0	0	0	0
41-44	0	0	0	0	0	0	0	0	0
45-48	0	0	0	0	0	0	0	0	0
49-52	0	0	0	0	0,1331033	0,1104671	0,0940444	0,0828997	0,0793538
53-56	0	0	0	0	0,1372846	0,0395727	0,0324291	0,0249865	0,0220167
57-60	0	0	0	0	0,1291811	0,0290528	0,0146051	0,0145169	0,0111814
61-64	0	0	0	0	0,1210012	0,0276093	0,0146055	0,0081404	0,0080811
65-68	0	0	0	0	0,0176468	0,0044237	0,0144836	0,0081851	0,0055594
69-72	0	0	0	0	0	0	0	0	0
73-76	0	0	0	0	0	0	0	0	0
77-80	0	0	0	0	0	0	0	0	0
81-84	0	0	0	0	0	0	0	0	0
85-88	0	0	0	0	0	0	0	0	0

Tableau 4.5. Premières vagues de dépistage du cancer du sein dans le Bas-Rhin. Dans les cases blanches numérotées, les nombres de femmes dépistées ont été égalés à 0 (scénario avec interruption du dépistage). Dans les autres cases , il s'agit de valeurs extraites des données.

Ag\An	1975-1976	1977-1980	1981-1984	1985-1988	1989-1992	1993-1996	1997-2000	2001-2004	2005-2008
25-28	0	0	0	0	0	0	0	0	0
29-32	0	0	0	0	0	0	0	0	0
33-36	0	0	0	0	0	0	0	0	0
37-40	0	0	0	0	0	0	0	0	0
41-44	0	0	0	0	0	0	0	0	0
45-48	0	0	0	0	0	0	0	0	0
49-52	0	0	0	0	10336	8653	0	0	0
53-56	0	0	0	0	10787	3041	0	0	0
57-60	0	0	0	0	10412	2254	0	0	0
61-64	0	0	0	0	9475	2176	0	0	0
65-68	0	0	0	0	1345	335	0	0	0
69-72	0	0	0	0	0	0	0	0	0
73-76	0	0	0	0	0	0	0	0	0
77-80	0	0	0	0	0	0	0	0	0
81-84	0	0	0	0	0	0	0	0	0
85-88	0	0	0	0	0	0	0	0	0

Tableau 4.6. Cancer du sein dans le Bas-Rhin. Valeurs de la variable de dépistage (proportion de femmes dépistées). Dans les cases encadrées, elle est égalée à 0 (scénario avec interruption du dépistage). Dans les autres cases, il s'agit de valeurs extraites des données.

Ag\An	1975-1976	1977-1980	1981-1984	1985-1988	1989-1992	1993-1996	1997-2000	2001-2004	2005-2008
25-28	0	0	0	0	0	0	0	0	0
29-32	0	0	0	0	0	0	0	0	0
33-36	0	0	0	0	0	0	0	0	0
37-40	0	0	0	0	0	0	0	0	0
41-44	0	0	0	0	0	0	0	0	0
45-48	0	0	0	0	0	0	0	0	0
49-52	0	0	0	0	0,1331033	0,1104671	0	0	0
53-56	0	0	0	0	0,1372846	0,0395727	0	0	0
57-60	0	0	0	0	0,1291811	0,0290528	0	0	0
61-64	0	0	0	0	0,1210012	0,0276093	0	0	0
65-68	0	0	0	0	0,0176468	0,0044237	0	0	0
69-72	0	0	0	0	0	0	0	0	0
73-76	0	0	0	0	0	0	0	0	0
77-80	0	0	0	0	0	0	0	0	0
81-84	0	0	0	0	0	0	0	0	0
85-88	0	0	0	0	0	0	0	0	0

4.1.3. Résultats

4.1.3.1. Analyse descriptive

L'effectif de la population féminine totale est passé de 450 101 en 1975 à 489 836 en 1990. Celui des femmes de 25 à 89 ans est passé de 265 246 en 1975 et à 316 560 en 1990.

Tableau 4.7. Tumeurs du sein in situ : incidences spécifiques et nombre de cas incidents chez les femmes dans le Bas-Rhin, 1975-1996.

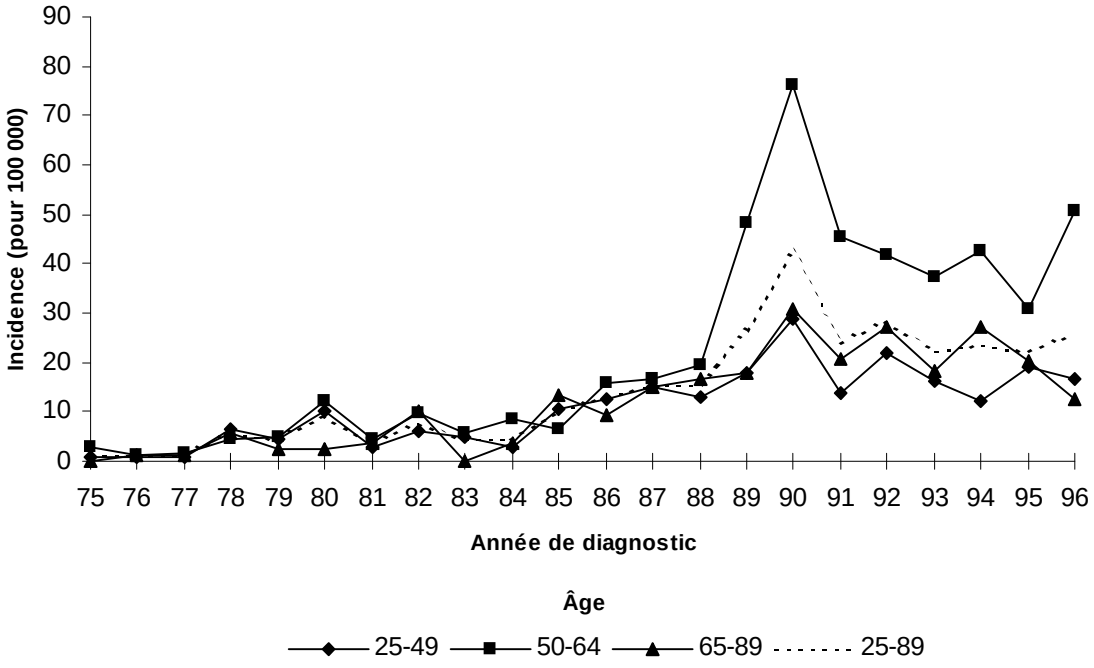
Âge \ Période	1975-1976	1977-1980	1981-1984	1985-1988	1989-1992	1993-1996
25-28	0,0 (0)	0,8 (1)	0,0 (0)	0,8 (1)	0,0 (0)	0,8 (1)
29-32	0,0 (0)	2,9 (3)	0,0 (0)	2,4 (3)	1,6 (2)	2,3 (3)
33-36	2,5 (1)	2,5 (2)	2,9 (3)	2,5 (3)	6,5 (8)	6,3 (8)
37-40	0,0 (0)	2,5 (2)	6,2 (5)	16,4 (17)	13,6 (16)	13,8 (17)
41-44	0,0 (0)	9,8 (8)	3,8 (3)	20,1 (16)	37,7 (39)	21,2 (25)
45-48	0,0 (0)	11,7 (10)	11,2 (9)	30,7 (24)	54,5 (43)	47,7 (49)
49-52	4,7 (2)	11,9 (10)	14,4 (12)	20,2 (16)	67,0 (52)	39,6 (31)
53-56	2,5 (1)	6,0 (5)	10,9 (9)	15,8 (13)	71,3 (56)	36,4 (28)
57-60	0,0 (0)	3,0 (2)	0,0 (0)	17,4 (14)	31,0 (25)	45,1 (35)
61-64	2,8 (1)	4,0 (2)	1,6 (1)	5,1 (4)	34,5 (27)	36,8 (29)
65-68	0,0 (0)	1,4 (1)	6,2 (3)	14,7 (9)	32,8 (25)	22,4 (17)
69-72	3,0 (1)	5,8 (4)	1,5 (1)	22,5 (10)	22,5 (13)	27,5 (20)
73-76	0,0 (0)	5,1 (3)	4,9 (3)	11,9 (7)	29,3 (12)	20,6 (11)
77-80	0,0 (0)	0,0 (0)	2,1 (1)	7,9 (4)	14,0 (7)	8,4 (3)
81-84	0,0 (0)	3,9 (1)	6,5 (2)	5,7 (2)	5,1 (2)	5,0 (2)
85-88	0,0 (0)	0,0 (0)	6,7 (1)	5,4 (1)	13,2 (3)	11,4 (3)
25-88	1,1 (6)	5,2 (54)	5,1 (53)	13,5 (144)	30,2 (330)	23,4 (282)

L'incidence est exprimée en nombre de cas pour 100 000. Entre parenthèses, figurent les nombres de cas incidents. Dans la dernière ligne figurent les incidences pour toute la population, standardisées selon la population européenne et, entre parenthèses, le nombre total de cas incidents.

L'incidence (standardisée selon la population européenne) des lésions ***in situ*** dans la population de 25 à 88 ans croît de 1,1 pour 10^5 pour la période 1975-1976 à 23,4 pour 10^5 durant la période 1993-1996 (Tableau 4.7.). Elle montre deux accélérations remarquables : la première, en 1985, correspond au « dépistage sauvage », la deuxième, en 1989, au début du dépistage organisé (Figure 4.2.). La pente atteint son maximum entre 1989 et 1990 puis l'incidence chute et la pente rejoint, en 1993, celle de la tendance générale (i.e. celle de la tendance antérieure à 1985). La Figure 4.2. montre que les trois classes d'âges (femmes dépistées, femmes plus jeunes, femmes plus âgées) présentent des évolutions

similaires, qualitativement et quantitativement, de 1975 à 1988 (croissance initiale et « dépistage sauvage »), mais l'augmentation d'incidence de 1989 concerne les femmes dépistées (50-64 ans).

Figure 4.2. Tumeurs du sein *in situ* chez les femmes dans le Bas-Rhin entre 1975 et 1996. Taux d'incidence standardisé selon la population européenne pour les trois groupes d'age 25-49, 50-64 et 65-89 ans et pour toute la population (25-89 ans).



Les incidences spécifiques des classes d'âge concernées par le dépistage (50 à 64 ans) augmentent très fortement durant les dépistages « sauvage » et organisé (pour la classe d'âge de 53 à 56 ans, par exemple, l'incidence est de 10,9 pour 100 000 en 1985-1988, 71,3 pour 100 000 en 1989-1992 et 36,4 pour 100 000 en 1993-1996) (Tableau 4.7.).

L'incidence des **tumeurs invasives**, standardisée selon la population européenne, croît régulièrement de 126,9 pour 10^5 au cours de la période 1975-1976 à 180,8 pour 10^5 durant la période 1993-1996 (Tableau 4.8.) ; la tendance suit une courbe oscillante et, contrairement aux tumeurs *in situ*, il n'y a pas de pic net correspondant au dépistage « sauvage » (1985) ou au dépistage organisé (1989) (Figure 4.3.). Pour les femmes âgées de 25 à 49 ans, l'incidence augmente très faiblement ; pour les femmes

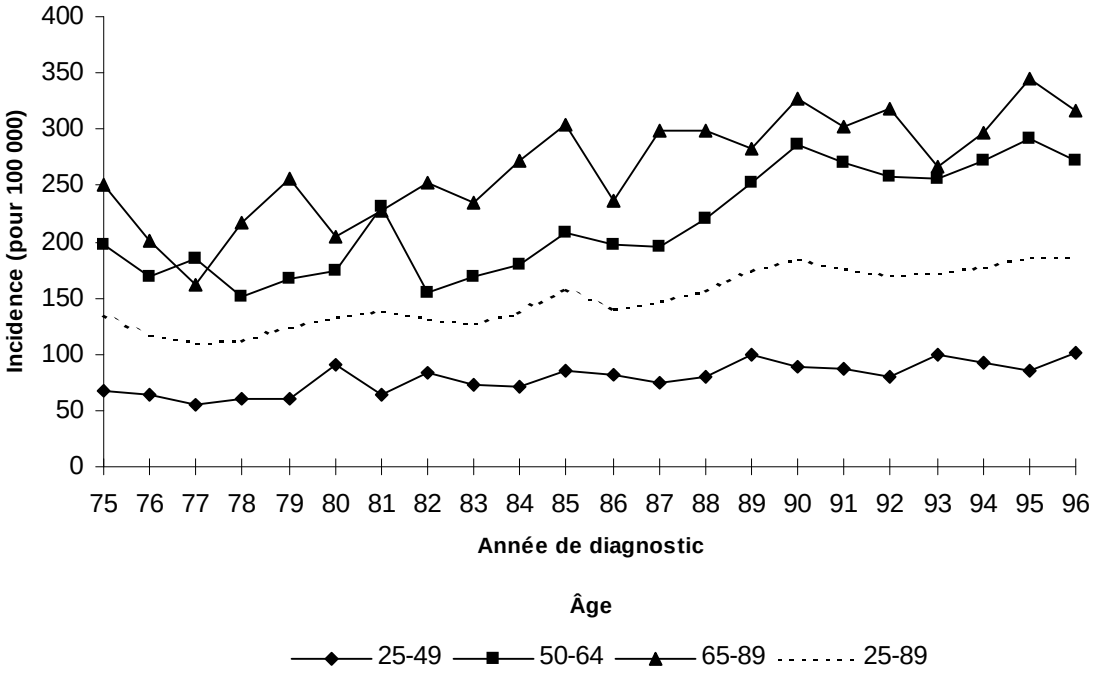
de 50 à 64 ans (femmes dépistées) et pour les femmes de plus de 64 ans, les tendances sont parallèles et nettement croissantes (les pentes sont, respectivement, de 5,83 et 5,73 cas / 100 000 personnes-années x année).

Tableau 4.8. Cancer du sein invasif : incidences spécifiques et nombre de cas incidents chez les femmes dans le Bas-Rhin, 1975-2008.

Âge\Période	1975-1976	1977-1980	1981-1984	1985-1988	1989-1992	1993-1996	1997-2000	2001-2004	2005-2008
25-28	12.4 (7)	2.5 (3)	8.2 (10)	9.8 (12)	4.7 (6)	10.8 (14)	10,3 (13)	11,9 (13)	13,0 (15)
29-32	19.7 (8)	17.2 (18)	19.4 (23)	17.8 (22)	20.6 (26)	23.3 (30)	24,2 (31)	25,6 (32)	27,7 (30)
33-36	24.9 (10)	33.5 (27)	39.6 (41)	39.1 (46)	47.8 (59)	56.3 (71)	51,1 (65)	54,1 (69)	58,7 (73)
37-40	76.5 (31)	71.3 (57)	69.9 (56)	64.8 (67)	76.3 (90)	84.1 (104)	93,8 (117)	100,2 (127)	107,9 (137)
41-44	124.1 (53)	100.4 (82)	107.2 (85)	142.1 (113)	144.1 (149)	141.8 (167)	159,9 (195)	171,1 (212)	183,5 (231)
45-48	124.4 (53)	152.7 (130)	161.2 (130)	192.8 (151)	210.2 (166)	223.0 (229)	229,8 (266)	246,5 (298)	264,5 (325)
49-52	165.5 (70)	154.6 (130)	188.8 (157)	176.6 (140)	229.2 (178)	250.2 (196)	266,4 (269)	283,8 (325)	304,2 (364)
53-56	149.2 (59)	147.0 (123)	175.1 (144)	203.5 (167)	248.2 (195)	285.0 (219)	287,8 (221)	309,1 (308)	328 (371)
57-60	189.8 (42)	184.3 (122)	173.1 (141)	185.5 (149)	287.8 (232)	283.6 (220)	310,7 (234)	340,5 (258)	364,9 (359)
61-64	233.8 (84)	204.7 (103)	202.4 (129)	252.1 (199)	303.9 (238)	251.2 (198)	335,9 (253)	377,2 (278)	412,1 (306)
65-68	237.2 (88)	215.0 (151)	237.4 (114)	272.5 (167)	304.4 (232)	286.6 (217)	352,9 (268)	398,3 (292)	446,1 (321)
69-72	240.4 (81)	198.5 (137)	249.6 (162)	308.4 (137)	340.4 (197)	315.2 (229)	362,1 (261)	407,1 (297)	459,6 (325)
73-76	217.5 (60)	217.2 (127)	253.0 (155)	311.2 (183)	327.2 (134)	364.8 (195)	368,9 (248)	404,7 (273)	455,3 (313)
77-80	244.9 (47)	213.5 (92)	260.6 (124)	288.2 (146)	296.0 (148)	308.5 (110)	355,2 (168)	384,6 (231)	421,9 (257)
81-84	216.7 (24)	243.9 (63)	237.2 (73)	260.2 (92)	239.4 (94)	315.5 (125)	334,8 (95)	361,7 (141)	388,5 (194)
85-88	175.8 (9)	167.7 (20)	260.0 (39)	219.8 (41)	273.2 (62)	273.8 (72)	305,8 (83)	337,3 (67)	365,0 (104)
25-88	126,9 (726)	119,9 (1385)	132,6 (1583)	150,0 (1832)	175,9 (2206)	180,8 (2396)	199,2 (2787)	217,3 (3221)	236,2 (3725)

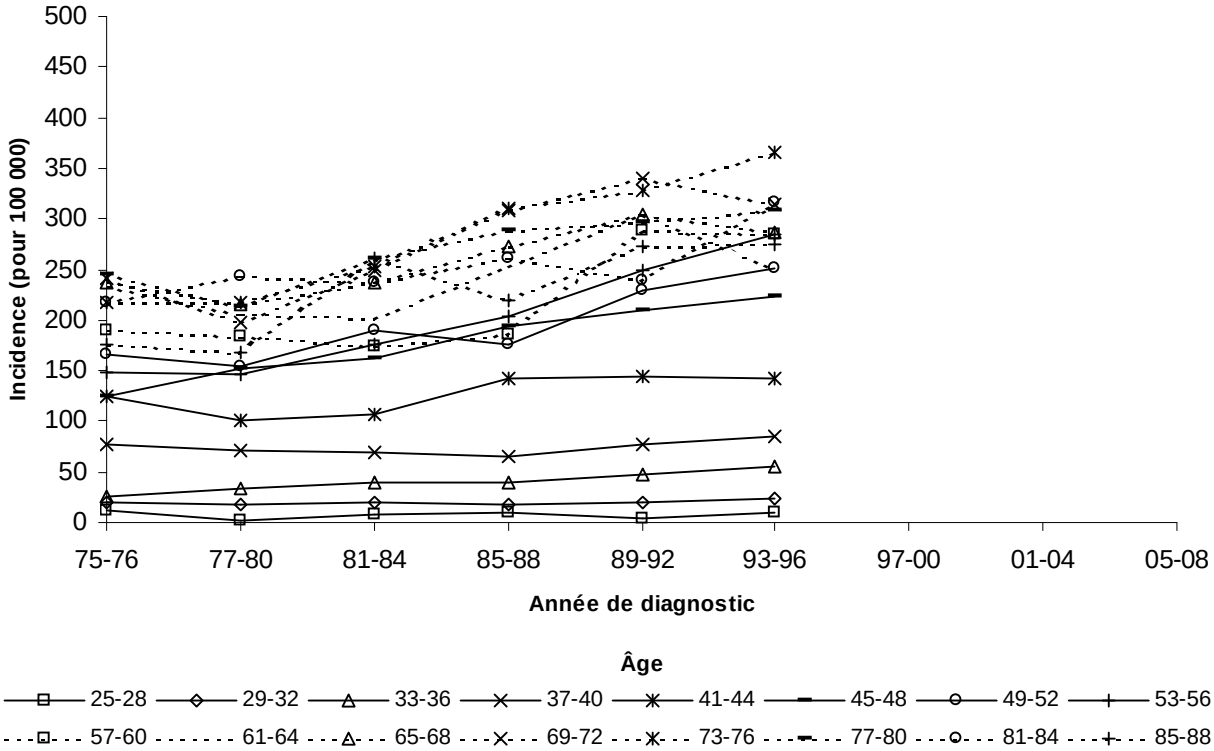
L'incidence est exprimée en nombre de cas pour 100 000. Entre parenthèses, figurent les nombres de cas incidents. Dans la dernière ligne figurent les incidences pour toute la population, standardisées selon la population européenne et, entre parenthèses, le nombre total de cas incidents. Pour les périodes 1975-76 à 1993-96, les incidences et les nombres de cas incidents sont extraits du registre ; pour les autres périodes, il s'agit de valeurs prédites. Les cases grisées contiennent les incidences modifiées dans l'analyse de sensibilité.

Figure 4.3. Cancers invasifs du sein chez les femmes dans le Bas-Rhin entre 1975 et 1996. Taux d'incidence standardisé selon la population européenne pour les trois groupes d'age 25-49, 50-64 et 65-89 ans et pour toute la population (25-89 ans).



Les incidences spécifiques des classes d'âge concernées par le dépistage (50 à 64 ans) augmentent relativement moins que pour les tumeurs *in situ* (pour la classe d'âge de 53 à 56 ans, par exemple, l'incidence est de 175,1 pour 10^5 en 1981-1984, 203,5 pour 10^5 en 1985-1988, 248,2 pour 10^5 en 1989-1992 et 285,0 pour 10^5 en 1993-1996) (Tableau 4.8., figure 4.4.).

Figure 4.4. Cancers invasifs du sein chez les femmes âgées de 25 à 88 ans dans Bas-Rhin : incidences spécifiques issues du registre, 1975-1996.



4.1.3.2. Analyse âge-période-cohorte et prédictions

En raison des fluctuations de l'incidence, dues aux dépistages, les projections réalisées pour les tumeurs in situ n'ont pas été retenues. En effet, les résultats obtenus à partir des données brutes étaient anormalement élevés et les tentatives de « lissage » des incidences mesurées au cours des dépistages ne sont pas parvenues à rendre les prédictions plus vraisemblables.

Le nombre total de cas incidents prédit pour les cancers invasifs croît de façon linéaire (Tableaux 4.8. et 4.9.) et double en 20 ans : 1832 en 1985-1988, 3725 en 2005-2008 (IC 95 % : [3291 – 4218]). En 1997-2000, et en 2001-2004, les nombres de cas incidents estimés sont, respectivement, de 2787 (IC 95 % : [2545 – 3053]) et de 3221 (IC 95 % : [2897 – 3587]).

Tableau 4.9. Cancer du sein invasif : nombre de cas incidents et intervalle de prédiction (95 %) chez les femmes dans le Bas-Rhin, 1997-2008.

Âge\Période	1997-2000	2001-2004	2005-2008
25-28	13 (10 - 18)	13 (8 - 20)	15 (7 – 31)
29-32	31 (25 - 38)	32 (24 - 44)	30 (20 - 47)
33-36	65 (56 - 75)	69 (57 - 88)	73 (54 - 100)
37-40	117 (104 - 131)	127 (110 – 149)	137 (112 - 172)
41-44	195 (176 - 213)	212 (189 – 240)	231 (198 - 268)
45-48	266 (242 - 289)	298 (266 – 329)	325 (284 - 366)
49-52	269 (246 - 292)	325 (292 – 358)	364 (320 - 405)
53-56	221 (203 - 240)	308 (278 – 336)	371 (328 - 408)
57-60	234 (217 - 253)	258 (237 – 281)	359 (324 - 395)
61-64	253 (236 - 274)	278 (256 – 302)	306 (280 - 336)
65-68	268 (249 - 290)	292 (268 – 318)	321 (294 - 353)
69-72	261 (241 - 284)	297 (272 – 326)	325 (296 - 357)
73-76	248 (228 - 271)	273 (248 – 303)	313 (282 - 348)
77-80	168 (155 - 185)	231 (209 – 257)	257 (231 - 289)
81-84	95 (85 - 105)	141 (126 – 159)	194 (173 - 222)
85-88	83 (72 - 95)	67 (57 - 77)	104 (88 - 121)
25-88	2787 (2545 - 3053)	3221 (2897 - 3587)	3725 (3291 - 4218)

Les classes d'âge 65-68 ans et 69-72 ans sont celles qui présentent les augmentations d'incidence les plus fortes durant les trois périodes de prédiction, soit, pour chacune d'elles, 360 cas pour 10^5 environ en 1997-2000, 400 cas pour 10^5 environ en 2001-2004 et 450 cas pour 10^5 environ en 2005-2008 (Tableaux 4.8 et 4.9., figure 4.5.).

L'incidence standardisée, prévue par le modèle pour les femmes de 25 à 89 ans, double entre 1977 (119,9 pour 10^5) et 2008 (236,2 pour 10^5) (Tableau 4.8., Figure 4.6.). Les incidences standardisées prédites par le modèle sont de 199,2 pour 10^5 en 1997-2000 (IC 95 % : [182,2 – 217,8]), de 217,3 pour 10^5 en 2001-2004 (IC 95 % : [195,6 – 241,6]) et de 236,2 pour 10^5 en 2005-2008 (IC 95 % : [208,4 – 267, 5]). Les âges les plus affectés en terme d'augmentation d'incidence sont compris entre 53 et 72 ans (100 à 130 % entre 1977-1980 et 2005-2008, soit 30 ans).

Figure 4.5.Cancers invasifs du sein chez les femmes âgées de 25 à 89 ans dans le Bas-Rhin : incidences spécifiques prédites (âges : 25-28 à 85-88), 1975-2008 (modèle âge-cohorte).

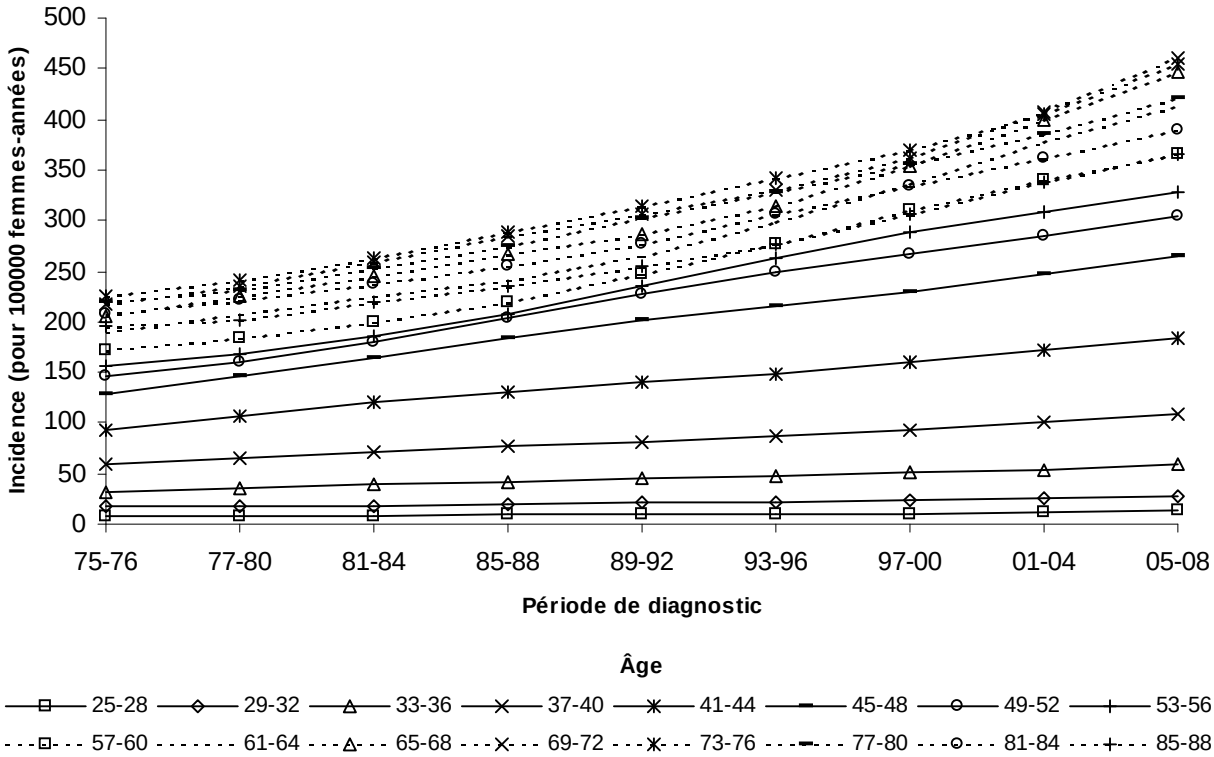
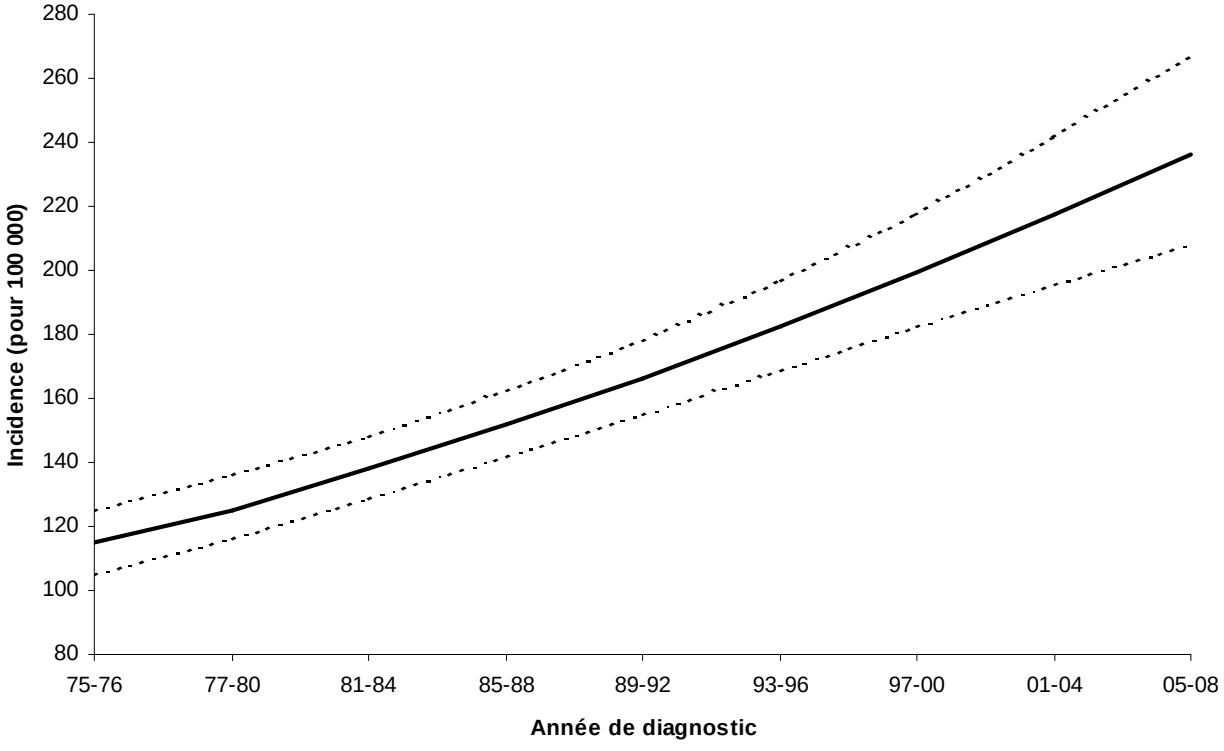


Figure 4.6. Cancers invasifs du sein chez les femmes âgées de 25 à 89 ans dans le Bas-Rhin. Prédiction de l'incidence (1975-2008) standardisée selon la population européenne et intervalle de prédiction à 95%.



4.1.3.3. Analyses de sensibilité

Cinq analyses de sensibilité ont été réalisées.

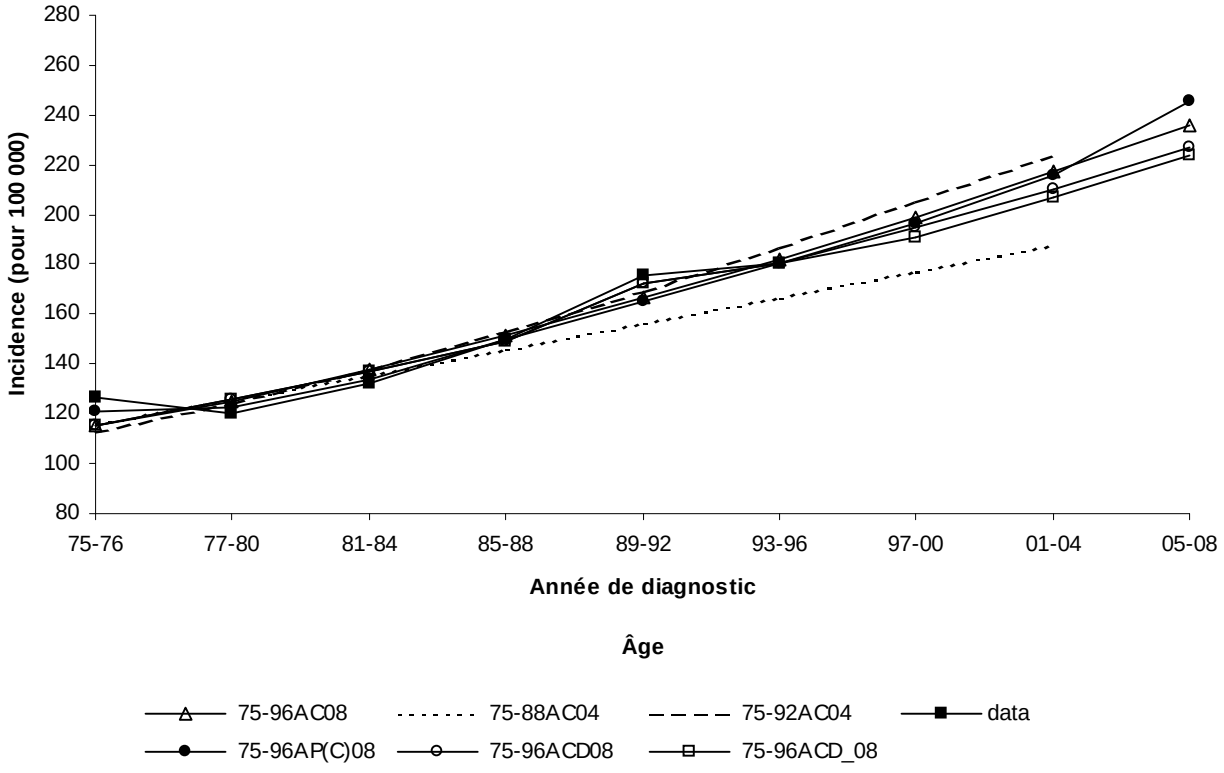
Afin de vérifier que le modèle n'a pas surestimé les incidences futures en raison de l'augmentation brutale des taux dans les classes d'âge dépistées lors du début du dépistage, les calculs ont été réalisés :

1°) A l'horizon 2004 sur la base des données enregistrées entre 1975 et 1988

2°) A l'horizon 2004 sur la base des données enregistrées entre 1975 et 1992

Les taux étaient en 2004, respectivement, de 187,6 pour 10^5 et 223,9 pour 10^5 (Figure 4.7.).

Figure 4.7. Cancers invasifs du sein chez les femmes âgées de 25 à 89 ans dans le Bas-Rhin. Prédiction de l'incidence (1975-2008) standardisée selon la population européenne à l'aide de différents modèles âge-cohorte.



Différents modèles : 1) base de prédiction 1975-1996, prédiction jusqu'en 2008 (75-96AC08) ; 2) base de prédiction 1975-1988, prédiction jusqu'en 2004 (75-88AC04) ; 3) base de prédiction 1975-1992, prédiction jusqu'en 2004 (75-92AC04). 4) données enregistrées ; 5) base de prédiction 1975-1996, prédiction jusqu'en 2008 après interpolation pour les cohortes concernées par le dépistage (75-96A(C)08) ; 6) base de prédiction 1975-1996, prédiction jusqu'en 2008 avec prise en compte du dépistage continu (75-96ACD08) ; 7) base de prédiction 1975-1996, prédiction jusqu'en 2008 avec prise en compte d'un dépistage interrompu en 1997 (75-96ACD08).

3°) Le Tableau 4.2. montre que deux cohortes présentent un pic d'incidence au moment du dépistage organisé (cases encadrées grisées et non grisées) : 303.9 pour 10^5, pour la cohorte de naissance 1925-1931 (classe d'âge 61-64 ans, période 1989-1992) et 287.8 pour 10^5, pour la cohorte de naissance 1929-1935 (classe d'âge 57-60 ans, période 1989-1992). L'analyse de sensibilité a été réalisée en remplaçant, pour chacune des deux cohortes, ces incidences par la moyenne des incidences des périodes juste avant (1985-1988) et juste après (1993-1996). Le résultat du calcul ne diffère pas beaucoup des prédictions précédentes : les incidences standardisées sont de 196,5 pour 10^5 en 1997-2000, de 215,8 pour 10^5 en 2001-2004 et de 245,3 pour 10^5 en 2005-2008 (Figure 4.7.). Cependant, les intervalles de confiance sont plus petits.

4°) L'introduction d'une variable dépistage donne également des résultats proches de l'analyse sans variable dépistage. Le modèle basé sur l'hypothèse d'une constance du recrutement affecte essentiellement les incidences spécifiques concernées par le dépistage (49 à 60 ans) (Figure 4.8.) et donne une prévision légèrement plus basse que le modèle qui ne tient pas compte du dépistage (Tableau 4.10.., figure 4.7.). Le modèle basé sur l'hypothèse d'un arrêt du dépistage (en 1997) prévoit une incidence très légèrement inférieure à celle du modèle avec continuation du dépistage avec, cependant, des intervalles de prévision de même amplitude (Tableau 4.10., figure 4.9.).

Figure 4.8. Cancers invasifs du sein chez les femmes âgées de 25 à 89 ans dans le Bas-Rhin : incidences spécifiques prédites (âges : 25-28 à 85-88), 1975-2008 (modèle âge-cohorte). Le modèle tient compte du dépistage avec l'hypothèse d'une intensité de recrutement constante (extrapolation).

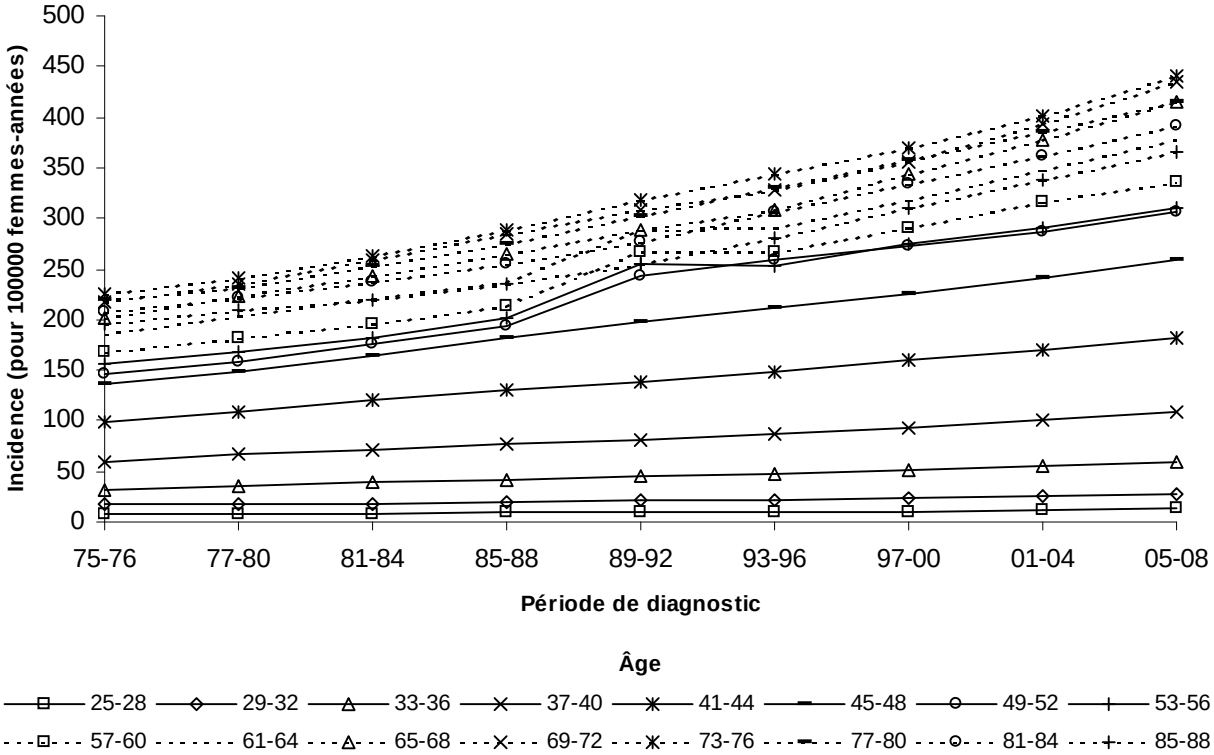
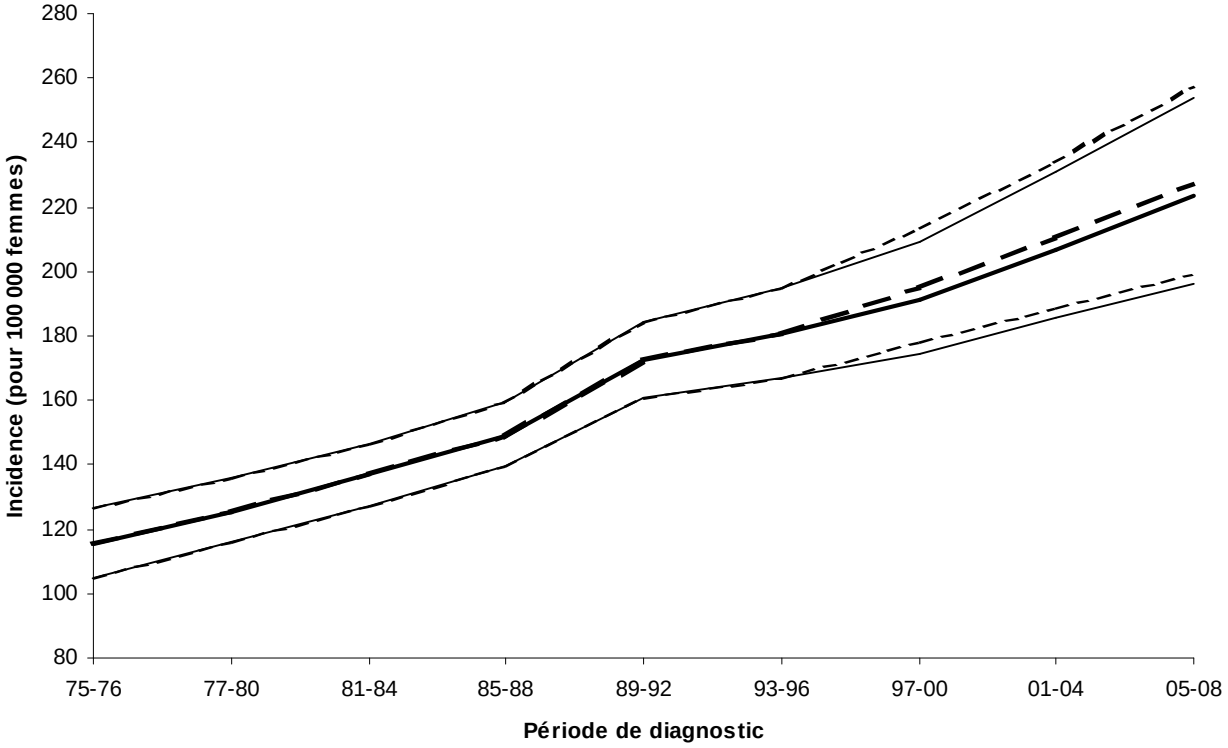


Tableau 4.10. Cancers invasifs du sein chez les femmes âgées de 25 à 89 ans dans le Bas-Rhin. Prédiction de l'incidence (1975-2008) standardisée selon la population européenne et intervalle de prédiction à 95 % grâce à différents modèles.

Période	1975-1976	1977-1980	1981-1984	1985-1988	1989-1992	1993-1996	1997-2000	2001-2004	2005-2008
A	115,0	125,3	138,0	151,7	166,5	182,4	199,2	217,3	236,2
	(104,7 - 125,3)	(116,1 - 136,0)	(128,5 - 148,1)	(141,6 - 162,2)	(155,2 - 178,2)	(168,8 - 196,9)	(182,2 - 217,8)	(195,6 - 241,6)	(208,4 - 267,5)
B	115,4	125,5	136,9	149,2	172,3	180,7	195	210,2	227
	(105,1 - 126,5)	(115,8 - 135,6)	(127,2 - 146,4)	(139,4 - 159,3)	(160,9 - 184,4)	(167,0 - 194,9)	(178,0 - 213,4)	(188,5 - 234,1)	(199,1 - 257,9)
C	115,4	125,5	136,9	149,2	172,3	180,7	191,1	207,0	223,7
	(105,1 - 126,5)	(115,8 - 135,6)	(127,2 - 146,4)	(139,4 - 159,3)	(160,9 - 184,4)	(167,0 - 194,9)	(174,6 - 209,3)	(185,5 - 230,7)	(196,3 - 254,2)

Trois modèles sont testés : 1°) Sans tenir compte du dépistage (A). 2°) En tenant compte du dépistage avec l'hypothèse d'une intensité de recrutement constante (extrapolation). 3°) En tenant compte du dépistage avec l'hypothèse d'un arrêt du recrutement en 1997 (C).

Figure 4.9. Cancers invasifs du sein chez les femmes âgées de 25 à 89 ans dans le Bas-Rhin. Prédiction de l'incidence (1975-2008) standardisée selon la population européenne et intervalle de prédiction à 95 % en tenant compte du dépistage. Deux hypothèses sont testées : 1°) Arrêt du dépistage en 1997 (trait continu). 2°) Continuation du dépistage avec une intensité de recrutement constante (extrapolation).



5°) Pour tester l'influence d'erreurs éventuelles d'enregistrement relatives à la mise en route du registre, les prédictions ont été réalisées sans tenir compte des années 1975 et 1976. Les résultats sont, là encore, proches des projections précédentes : pour les trois périodes de prédiction, les incidences calculées sont, respectivement, de 196.0 pour 10^5, 215.4 pour 10^5 et 235.4 pour 10^5.

4.1.4. Discussion

Les tendances respectives de l'incidence et du nombre de cas incidents ont été prédites pour les tumeurs invasives jusqu'en 2008. Les projections n'ont pas été retenues pour les tumeurs in situ en raison des perturbations dues aux dépistages. La méthode est basée sur un modèle âge-période-cohorte et impose un ensemble de contraintes afin de remédier à l'instabilité des paramètres correspondant aux cohortes extrêmes. L'incidence et le nombre de cas incidents augmentent pour les cancers invasifs.

L'analyse de la tendance et, par conséquent, sa prédiction, rencontre plusieurs difficultés liées aux données : le problème de la fiabilité de l'enregistrement des cas incidents (qui semble être résolu, à présent, grâce aux registres des cancers) et le problème lié à la précision de l'estimation (et de la prédiction) de la taille de la population, sont bien maîtrisés aujourd'hui.
Le début du dépistage peut provoquer une perturbation de la tendance de l'incidence, en augmentant artificiellement celle-ci par la révélation de cas de façon anticipée (lead time) puis, quand le nombre de femmes dépistées devient stable, par le retour de l'incidence à la tendance antérieure (Feuer, 1992 ; Wun, 1995 ; Kessler, 1991). La prédiction peut surestimer la tendance si elle ne prend pas en compte cette perturbation. L'analyse de sensibilité a montré que, pour les cancers invasifs, les projections étaient relativement stables lorsqu'on ôtait les pics d'incidence au moment du dépistage et lorsque la base de la prédiction recouvrait la période 1975-1992. Les analyses basées sur la période 1975-1978 obtenaient un taux plus faible en 2004 (187,6 pour 10^5) mais l'intervalle de prédiction (IC 95 % : [150,2 – 235,5] contenait le taux prédit sur la base de l'ensemble des données (217,3 pour 10^5 en 2004).

L'utilisation d'un modèle autorégressif appliqué aux effets âge, période et cohorte a un double intérêt : il permet de tenir compte de la dépendance existant entre les valeurs successives de la série longitudinale « nombre de cas » et sert de lissage tout en réduisant l'instabilité des cohortes extrêmes (Breslow, 1993). L'approche bayésienne a un double intérêt également : elle permet de résoudre des

problèmes inaccessibles à l'intégration numérique et offre une méthode élégante d'estimation des intervalles de confiance (Breslow, 1993).
Engeland et al. (Engeland, 1993) ont prédit l'incidence du cancer du sein en 2010 pour les cinq pays scandinaves sur la base des données des registres de 1958 à 1987. Les incidences sont plus faibles qu'ici mais toute la population a été prise en compte. Il prévoient, selon le pays, une augmentation d'incidence de 50 % (Suède) à près de 100 % (Finlande) entre 1990 et 2010. J. Vioque et al. (Vioque, 1993) prédisent une augmentation de 60 % de l'incidence du cancer du sein entre 1976 et 2000 sur la base des données du registre de Zaragoza (Espagne). En Bulgarie, L. Hristova et al. (Hristova, 1997) estiment l'augmentation de l'incidence à plus de 60 % entre 1968 et 2017. LG Kessler et al. (Kessler, 1991) partent du principe que le dépistage ne change pas la tendance de l'incidence à moyen terme et leur modèle met en évidence, aux Etats-Unis, une croissance de l'incidence d'environ 30% entre 1982 et 2000. K. Sigurdsson et al. (Sigurdsson, 1991) prévoient une élévation de 42 % de l'incidence, en Islande, de 1980 à 2000.

Les prévisions, obtenues à l'aide d'un modèle âge-période-cohorte ou fondées sur d'autres méthodes, concordent toutes sur la notion d'augmentation de l'incidence. L'importance de la croissance dépend de la population : les prédictions établies ici (augmentation de 2,5 % par an pour les cancers invasifs) présentent la même croissance que celles de Zaragoza (2,5% par an). Cette croissance est supérieure à celles de l'Islande et de la Bulgarie (2,1% par an), elle est nettement supérieure aux résultats de Kessler aux USA (1,7 % par an), mais est comprise entre les extrêmes prévus dans l'étude des 5 pays nordiques (1,7 à 3,4 % par an).

La base de prédiction utilisée dans cette étude ne tient pas compte du dépistage. Certains travaux tentent d'estimer la perturbation de l'incidence due au dépistage (Feuer, 1992 ; Wun, 1995 ; Miller, 1991 ; Miller, 1993), d'autres construisent un modèle de transition du stade non invasif au stade invasif (Day, 1984 ; Brookmeyer, 1987 ; Prorok, 1988 ; Van Oomarssen, 1995). Il serait intéressant d'intégrer ces résultats dans le modèle afin de l'améliorer. Ces résultats permettraient, d'autre part, d'estimer l'incidence future en tenant compte du « poids » du dépistage et, par là, d'évaluer le coût de la maladie et l'efficacité des différentes actions (dépistage, traitements).

4.2. Cancer du col de l'utérus

4.2.1. Introduction

Dans le monde, le cancer du col de l'utérus occupe la deuxième place des cancers de la femme pour l'incidence (après le cancer du sein) et représente 15% de l'ensemble des cancers (Coleman, 1993). Il se situe au premier rang dans les pays en voie de développement (20 à 30 % de l'ensemble des cancers féminins contre 4 à 6 % dans les pays développés) (Schaffer, 1997). Il participe pour 3% aux nouveaux cas de cancer en France : le taux d'incidence, standardisé selon la population européenne, était de 9,9 pour 100 000 en 1995 (Réseau Francim, 1998). Le taux spécifique par âges présente deux maxima : respectivement 20,0 et 23,0 pour 100 000, pour les tranches d'âge 40-45 ans et 75-79 ans. Sur la base des données des registres des cancers, le taux d'incidence standardisé selon la population européenne était compris, pour la période 1988-1992, entre 10,0 pour 100 000 (Tarn et Doubs) et 17,7 pour 100 000 (Héraut) ; dans le Bas-Rhin, l'incidence était de 12,9 pour 100 000 (Réseau Francim, 1998).
De façon générale, l'incidence décroît dans les pays où un dépistage a été mis en place (Coleman, 1993 ; Vizcaino, 1998 ; Vizcaino, 2000 ; Ciatto, 1995). En Europe, l'incidence diminue partout sauf en Angleterre (où elle est stable ou bien dépend de la région) (Gibson, 1997) et en Espagne (où elle augmente) (Coleman, 1993 ; Ciatto, 1995 ; Raymond, 1995 ; Levi, 1994). L'analyse de la tendance du cancer invasif dans dix registres du sud-ouest de l'Europe entre 1970 et 1990 a montré une diminution de l'incidence de 3% par an (Raymond, 1995). La décroissance n'était pas la même selon le registre et le groupe d'âge et concernait surtout les classes d'âge moyennes (ces dernières sont moins exposées aux papillomavirus que les jeunes et ont plus souvent recours au dépistage que les femmes les plus âgées). En France, le taux d'incidence, pour l'ensemble de la population féminine, est passé de 22.4 pour 100 000 en 1975 à 9.9 pour 100 000 en 1995 (Réseau Francim, 1998). Comme dans l'ensemble des registres, cette baisse, en France (3.5 % par an), concerne les femmes dont l'âge est compris entre 45 et 69 ans (Weidmann, 1998). En dessous de 45 ans et au dessus de 69 ans, l'incidence est stable. Dans le Bas-Rhin, l'incidence décroît de 5 % par an (Raymond, 1995).

Malgré la décroissance régulière de l'incidence, le cancer du col est encore trop fréquent. Ceci montre l'intérêt de prédire son évolution en terme d'incidence. D'autant plus que l'estimation du nombre des cas incidents futurs répond à la demande des pouvoirs publics qui ont pour mission de prévoir les

infrastructures nécessaires à la prise en charge de cette maladie. Cette estimation peut également s'inscrire utilement dans l'ensemble des connaissances nécessaires à la gestion et au développement du dépistage organisé du cancer du col.
Ici encore, les travaux qui ont été consacrés à ce domaine sont fort peu nombreux (Hristova, 1997 ; Sigurdsson, 1991 ; Engeland, 1993).
Différentes méthodes permettent de prédire l'évolution de l'incidence. Si l'on connaît les facteurs de risque du cancer, leur relation à la maladie et leur évolution future, Il est possible d'utiliser ceux-ci comme support de prévision. Dans le cas du cancer du col, l'hygiène et les comportements sexuels sont des facteurs de risque mais leur évolution est difficile à prédire. Aussi, est-il préférable d'utiliser une méthode de modélisation prenant en compte l'influence de l'âge (durée d'exposition aux facteurs de risque ou aux facteurs protecteurs), de la période (exposition simultanée de l'ensemble de la population à ces facteurs) et de la cohorte (exposition commune à des sujets nés au même moment).

La prévision des incidences a été établie, séparément pour les lésions in situ et les cancers invasifs, pour le Bas-Rhin jusqu'en 2014. En préliminaire, nous décrirons la tendance des taux d'incidence respectifs de ces deux ensembles de pathologies de 1975 à 1999.

4.2.2. Matériels et méthodes

La prévision a été réalisée pour les périodes 2000-2004, 2005-2009 et 2010-2014 à partir des périodes 1975-1979 à 1995-1997. La population prise en compte est composée des femmes bas-rhinoises dont les âges sont compris entre 20 et 89 ans.
Les données d'incidence (nombres de cas incidents de tumeurs in situ et de cancers invasifs, par âge et par année) ont été extraites de la base de données du registre des tumeurs du Bas-Rhin. À partir des données disponibles (1975 à 1997), les incidences et les nombres de cas incidents ont été calculés et rassemblés dans un tableau âge-période par tranches de cinq ans.
La prévision a été établie sur la base d'un modèle âge-période-cohorte autorégressif. L'analyse calcule les effets de ces trois facteurs à partir d'un tableau de données (nombre de cas incidents et nombre de personnes-années) réparties selon l'âge et la période (par tranches de 5 ans).
Le modèle âge-période-cohorte complet est testé tout d'abord. Il est conservé s'il est stable. S'il n'est pas stable, des modèles partiels sont testés (modèles âge-période, âge-cohorte ou cohorte-période).

4.2.3. Résultats

4.2.3.1. Description de la tendance de l'incidence au cours de la période 1975-1999.

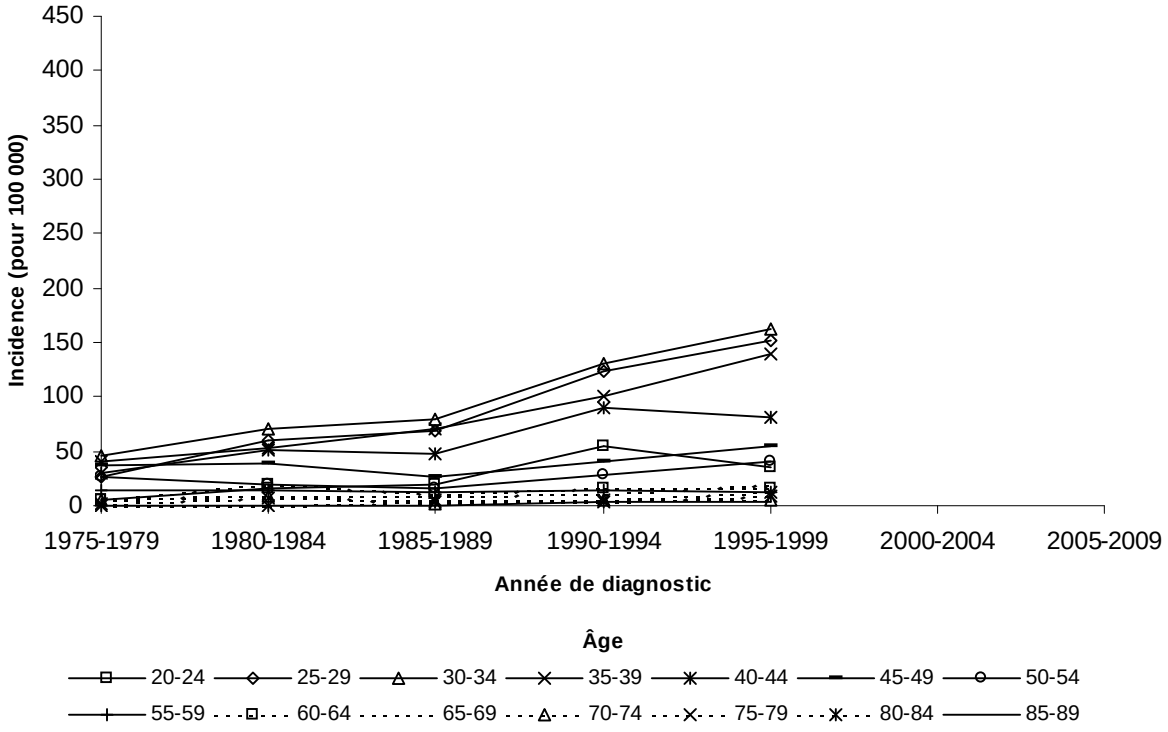
Quelle que soit la période, l'incidence des lésions *in situ* croît d'abord avec l'âge, passe par un maximum pour la tranche d'âge 30-34 ans puis décroît (Tableau 4.11., Figure 4.10.).

Tableau 4.11. Tumeurs *in situ* du col de l'utérus : incidences spécifiques et nombre de cas incidents chez les femmes dans le Bas-Rhin de 1975 à 2014.

Âge\Période	1975-1979	1980-1984	1985-1989	1990-1994	1995-1999	2000-2004	2005-2009	2010-2014
20-24	5,1(10)	15,5(32)	19,1(41)	54,8(114)	35,7(65)	54,7(101)	72,1(137)	125,3(243)
25-29	25,9(46)	59,1(112)	69,6(134)	123,0(247)	151,6(305)	188,0(332)	232,3(417)	306,4(566)
30-34	46,4(62)	69,9(124)	78,8(151)	130,3(258)	161,6(322)	258,1(515)	310,3(544)	383,2(683)
35-39	39,9(50)	53,6(71)	71,2(125)	101,3(194)	138,6(271)	206,5(409)	301,8(599)	363,0(634)
40-44	30,1(39)	50,7(63)	47,8(63)	89,9(159)	80,3(152)	137,2(266)	201,8(397)	294,7(581)
45-49	37,6(50)	38,3(49)	26,2(32)	41,2(54)	54,3(94)	77,2(144)	112,5(216)	165,2(322)
50-54	25,8(34)	19,3(25)	15,2(19)	28,0(34)	41,3(53)	43,9(75)	58,1(107)	85,5(162)
55-59	14,5(15)	14,0(18)	12,6(16)	14,5(18)	12,7(15)	21,4(27)	29,2(49)	38,0(69)
60-64	5,5(5)	20,2(20)	11,4(14)	15,4(19)	15,2(18)	16,4(19)	21,1(26)	28,0(46)
65-69	6,2(7)	8,2(7)	8,5(8)	11,0(13)	19,9(23)	13,0(15)	14,3(16)	18,3(22)
70-74	5,9(6)	8,9(9)	2,6(2)	5,7(5)	4,5(5)	9,9(11)	10,0(11)	11,2(12)
75-79	1,4(1)	7,3(6)	4,7(4)	6,0(4)	12,5(10)	9,1(9)	9,9(10)	9,9(10)
80-84	0,0(0)	0,0(0)	3,4(2)	3,1(2)	8,8(5)	6,4(4)	7,4(6)	8,3(7)
85-89	0,0(0)	0,0(0)	0,0(0)	2,9(1)	4,3(2)	3,1(1)	4,8(2)	5,4(3)
20-89	23,0(325)	33,9(536)	35,0(611)	59,5(1122)	69,4(1340)	99,7(1928)	132,7(2537)	177,1(3360)

L'incidence est exprimée en nombre de cas pour 100 000. Entre parenthèses figurent les nombres de cas incidents. De 1975 à 1999, les incidences et les nombres de cas incidents sont calculés directement à partir des données du registre. De 2000 à 2014, les incidences et les nombres de cas incidents sont projetés. La dernière ligne comporte les incidences standardisées selon la population européenne et, entre parenthèses, le nombre total de cas incidents.

Figure 4 .10. Cancers *in situ* du col de l'utérus chez les femmes âgées de 20 à 89 ans dans Bas-Rhin : incidences spécifiques issues du registre, 1975-1999.



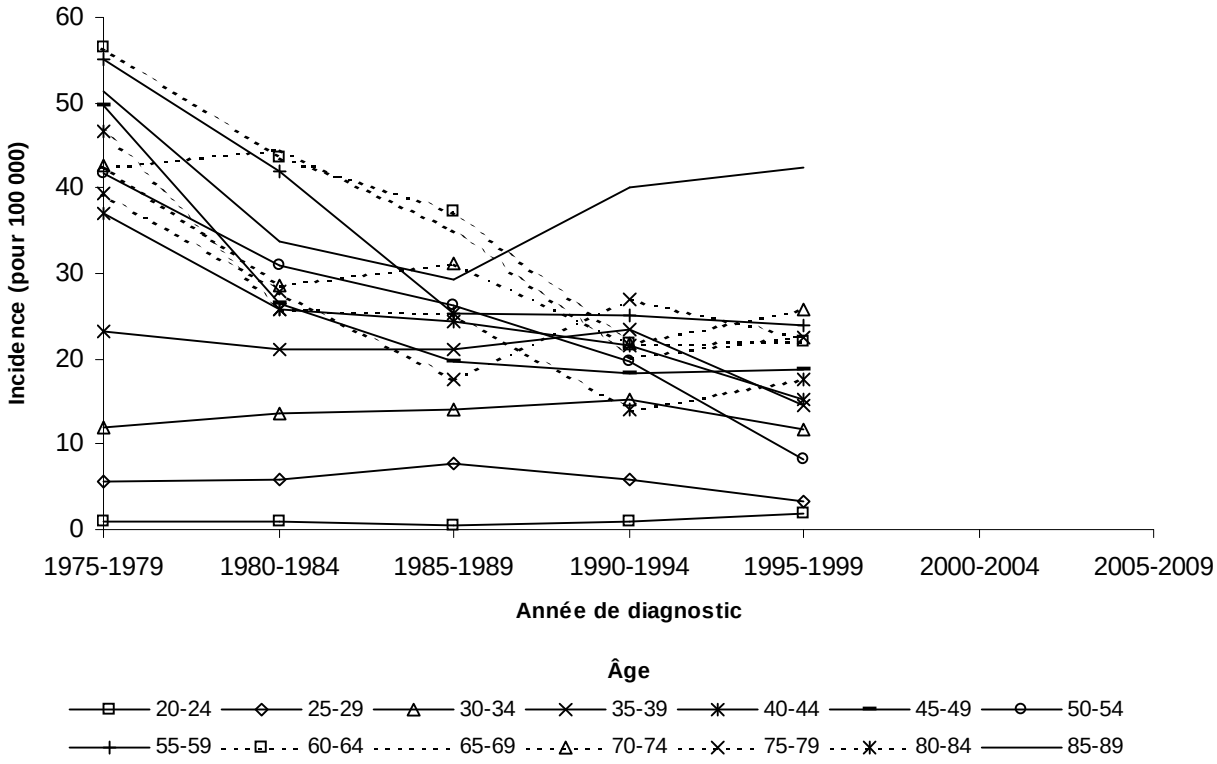
En ce qui concerne les cancers invasifs, la relation du taux d'incidence avec l'âge est moins nette (Tableau 4.12. , Figure 4.11.). Cette relation est tout d'abord croissante. Le taux passe alors par un maximum correspondant à une tranche d'âge comprise, selon la période, entre 50-54 ans et 65-69 ans (en 1990-1994, par exemple, l'incidence était maximale pour la tranche d'âge 55-59 ans et atteignait 25,0 pour 100 000). Puis le taux décroît, atteint un minimum et croît à nouveau.

Tableau 4.12. Cancers invasifs du col de l'utérus : incidences spécifiques et nombre de cas incidents chez les femmes dans le Bas-Rhin de 1975 à 2014.

Âge\Période	1975-1979	1980-1984	1985-1989	1990-1994	1995-1999	2000-2004	2005-2009	2010-2014
20-24	1(2)	1(2)	0,5(1)	1(2)	1,8(3)	1,1(2)	1,1(2)	1,5(3)
25-29	5,6(10)	5,8(11)	7,8(15)	6(12)	3,3(7)	4,0(7)	3,9(7)	3,8(7)
30-34	12(16)	13,5(24)	14,1(27)	15,2(30)	11,7(23)	10,0(20)	9,7(17)	9,5(17)
35-39	23,2(29)	21,1(28)	21,1(37)	23,5(45)	14,5(28)	15,6(31)	14,6(29)	14,3(25)
40-44	37(48)	25,7(32)	24,3(32)	21,5(38)	15,2(29)	17,5(34)	16,3(32)	15,7(31)
45-49	49,6(66)	26,6(34)	19,6(24)	18,3(24)	18,8(32)	16,6(31)	15,6(30)	14,9(29)
50-54	41,8(55)	30,9(40)	26,4(33)	19,8(24)	8,3(11)	14,6(25)	13,6(25)	13,2(25)
55-59	55,1(57)	42,1(54)	25,2(32)	25(31)	23,9(28)	16,7(21)	14,9(25)	14,3(26)
60-64	56,4(51)	43,5(43)	37,3(46)	21,9(27)	22,1(26)	17,3(20)	14,6(18)	13,4(22)
65-69	42,5(48)	44,6(38)	35,1(33)	20,3(24)	22,7(27)	16,5(19)	13,4(15)	11,7(14)
70-74	42,6(43)	28,6(29)	31,2(24)	21,7(19)	25,7(28)	15,3(17)	11,9(13)	10,3(11)
75-79	39,5(29)	27,8(23)	17,6(15)	27(18)	22,5(17)	15,2(15)	11,9(12)	9,9(10)
80-84	46,6(19)	25,7(13)	25,3(15)	14(9)	17,5(9)	16,1(10)	13,6(11)	10,7(9)
85-89	51,3(8)	33,8(7)	29,3(8)	40,1(14)	42,5(17)	24,6(8)	19,0(8)	16,2(9)
20-89	32,1(481)	24,3(378)	20,5(342)	17,6(317)	14,7(285)	13,0(260)	11,7(244)	11,1(238)

L'incidence est exprimée en nombre de cas pour 100 000. Entre parenthèses figurent les nombres de cas incidents. De 1975 à 1999, les incidences et les nombres de cas incidents sont calculés directement à partir des données du registre. De 2000 à 2014, les incidences et les nombres de cas incidents sont projetés. La dernière ligne comporte les incidences standardisées selon la population européenne et, entre parenthèses, le nombre total de cas incidents.

Figure 4 .11. Cancers invasifs du col de l'utérus chez les femmes âgées de 20 à 89 ans dans Bas-Rhin : incidence spécifiques issues du registre, 1975-1999.



Le taux d'incidence des tumeurs *in situ* augmente avec la période. En 1975-1979, l'incidence standardisée selon la population européenne était de 23,0 pour 100 000, en 1990-1994 elle était de 59,5 pour 100 000 (Figure 4.12.). L'incidence des cancers invasifs évolue en sens contraire : en 1975-1979, le taux était de 32,1 pour 100 000 et en 1990-1994, le taux était de 17,6 pour 100 000 (figure 4.13.).

Figure 4 .12. Cancers *in situ* du col de l'utérus chez les femmes âgées de 20 à 89 ans dans le Bas-Rhin. Prédiction de l'incidence (1975-2014) standardisée selon la population européenne et intervalle de prédiction à 95%. Pour les périodes 1975-79 à 1995-99, l'incidence est extraite du registre ; pour les autres périodes, il s'agit de valeurs prédites.

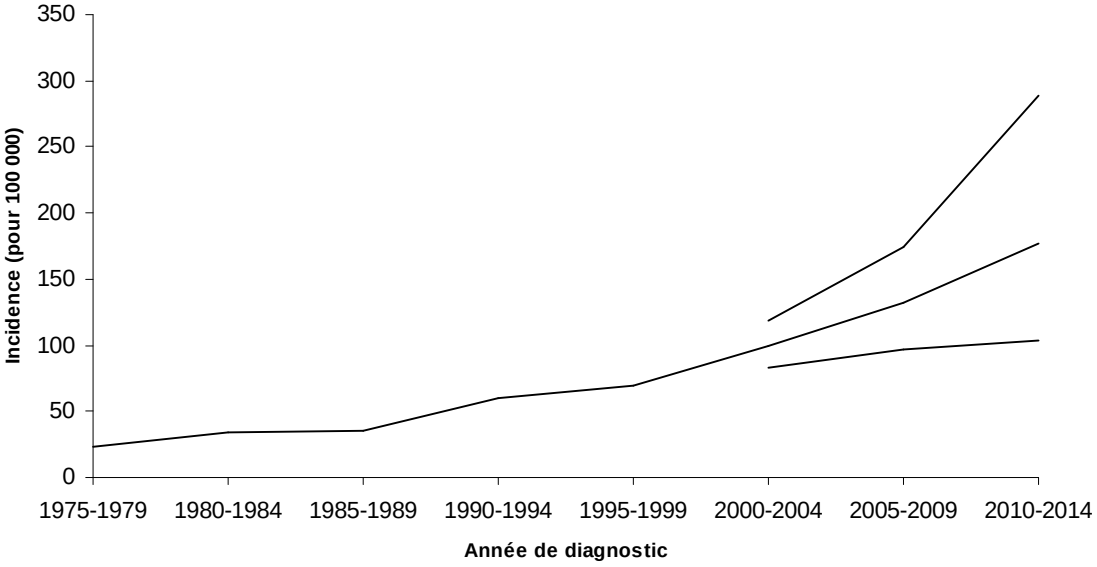
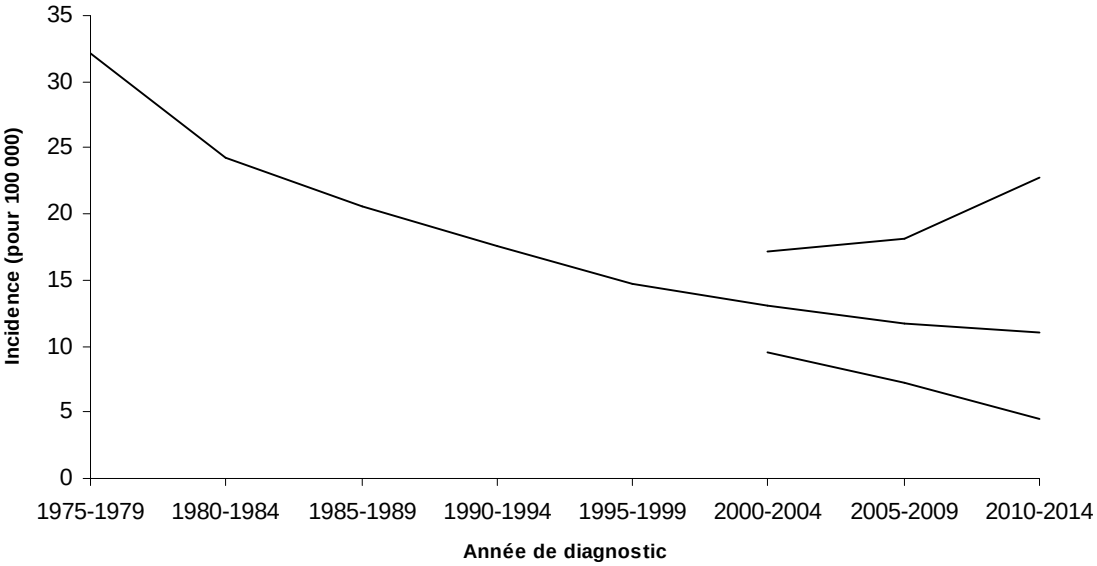


Figure 4 .13. Cancers invasifs du col de l'utérus chez les femmes âgées de 20 à 89 ans dans le Bas-Rhin. Prédiction de l'incidence (1975-2014) standardisée selon la population européenne et intervalle de prédiction à 95%. Pour les périodes 1975-79 à 1995-99, l'incidence est extraits du registre ; pour les autres périodes, il s'agit de valeurs prédites.

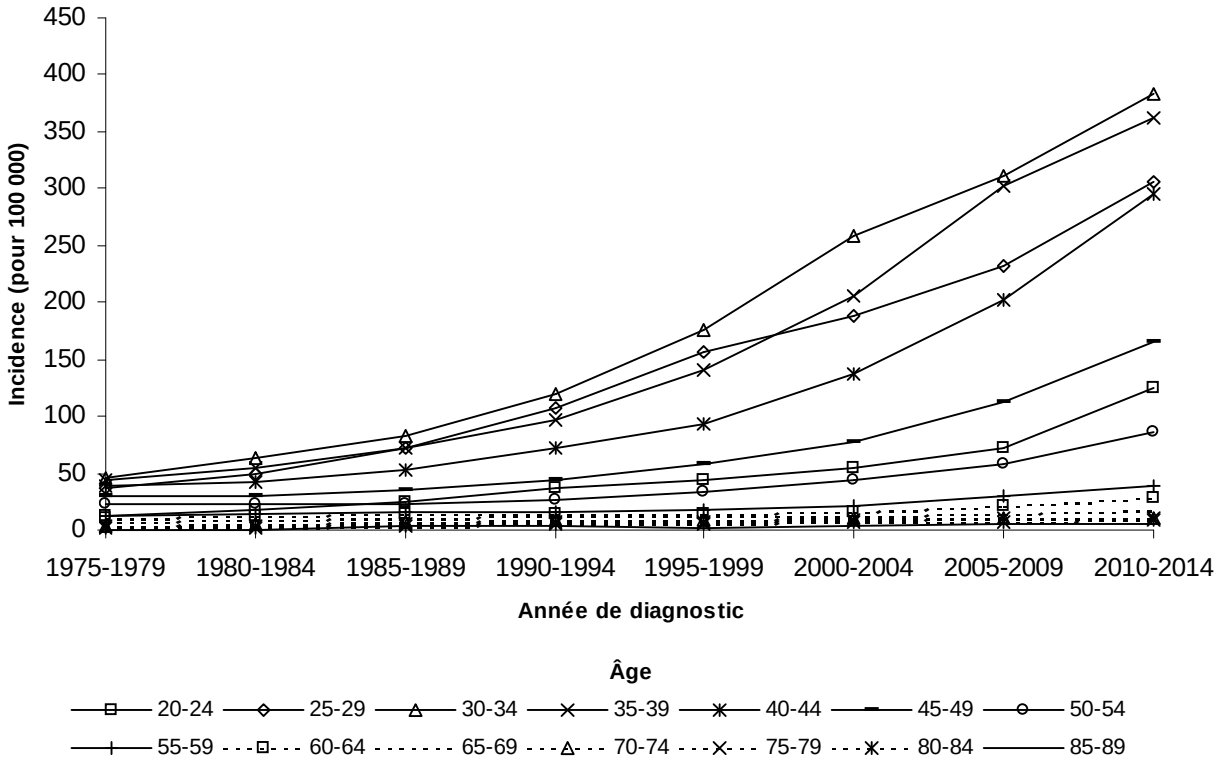


4.2.3.2. Prévision de la tendance de l'incidence au cours de la période 2000-2014.

Les taux d'incidence ont été estimés pour les périodes 2000-2004, 2005-2009 et 2010-2014.
Le modèle choisi pour la prévision des lésions *in situ* est un modèle âge–cohorte ; pour les cancers invasifs, c'est le modèle complet qui a été retenu.

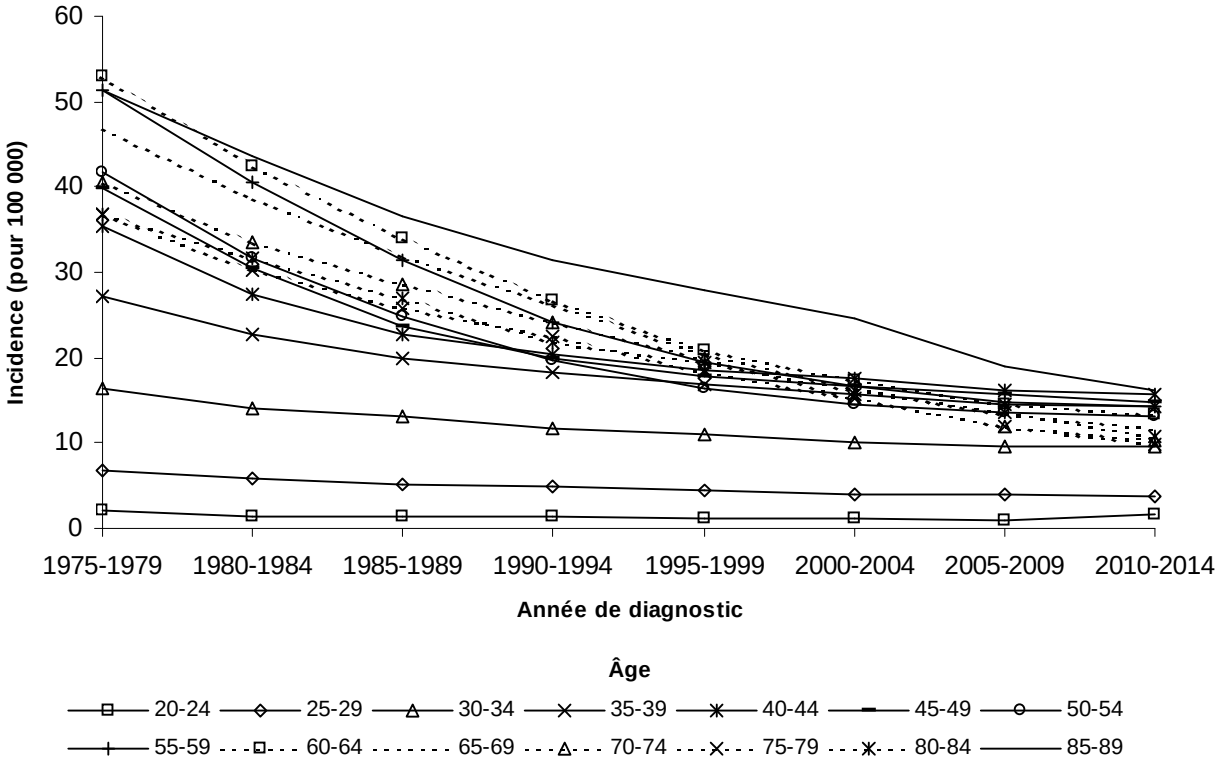
Les taux d'incidence standardisés sur la population européenne sont, en ce qui concerne les lésions *in situ*, respectivement pour les trois périodes de projection, de 99,7 (IC 95% : [82,7-118,5]), 132,7 (IC 95% : [97,2-174,5]) et 177,1 (IC 95% : [103,7-288,5]) pour 100 000 (Tableau 4.11., figure 4.12.). Pour les cancers invasifs, les taux sont respectivement de 13,0 (IC 95% : [9,5-17,2]), de 11,7 (IC 95% : [7,3-18,1]) et de 11,1 (IC 95% : [4,5-22,7]) pour 100 000 (Tableau 4.12., figure 4.13.). L'intervalle de prédiction (intervalle de confiance de la prédiction) augmente au fur et à mesure que l'on allonge le délai de prédiction (figures 4.12. et 4.13.).

Figure 4 .14. Cancers *in situ* du col de l'utérus chez les femmes âgées de 20 à 89 ans dans le Bas-Rhin : incidences spécifiques prédites, 1975-2014.



Les taux spécifiques des lésions *in situ* sont croissants avec la période quel que soit l'âge mais surtout pour les tranches d'âge 30-34 et 35-39 ans où ils doublent en l'espace de dix ans (Tableau 4.11., figure 4.14.). La croissance du taux est négligeable pour les classes d'âge supérieures à 55-59 ans. En ce qui concerne les cancers invasifs, les taux spécifiques déclinent quel que soit l'âge mais de façon inégale selon la tranche d'âge (Tableau 4.12., figure 4.15.). Cette variation est négligeable en dessous de 30-34 ans. Elle est maximale pour la tranche d'âge 60-64 ans (17,3 pour 100 000 en 200-2004, 13,4 pour 100 000 en 2010-1014). Le contraste entre les deux variétés de tumeurs persiste lorsque la variation de l'incidence est étudiée selon la cohorte de naissance : pour les lésions in situ, les cohortes les plus jeunes présentent les taux les plus élevés, pour les cancers invasifs, au contraire, l'incidence prévue par le modèle est plus faible pour les cohortes les plus récentes.

Figure 4 .15. Cancers invasifs du col de l'utérus chez les femmes âgées de 20 à 89 ans dans le Bas-Rhin : incidences spécifiques prédites, 1975-2014.



Le nombre prévu de nouveaux cas de tumeurs in situ augmente de façon importante entre 2000-2004 (1928 cas) et 2010-2014 (3360 cas) (tableau 4.11. et 4.13.). En ce qui concerne les cancers invasifs, la variation est faible : le nombre de cas incidents passe de 260 en 2000-2004 à 238 en 2010-2014 (tableau 4.12. et 4.14.).

Tableau 4.13. Cancer du col *in situ* : nombre de cas incidents et intervalle de prédiction (95 %) chez les femmes dans le Bas-Rhin, 1995-2014.

Âge\Période	1995-1999	2000-2004	2005-2009	2010-2014
20-24	80 (63 - 97)	101 (52 – 152)	137 (35 - 291)	243 (13 - 914)
25-29	314 (285 - 346)	332 (249 - 417)	417 (210 - 650)	566 (138 - 1213)
30-34	351 (323 - 382)	515 (453 - 589)	544 (404 - 691)	683 (338 - 1054)
35-39	275 (252 - 299)	409 (367 - 458)	599 (519 - 697)	634 (470 - 809)
40-44	177 (159 - 196)	266 (235 - 299)	397 (345 - 456)	581 (493 - 686)
45-49	101 (88 - 114)	144 (123 - 166)	216 (184 - 251)	322 (273 - 380)
50-54	42 (35 - 50)	75 (62 – 90)	107 (87 - 130)	162 (131 - 196)
55-59	20 (16 - 26)	27 (21 – 35)	49 (36 - 62)	69 (51 - 89)
60-64	17 (13 - 22)	19 (14 – 25)	26 (19 - 33)	46 (34 - 59)
65-69	15 (11 - 19)	15 (11 – 20)	16 (11 - 23)	22 (16 - 30)
70-74	10 (7 - 14)	11 (8 - 15)	11 (7 - 16)	12 (8 - 18)
75-79	6 (4 - 9)	9 (6 - 13)	10 (6 - 15)	10 (6 - 15)
80-84	3 (2 - 5)	4 (2 - 7)	6 (3 - 10)	7 (3 - 12)
85-89	1 (1 - 3)	1 (0 - 3)	2 (1 - 5)	3 (1 - 8)
20-89	1412 (1259 - 1582)	1928 (1603 - 2289)	2537 (1867 - 3330)	3360 (1975 - 5483)

Les valeurs de la période 1995-1999 (colonne grisée) ont été estimées par extrapolation (voir texte) et servent de base à la prédiction. Les valeurs des autres périodes (2000 à 2014) ont été prédites par le modèle.

Tableau 4.14. Cancer du col invasif : nombre de cas incidents et intervalle de prédiction (95 %) chez les femmes dans le Bas-Rhin, 1995-2014.

Âge\Période	1995-1999	2000-2004	2005-2009	2010-2014
20-24	2 (1 - 4)	2 (1 - 4)	2 (1 - 6)	3 (0 - 9)
25-29	9 (6 - 12)	7 (4 – 12)	7 (3 - 14)	7 (2 - 20)
30-34	22 (17 - 28)	20 (13 - 29)	17 (9 - 29)	17 (5 - 40)
35-39	33 (27 - 40)	31 (22 - 41)	29 (17 - 47)	25 (9 - 53)
40-44	35 (29 - 42)	34 (25 - 43)	32 (19 - 48)	31 (13 - 60)
45-49	31 (26 - 37)	31 (24 - 39)	30 (19 - 44)	29 (12 - 58)
50-54	21 (17 - 26)	25 (19 - 32)	25 (17 - 36)	25 (11 - 49)
55-59	23 (19 - 28)	21 (16 - 27)	25 (16 - 36)	26 (12 - 49)
60-64	25 (21 - 30)	20 (15 - 26)	18 (12 - 26)	22 (10 - 43)
65-69	24 (20 - 29)	19 (14 - 24)	15 (10 - 22)	14 (6 - 27)
70-74	22 (17 - 26)	17 (13 - 22)	13 (9 - 20)	11 (5 - 21)
75-79	14 (11 - 18)	15 (11 - 19)	12 (8 - 18)	10 (4 - 19)
80-84	10 (8 - 13)	10 (7 - 14)	11 (7 - 17)	9 (4 - 18)
85-89	11 (8 - 15)	8 (5 - 11)	8 (5 - 14)	9 (4 - 19)
Total	282 (227 - 348)	260 (189 - 343)	244 (152 - 377)	238 (97 - 485)

Les valeurs de la période 1995-1999 (colonne grisée) ont été estimées par extrapolation (voir texte) et servent de base à la prédiction. Les valeurs des autres périodes (2000 à 2014) ont été prédites par le modèle.

4.2.4. Discussion

Les projections basées sur le modèle âge-période-cohorte montrent une augmentation de près de 80 % de l'incidence des lésions in situ et une diminution de 15 % de l'incidence des cancers invasifs en dix ans.

Ces résultats ont été obtenus par application d'un modèle âge-période-cohorte autorégressif à un ensemble de données d'incidence extraites d'un registre et à un ensemble d'effectifs de population estimés ou prédits à partir de trois recensements. Les données du registre ont l'intérêt d'être fiables et exhaustives. Les méthodes de calculs utilisées pour l'estimation et la prévision des effectifs de population sont de plus en plus performantes. Enfin, le modèle autorégressif bayésien a plusieurs avantages : il permet de tenir compte de l'autocorrélation présente dans la série des incidences (le nombre de cas incidents survenus au cours d'une période donnée n'est pas indépendant du nombres de

cas survenu lors des périodes précédentes) ; il réduit l'instabilité du modèle due au faible nombre de données dont disposent les cohortes extrêmes ; il permet, enfin, de s'affranchir d'un ensemble de difficultés liées aux calculs d'intégration numérique (Breslow, 1993).

Le Bas-Rhin bénéficie, depuis 1994, d'une campagne de dépistage du cancer du col de l'utérus s'adressant aux femmes âgées de 25 à 64 ans (campagne EVE). Il arrive fréquemment que le début d'un dépistage fasse apparaître un certain nombre de cas incidents de façon anticipée, augmentant ainsi brutalement l'incidence enregistrée. La qualité de la prévision peut s'en trouver alors relativement perturbée. En fait, l'observation de l'ensemble des graphes représentatifs des variations des incidences spécifiques année par année (et plus particulièrement pour les âges concernés par le dépistage) n'a pas montré ici de modification brutale de la tendance en 1994 ou au cours des périodes voisines. De plus la base de la prédiction est constituée de données regroupées par périodes de cinq ans. Or, l'année 1994 et les deux ou trois années suivantes susceptibles de perturber l'incidence par la révélation anticipée de cas de cancer n'appartiennent pas à la même période de cinq ans. Ce découpage opère comme un lissage et amortit les accidents éventuels qui auraient pu affecter la tendance. Enfin, un dépistage spontané préexistait au dépistage organisé lorsque ce dernier a été mis en place, atténuant ainsi l'effet de discontinuité attendu.

Le nombre de travaux consacrés à la prévision de l'incidence du cancer du col est relativement faible. Hristova et al. (1997) ont réalisé une projection du taux d'incidence en Bulgarie pour la période 1993-2017 : les incidences spécifiques augmentent chez les femmes de 40 à 69 ans et restent constantes chez les femmes plus âgées. L'incidence standardisée selon la population mondiale (Waterhouse, 1976) augmente et atteint 19,30 pour 100 000 en 2013-2017. Dans le Bas-Rhin, les incidences du cancer du col standardisées selon la population mondiale sont du même ordre bien que variant en sens inverse (12,1 pour 100 000 en 2000-2004, 10,9 pour 100 000 en 2005-2009 et 10,4 pour 100 000 en 2010-2014). Sigurdsson et al. (1991) ont prédit une décroissance de l'incidence du cancer du col en Islande (standardisée selon la population mondiale) de 10,0 à 7,0 pour 100 000 au cours de la période 1990-2000 à partir de l'estimation de la tendance durant la période 1955-1989. Cette projection tenait compte de la distribution des incidences selon le grade, du suivi et de la fréquentation du dépistage. Dans les pays scandinaves, les projections montrent également une diminution de l'incidence (Engeland, 1993). Cette décroissance s'effectue de façon parallèle dans tous les pays concernés. Les taux maximaux sont retrouvés au Danemark : le taux standardisé selon la population européenne passe de 19.8 pour 100 000 en 1983-1987 à 12.7 pour 100 000 en 2008-2012. Les taux minimaux sont retrouvés en Finlande : le taux standardisé selon la population européenne passe de 5.6 pour 100 000 en 1983-1987 à 2,5 pour 100 000 en 2008-2012. Les taux d'incidence du cancer invasif dans le Bas-Rhin sont, par conséquent, relativement proches des taux prévus dans le reste de l'Europe, mise à part la Finlande. Ils présentent, de plus, le même type d'évolution dans le temps.

Les variations de sens contraires, propres aux deux variétés de lésions (augmentation de l'incidence des tumeurs in situ et diminution de l'incidence des cancers), trouvent en général leur justification dans l'existence d'un dépistage organisé. Ce dernier permettrait de découvrir des tumeurs à un stade précoce (lésions in situ) (Sigursson, 1991) et la diminution des cancers pourrait être attribuée au traitement de ces lésions pré-invasives (Réseau Francim, 1998 ; Ciatto, 1995 ; Weidmann, 1998 ; Sigursson, 1991). Or, dans le Bas-Rhin, les tendances respectives des deux types de tumeurs préexistent au dépistage organisé. Il est probable que le frottis vaginal, déjà largement utilisé à titre individuel avant la mise en place de la campagne de dépistage EVE, ait déjà produit l'effet constaté. Il est fort probable également que, comme ailleurs, la diminution de la fréquence du cancer soit due en partie aussi à l'amélioration des pratiques sexuelles (Réseau Francim, 1998). Il peut-être intéressant, cependant, de reconsidérer l'effet du dépistage à la lueur de la différence constatée précédemment entre le Bas-Rhin et la Finlande. Dans ce pays où il existe un dépistage organisé depuis les années soixante, le taux d'incidence est l'un des plus faibles du monde. Les qualités respectives du dépistage et de la prise en charge des lésions in situ expliquent très probablement cette différence.
Il sera intéressant de comparer dans l'avenir les incidences prédites dans cette étude et les incidences futures réellement enregistrées. Les éventuelles différences constatées pourraient être dues au dépistage. En tout état de cause, même si l'incidence ne devait pas être modifiée, l'utilité du dépistage n'est aucunement remise en cause en raison de son effet certain sur la diminution de la mortalité par cancer du col de l'utérus.

4.3. Cancer colorectal

4.3.1. Introduction

Le cancer colorectal se situe au troisième rang des cancers chez l'homme (après la prostate et le poumon) et au deuxième rang des cancers chez la femme (après le sein) pour l'incidence comme pour la mortalité en France. Il devient plus fréquent après 50 ans. Ce cancer a atteint 33 405 personnes en France en 1995 (Réseau Francim, 1998). Dans le Bas-Rhin, pour la période 1988-1992, son incidence standardisée selon la population européenne de référence est de 74,1 pour 10^5 hommes et de 41,2 pour 10^5 femmes (ce qui place ce département, parmi ceux qui disposent d'un registre, en seconde place après le Haut-Rhin) (Réseau Francim, 1998). En France, le taux d'incidence est croissant (l'incidence standardisée selon la population européenne de référence est passée de 53,2 pour 10^5 chez l'homme et

de 35,3 pour 10^5 chez la femme en 1975 à 62,3 pour 10^5 chez l'homme et à 37,4 pour 10^5 chez la femme en 1995) (Réseau Francim, 1998). De même, en Europe (pays de l'UE, d'Europe de l'Est et pays scandinaves), l'incidence est en augmentation régulière comme l'atteste le travail mené en 1996 par le Centre International de Recherche sur le Cancer (Coleman, 1993) ou l'étude de Pollock et al. en 1995 (Pollock, 1995). Pour les USA, l'accroissement du taux d'incidence est moins net et dépend des études (Coleman, 1993 ; Beard, 1995 ; Wingo, 1998 ; Berman, 1993 ; Garfinkel, 1994 b) ; l'incidence peut même être stable comme en Alaska (Bowerman, 1998) ; au Japon, le taux d'incidence est croissant (Koyama, 1997).

Cette étude s'est fixé comme objectif de prédire, pour les femmes et les hommes séparément, l'évolution de l'incidence du cancer du côlon et du rectum dans le Bas-Rhin jusqu'en 2009.

4.3.2. Matériel et méthodes

4.3.2.1. Données

La période de disponibilité des données d'incidence était de 20 ans : de 1975 à 1994.
Le nombre de cas incidents de cancer colique et de cancer rectal, par âge et par année, est fourni par le registre des tumeurs du Bas-Rhin.
La population étudiée était constituée des femmes et des hommes de 25 à 89 ans.

4.3.2.2. Analyse

La prédiction est réalisée, séparément pour le cancer du côlon et le cancer du rectum, pour les périodes 1995-1999, 2000-2004 et 2005-2009 à partir des périodes 1975-1979, 1980-1984, 1985-1989 et 1990-1994

La prévision se base sur un modèle âge-période-cohorte et sur la mise en évidence des effets de ces trois facteurs à partir d'un tableau de données (nombre de cas incidents et nombre de personnes-années) réparties selon l'âge et la période (par tranches de 5 ans) et la cohorte.

Les incidences, quand elles concernaient l'ensemble des classes d'âge, ont été exprimées sous forme de taux standardisés selon la population mondiale.

4.3.3. Résultats

4.3.3.1. Analyse descriptive de la période 1975-1994 (base de la prédiction)

L'incidence du cancer colo-rectal augmente avec l'âge. Au cours de la période 1990-1994, elle est quasiment nulle dans les deux sexes pour la tranche d'âge de 25 à 29 ans et pour les deux localisations (Tableaux 4.15., 4.16., 4.17., et 4.18., figures 4.16, 4.17., 4.18., et 4.19.). Pour la tranche d'âge 85-89 ans, l'incidence du cancer du côlon est égale à 229,2 pour 100 000 chez les femmes et 468,0 pour 100 000 chez les hommes (Tableaux 4.15. et 4.16.). Pour la même tranche d'âge, l'incidence du cancer du rectum est égale à 106,0 pour 100 000 chez les femmes et 200,6 pour 100 000 chez les hommes (Tableaux 4.17., et 4.18.).

Tableau 4.15. Cancer du côlon : incidences spécifiques et nombre de cas incidents chez les femmes dans le Bas-Rhin, 1975-2009.

Âge\Période	1975-1979	1980-1984	1985-1989	1990-1994	1995-1999	2000-2004	2005-2009
25-29	1,7 (3)	0,5 (1)	0,5 (1)	0,0 (0)	1,0 (2)	1,1 (2)	1,1 (2)
30-34	1,5 (2)	1,1 (2)	3,1 (6)	2,0 (4)	2,0 (4)	2,5 (5)	2,9 (5)
35-39	5,6 (7)	3,8 (5)	4,0 (7)	4,7 (9)	4,6 (9)	5,0 (10)	5,5 (11)
40-44	8,5 (11)	5,6 (7)	12,1 (16)	8,5 (15)	9,5 (18)	10,3 (20)	11,2 (22)
45-49	19,5 (26)	17,2 (22)	12,3 (15)	18,3 (24)	18,5 (32)	19,8 (37)	21,4 (41)
50-54	27,3 (36)	23,2 (30)	38,3 (48)	34,6 (42)	32,8 (42)	34,6 (59)	37,5 (69)
55-59	48,4 (50)	42,1 (54)	44,1 (56)	53,3 (66)	51,5 (61)	53,9 (68)	58,3 (98)
60-64	58,6 (53)	64,8 (64)	68,1 (84)	64,0 (79)	73,6 (88)	77,0 (89)	83,5 (103)
65-69	76,1 (86)	91,6 (78)	85,2 (80)	90,3 (107)	99,5 (117)	106,5 (123)	115,9 (130)
70-74	80,3 (81)	99,6 (101)	145,6 (112)	125,9 (110)	135,2 (149)	143,5 (159)	157,8 (173)
75-79	122,5 (90)	156,0 (129)	147,9 (126)	170,9 (114)	180,6 (139)	190,6 (188)	208,8 (210)
80-84	110,5 (45)	146,4 (74)	163,6 (97)	187,8 (121)	211,9 (109)	230,1 (143)	251,6 (203)
85-89	141,0 (22)	130,5 (27)	190,5 (52)	229,2 (80)	242,4 (96)	273,3 (89)	304,0 (128)
25-89	27,4 (512)	28,5 (594)	32,8 (700)	33,6 (771)	35,7 (866)	38,1 (992)	41,5 (1195)

L'incidence est exprimée en nombre de cas pour 100 000. Entre parenthèses, figurent les nombres de cas incidents. Dans la dernière ligne figurent les incidences pour toute la population, standardisées selon la population mondiale et, entre parenthèses, le nombre total de cas incidents. Pour les périodes 1975-79 à 1990-94, les incidences et les nombres de cas incidents sont extraits du registre ; pour les autres périodes, il s'agit de valeurs prédites.

Figure 4.16. Cancer du côlon : incidences spécifiques chez les femmes de 25 à 89 ans dans le Bas-Rhin, 1975-1994 (données du registre).

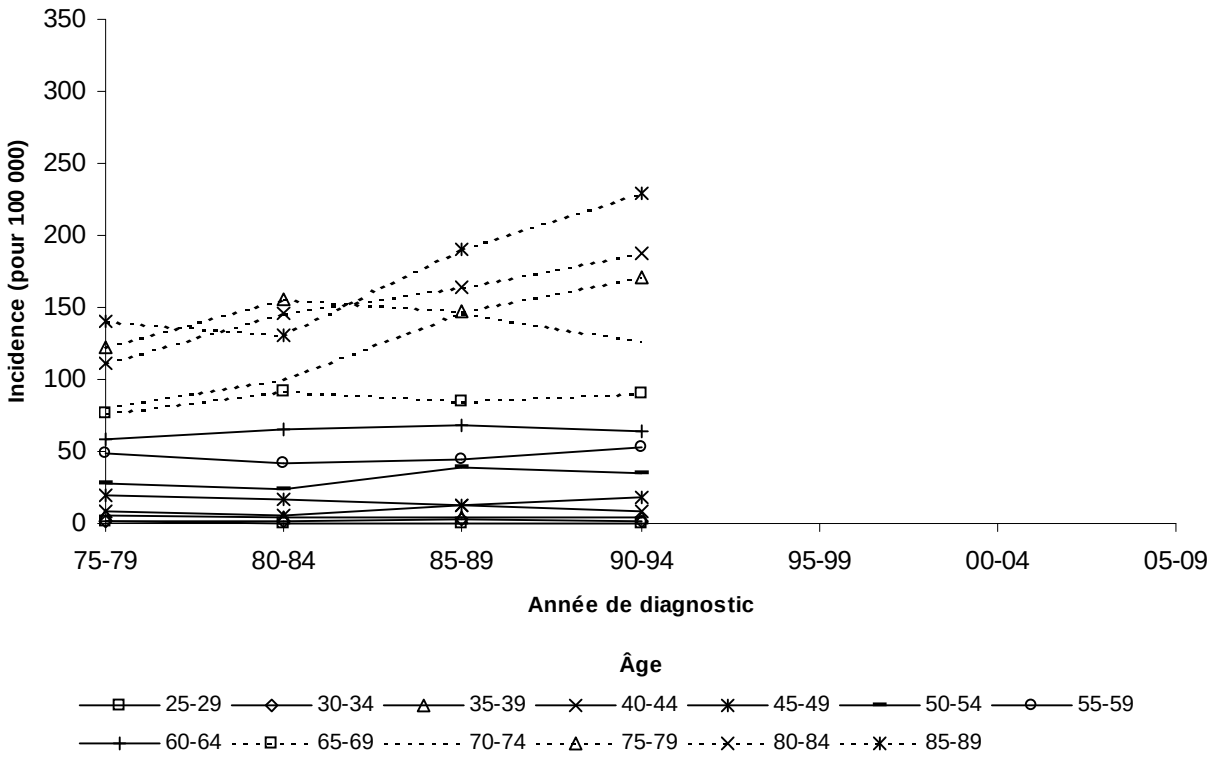


Tableau 4.16. Cancer du côlon : incidences spécifiques et nombre de cas incidents chez les hommes dans le Bas-Rhin, 1975-2009.

Âge\Période	1975-1979	1980-1984	1985-1989	1990-1994	1995-1999	2000-2004	2005-2009
25-29	1,0 (2)	1,0 (2)	2,0 (4)	1,0 (2)	1,5 (3)	1,7 (3)	2,2 (4)
30-34	2,0 (3)	1,6 (3)	1,5 (3)	2,5 (5)	3,0 (6)	3,1 (6)	4,0 (7)
35-39	1,4 (2)	2,8 (4)	2,2 (4)	4,7 (9)	5,8 (11)	6,7 (13)	8,4 (16)
40-44	5,0 (7)	11,8 (16)	11,5 (16)	12,0 (22)	12,3 (23)	14,5 (27)	18,0 (34)
45-49	14,7 (20)	11,1 (15)	13,0 (17)	25,6 (35)	25,2 (45)	29,6 (54)	36,3 (66)
50-54	38,2 (42)	34,8 (45)	34,9 (45)	41,7 (53)	50,2 (67)	59,5 (102)	72,9 (128)
55-59	57,4 (44)	55,9 (57)	62,9 (75)	83,8 (102)	91,6 (109)	108,3 (137)	132,3 (215)
60-64	92,2 (67)	97,4 (67)	113,2 (104)	124 (136)	151,5 (169)	178,5 (197)	219,1 (258)
65-69	105,7 (88)	129,5 (80)	169,2 (102)	218,8 (179)	230,1 (226)	269,9 (271)	330,3 (331)
70-74	165,0 (110)	216,5 (141)	216,4 (107)	260,2 (131)	321,1 (225)	377,3 (318)	459,2 (401)
75-79	174,0 (69)	244,3 (112)	243,1 (112)	305,9 (113)	395,3 (162)	475,3 (262)	578,4 (391)
80-84	220,4 (39)	270,1 (60)	353,1 (94)	399,8 (116)	477,1 (108)	581,5 (161)	725,8 (276)
85-89	124,3 (8)	399,7 (29)	219,8 (20)	468,0 (56)	538,0 (78)	655,8 (76)	833,9 (127)
25-89	37,1 (501)	45,4 (631)	48,7 (703)	62,2 (959)	72,1 (1232)	85,4 (1627)	104,9 (2254)

L'incidence est exprimée en nombre de cas pour 100 000. Entre parenthèses, figurent les nombres de cas incidents. Dans la dernière ligne figurent les incidences pour toute la population, standardisées selon la population mondiale et, entre parenthèses, le nombre total de cas incidents. Pour les périodes 1975-79 à 1990-94, les incidences et les nombres de cas incidents sont extraits du registre ; pour les autres périodes, il s'agit de valeurs prédites.

Figure 4.17. Cancer du côlon : incidences spécifiques chez les hommes de 25 à 89 ans dans le Bas-Rhin, 1975-1994 (données du registre).

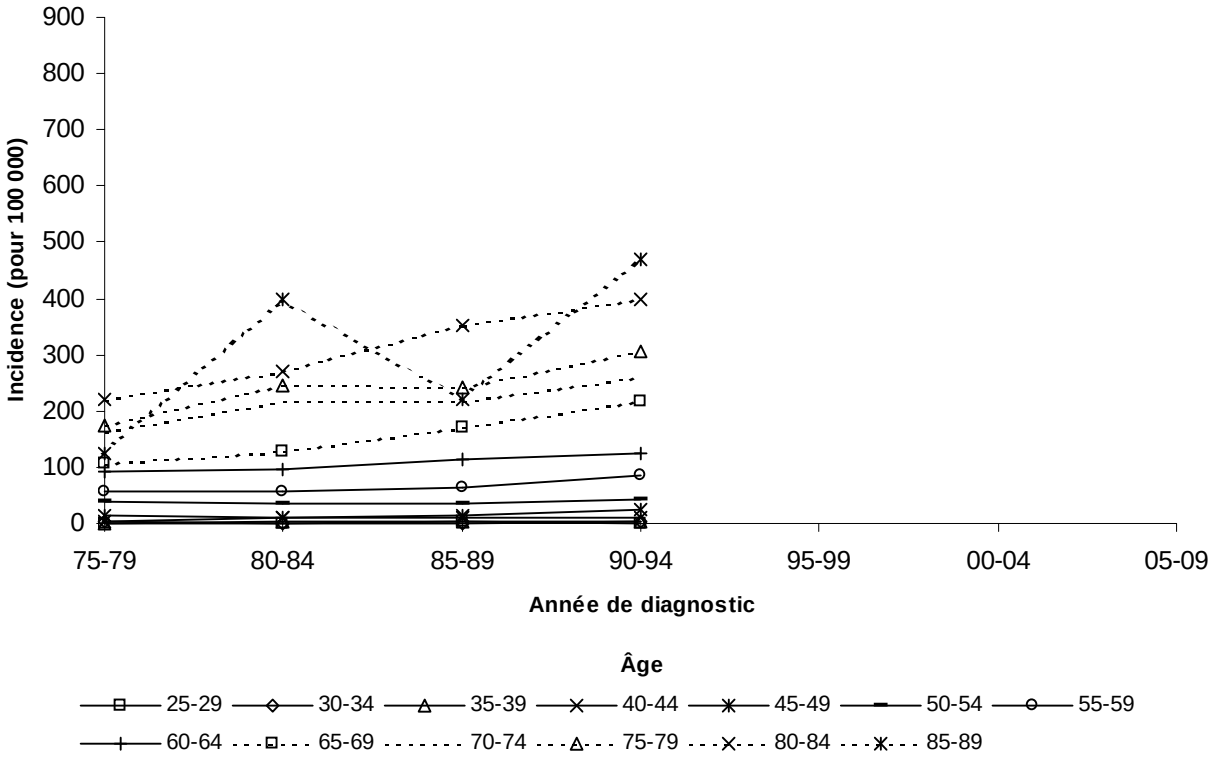


Tableau 4.17. Cancer du rectum : incidences spécifiques et nombre de cas incidents chez les femmes dans le Bas-Rhin, 1975-2009.

Âge\Période	1975-1979	1980-1984	1985-1989	1990-1994	1995-1999	2000-2004	2005-2009
25-29	0,6 (1)	1,1 (2)	0,0 (0)	0,0 (0)	0,5 (1)	0,6 (1)	0,6 (1)
30-34	2,2 (3)	1,1 (2)	1,0 (2)	2,0 (4)	1,5 (3)	1,5 (3)	1,7 (3)
35-39	1,6 (2)	0,8 (1)	2,8 (5)	2,1 (4)	2,6 (5)	2,5 (5)	3,0 (6)
40-44	3,1 (4)	6,4 (8)	3,8 (5)	9,1 (16)	5,3 (10)	5,7 (11)	6,1 (12)
45-49	7,5 (10)	11,7 (15)	10,6 (13)	9,9 (13)	9,8 (17)	10,2 (19)	11,5 (22)
50-54	19,0 (25)	17,0 (22)	19,2 (24)	11,5 (14)	17,9 (23)	18,2 (31)	20,1 (37)
55-59	30,0 (31)	32,7 (42)	28,4 (36)	30,7 (38)	28,7 (34)	30,1 (38)	33,3 (56)
60-64	31,0 (28)	42,5 (42)	43,8 (54)	35,6 (44)	42,6 (51)	44,1 (51)	48,7 (60)
65-69	54,9 (62)	67,0 (57)	58,6 (55)	55,7 (66)	60,4 (71)	62,4 (72)	68,7 (77)
70-74	64,4 (65)	54,2 (55)	79,3 (61)	79,0 (69)	77,1 (85)	79,4 (88)	88,5 (97)
75-79	55,8 (41)	95,6 (79)	68,1 (58)	88,5 (59)	92,2 (71)	96,3 (95)	106,4 (107)
80-84	71,2 (29)	69,2 (35)	86,0 (51)	107,1 (69)	105,0 (54)	111,0 (69)	125,2 (101)
85-89	89,7 (14)	91,8 (19)	80,6 (22)	106,0 (37)	113,6 (45)	125,9 (41)	144,9 (61)
25-89	16,5 (315)	19,2 (379)	18,8 (386)	18,9 (433)	19,8 (470)	20,6 (524)	22,9 (640)

L'incidence est exprimée en nombre de cas pour 100 000. Entre parenthèses, figurent les nombres de cas incidents. Dans la dernière ligne figurent les incidences pour toute la population, standardisées selon la population mondiale et, entre parenthèses, le nombre total de cas incidents. Pour les périodes 1975-79 à 1990-94, les incidences et les nombres de cas incidents sont extraits du registre ; pour les autres périodes, il s'agit de valeurs prédites.

Figure 4.18. Cancer du rectum : incidences spécifiques chez les femmes de 25 à 89 ans dans le Bas-Rhin, 1975-1994 (données du registre).

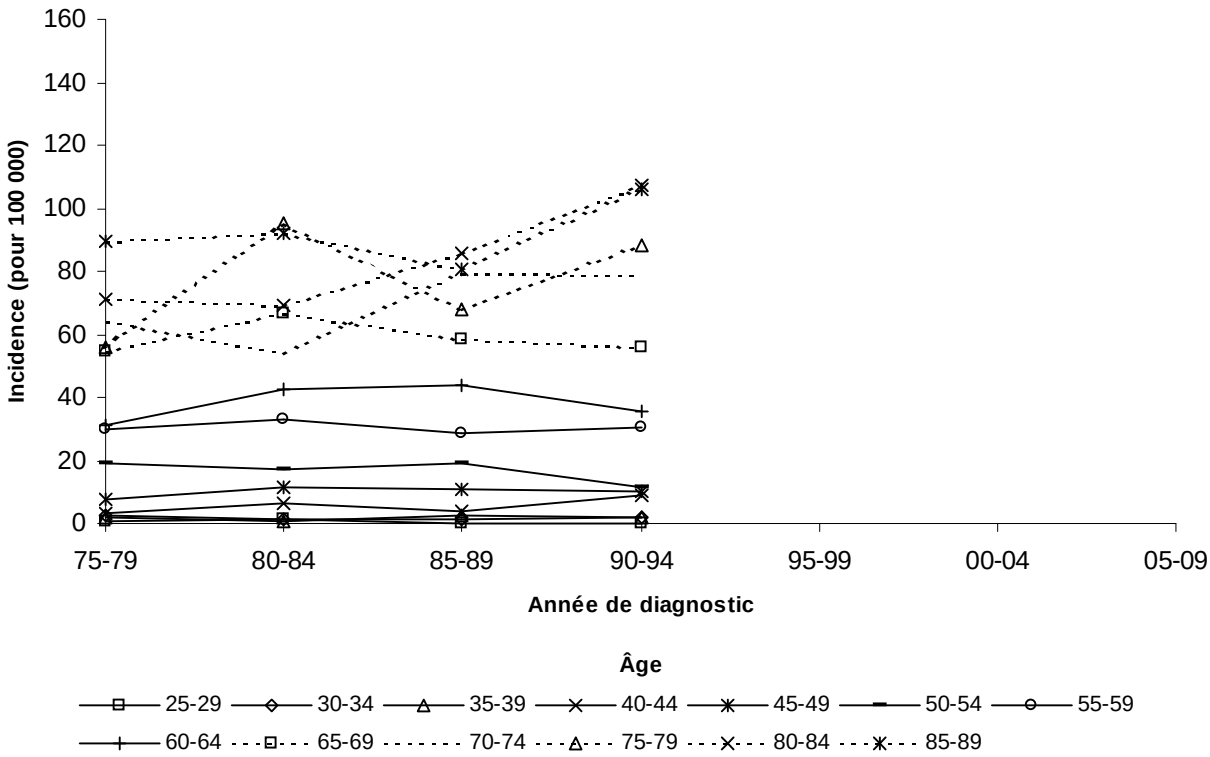
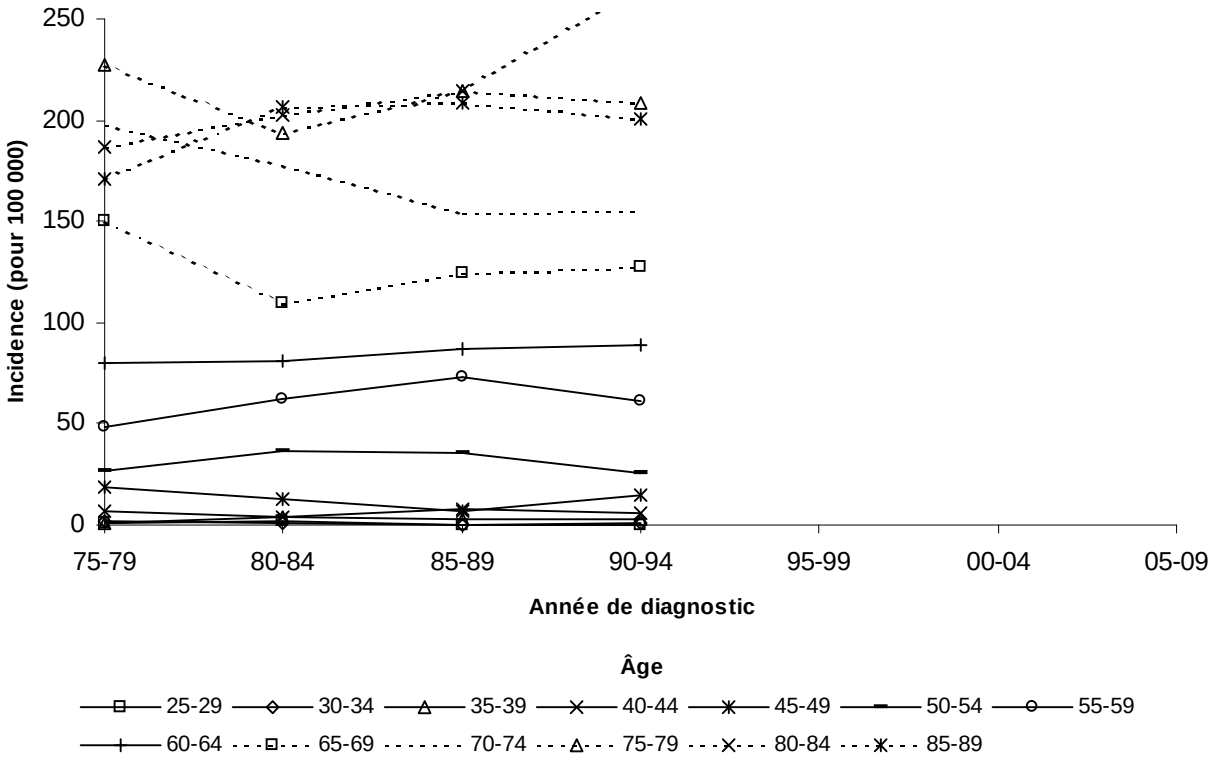


Tableau 4.18. Cancer du rectum : incidences spécifiques et nombre de cas incidents chez les hommes dans le Bas-Rhin, 1975-2009.

Âge\Période	1975-1979	1980-1984	1985-1989	1990-1994	1995-1999	2000-2004	2005-2009
25-29	0,5 (1)	1,5 (3)	0,0 (0)	0,0 (0)	0,5 (1)	0,6 (1)	0,6 (1)
30-34	2,0 (3)	0,5 (1)	0,0 (0)	0,5 (1)	1,0 (2)	0,5 (1)	0,6 (1)
35-39	1,4 (2)	3,5 (5)	2,7 (5)	3,1 (6)	2,1 (4)	1,5 (3)	1,6 (3)
40-44	7,1 (10)	4,4 (6)	7,9 (11)	5,5 (10)	4,8 (9)	4,3 (8)	4,2 (8)
45-49	18,3 (25)	13,3 (18)	6,9 (9)	14,6 (20)	11,8 (21)	11,0 (20)	9,9 (18)
50-54	26,4 (29)	36,4 (47)	35,6 (46)	26,0 (33)	27 (36)	25,1 (43)	23,9 (42)
55-59	48,3 (37)	62,7 (64)	72,9 (87)	60,8 (74)	53,8 (64)	50,6 (64)	48,6 (79)
60-64	79,8 (58)	81,4 (56)	87,1 (80)	89,4 (98)	88,7 (99)	82,4 (91)	80,7 (95)
65-69	150,2 (125)	110,1 (68)	124,4 (75)	127,1 (104)	133,4 (131)	129,5 (130)	125,8 (126)
70-74	198,0 (132)	178,1 (116)	153,7 (76)	154,9 (78)	172,7 (121)	179,2 (151)	180,9 (158)
75-79	226,9 (90)	194,1 (89)	214,9 (99)	208,4 (77)	202,5 (83)	210,4 (116)	226,3 (153)
80-84	186,5 (33)	202,6 (45)	214,1 (57)	265,4 (77)	216,4 (49)	216,7 (60)	234,0 (89)
85-89	170,9 (11)	206,8 (15)	208,8 (19)	200,6 (24)	227,6 (33)	215,7 (25)	223,3 (34)
25-89	39,9 (556)	38,2 (533)	39,0 (564)	38,6 (602)	38,4 (653)	37,3 (713)	37,2 (807)

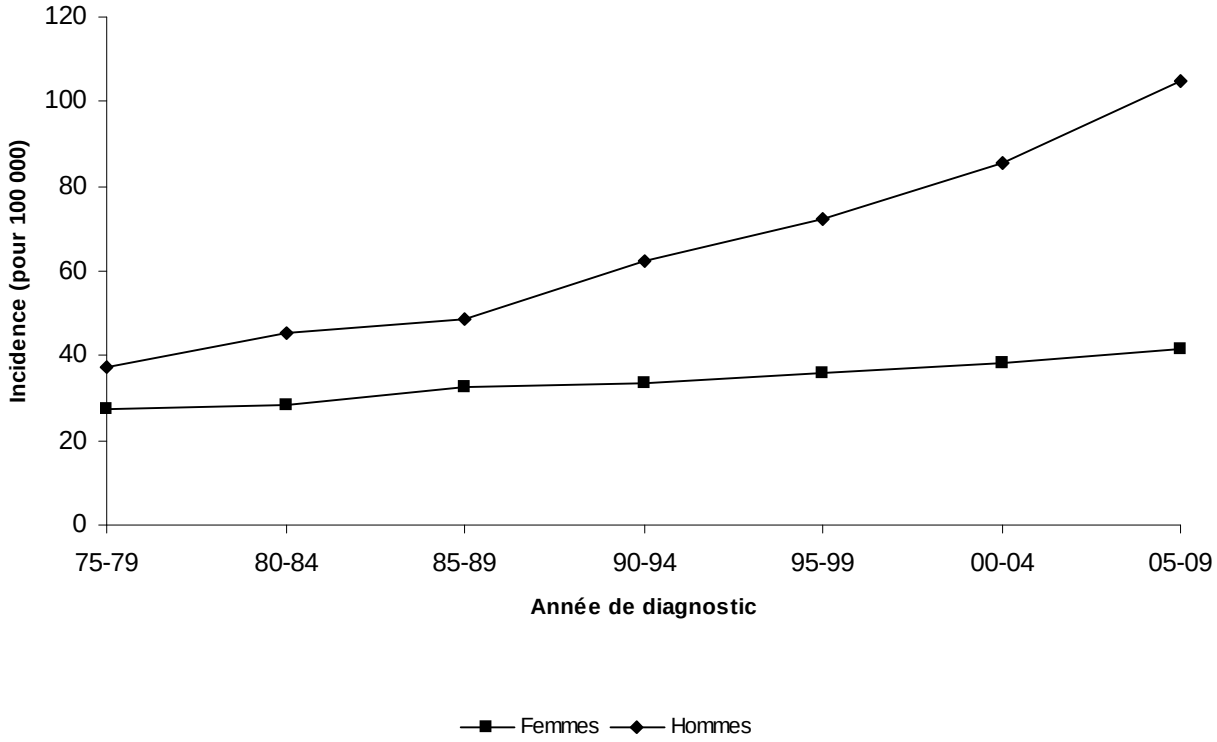
L'incidence est exprimée en nombre de cas pour 100 000. Entre parenthèses, figurent les nombres de cas incidents. Dans la dernière ligne figurent les incidences pour toute la population, standardisées selon la population mondiale et, entre parenthèses, le nombre total de cas incidents. Pour les périodes 1975-79 à 1990-94, les incidences et les nombres de cas incidents sont extraits du registre ; pour les autres périodes, il s'agit de valeurs prédites.

Figure 4.19. Cancer du rectum : incidences spécifiques chez les hommes de 25 à 89 ans dans le Bas-Rhin, 1975-1994 (données du registre).



De 1975 à 1994, l'incidence du cancer du côlon standardisée selon la population mondiale de référence augmente dans les deux sexes mais nettement plus chez les hommes (Figure 4.20.) : les taux correspondants (respectivement en 1975-1979 et 1990-1994) sont de 27,4 et de 33,6 pour 10^5 chez les femmes (Tableau 4.15.) et de 37,1 et de 62,2 pour 10^5 chez les hommes (Tableau 4.16.).

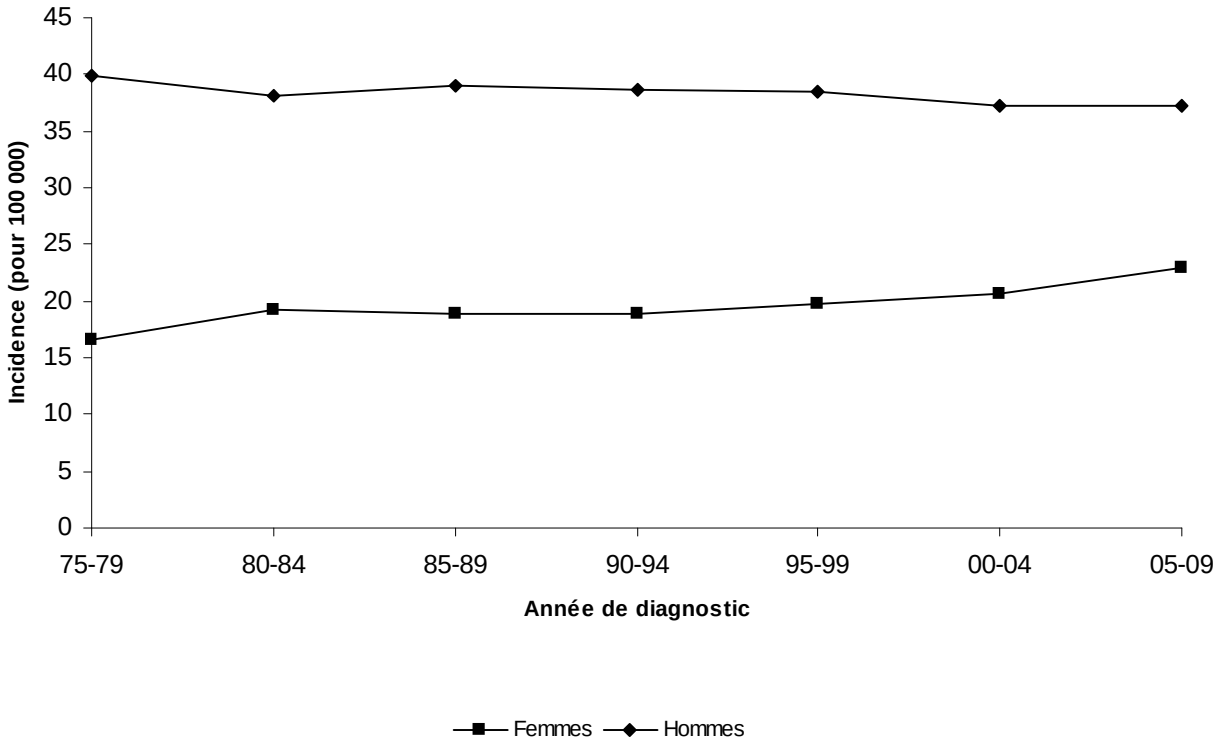
Figure 4.20. Incidence du cancer du côlon chez les femmes et les hommes dans le Bas-Rhin de 1975 à 2009 (standardisation selon la population mondiale).



Les taux sont exprimés pour 100 000 personnes-années. De 1975 à 1994 l'incidence est calculée à partir des données du registre ; elle est prédite pour la période 1995 à 2009.

En ce qui concerne le cancer du rectum, les taux évoluent en sens contraire selon le sexe (Figure 4.21.) : chez les femmes, l'incidence évolue peu (de la période 1975-1979 à la période 1990-1994, l'incidence passe de 16,5 à 18,9 pour 10^5) (Tableau 4.17.), chez les hommes elle est légèrement décroissante (de 39,9 à 38,6 pour 10^5) (Tableau 4.18.).

Figure 4.21. Incidence du cancer du rectum chez les femmes et les hommes dans le Bas-Rhin de 1975 à 2009 (standardisation selon la population mondiale).



Les taux sont exprimés pour 100 000 personnes-années. De 1975 à 1994 l'incidence est calculée à partir des données du registre ; elle est prédite pour la période 1995 à 2009.

4.3.3.2. Analyse âge-période-cohorte et prédictions

Les taux d'incidence standardisés selon la population mondiale ont été estimés pour les périodes 1995-1999, 2000-2004 et 2005-2009. Le modèle prévoit, pour ces trois périodes, des taux d'incidence du cancer du côlon de 35,7, 38,1 et 41,5 pour 10^5 chez les femmes, et de 72,1, 85,4 et 104,9 pour 10^5 chez les hommes (Tableaux 4.15. et 4.16., figure 4.20.). Quant au cancer du rectum, chez les femmes, les taux estimés sont respectivement de 19,8, de 20,6 et de 22,9 pour 10^5 et chez les hommes de 38,4 puis 37,3 et 37,2 pour 10^5 (Tableaux 4.17. et 4.18., figure 4.21.).

Le nombre de cas incidents correspondant est croissant quel que soit le sexe et le type de cancer (Tableaux 4.19, 4.20., 4.21. et 4.22.).

Tableau 4.19. Cancer du côlon : nombre de cas incidents et intervalle de prédiction (95 %) chez les femmes dans le Bas-Rhin, 1995-2009.

Âge\période	1995-1999	2000-2004	2005-2009
25-29	2 (1 - 4)	2 (1 - 4)	2 (0 – 7)
30-34	4 (2 - 7)	5 (2 - 8)	5 (1 - 11)
35-39	9 (6 - 13)	10 (5 - 16)	11 (3 - 23)
40-44	18 (12 - 26)	20 (11 - 32)	22 (8 - 46)
45-49	32 (24 - 43)	37 (22 - 56)	41 (16 - 81)
50-54	42 (33 - 53)	59 (37 - 87)	69 (29 - 134)
55-59	61 (49 - 74)	68 (45 - 96)	98 (44 - 183)
60-64	88 (71 - 105)	89 (60 - 123)	103 (48 - 190)
65-69	117 (97 – 140)	123 (86 - 168)	130 (61 - 239)
70-74	149 (123 – 179)	159 (111 - 220)	173 (84 - 317)
75-79	139 (116 – 167)	188 (132 - 261)	210 (102 - 391)
80-84	109 (91 – 131)	143 (100 - 196)	203 (98 - 370)
85-89	96 (77 - 118)	89 (60 - 123)	128 (60 - 231)
25-89	866 (702 – 1060)	992 (672 - 1390)	1195 (554 - 2223)

Tableau 4.20. Cancer du côlon : nombre de cas incidents et intervalle de prédiction (95 %) chez les hommes dans le Bas-Rhin, 1995-2009.

Âge\période	1995-1999	2000-2004	2005-2009
25-29	3 (1 - 5)	3 (1 - 6)	4 (1 - 12)
30-34	6 (3 - 8)	6 (3 - 12)	7 (2 - 16)
35-39	11 (8 - 15)	13 (8 - 22)	16 (6 - 35)
40-44	23 (17 - 30)	27 (17 - 42)	34 (14 - 70)
45-49	45 (35 - 57)	54 (35 - 81)	66 (29 - 132)
50-54	67 (54 - 85)	102 (68 - 150)	128 (58 - 247)
55-59	109 (89 - 133)	137 (94 - 197)	215 (100 - 408)
60-64	169 (140 - 202)	197 (138 - 278)	258 (122 - 490)
65-69	226 (190 - 270)	271 (192 - 377)	331 (160 - 614)
70-74	225 (188 - 268)	318 (224 - 444)	401 (198 - 738)
75-79	162 (136 - 194)	262 (185 - 367)	391 (188 - 734)
80-84	108 (90 - 129)	161 (112 - 230)	276 (133 - 514)
85-89	78 (62 - 97)	76 (52 - 109)	127 (59 - 240)
25-89	1232 (1013 – 1493)	1627 (1129 - 2315)	2254 (1070 - 4250)

Tableau 4.21. Cancer du rectum : nombre de cas incidents et intervalle de prédiction (95 %) chez les femmes dans le Bas-Rhin, 1995-2009.

Âge\période	1995-1999	2000-2004	2005-2009
25-29	1 (1 - 2)	1 (0 – 2)	1 (0 - 4)
30-34	3 (2 - 4)	3 (1 – 5)	3 (1 - 6)
35-39	5 (4 - 7)	5 (3 – 9)	6 (2 - 12)
40-44	10 (7 - 13)	11 (6 - 16)	12 (4 - 24)
45-49	17 (13 - 23)	19 (12 - 29)	22 (9 - 44)
50-54	23 (17 - 29)	31 (19 - 46)	37 (14 - 74)
55-59	34 (26 - 43)	38 (23 - 56)	56 (21 - 111)
60-64	51 (40 - 64)	51 (32 - 75)	60 (23 - 115)
65-69	71 (57 - 87)	72 (47 - 102)	77 (30 - 145)
70-74	85 (70 - 105)	88 (59 - 127)	97 (39 - 183)
75-79	71 (57 - 85)	95 (64 - 134)	107 (43 - 204)
80-84	54 (43 - 65)	69 (45 - 99)	101 (41 - 193)
85-89	45 (35 - 58)	41 (26 - 59)	61 (25 - 118)
25-89	470 (372 - 585)	524 (337 - 759)	640 (252 - 1233)

Tableau 4.22. Cancer du rectum : nombre de cas incidents et intervalle de prédiction (95 %) chez les femmes dans le Bas-Rhin, 1995-2009.

Âge\période	1995-1999	2000-2004	2005-2009
25-29	1 (0 - 1)	1 (0 - 2)	1 (0 - 2)
30-34	2 (1 - 3)	1 (0 - 3)	1 (0 - 3)
35-39	4 (2 - 6)	3 (1 - 7)	3 (1 - 8)
40-44	9 (5 - 13)	8 (4 - 14)	8 (2 - 17)
45-49	21 (14 - 28)	20 (10 - 31)	18 (6 - 39)
50-54	36 (27 - 46)	43 (25 - 65)	42 (15 - 84)
55-59	64 (50 - 80)	64 (41 - 93)	79 (32 - 157)
60-64	99 (79 - 120)	91 (60 - 128)	95 (40 - 185)
65-69	131 (107 - 161)	130 (89 - 183)	126 (55 - 245)
70-74	121 (98 - 145)	151 (103 - 212)	158 (69 - 306)
75-79	83 (67 - 100)	116 (79 - 161)	153 (68 - 296)
80-84	49 (39 - 60)	60 (40 - 84)	89 (39 - 172)
85-89	33 (24 - 43)	25 (16 - 38)	34 (14 - 66)
25-89	653 (513 - 806)	713 (468 - 1021)	807 (341 - 1580)

Les taux d'incidence spécifiques pour l'âge, entre 1975-1979 et 2005-2009, augmentent pour le cancer du côlon dans les deux sexes ainsi que dans toutes les tranches d'âge et ce, d'autant plus vite que l'âge est important (Tableaux 4.15., 4.16., figures 4.22. et 4.23.) ; pour les femmes, l'augmentation est faible, voire négligeable jusqu'à 44 ans, au delà elle est nette (doublement de l'incidence de 1975-1979 à 2005-2009 pour la tranche d'âge 75-79 ans) ; chez les hommes, la croissance des taux en fonction de la période est beaucoup plus rapide (triplement de l'incidence de 1975-1979 à 2005-2009 pour la tranche d'âge 75-79 ans).

Figure 4.22. Cancer du côlon : incidences spécifiques chez les femmes de 25 à 89 ans dans le Bas-Rhin, 1975-2009 (modèle âge-période-cohorte).

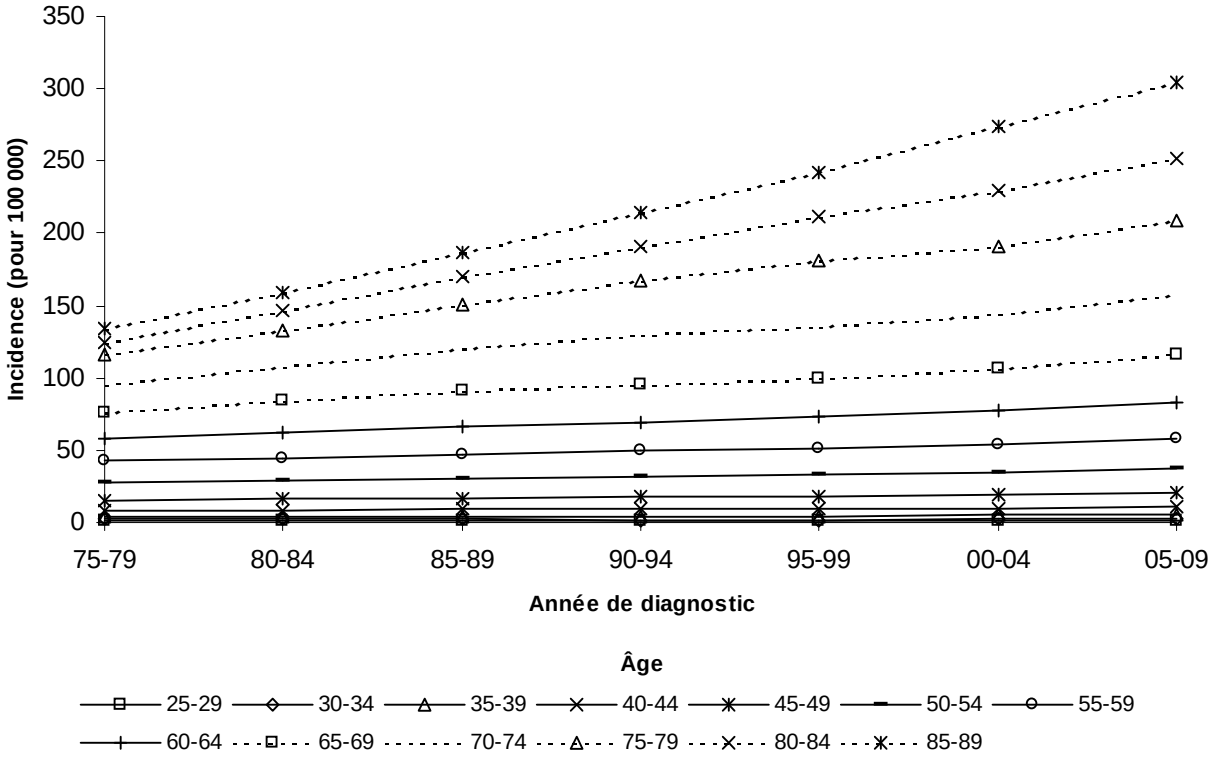
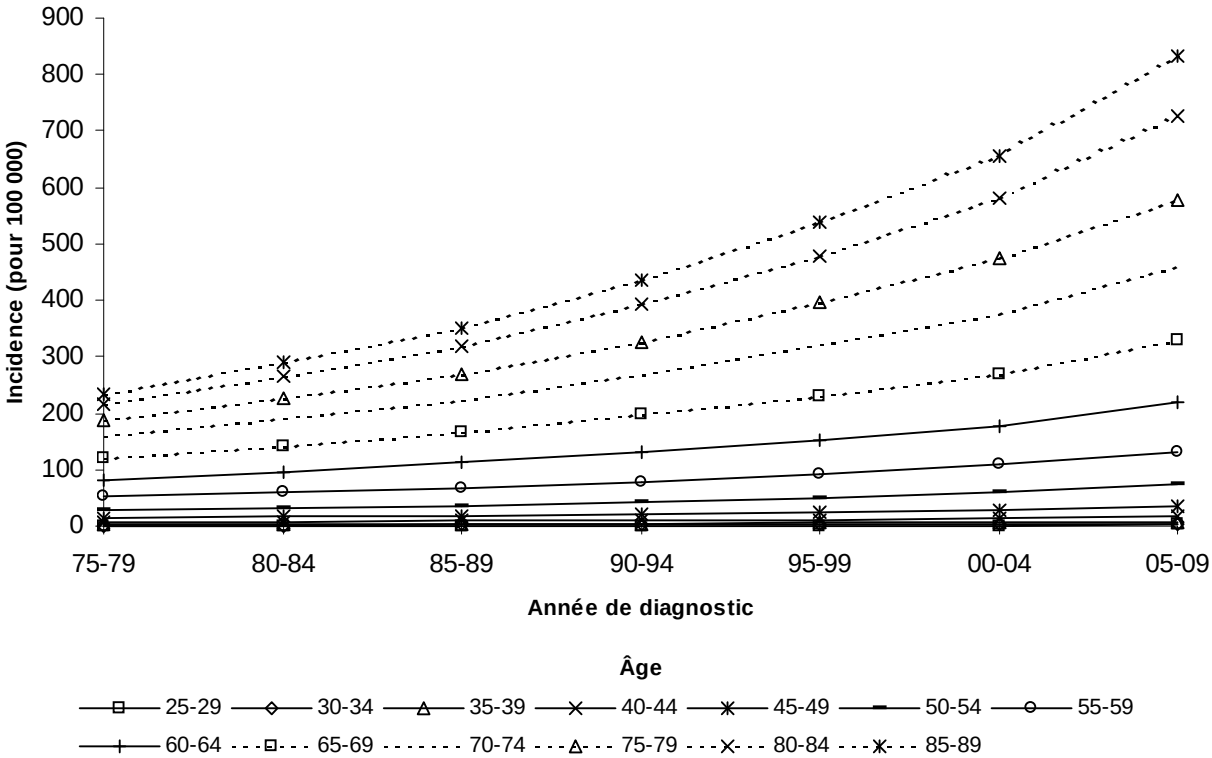
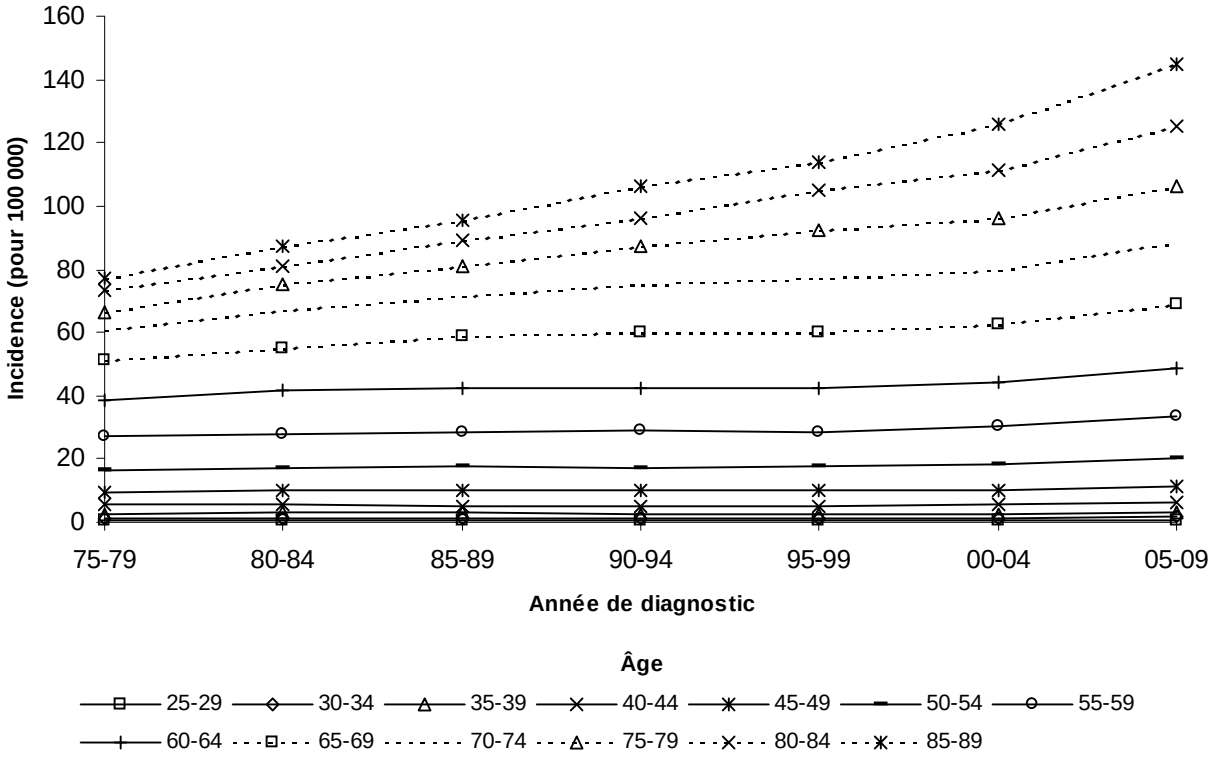


Figure 4.23. Cancer du côlon : incidences spécifiques chez les hommes de 25 à 89 ans dans le Bas-Rhin, 1975-2009 (modèle âge-période-cohorte).



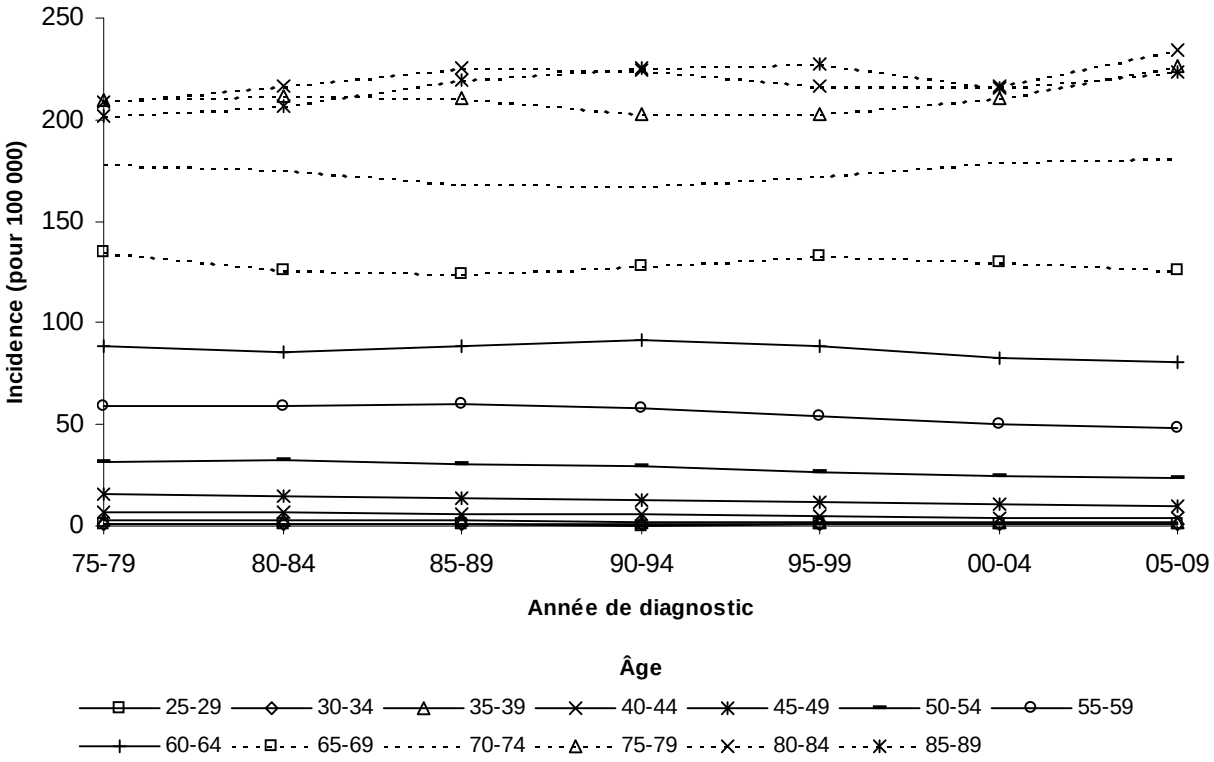
Pour le cancer du rectum, l'évolution de l'incidence est différente selon le sexe. Pour les femmes, l'augmentation du taux d'incidence en fonction du temps est, comme pour le côlon, d'autant plus importante que l'âge est plus avancé (Figure 4.24.) : jusqu'à 65-69 ans l'augmentation est faible, puis elle est franche au delà (l'incidence, dans la tranche d'âge 85-89 ans, double entre 1975-1979 et 2005-2009).

Figure 4.24. Cancer du rectum : incidences spécifiques chez les femmes de 25 à 89 ans dans le Bas-Rhin, 1975-2009 (modèle âge-période-cohorte).



Chez l'homme, le taux d'incidence du cancer du rectum en fonction du temps diminue dans les tranches d'âge comprises entre 25-29 ans et 65-69 ans ; au dessus de 69 ans, l'incidence augmente mais faiblement (Figure 4.25.).

Figure 4.25. Cancer du rectum : incidences spécifiques chez les hommes de 25 à 89 ans dans le Bas-Rhin, 1975-2009 (modèle âge-période-cohorte).



Le nombre estimé de cas incidents est présenté dans les tableaux 4.12., 4.13., 4.14. et 4.15.. Le nombre de cancers du côlon, pour la dernière des périodes prédites (2005-2009), est de 1195 pour les femmes et 2254 pour les hommes (population de 25 à 89 ans). Ce nombre, pour le cancer du rectum est, respectivement, de 640 femmes et 807 hommes.

4.3.4. Discussion

Le modèle utilisé dans cette étude montre un accroissement du taux d'incidence du cancer du côlon en fonction du temps dans les deux sexes ; cet accroissement est important dans les tranches d'âge élevées. Pour le cancer du rectum chez la femme, les prédictions vont dans le même sens ; chez l'homme, par contre, le modèle prévoit une diminution du taux avant 69 ans et au delà un comportement plus irrégulier. La précision de la valeur prédite diminue avec l'éloignement par rapport à la période servant de base à la prédiction : ainsi, par exemple, pour le cancer du côlon, l'amplitude des intervalles de prédiction (à 95%) est, pour les périodes 1995-1999, 2000-20004 et 2005-2009 respectivement, pour les femmes 118, 251 et 651 et pour les hommes 165, 437 et 1320, ce qui représente, relativement à la valeur prédite, respectivement, pour les femmes 14, 25 et 54 % et pour les

hommes 13, 27 et 59 %. La valeur prédite par le modèle pour la dernière période est imprécise et doit donc être considérée comme indicative.

Le nombre de travaux consacrés à la prévision de l'incidence du côlon est relativement faible. Capocaccia et al. (1997) ont réalisé, en Italie, un calcul de prédiction d'incidence au sein d'une population tronquée (25-84 ans) comparable à celle qui est étudiée ici pour la période 1991-2000 à partir de la période 1970-1990 en supposant que les effets âge et cohortes restaient constants. Ils obtiennent, respectivement pour les femmes et les hommes (incidence du cancer colorectal standardisée selon la population mondiale), 18 et 23 pour 10^5 en 1970, 22 et 29 pour 10^5 en 1980, 28 et 40 pour 10^5 en 1990, 33 et 50 pour 10^5 en 2000.

Ici, après sommation des incidences du cancer du côlon et du cancer du rectum puis interpolation (pour rendre les périodes d'étude et les données comparables) on trouve, pour les femmes et les hommes, 46 et 80 pour 10^5 en 1980, 52 et 94 pour 10^5 en 1990, 57 et 117 pour 10^5 en 2000 ; les taux d'incidences dans le Bas-Rhin sont donc deux fois plus élevés que les taux italiens pour les femmes et deux fois et demi plus importantes pour les hommes avec, toutefois un accroissement relatif plus important en Italie ; le rapport entre les taux d'incidence chez l'homme et chez la femme augmente dans les deux zones ; il est plus important dans le Bas-Rhin qu'en Italie (2,1 pour 1,5 en 2000) mais il évolue de façon parallèle dans les deux zones. Engeland et al. (1993) ont prédit les incidences du cancer du côlon et du rectum dans les pays scandinaves pour la période 1998-2012 à partir de la période 1958-1987. Ils utilisent un modèle âge-période-cohorte adapté à chaque pays. Les résultats sont relativement disparates mais l'incidence augmente globalement quoique faiblement. Pour le côlon, dans toute la population, les taux standardisés sur la population mondiale sont, en 2008, compris entre 17 pour 10^5 (Suède) et 25 pour 10^5 (Norvège) pour les hommes et entre 14 pour 10^5 (Finlande) et 22 pour 10^5 (Norvège). Pour le rectum, en 2008, les incidences sont comprises entre 9 pour 10^5 (Finlande) et 21 pour 10^5 (Norvège) chez l'homme et 6 pour 10^5 (Finlande) et 12 pour 10^5 (Norvège). Ces taux ainsi que leur croissance sont donc nettement inférieurs à ceux qui ont été prédits dans le cas présent. Les différences constatées entre les taux projetés dans le Bas-Rhin et les taux projetés dans les deux autres régions sont le reflet des différences observées entre les tendances actuelles des cancers : l'incidence du cancer colorectal est, en effet, nettement plus élevée dans le Bas-Rhin que dans les autres régions (Coleman, 1993). Ces dissemblances sont vraisemblablement en relation avec des disparités d'ordre alimentaire (rôle protecteur des légumes, rôle aggravant des graisses et d'un apport calorique élevé) (Boutron-Ruault, 1998 ; Steenland, 1995). En ce qui concerne l'étude scandinave, les différences sont dues aussi au fait que les tranches d'âges concernées ne sont pas semblables (dans l'étude scandinave, les taux ont été estimés pour la population entière alors qu'ici ils ont été estimés pour une population tronquée).

La différence des tendances des cancers du côlon et du rectum (croissante pour le côlon, faiblement croissante ou stable pour le rectum) est souvent retrouvée (Launois, 1995 ; Loffeld, 1996 ; Benhamiche, 1997 ; Nelson, 1998 ; Obrand, 1998). Cette différence est expliquée par un effet de

l'alimentation, non homogène selon le segment intestinal concerné (Faivre, 1997) et par un mode d'altération génétique de la cellule cancéreuse, différent selon la localisation de la tumeur (Benhamiche, 1998 ; Elsaleh, 2000). Les cancers des côlons droit et gauche pourraient ainsi correspondre à des facteurs étiologiques différents. Enfin, les tumeurs du côlon gauche, plus facilement accessibles à la coloscopie que celles du côlon droit, seraient détectées plus tôt (Lichtman, 1994).

La différence de tendance de l'incidence selon le sexe est attribuée à des facteurs hormonaux (effet protecteur de la contraception orale et de l'apparition tardive des règles) (Dos-Santos-Silva, 1996 ; Martinez, 1997). La susceptibilité des tumeurs intestinales aux effets protecteurs ou aggravants de l'alimentation est différente selon le sexe et serait également impliquée dans ces disparités hommes-femmes (Faivre, 1997).

La croissance inquiétante de l'incidence du cancer du côlon conduit tout naturellement au débat concernant le dépistage organisé (Hémoccult) qui, en découvrant des lésions intestinales peu évoluées, diminuerait la mortalité liée à ce cancer (Tazi, 1999). Les différences observées quant à la fréquence des tumeurs ou à leur sensibilité vis à vis de l'environnement (selon la région, le sexe, le segment intestinal) incitent à développer la connaissance biomoléculaire de ces lésions afin d'adapter leur traitement.

4.4. Cancer du poumon

4.4.1. Introduction

En terme d'incidence, le cancer du poumon est le premier cancer, chez l'homme, dans le monde et le deuxième cancer en France (après le cancer de la prostate). En Europe, c'est l'Écosse qui présente le plus fort taux avec, en 1985, une incidence standardisée selon la population mondiale de 166,9 cas pour 100 000 (personnes-années) chez l'homme et de 63,5 cas pour 100 000 chez la femme (Coleman, 1993).

En France, le taux standardisé selon la population mondiale était, en 1995, de 47,1 pour 10^5 chez l'homme et de 6,4 pour 10^5 chez la femme (Réseau Francim, 1998). La comparaison réalisée par le réseau Francim (1998) entre les différents registres français sur la base des données de 1988 à 1992, place le Bas-Rhin en première position pour l'incidence du cancer du poumon chez l'homme (taux standardisé à l'Europe de 96,6 pour 10^5) et chez la femme (taux standardisé à l'Europe de 10,2

pour 10^5). Ces taux sont nettement plus importants que ceux qui sont attribués à la France (le taux standardisé à l'Europe était, pour la France, en 1995, de 75,2 pour 10^5 chez l'homme et de 7,9 pour 10^5 chez la femme).

Depuis une vingtaine années, cependant, la tendance du taux d'incidence du cancer du poumon amorce une stabilisation, voire une décroissance chez l 'homme. Chez la femme, au contraire, la tendance est croissante, quel que soit le pays (Coleman, 1993). Cette évolution est attribuée à la forte augmentation du tabagisme féminin dans le monde.

En France, comme dans le monde, une stabilisation de la tendance est observée depuis le début des années 80 chez l'homme alors que l'incidence continue à croître chez la femme et plus particulièrement chez les jeunes femmes : ainsi, entre 1985 et 1995, l'accroissement de l'incidence est de 5 % chez l'homme alors qu'il atteint 56 % chez les femmes de moins de 65 ans (Réseau Francim, 1998).

Dans le Bas-Rhin l'incidence augmente de 9,3 % tous les 5 ans chez l'homme et 25,1 % chez la femme (Coleman, 1993).

La tendance de la mortalité suit celle de l'incidence, de façon relativement fidèle chez la femme : les taux sont restés très voisins entre 1975 et 1995 (Réseau Francim, 1998). Chez l'homme, le taux de mortalité continue à croître, contrairement à l'incidence.

La connaissance du développement du cancer du poumon est essentielle pour la politique de santé d'un pays (Aareleid, 1994). L'objectif de l'étude présente est de prédire l'évolution de l'incidence du cancer du poumon, pour les femmes et les hommes séparément, dans le Bas-Rhin, jusqu'en 2009.

4.4.2. Matériel et méthodes

4.4.2.1. Données

La période de disponibilité des données d'incidence s'étend de 1975 à 1994.

Le nombre de cas incidents de cancer du poumon, par âge et par année, est extrait du registre des tumeurs du Bas-Rhin.

La population étudiée est constituée des femmes et des hommes de 25 à 94 ans.

4.4.2.2. Analyse

La prédiction est réalisée pour les périodes 1995-1999, 2000-2004 et 2005-2009 à partir des périodes 1975-1979, 1980-1984, 1985-1989 et 1990-1994.

La prévision se base sur un modèle âge-période-cohorte à partir d'un tableau de données (nombre de cas incidents et nombre de personnes-années) réparties selon l'âge et la période (par tranches de 5 ans) et la cohorte.

Les incidences, quand elles concernaient l'ensemble des classes d'âge, ont été exprimées sous forme de taux standardisés selon la population mondiale.

4.4.3. Résultats

4.4.3.1. Analyse descriptive de la période 1975-1994 (base de la prédiction)

L'incidence du cancer du poumon est proche de zéro dans les deux sexes pour la tranche d'âge de 25 à 29 ans (Tableaux 4.23. et 4.24.). Elle augmente avec l'âge jusqu'à atteindre un maximum pour la tranche d'âge 75-79 ans chez les femmes (avec une incidence de 36,4 pour 100 000 puis elle diminue. Pour les hommes, l'incidence atteint son maximum pour la tranche d'âge de 70 à 75 ans (467,0 pour 100 000) puis décroît.

Tableau 4.23. Cancer du poumon : incidences spécifiques et nombre de cas incidents chez les femmes dans le Bas-Rhin, 1975-2009.

Âge\Période	1975-1979	1980-1984	1985-1989	1990-1994	1995-1999	2000-2004	2005-2009
25-29	0,6 (1)	0,0 (0)	1,0 (2)	0,5 (1)	1,0 (2)	1,1 (2)	1,7 (3)
30-34	0,7 (1)	1,7 (3)	2,1 (4)	0,5 (1)	2,0 (4)	2,0 (4)	2,3 (4)
35-39	1,6 (2)	3,0 (4)	2,3 (4)	3,1 (6)	3,6 (7)	4,0 (8)	4,5 (9)
40-44	3,1 (4)	3,2 (4)	6,8 (9)	4,5 (8)	6,4 (12)	7,2 (14)	8,6 (17)
45-49	6,0 (8)	7,0 (9)	9,8 (12)	9,2 (12)	11,0 (19)	12,9 (24)	15,1 (29)
50-54	8,4 (11)	11,6 (15)	16,0 (20)	13,2 (16)	18,7 (24)	21,7 (37)	25,0 (46)
55-59	10,6 (11)	13,2 (17)	22,8 (29)	25,0 (31)	28,7 (34)	33,3 (42)	38,7 (65)
60-64	24,3 (22)	12,1 (12)	30,8 (38)	34,8 (43)	40,1 (48)	47,6 (55)	56,0 (69)
65-69	23,9 (27)	25,8 (22)	39,4 (37)	40,5 (48)	51,0 (60)	61,5 (71)	72,2 (81)
70-74	24,8 (25)	27,6 (28)	41,6 (32)	42,4 (37)	57,2 (63)	69,5 (77)	83,9 (92)
75-79	42,2 (31)	16,9 (14)	45,8 (39)	42,0 (28)	59,8 (46)	73,0 (72)	89,5 (90)
80-84	9,8 (4)	35,6 (18)	40,5 (24)	48,1 (31)	58,3 (30)	70,8 (44)	86,8 (70)
85-89	0,0 (0)	24,2 (5)	62,3 (17)	34,4 (12)	53,0 (21)	64,5 (21)	80,8 (34)
90-94	26,6 (1)	20,2 (1)	13,9 (1)	18,5 (2)	48,7 (7)	58,5 (10)	76,3 (11)
25-94	8,1 (148)	8,2 (152)	14,0 (268)	13,5 (276)	17,3 (377)	20,4 (481)	24,2 (620)

L'incidence est exprimée en nombre de cas pour 100 000. Entre parenthèses, figurent les nombres de cas incidents. Dans la dernière ligne figurent les incidences pour toute la population, standardisées selon la population mondiale et, entre parenthèses, le nombre total de cas incidents. Pour les périodes 1975-79 à 1990-94, les incidences et les nombres de cas incidents sont extraits du registre ; pour les autres périodes, il s'agit de valeurs prédites.

Tableau 4.24. Cancer du poumon : incidences spécifiques et nombre de cas incidents chez les hommes dans le Bas-Rhin, 1975-2009.

Âge\Période	1975-1979	1980-1984	1985-1989	1990-1994	1995-1999	2000-2004	2005-2009
25-29	0,0 (0)	1,0 (2)	2,0 (4)	0,5 (1)	1,0 (2)	1,1 (2)	1,1 (2)
30-34	0,7 (1)	2,1 (4)	2,0 (4)	4,0 (8)	3,0 (6)	3,1 (6)	3,5 (6)
35-39	7,2 (10)	10,3 (15)	13,3 (25)	7,7 (15)	10,5 (20)	10,8 (21)	12,1 (23)
40-44	28,5 (40)	18,5 (25)	33,7 (48)	34,5 (64)	32,0 (60)	33,3 (62)	35,9 (68)
45-49	66,3 (91)	75,6 (102)	75,1 (98)	76,9 (108)	79,5 (142)	82,7 (151)	88,0 (160)
50-54	136,8 (153)	133,1 (173)	135,3 (174)	136,1 (173)	149,0 (199)	156,2 (268)	166,4 (292)
55-59	216,9 (175)	229,8 (238)	262,1 (315)	257,4 (313)	257,1 (306)	268,1 (339)	287,3 (467)
60-64	291,5 (201)	341,2 (247)	338,2 (316)	327,1 (362)	349,5 (390)	357,8 (395)	382,1 (450)
65-69	380,4 (312)	379,6 (223)	407,3 (259)	436,7 (364)	440,8 (433)	451,2 (453)	474,1 (475)
70-74	456,5 (306)	432,4 (278)	485,6 (229)	505,5 (269)	519,5 (364)	532,8 (449)	557,7 (487)
75-79	311,5 (126)	398,7 (184)	480,7 (220)	463,6 (164)	478,3 (196)	495,2 (273)	519,2 (351)
80-84	243,4 (44)	304,3 (69)	384,5 (104)	355,3 (103)	393,1 (89)	408,1 (113)	433,9 (165)
85-89	184,4 (12)	257,5 (19)	244,8 (23)	301,5 (37)	317,3 (46)	327,9 (38)	354,6 (54)
90-94	69,9 (1)	249,5 (4)	284,5 (6)	136,2 (4)	249,3 (10)	270,4 (13)	279,4 (11)
25-94	113,7 (1472)	120,8 (1583)	131,3 (1825)	132,2 (1985)	137,0 (2263)	141,5 (2583)	150,1 (3011)

L'incidence est exprimée en nombre de cas pour 100 000. Entre parenthèses, figurent les nombres de cas incidents. Dans la dernière ligne figurent les incidences pour toute la population, standardisées selon la population mondiale et, entre parenthèses, le nombre total de cas incidents. Pour les périodes 1975-79 à 1990-94, les incidences et les nombres de cas incidents sont extraits du registre ; pour les autres périodes, il s'agit de valeurs prédites.

De 1975 à 1989, l'incidence du cancer du poumon standardisée selon la population mondiale augmente dans les deux sexes (Tableaux 4.23. et 4.24.). Pour la période 1990-1994, l'incidence standardisée stagne chez les femmes (13,5 pour 10^5) et chez les hommes (132,2 pour 10^5).

L'augmentation la plus forte survient dans la tranche d'âge 85-89 ans, chez les femmes (Tableau 4.23. et figure 4.26.). Chez l'homme, l'incidence s'accroît en proportion équivalente dans toutes les classes d'âge (Tableau 4.24. et figure 4.27.).

Figure 4.26. Cancer du poumon : incidences spécifiques chez les femmes de 25 à 94 ans dans le Bas-Rhin, 1975-1994 (données du registre).

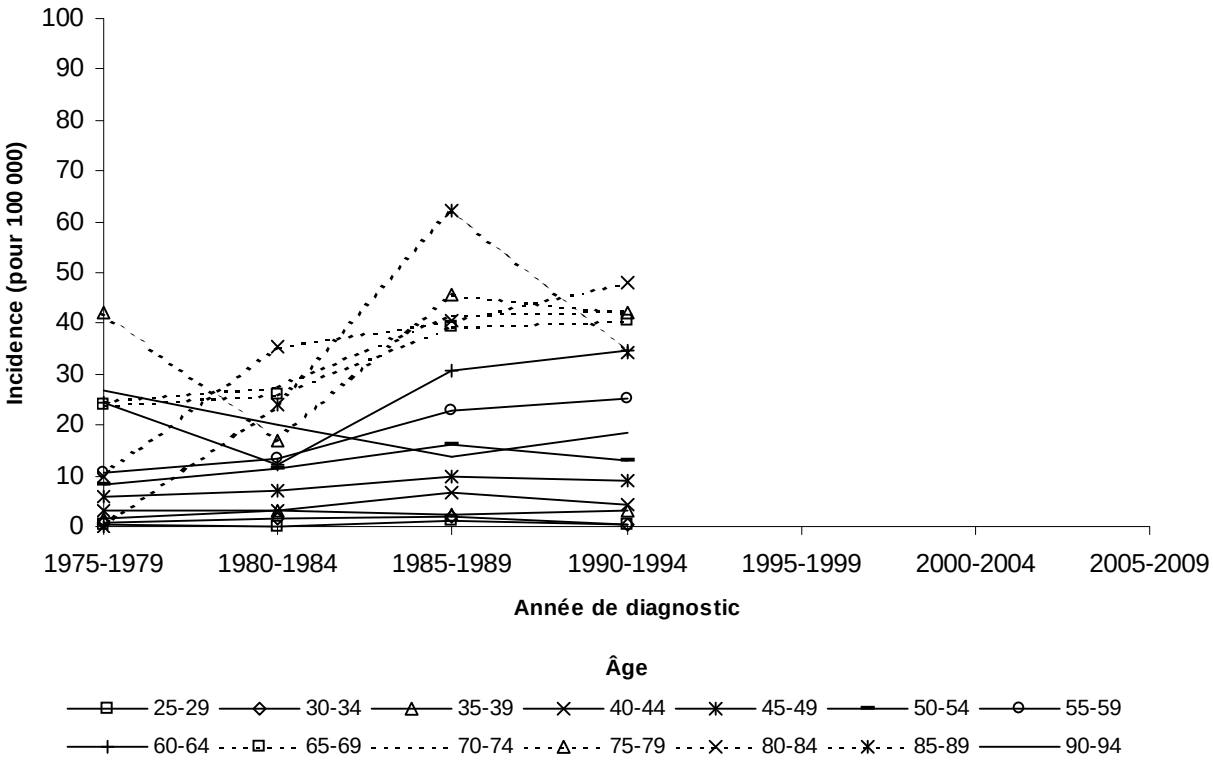
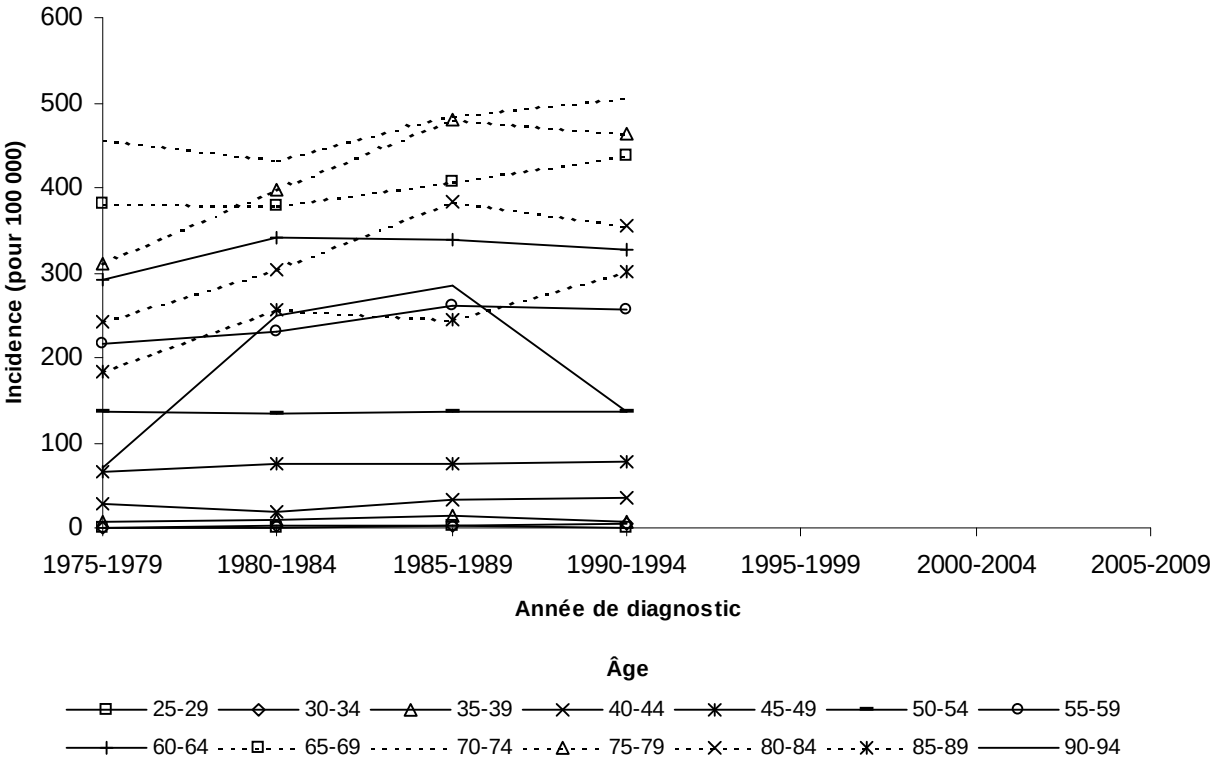


Figure 4.27. Cancer du poumon : incidences spécifiques chez les hommes de 25 à 94 ans dans le Bas-Rhin, 1975-1994 (données du registre).



4.4.3.2. Prévisions

Les taux d'incidence standardisés selon la population mondiale ont été estimés pour les périodes 1995-1999, 2000-2004 et 2005-2009. Le modèle prévoit, pour ces trois périodes, des taux d'incidence de 17,3, 20,4 et 24,2 pour 10^5 chez les femmes et de 137,0, 141,5 et 150,1 pour 10^5 (Tableaux 4.23. et 4.24., figures 4.28. et 4.29). La variation relative annuelle est de + 3,8 % chez les femmes et de + 1,0 % chez les hommes.

Figure 4.28. Cancers du poumon chez les femmes âgées de 25 à 94 ans dans le Bas-Rhin. Prédiction de l'incidence (1975-2014) standardisée selon la population mondiale et intervalle de prédiction à 95%. Pour les périodes 1975-79 à 1990-94, l'incidence est extraites du registre ; pour les autres périodes, il s'agit de valeurs prédites.

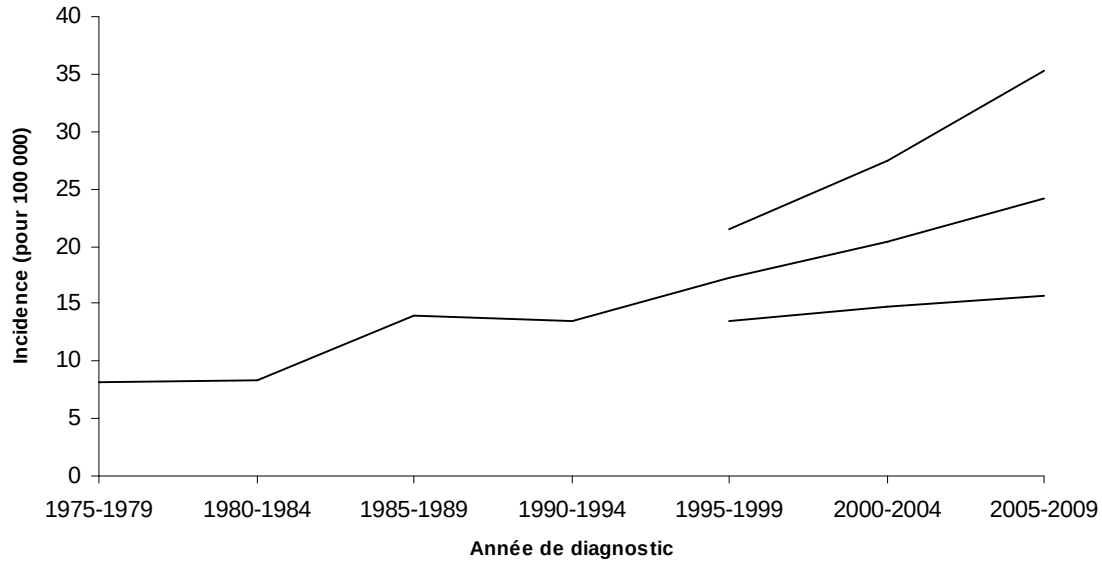
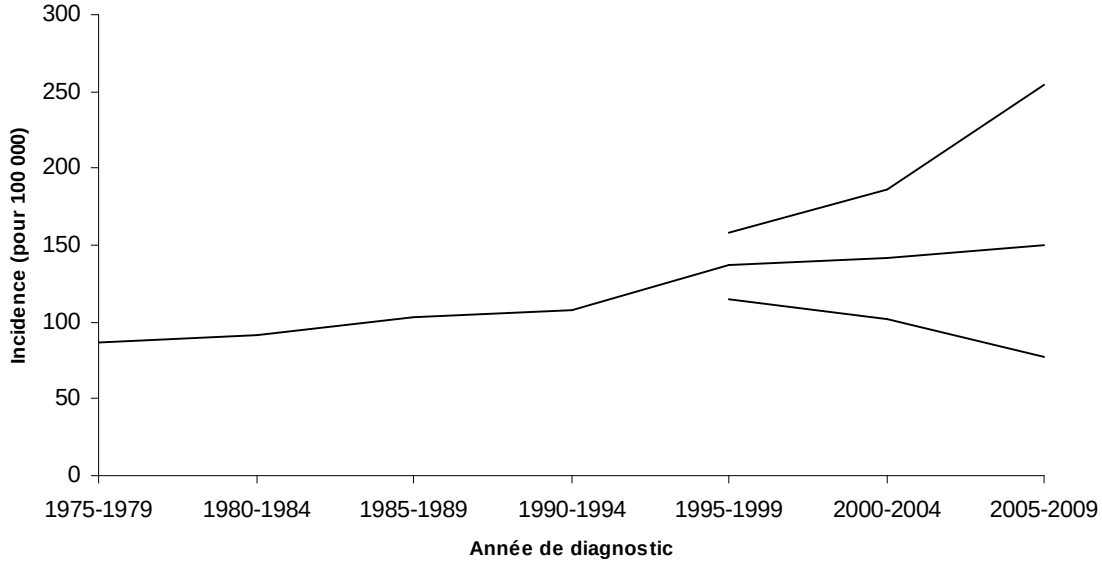


Figure 4.29. Cancers du poumon chez les hommes âgées de 25 à 94 ans dans le Bas-Rhin. Prédiction de l'incidence (1975-2014) standardisée selon la population mondiale et intervalle de prédiction à 95%. Pour les périodes 1975-79 à 1990-94, l'incidence est extraites du registre ; pour les autres périodes, il s'agit de valeurs prédites.



Le nombre de cas incidents est croissant quelque soit la tranche d'âge et dans les deux sexes mais chez la femme, l'augmentation relative du nombre de cas est plus importante que chez l'homme (30 % tous les cinq ans contre 15 % tous les cinq ans) (Tableaux 4.25. et 4.26.).

Tableau 4.25. Cancer du poumon : nombre de cas incidents et intervalle de prédiction (95 %) chez les femmes dans le Bas-Rhin, 1995-2009.

Âge\Période	1995-1999	2000-2004	2005-2009
25-29	2 (1 - 4)	2 (1 - 4)	3 (0 - 7)
30-34	4 (2 - 6)	4 (1 - 8)	4 (1 - 9)
35-39	7 (4 - 10)	8 (3 - 13)	9 (3 - 18)
40-44	12 (7 - 17)	14 (7 - 22)	17 (6 - 30)
45-49	19 (14 - 26)	24 (14 - 37)	29 (13 - 50)
50-54	24 (18 - 31)	37 (25 - 53)	46 (26 - 74)
55-59	34 (27 - 42)	42 (30 - 57)	65 (41 - 98)
60-64	48 (40 - 57)	55 (42 - 71)	69 (48 - 96)
65-69	60 (50 - 71)	71 (56 - 88)	81 (60 - 107)
70-74	63 (53 - 74)	77 (62 - 94)	92 (71 - 117)
75-79	46 (38 - 54)	72 (58 - 88)	90 (70 - 113)
80-84	30 (24 - 36)	44 (35 - 55)	70 (55 - 90)
85-89	21 (16 - 27)	21 (16 - 28)	34 (24 - 46)
90-94	7 (4 - 10)	10 (6 - 15)	11 (6 - 16)
25-94	377 (298 – 465)	481 (356 - 633)	620 (424 - 871)

Tableau 4.26. Cancer du poumon : nombre de cas incidents et intervalle de prédiction (95 %) chez les hommes dans le Bas-Rhin, 1995-2009.

Âge\Période	1995-1999	2000-2004	2005-2009
25-29	2 (1 - 3)	2 (1 - 4)	2 (1 - 6)
30-34	6 (4 - 9)	6 (4 - 11)	6 (2 - 13)
35-39	20 (15 - 27)	21 (13 - 33)	23 (10 - 44)
40-44	60 (47 - 77)	62 (42 - 91)	68 (32 - 124)
45-49	142 (116 - 172)	151 (105 - 209)	160 (79 - 286)
50-54	199 (166 - 233)	268 (191 - 353)	292 (149 - 502)
55-59	306 (258 - 351)	339 (243 - 441)	467 (241 - 787)
60-64	390 (330 - 443)	395 (285 - 508)	450 (236 - 757)
65-69	433 (370 - 491)	453 (327 - 588)	475 (248 - 794)
70-74	364 (311 - 416)	449 (329 - 579)	487 (254 - 814)
75-79	196 (166 - 225)	273 (197 - 354)	351 (183 - 580)
80-84	89 (75 - 104)	113 (82 - 149)	165 (85 - 278)
85-89	46 (36 - 56)	38 (27 - 52)	54 (28 - 92)
90-94	10 (6 - 14)	13 (8 - 20)	11 (5 - 21)
25-94	2263 (1901 - 2621)	2583 (1854 - 3392)	3011 (1553 - 5098)

Les taux d'incidences spécifiques pour l'âge augmentent nettement plus rapidement chez les femmes (15 à 30 % selon la classe d'âge tous les 5 ans) que chez les hommes (5 à 10 % selon la classe d'âge, tous les 5 ans) (Tableaux 4.23. et 4.24., figures 4.30. et 4.31.). Chez les femmes les tranches d'âge les plus menacées en terme d'augmentation de l'incidence sont les tranches de 75 à 94 ans (23 à 27 % d'augmentation en moyenne tous les 5 ans). Chez les hommes, ce sont les tranches d'âges de 85 à 94 ans qui sont les plus menacées (7 à 11 % d'augmentation en moyenne tous les 5 ans) (Tableaux 4.23. et 4.24.).

Figure 4.30. Cancer du poumon : incidences spécifiques chez les femmes de 25 à 94 ans dans le Bas-Rhin, 1975-2009 (modèle âge-cohorte).

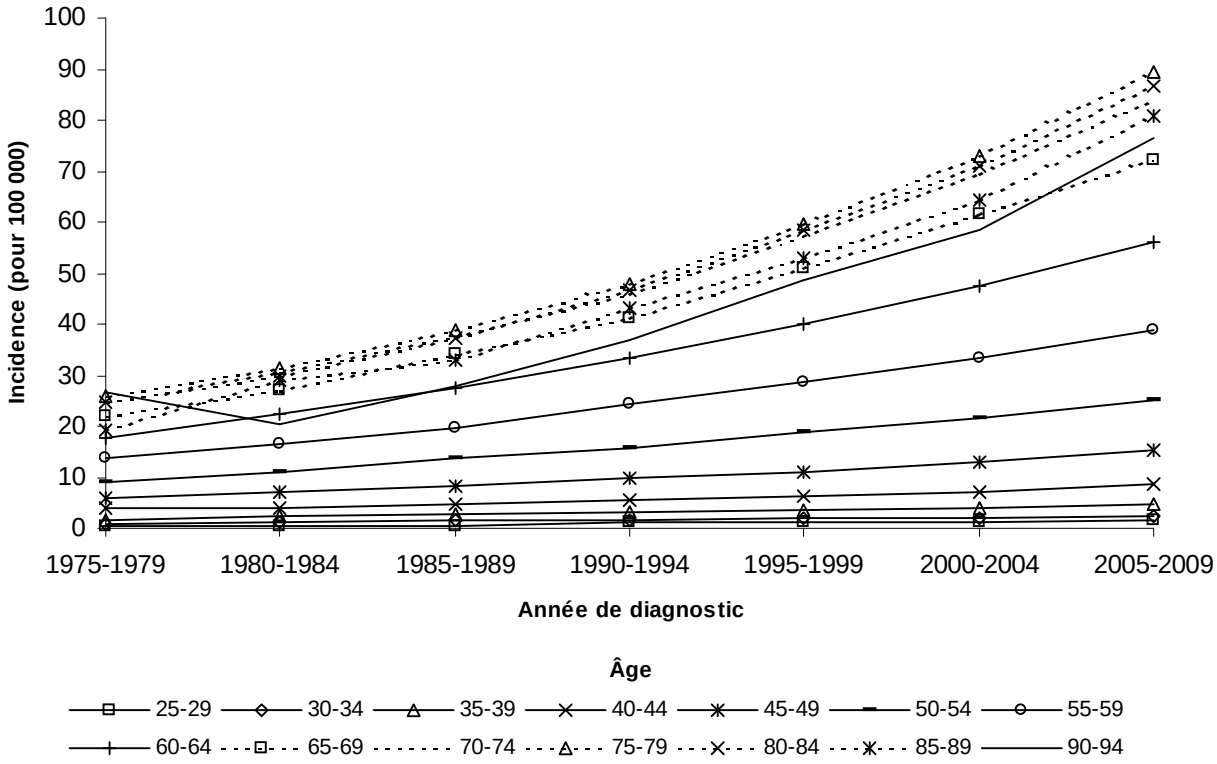
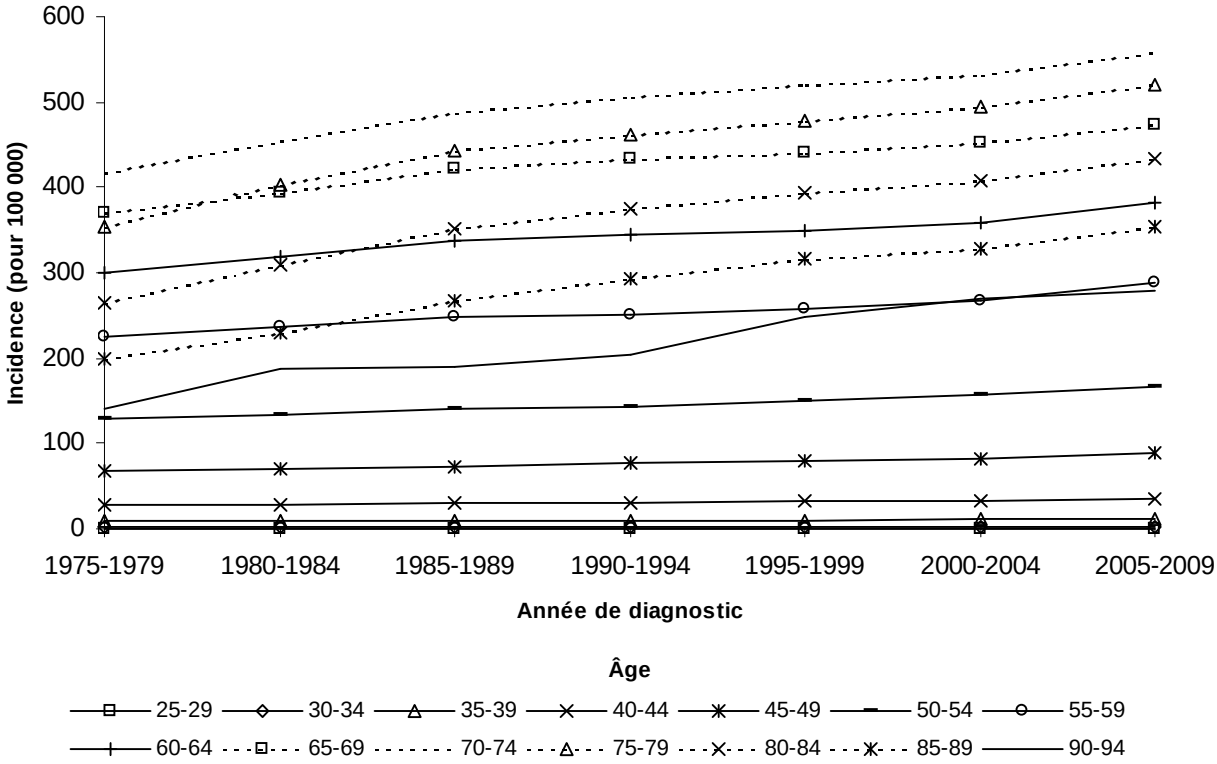


Figure 4.31. Cancer du poumon : incidences spécifiques chez les hommes de 25 à 94 ans dans le Bas-Rhin, 1975-2009 (modèle âge-période-cohorte).



4.4.4. Discussion

L'analyse a prédit une croissance du taux d'incidence et du nombre de cas incidents de cancer du poumon, dans les deux sexes et pour toutes les tranches d'âge. Dans tous les cas (incidence standardisée, incidences spécifiques, nombre de cas) les valeurs obtenues sont plus importantes chez l'homme que chez la femme. En revanche, dans tous les cas également le taux de variation de ces indicateurs est plus important chez la femme que chez l'homme.

Gregor *et al.* (Gregor, 2001) on fait un ensemble de prédiction de l'incidence et du nombre de cas incidents de cancer (du poumon, entre autres) pour l'Écosse, jusqu'en 2010-2014. Il s'agit d'un rapport destiné à planifier la prise en charge des cancers en Écosse. Les données d'incidence sont extraites d'un registre (Scottish Cancer Registry). La méthode utilisée (Alexander, 2001) est basée sur le modèle âge-période-cohorte de Clayton et Schifflers (Clayton, 1987 a et b) en laissant les effets non linéaires inchangés dans les séries prédites et en incluant ou pas, une dérive. Pour les hommes, il trouvent, selon le modèle, une stabilisation ou une diminution de l'incidence. Les taux n'apparaissent pas de façon précise dans ce travail (il sont présentés sous forme de graphes) mais dans le cas ou le taux d'incidence est stable il est voisin de 125 cas pour 100 000 personnes-années durant toute la période de prédiction et dans le cas où le modèle prévoit une décroissance l'incidence passe d'environ 135 cas pour 100 000 en 1980-1984 à environ 60 cas pour 100 000 en 2010-2014. Pour les femmes, dans tous les modèles, l'incidence reste quasiment stable ou augmente légèrement et avoisine 50 cas pour 100 000 personnes-années en 20010-2014. Le nombre annuel de cas de cancer pulmonaire serait, dans le modèle le plus favorable aux hommes, plus élevé, en 2014, chez les femmes que chez les hommes ! Les taux prédits pour les hommes dans le Bas-Rhin sont légèrement supérieurs à ceux de l'Écosse mais présentent une tendance croissante alors que les taux écossais, plus élevés au départ, sont stables voire diminuent. Pour les femmes, le taux prévu en Écosse est deux fois plus élevé que dans le Bas-Rhin. Ainsi, même si les taux d'incidence ne sont pas semblables, on retrouve une inversion du rapport des taux d'incidence entre les deux sexes.

Un travail de prédiction d'incidence du cancer du poumon a été réalisé en Australie dans le cadre d'un rapport sur le cancer également, paru en 1998 (Australian Institute oh Health, 1998). Les données d'incidence sont extraites de plusieurs registres, réunis au sein de l'Australasian Association of Cancer Registries (AACR). Les projections extrapolent jusqu'en 1999 un ensemble de régressions linéaires calculées, pour les deux sexes séparément, sur les incidences spécifiques enregistrées de 1990 à 1994. Ce travail estime une décroissance de l'incidence du cancer du poumon chez l'homme de 1,5 % par an

(entre 1994 et 1999) et une croissance du taux chez la femme de 1,6 % par an. Chez l'homme, l'incidence, standardisée sur la population australienne de 1991, passe de 61,1 pour 10^5 en 1995 (5150 cas) à 58,6 pour 10^5 en 1999 (5440 cas). Chez la femme, l'incidence, standardisée sur la population australienne de 1991, passe de 21,4 pour 10^5 en 1995 (2200 cas) à 22,8 pour 10^5 en 1999 (2590 cas). L'antagonisme apparent des incidences et des nombres de cas incidents chez l'homme est dû à la croissance de l'effectif de la population australienne. Le mode de standardisation utilisé dans cette étude (population australienne) ne permet pas de procéder à des comparaisons des taux d'incidence avec le Bas-Rhin, de façon absolue. Il est possible de comparer, par contre, les variations relatives annuelles : pour l'homme, diminution de 1,5 % par an en Australie contre 1,0 % d'augmentation dans le Bas-Rhin ; pour les femmes, augmentation de 1,6 % par an en Australie contre 3,8 % d'augmentation dans le Bas-Rhin. Les durées des périodes de projection ne sont pas comparables non plus : l'étude australienne a réalisé une projection à court terme (5 ans) ; dans le Bas-Rhin la période de projection s'étend sur 15 ans.

Hakulinen et Pukkala (Hakulinen, 1981) ont effectué un certain nombre de prédictions d'incidence en Finlande en tenant compte de différents scénarios de comportement tabagique. Les divers modèles identifient les groupes d'âges concernés par le début de la consommation de tabac (de 10 à 24 ans) et testent un ensemble d'hypothèses quant au nombre de cigarettes journalier, quant à l'âge de début et d'arrêt du tabagisme. Les prédictions sont établies pour 2050 ! Il est là aussi difficile de comparer les résultats avec ceux du Bas-Rhin. Mais cette étude montre à quel point la tendance de l'incidence est sensible à la nature des différents scénarios et plus particulièrement à la proportion de fumeurs « repentants » : 10 % d'interruption de tabagisme dans chaque groupe concerné inversent déjà la tendance.

Dans l'avenir il serait souhaitable de compléter les modèles bayésiens utilisés ici en incluant les paramètres et hypothèses liés au tabagisme (Hakulinen, 1981 ; Janssen-Heitjnen, 1995 ; Zaridze, 1986). L'hypothèse de l'existence d'autres facteurs environnementaux éventuels, évoquée par le travail de Schwartz (1992) et la possibilité d'impliquer la qualité de l'air dans la genèse d'un certains nombre de cancer du poumon comme l'ont suggérée les résultats de deux études de cohortes américaines (Dockery, 1993 ; Pope, 1995), devraient permettre d'inclure d'autres variables dans les modèles utilisés ici.

En ce qui concerne la consommation de tabac, les études de prévalence sont peu nombreuses dans le Bas-Rhin en dehors d'une enquête téléphonique réalisée par le Laboratoire d'épidémiologie de l'Université Louis Pasteur, en 1994 dans le Bas-Rhin et d'un travail effectué dans le cadre des centre d'examens de santé de la Mutuelle Générale des Enseignants (MGEN) (Wertenschlag, 1996).

5. AUTRES MÉTHODES

« The weird sisters, hand in hand,
Posters of the sea and land,
Thus do go about, about:
Thrice to thine, and thrice to mine,
And thrice again, to make up nine:
Peace !—the charm's wound up. »[9]

[9] William Shakespeare. Macbeth. The compete works of Shakespeare. London, Spring Books, 1964.

À titre comparatif, d'autres méthodes, basées sur le modèle linéaire généralisé, ont été testées (voir chapitre 2.2.2.2.3.2. Différentes approches et différents modèles)

Ce sont les méthodes suivantes :

- Méthode de Decarli et La Vecchia
- Méthode du CIRC
- Modèle additif généralisé

5.1. Méthode de Decarli et La Vecchia

La méthode de Decarli et La Vecchia (cf. 2.2.2.2.3.2. (b) (α)) a été appliquée aux données du cancer du sein invasif dans le Bas-Rhin. La base de prédiction était constituée de quatre périodes de 5 ans : 1975-1979 à 1990-1994. La période sur laquelle portait la prédiction s'étend de 1995 à 2009, par tranches de 5 ans (trois périodes). La méthode de prévision consistait à extraire les effets âge, période et cohorte de l'analyse des données de 1975 à 1994 à l'aide de la méthode de Decarli et La Vecchia puis d'extrapoler les trois effets constitutifs du modèle (en fait, uniquement la période et la cohorte, les classes d'âge restant constantes) par une régression linéaire ou logarithmique. L'incidence spécifique est obtenue dans chaque cas du tableau âge-période en calculant le produit des trois effets correspondants.

Les extrapolations par régression linéaire et par régression logarithmique donnent des résultats très proches. Aussi, ne seront présentés que les résultats de la régression logarithmique. L'augmentation est régulière pour les cancers invasifs avec, *si l'on ajoute foi à la méthode*, un « effet âge » élevé, un « effet cohorte » fonction lentement croissante de la cohorte, un « effet période » négligeable (Tableau 5.1.).

Tableau 5.1. Cancer du sein invasif : « effets » âge, période et cohorte calculés par la méthode de Decarli et La Vecchia (extrapolation par régression logarithmique).

Numéro	Âge	Effet âge	Période	Effet période	Cohorte	Effet cohorte
1	**25-29**	4,961	**1975-1979**	0,9750	**1886-1894**	0,5023
2	**30-34**	14,380	**1980-1984**	0,9680	**1891-1899**	0,6211
3	**35-39**	36,680	**1985-1989**	1,0140	**1896-1904**	0,5860
4	**40-44**	75,940	**1990-1994**	1,0270	**1901-1909**	0,6469
5	**45-49**	129,400	**1995-1999**	1,0473	**1906-1914**	0,7071
6	**50-54**	160,200	**2000-2004**	1,0687	**1911-1919**	0,8940
7	**55-59**	193,700	**2005-2009**	1,0906	**1916-1924**	0,7770
8	**60-64**	259,600			**1921-1929**	0,9603
9	**65-69**	305,600			**1926-1934**	1,0860
10	**70-74**	350,000			**1931-1939**	1,2880
11	**75-79**	385,600			**1936-1944**	1,5190
12	**80-84**	358,800			**1941-1949**	1,5430
13	**85-89**	324,200			**1946-1954**	1,6350
14					**1951-1959**	1,7940
15					**1956-1964**	2,1430
16					**1961-1969**	1,4600
17					**1966-1974**	2,2563
18					**1971-1979**	2,4736
19					**1976-1984**	2,7117

En grisé : effet période ou effet cohorte prédit.

Quant aux prévisions, elles s'établissent, pour l'incidence standardisée sur la population européenne et pour les périodes 1995-1999, 2000-2004 et 2005-2009, respectivement à 194,1 pour 10^5 (IC 95 % : [174,4 – 216,2]), 218,3 pour 10^5 (IC 95 % : [190,8 – 250,4]) et 241,9 pour 10^5 (IC 95 % : [203,9 – 288,8]) (Tableau 5.2. et figure 5.1).

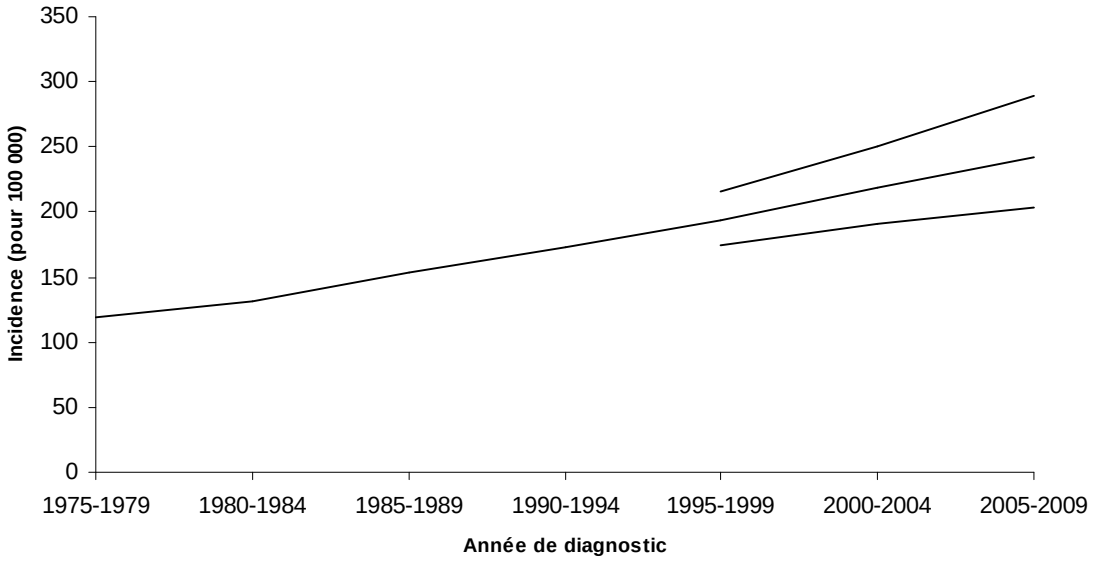
Ces résultats sont comparables aux incidences standardisées prédites par l'analyse bayésienne qui sont de 199,2 pour 10^5 en 1997-2000 (IC 95 % : [182,2 – 217,8]), de 217,3 pour 10^5 en 2001-2004 (IC 95 % : [195,6 – 241,6]) et de 236,2 pour 10^5 en 2005-2008 (IC 95 % : [208,4 – 267, 5]) (voir chapitre 4, paragraphe 4.1.). Il est à noter, cependant, que les intervalles de prédiction sont légèrement plus petits dans l'estimation bayésienne.

Tableau 5.2. Cancer du sein invasif : incidences spécifiques et nombre de cas incidents chez les femmes dans le Bas-Rhin, 1975-2009. calculées par la méthode de Decarli et La Vecchia (extrapolation par régression logarithmique).

Âge\Période	1975-1979	1980-1984	1985-1989	1990-1994	1995-1999	2000-2004	2005-2009
25-29	8,7 (16)	8,6 (17)	10,8 (22)	7,4 (15)	11,7 (23)	13,1 (23)	14,7 (27)
30-34	21,6 (31)	22,8 (42)	26,2 (50)	31,7 (62)	22,0 (44)	34,7 (68)	38,8 (67)
35-39	54,3 (68)	54,8 (77)	60,8 (110)	67,6 (130)	82,3 (161)	57,2 (114)	90,3 (177)
40-44	95,4 (122)	111,7 (139)	118,8 (166)	127,5 (230)	142,7 (272)	173,9 (339)	120,9 (239)
45-49	137,0 (183)	161,3 (205)	199,3 (244)	205,1 (284)	221,6 (395)	248,1 (468)	302,4 (584)
50-54	150,0 (197)	168,4 (220)	209,2 (260)	249,9 (301)	258,9 (353)	279,9 (492)	313,4 (584)
55-59	146,7 (168)	180,1 (231)	213,3 (271)	256,2 (312)	308,2 (365)	319,4 (429)	345,4 (598)
60-64	226,3 (188)	195,3 (215)	252,8 (313)	289,5 (357)	350,2 (417)	421,4 (489)	436,8 (575)
65-69	210,7 (235)	264,5 (209)	240,8 (252)	301,4 (358)	347,6 (413)	420,7 (484)	506,3 (570)
70-74	220,8 (230)	239,6 (244)	317,3 (229)	279,3 (274)	352,0 (394)	406,2 (458)	491,6 (540)
75-79	220,3 (172)	241,5 (210)	276,5 (239)	354,0 (226)	313,8 (276)	395,7 (401)	456,7 (473)
80-84	217,3 (96)	203,5 (114)	235,4 (149)	260,6 (178)	336,0 (171)	298,0 (217)	375,8 (320)
85-89	158,8 (29)	194,9 (47)	192,6 (62)	215,4 (89)	240,1 (107)	309,8 (106)	274,7 (142)
25-89	118,8 (1736)	131,4 (1972)	153,4 (2370)	172,4 (2820)	194,1 (3396)	218,3 (4094)	241,9 (4903)

L'incidence est exprimée en nombre de cas pour 100 000. Entre parenthèses, figurent les nombres de cas incidents. Dans la dernière ligne figurent les incidences pour toute la population, standardisées selon la population européenne et, entre parenthèses, le nombre total de cas incidents. Pour les périodes 1975-79 à 1990-964, les incidences et les nombres de cas incidents sont extraits du registre ; pour les autres périodes, il s'agit de valeurs prédites.

Figure 5.1. Cancers invasifs du sein chez les femmes âgées de 25 à 89 ans dans le Bas-Rhin. Prédiction de l'incidence (1975-2009) standardisée selon la population européenne et intervalle de prédiction à 95%. Méthode de Decarli et La Vecchia.



Les nombres de cas incidents ne sont pas comparables car ils ne concernent pas des périodes de durées équivalentes (5 ans pour la méthode de De Carli et La Vecchia, 4 ans pour la méthode bayésienne). En rapportant les nombres de cas incidents à l'année, ceux-ci ne sont tout de même pas comparables (981 cas par an pour la méthode de Decarli et La Vecchia, 897 cas pour la méthode bayésienne). Cette différence est peut-être due au fait que l'année supplémentaire dans la méthode de Decarli et La Vecchia est la dernière année et par conséquent la plus riche en cas incidents.
La comparaison des incidences spécifiques n'est pas possible de façon rigoureuse non plus car les périodes et les classes d'âge ne correspondent pas. Il est cependant aisé de constater que pour des cases « presque comparables » (classes d'âge 70-74, 75-79, 80-84 et 85-89, périodes 1995-1999, 2000-2004 et 2005-2009, par exemple), les incidences spécifiques sont en général du même ordre de grandeur (Tableaux 4.2 et 5.2.).

5. 2. Méthode du CIRC

La méthode du CIRC (cf. 2.2.2.2.3.2 (a) (α)) a été appliquée aux données du cancer invasif du sein dans le Bas-Rhin.

Le modèle est un modèle âge-période-cohorte auquel une composante « dépistage » a été ajoutée.

Les quatre facteurs sont considérés comme variables quantitatives.
Les valeurs prises par les variables âge, période et cohorte sont leur numéro d'ordre (cf. Tableau 5.1.). Par exemple, la classe d'âge 25-29 ans prend la valeur 1. Il en est de même pour la période 1975-1979 et la cohorte 1886-1894. La classe d'âge 30-34 ans prend la valeur 2, etc.
La variable « dépistage » est égale à la proportion de femmes ayant bénéficié d'une mammographie dans le cadre du dépistage ADEMAS (voir paragraphe 4.1. Introduction).

L'analyse détermine le degré du polynôme pour chacune des deux variables âge et cohorte, les variables période et « dépistage » étant imposées sous forme linéaire dans le modèle. Lorsque le meilleur modèle a été trouvé, différentes projections ont été réalisées :

- Une régression logarithmique sur la cohorte en laissant la variable « dépistage » constante ;
- Une régression logarithmique sur la cohorte et une régression linéaire sur la variable « dépistage » ;
- Une régression polynomiale sur la cohorte en laissant la variable « dépistage » constante ;
- Une régression polynomiale sur la cohorte et une régression linéaire sur la variable « dépistage »

Le degré du polynôme déterminé par l'analyse pour l'âge est 6, le degré du polynôme pour la cohorte est 8.

Le type de régression sur la cohorte (logarithmique ou polynomiale) n'influe pas sur les valeurs des incidences spécifiques prédites, mises à part celles qui correspondent aux femmes les plus jeunes (Tableaux 5.3.à 5.6.). Les incidences standardisées sont légèrement plus élevées pour la période prédite dans le cas de la régression logarithmique.

Tableau 5.3. Cancer du sein invasif : incidences spécifiques chez les femmes dans le Bas-Rhin, 1975-2009, calculées par la méthode du CIRC et prédites par une régression logarithmique sur la cohorte en laissant la variable « dépistage » constante.

Âge\Période	1975-1979	1980-1984	1985-1989	1990-1994	1995-1999	2000-2004	2005-2009
25-29	4,6	4,6	5,3	3,2	5,3	5,5	5,8
30-34	13,1	12,4	12,9	14,7	9,0	14,8	15,5
35-39	37,1	36,5	34,5	35,8	40,9	25,0	41,1
40-44	73,5	78,6	77,5	73,3	76,1	86,9	53,1
45-49	105,5	116,9	125,2	123,3	116,7	121,1	138,3
50-54	129,2	138,8	155,8	180,6	177,9	168,4	174,7
55-59	154,3	158,4	172,3	208,0	222,7	219,3	207,6
60-64	185,9	191,1	198,2	230,9	255,9	274,0	269,8
65-69	215,5	235,7	242,6	258,2	277,5	307,6	329,3
70-74	230,9	266,5	291,5	299,9	307,8	330,7	366,6
75-79	236,5	258,5	298,4	326,3	335,6	344,4	370,1
80-84	243,2	232,9	254,6	293,8	321,3	330,5	339,2
85-89	189,3	237,3	227,3	248,5	286,8	313,7	322,6
25-89	107,1	114,9	122,3	133,5	139,2	143,6	147,1

L'incidence est exprimée en nombre de cas pour 100 000. Dans la dernière ligne figurent les incidences pour toute la population, standardisées selon la population européenne. Les valeurs sont prédites par le modèle pour toutes les périodes. Les cellules encadrées correspondent aux femmes ayant bénéficié du dépistage organisé.

Le régression linéaire sur la variable « dépistage » élève le niveau des incidences spécifiques correspondant aux âges et périodes du dépistage (Tableaux 5.3. à 5.6., cellules encadrées). Les incidences standardisées prédites sont, par conséquent, plus élevées aussi.

Tableau 5.4. Cancer du sein invasif : incidences spécifiques chez les femmes dans le Bas-Rhin, 1975-2009, calculées par la méthode du CIRC et prédites par une régression logarithmique sur la cohorte et une régression linéaire sur la variable « dépistage ».

Âge\Période	1975-1979	1980-1984	1985-1989	1990-1994	1995-1999	2000-2004	2005-2009
25-29	4,6	4,6	5,3	3,2	5,3	5,5	5,8
30-34	13,1	12,4	12,9	14,7	9,0	14,8	15,5
35-39	37,1	36,5	34,5	35,8	40,9	25,0	41,1
40-44	73,5	78,6	77,5	73,3	76,1	86,9	53,1
45-49	105,5	116,9	125,2	123,3	116,7	121,1	138,3
50-54	129,2	138,8	155,8	180,6	183,0	178,5	191,0
55-59	154,3	158,4	172,3	208,0	232,4	239,3	236,8
60-64	185,9	191,1	198,2	230,9	269,3	303,7	315,1
65-69	215,5	235,7	242,6	258,2	297,4	352,9	404,4
70-74	230,9	266,5	291,5	299,9	308,2	331,6	367,9
75-79	236,5	258,5	298,4	326,3	335,6	344,4	370,1
80-84	243,2	232,9	254,6	293,8	321,3	330,5	339,2
85-89	189,3	237,3	227,3	248,5	286,8	313,7	322,6
25-89	107,1	114,9	122,3	133,5	143,0	151,8	159,9

L'incidence est exprimée en nombre de cas pour 100 000. Dans la dernière ligne figurent les incidences pour toute la population, standardisées selon la population européenne. Les valeurs sont prédites par le modèle pour toutes les périodes. Les cellules encadrées correspondent aux femmes ayant bénéficié du dépistage organisé.

Tableau 5.5. Cancer du sein invasif : incidences spécifiques chez les femmes dans le Bas-Rhin, 1975-2009, calculées par la méthode du CIRC et prédites par une régression polynomiale sur la cohorte en laissant la variable « dépistage » constante.

Âge\Période	1975-1979	1980-1984	1985-1989	1990-1994	1995-1999	2000-2004	2005-2009
25-29	4,6	4,6	5,3	3,2	0,1	0,0	0,0
30-34	13,1	12,4	12,9	14,7	9,0	0,2	0,0
35-39	37,1	36,5	34,5	35,8	40,9	25,0	0,6
40-44	73,5	78,6	77,5	73,3	76,1	86,9	53,1
45-49	105,5	116,9	125,2	123,3	116,7	121,1	138,3
50-54	129,2	138,8	155,8	180,6	177,9	168,4	174,7
55-59	154,3	158,4	172,3	208,0	222,7	219,3	207,6
60-64	185,9	191,1	198,2	230,9	255,9	274,0	269,8
65-69	215,5	235,7	242,6	258,2	277,5	307,6	329,3
70-74	230,9	266,5	291,5	299,9	307,8	330,7	366,6
75-79	236,5	258,5	298,4	326,3	335,6	344,4	370,1
80-84	243,2	232,9	254,6	293,8	321,3	330,5	339,2
85-89	189,3	237,3	227,3	248,5	286,8	313,7	322,6
25-89	107,1	114,9	122,3	133,5	138,6	141,4	140,3

L'incidence est exprimée en nombre de cas pour 100 000. Dans la dernière ligne figurent les incidences pour toute la population, standardisées selon la population européenne. Les valeurs sont prédites par le modèle pour toutes les périodes. Les cellules encadrées correspondent aux femmes ayant bénéficié du dépistage organisé.

La comparaison des résultats obtenus ici avec les projections de la méthode bayésienne (Tableau 4.2) et les projections de la méthode de Decarli et La Vecchia (Tableau 5.2.) montre de grandes différences, quels que soient l'âge et la période. La méthode bayésienne (et la méthode de Decarli et La Vecchia, puisque leurs résultats sont voisins) donne des prévisions d'incidence spécifiques (et par conséquent standardisées) nettement plus grandes (au moins 75 % de plus) que la méthode du CIRC. Les incidences « prédites » par cette dernière méthode, pour la période de disponibilité des données (base de la prévision), sont relativement proches, cependant, des estimations des autres méthodes.

Tableau 5.6. Cancer du sein invasif : incidences spécifiques chez les femmes dans le Bas-Rhin, 1975-2009, calculées par la méthode du CIRC et prédites par une régression polynomiale sur la cohorte et une régression linéaire sur la variable « dépistage ».

Âge\Période	1975-1979	1980-1984	1985-1989	1990-1994	1995-1999	2000-2004	2005-2009
25-29	4,6	4,6	5,3	3,2	0,1	0,0	0,0
30-34	13,1	12,4	12,9	14,7	9,0	0,2	0,0
35-39	37,1	36,5	34,5	35,8	40,9	25,0	0,6
40-44	73,5	78,6	77,5	73,3	76,1	86,9	53,1
45-49	105,5	116,9	125,2	123,3	116,7	121,1	138,3
50-54	129,2	138,8	155,8	180,6	183,0	178,5	191,0
55-59	154,3	158,4	172,3	208,0	232,4	239,3	236,8
60-64	185,9	191,1	198,2	230,9	269,3	303,7	315,1
65-69	215,5	235,7	242,6	258,2	297,4	352,9	404,4
70-74	230,9	266,5	291,5	299,9	308,2	331,6	367,9
75-79	236,5	258,5	298,4	326,3	335,6	344,4	370,1
80-84	243,2	232,9	254,6	293,8	321,3	330,5	339,2
85-89	189,3	237,3	227,3	248,5	286,8	313,7	322,6
25-89	107,1	114,9	122,3	133,5	142,4	149,6	153,1

L'incidence est exprimée en nombre de cas pour 100 000. Dans la dernière ligne figurent les incidences pour toute la population, standardisées selon la population européenne. Les valeurs sont prédites par le modèle pour toutes les périodes. Les cellules encadrées correspondent aux femmes ayant bénéficié du dépistage organisé.

5.3. Modèle additif généralisé

Le modèle additif généralisé (abrégé GAM, pour *Generalized linear model*) est une extension du modèle linéaire généralisé (GLM). Le prédicteur linéaire du GLM est remplacé ici par un prédicteur dit additif. Celui-ci comprend, entre autres, des fonctions de lissage non paramétriques dont un avantage est d'approcher plus précisément la forme de la relation entre la variable expliquée et les variables explicatives.

Le paragraphe 7.2.4. présente de façon plus détaillée la structure du modèle additif généralisé.

Les données d'incidence du cancer du sein invasif ont été analysées à l'aide d'un modèle additif généralisé (le détail des instructions S-Plus figurent en Annexe 3). Les calculs ont été réalisés sous le logiciel S-Plus (MathSoft 1997 a, b et c ; MathSoft 1998 ; Krause, 1997 ; Venables, 1997).

En préliminaire une approche graphique a été réalisée pour visualiser les relations entre l'incidence, l'âge, la période et la cohorte (Jolley, 1992 ; Robertson, 1998 b). Une interaction âge-période apparaît (Figure 5.2.) alors qu'il ne semble pas exister d'interaction importante entre l'âge et la cohorte (Figure 5.3.).

Figure 5.2. Cancer du sein invasif dans le Bas-Rhin, 1975-1994. Incidence en fonction de l'âge et la période (lissage par fonction *spline*).

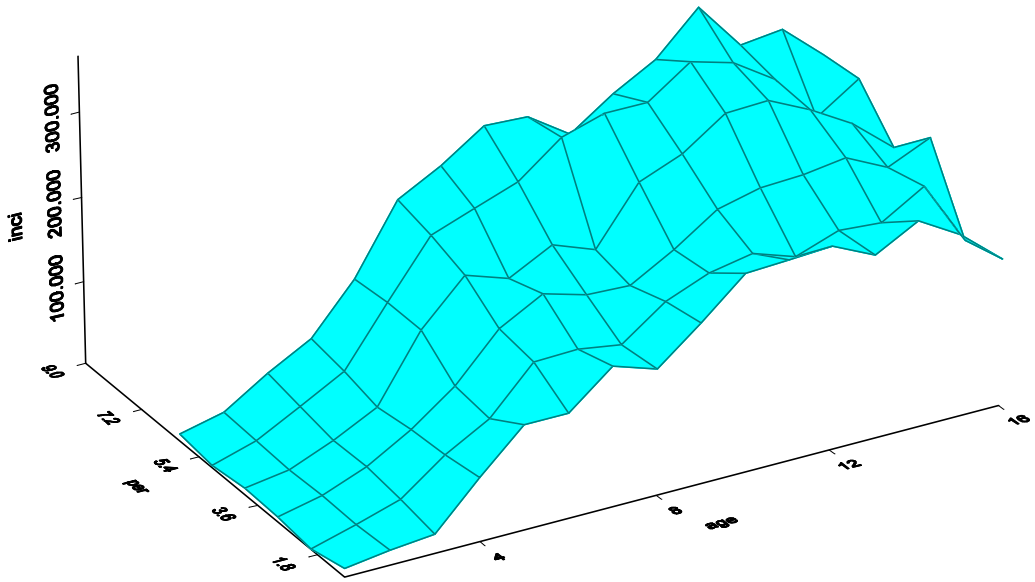
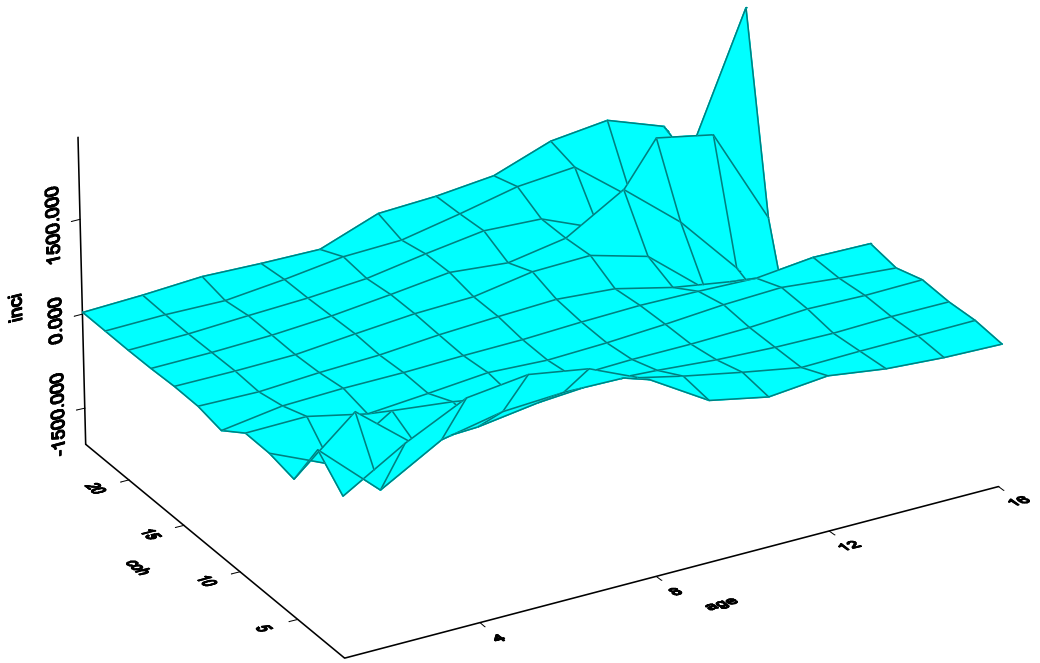
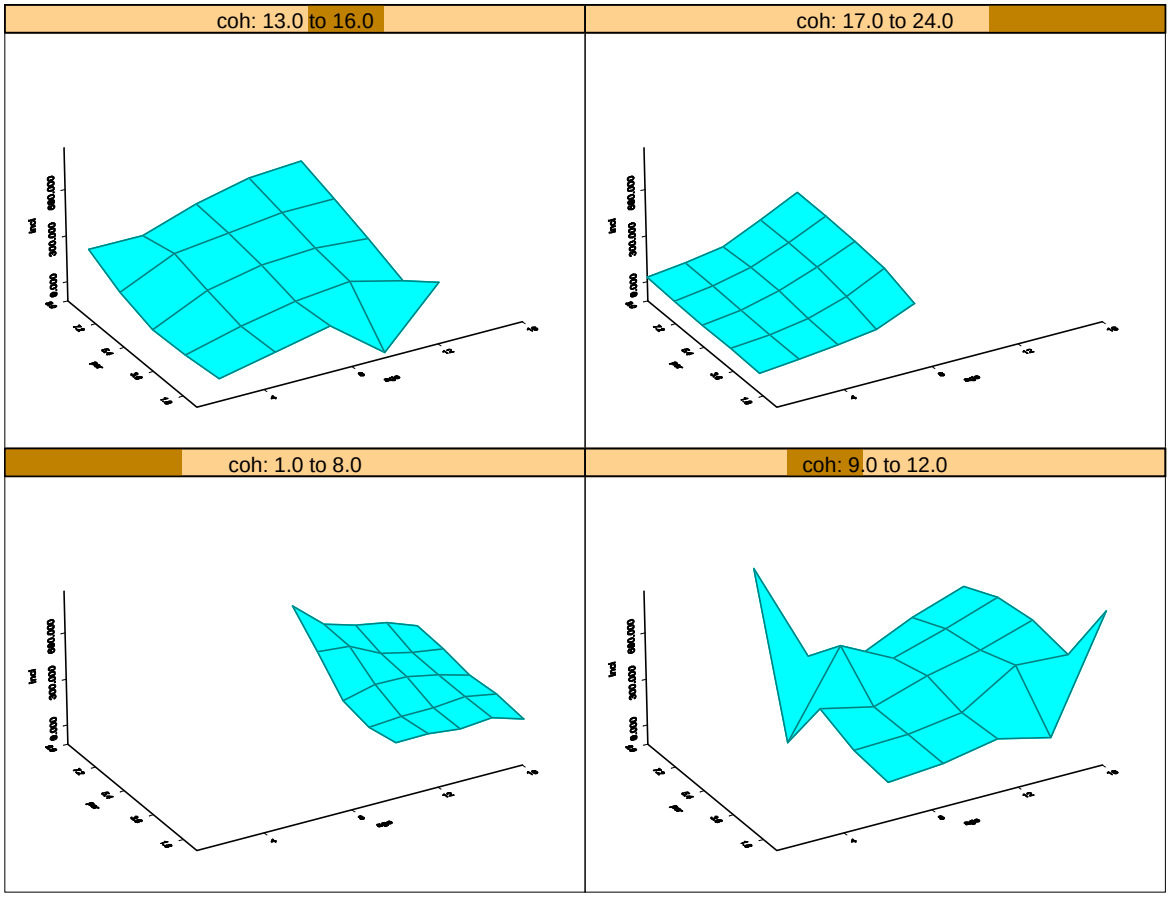


Figure 5.3. Cancer du sein invasif dans le Bas-Rhin, 1975-1994. Incidence en fonction de l'âge et la cohorte (lissage par fonction *spline*).



La Figure 5.4. représente l'interaction âge-période selon le groupe de cohortes. Cette interaction est plus manifeste pour les cohortes 9 à 12 (1917-1920, 1921-1924, 1925-1928 et 1929-1932).

Figure 5.4. Cancer du sein invasif dans le Bas-Rhin, 1975-1994. Incidence en fonction de l'âge et la période selon les classes de cohorte (lissage par fonction *spline*).



Un premier modèle a été testé avec, comme seul facteur explicatif, l'âge sous forme de variable qualitative.

Puis, la période a été introduite sous forme de fonction de régression (voir fonction *loess* de S-Plus à l'Annexe 3.). La forme de la relation entre le nombre de cas, ajusté sur l'âge et la période est de type sigmoïde (fonction logistique) mais pour les valeurs les plus élevées de la période approchent une courbe de forme logarithmique (Figure 5.5.). Un nouveau modèle, GLM celui-là, est testé avec le logarithme de la période (Figure 5.5). L'adéquation du modèle aux données est de bonne qualité (Figure 5.6.).

Figure 5.5. Modélisation du nombre de cas incidents de cancers invasifs dans le Bas-Rhin. Les facteurs sont l'âge sous forme de variable qualitative et la période sous forme de variable quantitative. Les deux figures du haut représentent le modèle GAM avec une fonction de régression locale de la période. Les deux figures du bas représentent le modèle GLM avec le logarithme de la période.

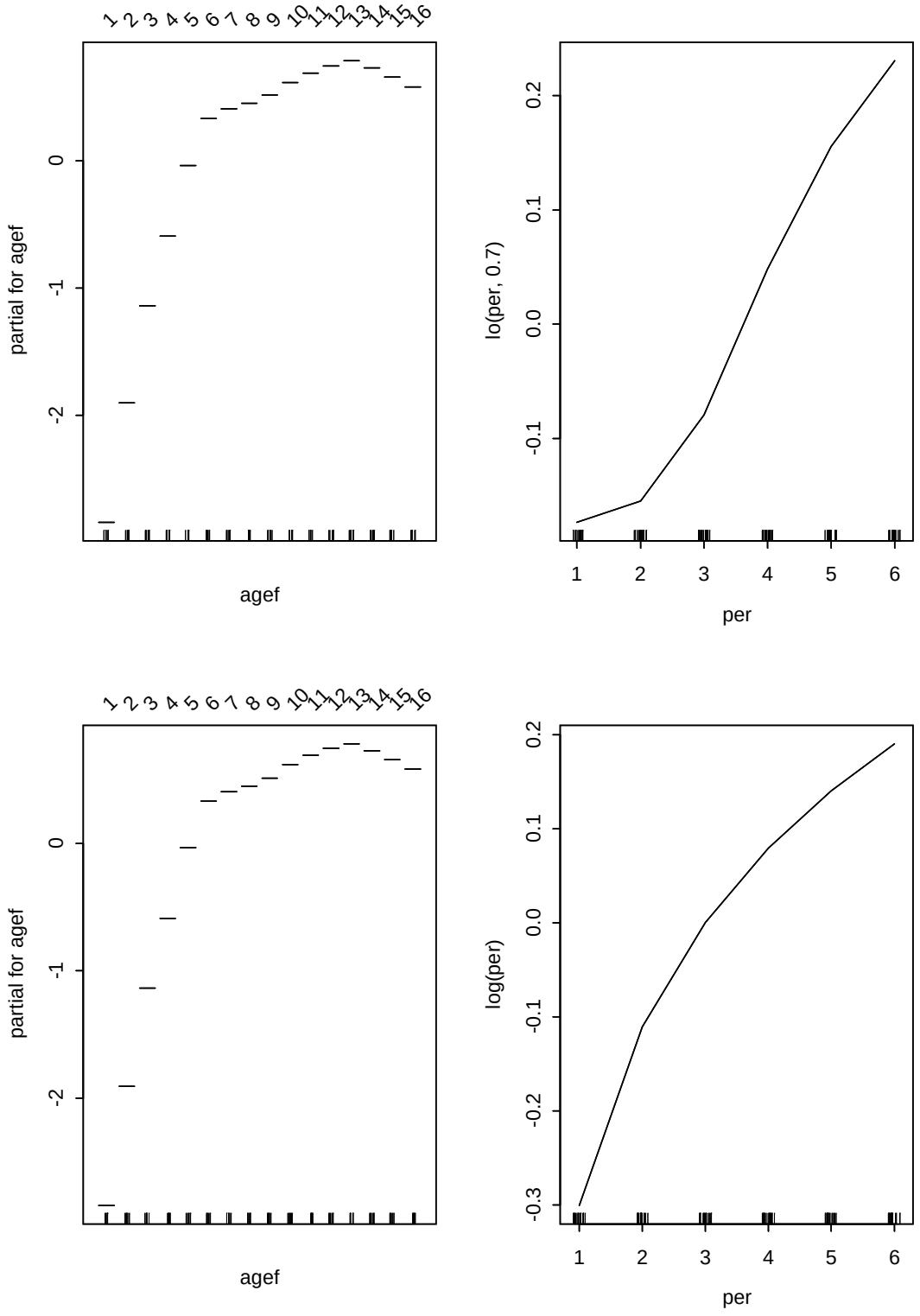
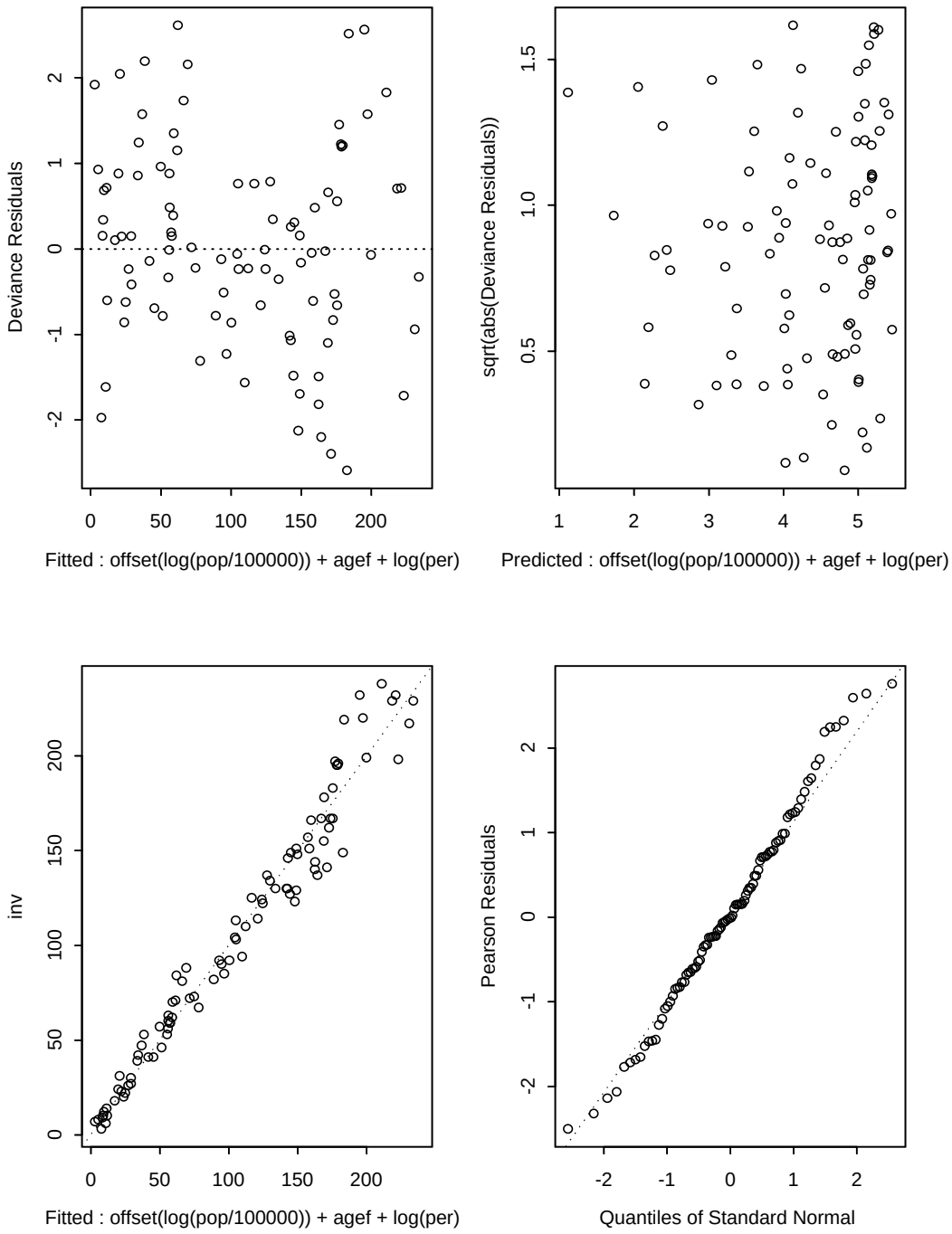
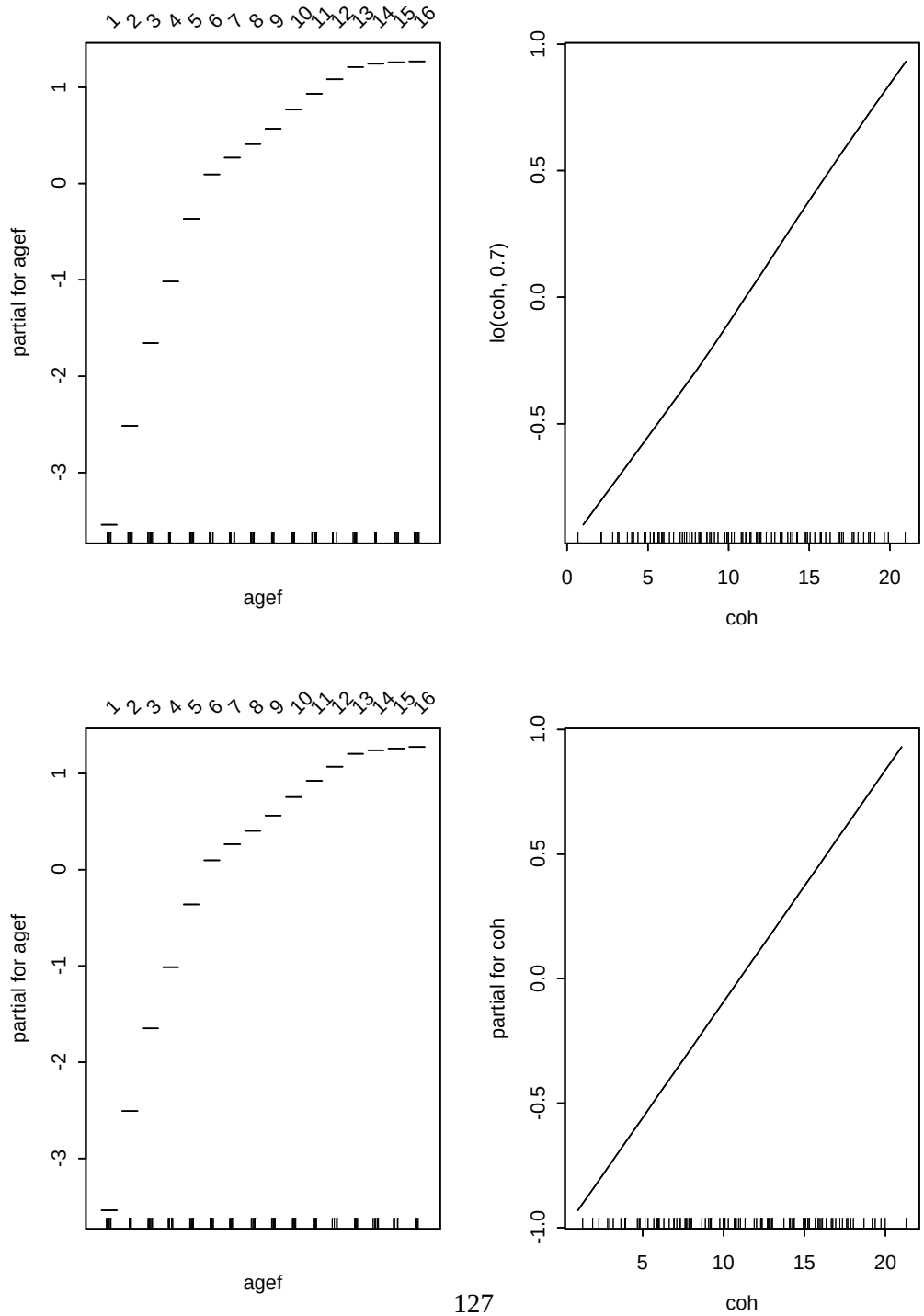


Figure 5.6. Adéquation du modèle GLM avec l'âge sous forme de variable qualitative et le logarithme de la période.



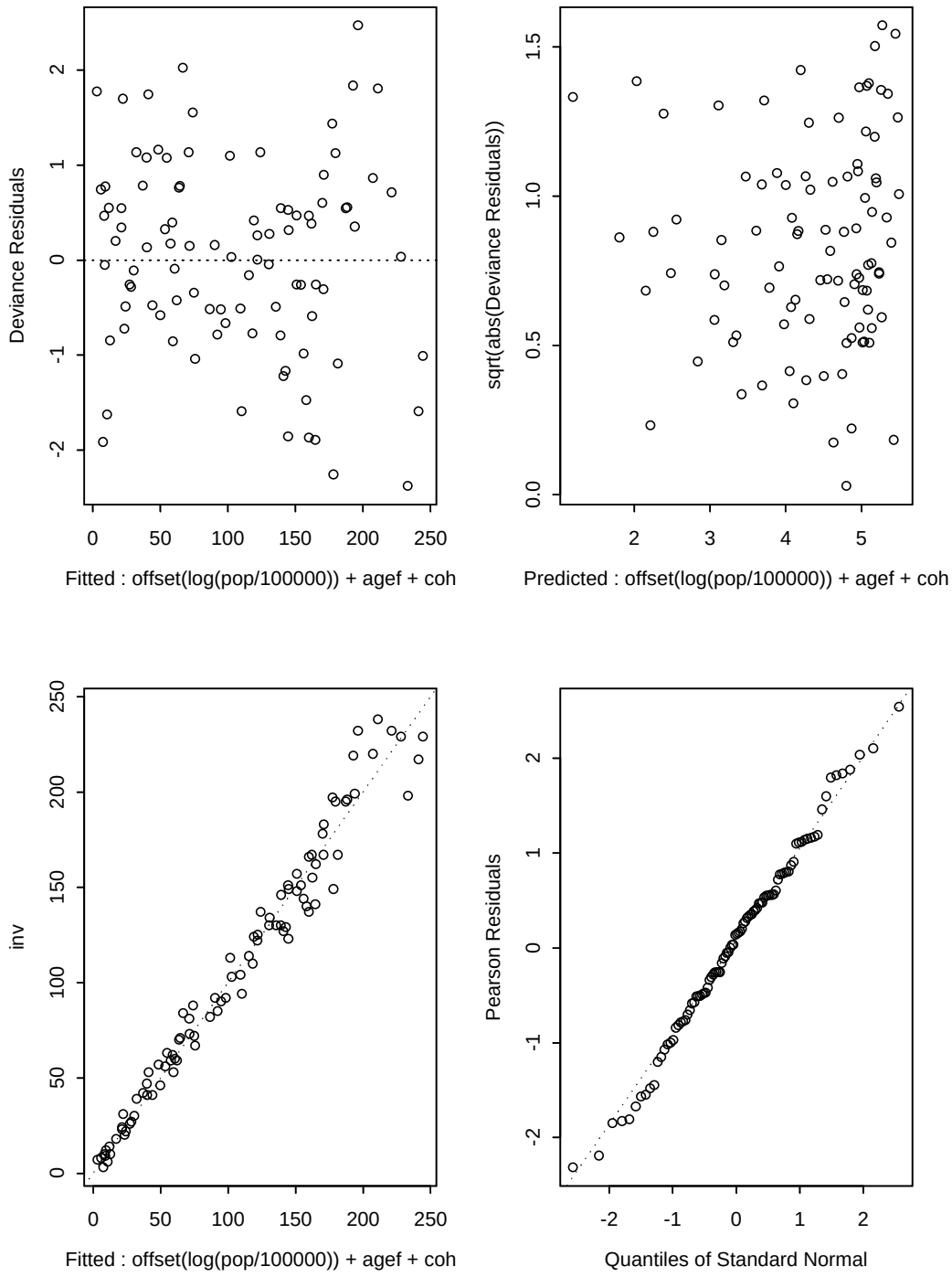
Un deuxième modèle GAM à deux facteurs a été testé en remplaçant la période par la cohorte (sous forme de régression locale). La forme de la relation du nombre de cas avec la cohorte est manifestement linéaire (Figure 5.7.). Aussi, un modèle GLM est testé avec la cohorte apparaissant de façon linéaire (Figure 5.7.).

Figure 5.7. Modélisation du nombre de cas incidents de cancers invasifs dans le Bas-Rhin. Les facteurs sont l'âge sous forme qualitative et la cohorte sous forme quantitative. Les figures du haut représentent le modèle GAM avec une fonction de régression locale de la cohorte. Les figures du bas représentent le modèle GLM avec la cohorte sous forme linéaire.



L’adéquation du modèle est de bonne qualité (Figure 5.8.).

Figure 5.8. Adéquation du modèle GLM avec l’âge sous forme de variable qualitative et la cohorte sous forme linéaire.



Le modèle final contient l'âge sous forme de variable factorielle, la période sous forme logarithmique et la cohorte sous forme linéaire (Figure 5.9.). Là aussi, le modèle représente correctement les données (Figure 5.10.).

Figure 5.9. Modélisation du nombre de cas incidents de cancers invasifs dans le Bas-Rhin. Les facteurs sont l'âge (forme qualitative), le logarithme de la période et la cohorte sous forme linéaire.

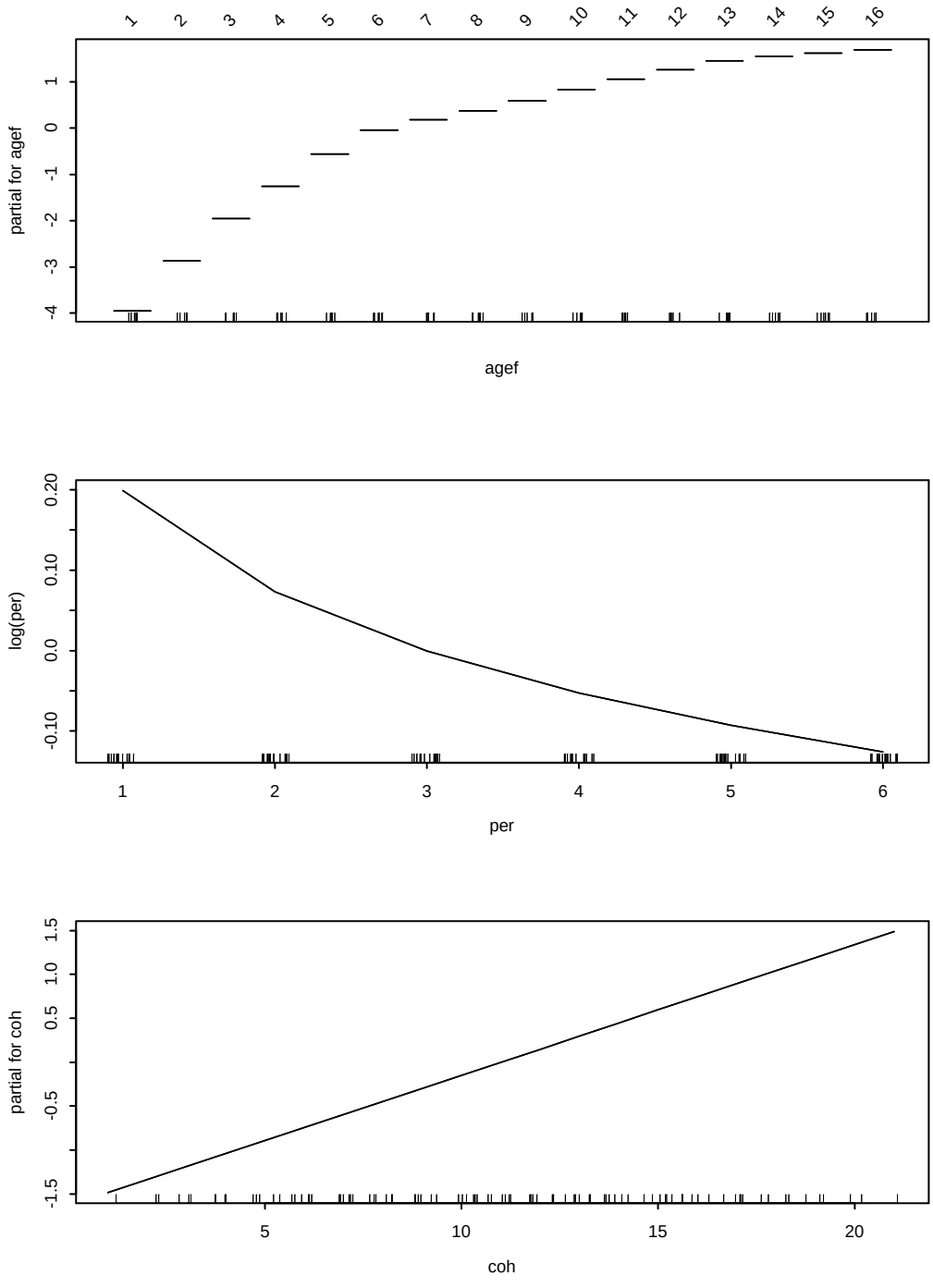
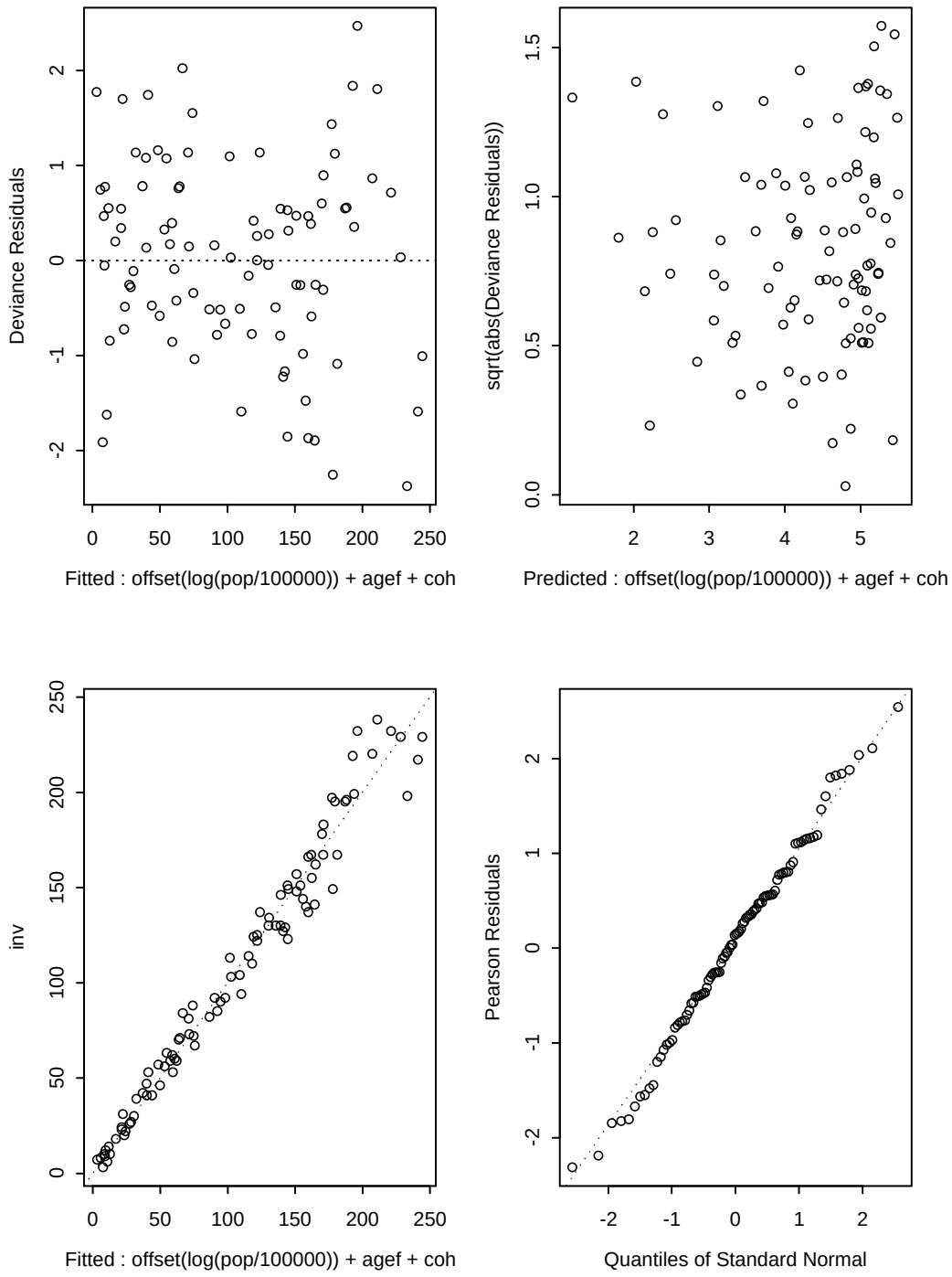


Figure 5.10. Adéquation du modèle GLM avec l'âge sous forme de variable qualitative, le logarithme de la période et la cohorte sous forme linéaire.



Les trois modèles, âge-période, âge-cohorte et âge-période-cohorte, sont testés grâce au critère d'Akaïke (voir Annexe 3.). Le meilleur des trois est le modèle complet.

Sur la base de cette modélisation, les nombres de cas incidents ont été prédits pour l'ensemble des périodes et, en particulier, pour la période 1997-2005 (Tableau 5.7.).

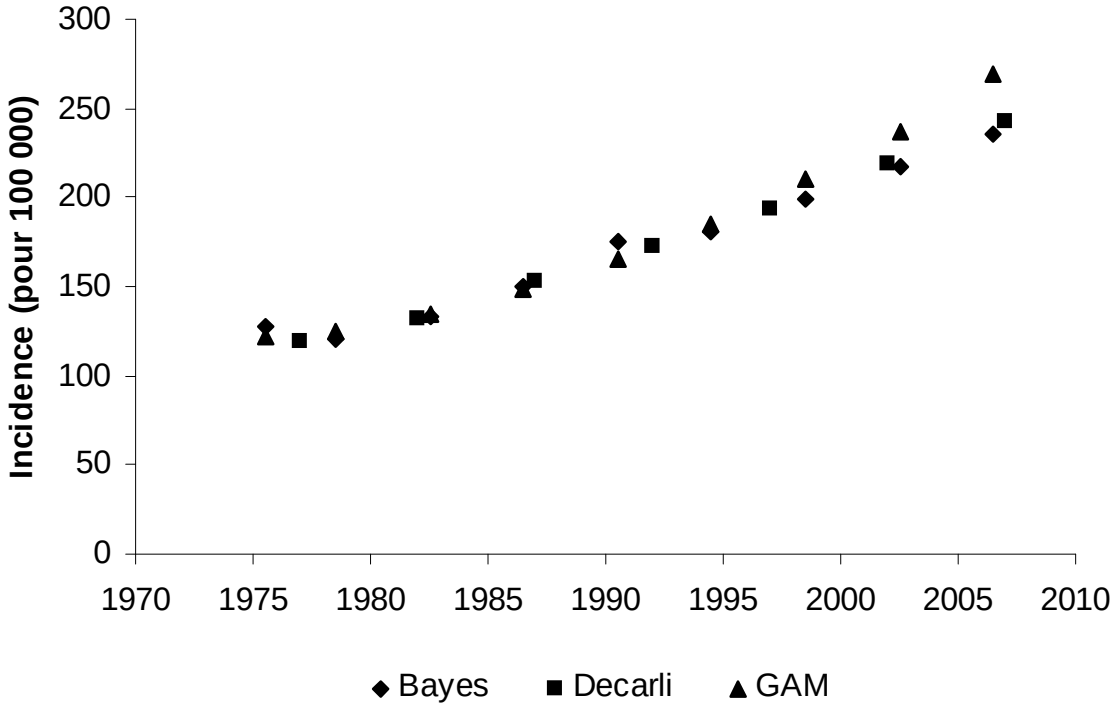
Tableau 5.7. Cancer du sein invasif : incidences spécifiques et nombre de cas incidents chez les femmes dans le Bas-Rhin, 1975-2008. Prédiction par modèle GAM-GLM complet.

Âge\Période	1975-1976	1977-1980	1981-1984	1985-1988	1989-1992	1993-1996	1997-2000	2001-2004	2005-2008
25-28	7,1 (4)	6,7 (8)	6,6 (8)	7,3 (9)	8,6 (11)	9,3 (12)	11,1 (14)	11,9 (13)	13,8 (16)
29-32	17,2 (7)	16,3 (17)	17,7 (21)	19,5 (24)	21,4 (27)	24,1 (31)	27,3 (35)	31,2 (39)	35,1 (38)
33-36	34,9 (14)	34,7 (28)	37,7 (39)	41,6 (49)	46,2 (57)	52,4 (66)	59 (75)	66,7 (85)	75,6 (94)
37-40	59,2 (24)	60,1 (48)	64,9 (52)	71,6 (74)	80,6 (95)	89,8 (111)	101,8 (127)	115,2 (146)	130,7 (166)
41-44	103,1 (44)	105,3 (86)	113,5 (90)	125,8 (100)	139,3 (144)	157,1 (185)	177,1 (216)	200,2 (248)	227,3 (286)
45-48	147,9 (63)	152,7 (130)	164,9 (133)	181,4 (142)	201,4 (159)	225,9 (232)	255,7 (296)	289,5 (350)	328,7 (404)
49-52	160,8 (68)	165,4 (139)	178 (148)	195,5 (155)	217,6 (169)	245,1 (192)	276,3 (279)	313,4 (359)	355,2 (425)
53-56	169,5 (67)	172,1 (144)	186,1 (153)	204,7 (168)	227,8 (179)	256,4 (197)	289,1 (222)	327,1 (326)	371,4 (420)
57-60	180,8 (40)	184,3 (122)	197,7 (161)	217,9 (175)	243,2 (196)	273,3 (212)	308 (232)	348,4 (264)	396,4 (390)
61-64	197,6 (71)	202,7 (102)	218,1 (139)	240,7 (190)	268,2 (210)	302 (238)	339,9 (256)	385,3 (284)	437,7 (325)
65-68	212,9 (79)	217,8 (153)	235,3 (113)	259,5 (159)	288,6 (220)	323,5 (245)	366 (278)	413,3 (303)	469,8 (338)
69-72	225,6 (76)	230,4 (159)	248 (161)	274,6 (122)	304,1 (176)	342,7 (249)	385,7 (278)	437,3 (319)	496,4 (351)
73-76	235,7 (65)	241,1 (141)	259,5 (159)	285,7 (168)	317,4 (130)	357,3 (191)	403,2 (271)	456,6 (308)	519,3 (357)
77-80	224 (43)	227,4 (98)	245,8 (117)	270,5 (137)	302 (151)	339,4 (121)	382,7 (181)	432,8 (260)	492,4 (300)
81-84	207,7 (23)	213 (55)	227,5 (70)	251,7 (89)	280,1 (110)	315,5 (125)	355,9 (101)	402,8 (157)	456,5 (228)
85-88	195,4 (10)	192,9 (23)	213,3 (32)	230,5 (43)	259,9 (59)	289 (76)	327,9 (89)	372,6 (74)	421,2 (120)
25-88	122,4 (698)	124,9 (1453)	134,6 (1596)	148,4 (1804)	165,1 (2093)	185,7 (2483)	209,6 (2950)	237,2 (3535)	269,3 (4258)

L'incidence est exprimée en nombre de cas pour 100 000. Entre parenthèses, figurent les nombres de cas incidents. Dans la dernière ligne figurent les incidences pour toute la population, standardisées selon la population européenne et, entre parenthèses, le nombre total de cas incidents. Il s'agit de valeurs prédites pour toutes les périodes.

Les valeurs des incidences spécifiques et les nombres de cas incidents prédits par l'approche GAM / GLM sont plus importants que ceux qui ont été estimés par la méthode bayésienne de référence (Tableau 5.7.). Pour les incidences standardisées, la différence est moins importante mise à part pour la dernière période de projection (2005-2008) (Figure 5.11).

Figure 5.11. Comparaison des prévisions issues de l'approche bayésienne, de l'approche de Decarli et La Vecchia et du modèle GAM.



6. CONCLUSION

« Nous sommes plus intensément affectés à l'égard d'un objet que nous imaginons dans un futur rapproché que si nous l'imaginions dans un futur très lointain ; et le souvenir d'une chose que nous imaginons dans un passé récent nous affecte aussi plus intensément que si nous l'imaginions dans un passé plus lointain. »[10]

[10] Baruch Spinoza. L'éthique. De la servitude humaine. Proposition 10. Introduction, traduction, notes et commentaires, index de R Misrahi. 2ème éd. Paris, Presses universitaires de France, 1990.

La méthode retenue pour réaliser la prévision de l'incidence des cancers est fondée sur un modèle âge-période-cohorte, une approche bayésienne et un calcul par échantillonnage de Gibbs.

Le modèle âge-période-cohorte a été choisi parce qu'il permettait – dans presque tous les cas – de réaliser la prévision de l'incidence du cancer sans préjuger de l'existence de facteurs de risque et, dans le cas où ces facteurs sont connus, sans être contraint de prévoir l'évolution de ceux-ci. Une double prévision augmenterait, en effet, les risque d'estimer les taux futurs avec peu de précision et avec beaucoup de risque de se tromper. La modélisation de l'incidence sur la base du trépied « âge » (qui rend compte de la durée de l'exposition), « période » (qui rend compte de la variation – à court terme – de l'exposition de la population à des facteurs de risques ou protecteurs) et la « cohorte» (qui rend compte de l'exposition des sujets à un ou des moments clefs de leur vie) permet de prendre en compte les variations passées et, par conséquent, futures pour la prévision, des différents facteurs de risque. Les variables « âge », « période » et « cohorte » jouent un rôle de *proxi* de ces expositions.
Dans les travaux portant sur la tendance de l'incidence, le problème lié à la difficulté d'identifier les effets respectifs des trois facteurs du modèle est une limite de l'analyse et nécessitent la mise en place de contraintes. Dans le domaine de la prévision, le problème de l'identification ne perturbe pas le calcul puisque chaque incidence est le produit des trois facteurs et que seul ce produit est retenu sans chercher à le décomposer.

L'utilisation d'un modèle autorégressif appliqué aux effets âge, période et cohorte a un double intérêt : il permet de tenir compte de la dépendance existant entre les valeurs successives de la série longitudinale « nombre de cas » et sert de lissage (Mezzetti, 1999) tout en réduisant l'instabilité des cohortes extrêmes (Breslow, 1993 ; Berzuini, 1994 ; Bashir, 2001). L'approche bayésienne a un double intérêt également : elle permet de résoudre des problèmes inaccessibles à l'intégration numérique et offre une méthode élégante d'estimation (Breslow, 1993) et de diminution de l'amplitude (Richardson, 1993 ; Liu, 1994) des intervalles de confiance.
Un ensemble de difficultés éventuelles pourrait affecter la qualité de la prévision. Ce sont des incertitudes de mesures liées aux données d'incidence et d'effectif de population. La fiabilité de l'enregistrement des cas incidents semble être résolu, aujourd'hui, grâce à la qualité des méthodes d'enregistrement des registres des cancers (MacLennan, 1978 ; Powell, 1991 ; Skeet, 1991 ; Hédelin, 1992). La précision de l'estimation et de la prédiction de la taille de la population est relativement bien maîtrisé à ce jour par l'INSEE.

La prévision a été appliquée aux incidences de quelques cancers considérés comme étant de poids non négligeable en tant que « pourvoyeurs » de cas incidents mais aussi comme étant affectés d'une mortalité importante (chapitre 4. Exemples). Les cancers invasifs du sein, les tumeurs *in situ* du col de

l'utérus, du côlon chez l'homme et du poumon dans les deux sexes augmenteront nettement dans les huit à dix années prochaines. Les cancers du côlon et du rectum, chez la femme, augmenteront plus faiblement. Le cancer du rectum chez l'homme restera relativement stable. Quant au cancer invasif du col de l'utérus, il devrait voir sa tendance décroître.
À l'exception du cancer invasif du col de l'utérus, le nombre des cas incidents augmente dans tous les cancers étudiés et quel que soit le sexe (de façon plus ou moins importante selon la nature du cancer).

La méthode de référence (méthode bayésienne) a été comparée, pour le cas particulier du cancer invasif du sein, avec d'autres méthodes basées sur le modèle âge-période-cohorte : la méthode de Decarli et La Vecchia, la méthode du CIRC et l'analyse par modèle additif généralisé ou GAM (chapitre 5 : Autres méthodes). De façon générale, l'intervalle de confiance « produit » par la méthode bayésienne est plus petit que celui qui est issu des autres méthodes. Quant à l'estimation de l'incidence future (incidence standardisée et incidences spécifiques), la méthode de Decarli et La Vecchia donne des résultats équivalents alors que la méthode du CIRC donne des valeurs plus basses tant pour les incidences spécifiques que pour l'incidence standardisée (le modèle polynomial pourrait être vraisemblablement amélioré). L'analyse par GAM donne des niveaux d'incidences spécifiques plus élevés que la méthode de référence mais, globalement, la différence est moins nette pour l'incidence standardisée.

Pour certains cancers, un dépistage a été mis en place au cours de la période d'étude. Or, la première vague du dépistage peut induire une augmentation artificielle par la révélation de cas incidents de façon anticipée puis, quand le nombre de femmes dépistées devient stable, par le retour de l'incidence à la tendance antérieure (Kessler, 1991 ; Feuer, 1992 ; Wun, 1995). Si l'analyse ne prend pas en compte cette perturbation, les incidences prévues peuvent être surestimées. Un ensemble d'analyses de sensibilité a été pratiqué dans chaque cas et n'a pas montré de grandes perturbations des incidences prédites.

La variable aléatoire représentant le nombre de cas incidents est supposée suivre une loi de Poisson. Ceci implique l'hypothèse d'une égalité entre sa variance et son espérance. Il peut être utile de tenir compte de l'existence éventuelle d'une surdispersion (Cox, 1983 ; Breslow, 1984). Certaines méthodes ont été utilisées pour en tenir compte (Wedderburn, 1974) et parmi celles-ci, des approches bayésiennes (Hartigan, 1969). Il serait intéressant d'intégrer cette dimension supplémentaire au modèle utilisé ici.
Les prévisions réalisées dans le cadre de ce travail ont tenu compte d'une ensemble de facteurs purement temporels (âge, période et cohorte de naissance). L'algorithme de Gibbs a été utilisé par ailleurs pour explorer des données spatio-temporelles (Sun, 2000 ; Knorr-Held, 2000). À l'avenir, il

devrait être possible d'intégrer des mesures complémentaires de type géographique aux données passées et futures d'incidence.

Enfin, les prévisions réalisées ici concernent uniquement les données du Bas-Rhin. Il est envisagé d'associer, par la suite, l'ensemble des registres français à l'analyse en incluant une variable « centre ».

7. BASES STATISTIQUES

« L'avenir, c'est du passé en préparation. »[11]

[11] Pierre Dac. L'os à moelle. Textes réunis et présentés par Michel Laclos. Paris, René Julliard, 1963.

7.1. Lois de probabilité

Les lois de probabilité utilisées dans les modèles bayésiens utilisés dans l'analyse sont essentiellement les lois de Poisson et les lois gamma (Monfort, 1980 ; Saporta, 1990 ; Simar, 1998 a).

7.1.1. Loi de Poisson

$X \sim P(\lambda)$

$$f(X = x) = e^{-\lambda} \frac{\lambda^{x}}{x!}$$

$x \in N, \lambda > 0$

$$E(X) = V(X) = \lambda$$

7.1.2. Densité gamma

7.1.2.1. Définition préliminaire : fonction gamma

Elle est définie pour $x > 0$

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$$

7.1.2.2. Densité gamma

$$X \sim \Gamma(p,\theta)$$

$$f(x) = \frac{\theta^{p} x^{p-1} e^{-\theta x}}{\Gamma(p)}$$

$x \geq 0, p > 0, \theta > 0$

$$E(X) = \frac{p}{\theta} \; ; \; V(X) = \frac{p}{\theta^2}$$

(Cette loi, bien qu'appliquée à une variable continue, présente une analogie avec la loi de Poisson, plus particulièrement quand θ vaut 1.

7.2. Modèle linéaire généralisé et modèle additif généralisé

Le modèle linéaire généralisé (GLM) est une structure mixte, mathématique et statistique permettant de modéliser les relations entre une grandeur mesurée d'intérêt et un ensemble de facteurs dits explicatifs.

7.2.1. Problématique

En général, les données se présentent de la façon suivante :
Soit y une grandeur étudiée et x_1, x_2, …, x_j, …, x_p, p grandeurs explicatives ou cofacteurs.
n observations sont réalisées qui peuvent être résumées dans le tableau suivant :

Tableau 7.1. Tableau des données

n° observ	**y**	$\mathbf{x_1}$	$\mathbf{x_2}$	…	$\mathbf{x_j}$	…	$\mathbf{x_p}$
1	y_1	x_{11}	x_{12}	…	x_{1j}	…	x_{1p}
2	y_2	x_{21}	x_{22}	…	x_{2j}	…	x_{2p}
…	…	…	…	…	…	…	…
i	y_i	x_{i1}	x_{i2}	…	x_{ij}	…	x_{ip}
…	…	…	…	…	…	…	…
n	y_n	x_{n1}	x_{n2}	…	x_{nj}	…	x_{np}

Soient

$\mathbf{y} = (y_1, y_2, \ldots, y_i, \ldots, y_n)'$ le vecteur des observations à expliquer ;

$\mathbf{Y} = (Y_1, Y_2, \ldots, Y_i, \ldots, Y_n)'$ le vecteur des variables aléatoires correspondantes ;

$\mathbf{x_j} = (x_{1j}, x_{2j}, \ldots, x_{ij}, \ldots, x_{nj})'$, j = 1, 2, …p, les p vecteurs cofacteurs que l'on peut résumer sous la forme de la matrice **X** :

$$\mathbf{X} = \begin{bmatrix} x_{11} & \ldots & x_{1j} & \ldots & x_{1p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ x_{i1} & \ldots & x_{ij} & \ldots & x_{ip} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ x_{n1} & \ldots & x_{nj} & \ldots & x_{np} \end{bmatrix}$$

La modélisation de ces observations consiste à considérer les n mesures y_i comme les réalisations de n variables aléatoires Y_i

7.2.2. Le modèle linéaire général.

Il met en relation l'espérance des Y_i avec les valeurs prises par les p cofacteurs de la façon suivante :

$$E[Y_i] = \mu_i = \sum_{j=1}^{p} x_{ij}\beta_j \quad i = 1, 2, \ldots, n$$

Où les x_{ij} sont les valeurs prises par les covariables (voir tableau ci-dessus) et les β_j, p coefficients à déterminer.

Ceci peut s'écrire aussi de façon vectorielle :

$$E[\mathbf{Y}] = \boldsymbol{\mu} = \sum_{j=1}^{p} \mathbf{x}_j\beta_j$$

avec $\mu = (\mu_1, \mu_2, \ldots, \mu_i, \ldots, \mu_n)'$

Et de façon matricielle :

$$E(\mathbf{Y}) = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$$

avec $\beta = (\beta_1, \beta_2, \ldots, \beta_j, \ldots, \beta_p)'$

Les variables Y_i sont supposées être normales indépendantes de variance constante σ^2.

Ceci peut s'exprimer de façon équivalente : les erreurs ε_i (telles que $\varepsilon_i = Y_i - \mu_i$, i = 1, 2, …, n) sont des variables aléatoires indépendantes et présentant les propriétés suivantes :

$E(\varepsilon_i) = 0$ et $var(\varepsilon_i) = \sigma_i^2$, $\forall\, i \in [1,n]$

$Cov(\varepsilon_i,\varepsilon_k) = 0, \forall (i,k) \in [1,n]^2$

On suppose aussi en général que ε_i suit une loi normale donc que :

$\varepsilon_i \sim N(0, \sigma^2)$, $\forall i \in [1,n]$

7.2.3. Le modèle linéaire généralisé

Le modèle linéaire généralisé (GLM) est une extension du modèle linéaire général. La formalisation et les hypothèses ont été formulés par Nelder et Wedderburn (Nelder, 1972) puis McCullagh et Nelder (McCullagh, 1989).

Outre les notations et définitions précédentes :

g désignera, dans la suite, une fonction ;

ηi, η sont appelés prédicteurs ;

L_{exp} est une loi quelconque de la famille exponentielle.

Le GLM s'écrit alors :

$$Y_i \sim L_{exp} \quad \text{et} \quad \mu_i = E[Y_i]$$
$$\eta_i = g(\mu_i)$$
$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$$

Ou encore, en utilisant la notation vectorielle et matricielle :

$$\mathbf{Y} \sim L_{exp} \quad \text{et} \quad \boldsymbol{\mu} = E[\mathbf{Y}]$$
$$\boldsymbol{\eta} = g(\boldsymbol{\mu})$$
$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

De plus l'écriture générale de la densité d'une loi exponentielle s'écrit :

$$f(y|\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)$$

θ est le « paramètre naturel » de la famille exponentielle appelé paramètre « canonique ».

ϕ est le paramètre de dispersion.

θ est une fonction de μ : $\theta(\mu)$, définie par la relation

$$E(Y) = \mu = b'(\theta) = \frac{\partial b(\theta)}{\partial \theta}$$

La variance de Y est :

$$\text{var}(Y) = b''(\theta)a(\phi) = \frac{\partial^2 b(\theta)}{\partial \theta^2} a(\phi)$$

La variance de Y est donc composée de deux termes distincts. L'un est une fonction de θ, l'autre est une fonction de ϕ.

Puisque θ dépend de μ, b″(θ) dépend de μ.

b″(θ) et est appelé fonction de variance. La fonction de variance est définie par la relation :

$$v(\mu) = b''(\theta) = \frac{\partial^2 b(\theta)}{\partial \theta^2}$$

Quant à la fonction a(ϕ), elle se présente sous la forme :

$$a(\phi) = \frac{\phi}{\omega}$$

Le paramètre de dispersion ϕ qui s'écrit aussi σ^2, est constant alors que ω est un poids propre à chaque observation.

Le choix de la fonction de lien dépend de la loi exponentielle : pour chaque famille exponentielle, il existe une fonction de lien « naturelle » dite « fonction de lien canonique » telle que :

$$\theta = \theta(\mu) = \eta = X\beta$$

Donc telle que

$$g(\mu) = \theta(\mu)$$

Les modalités relatives aux différentes familles exponentielles (paramètre de dispersion, fonction de lien, fonction de variance, etc.) sont présentées dans de nombreux ouvrages (Lindsey, 1997).

7.2.4. Le modèle additif généralisé

Le modèle additif généralisé (GAM) est une extension du modèle linéaire généralisé. Sa structure n'est pas fondamentalement différente du GLM. Dans son expression, le prédicteur linéaire est remplacé par un prédicteur dit additif (Hastie, 1990 ; Schwartz, 1994 a ; Schwartz, 1996).

$$Y_i \sim L_{exp} \qquad \text{et} \qquad \mu_i = E[Y_i]$$

$$\eta_i = g(\mu_i)$$

$$\eta_i = \sum_{j=1}^{p} f_j(x_{ij})$$

Les fonctions fj du prédicteur additif sont des fonctions quelconques d'une ou plusieurs variables. Elles peuvent être paramétriques (polynomiales, trigonométriques, etc.), semi-paramétriques ou non paramétriques (fonctions *splines*, fonctions de lissage non paramétriques comme les *locally-weighted running-line smoother* ou fonction *loess* du logiciel S-Plus qui disposent d'un « réglage » de la largeur de la fenêtre de lissage et permettent la prise en compte plus ou moins fine des variations temporelles de la variable) (Cleveland, 1988 ; Fahrmeir, 1996 ; Venables, 1997). L'intérêt des fonctions de lissage est de pouvoir s'adapter plus fidèlement à la forme des relations entre variables expliquées (ici, le nombre de cas incident) et les variables explicatives. La modélisation n'impose pas de forme *a priori* à la relation et n'est pas affectée par la rigidité des fonctions paramétriques. Ce modèle est largement utilisé pour l'analyse des relations entre la pollution atmosphérique et la santé (Schwartz, 1994 b ; Kelsall, 1997, Eilstein, 2001).

Les GAM ont ainsi été utilisés comme technique d'analyse de la tendance de l'incidence par modèle âge-période-cohorte dans le domaine du cancer (Schwartz, 1990 a et b ; Heuer, 1997).

Il est de coutume d'utiliser le critère d'Akaike ou AIC (Akaike Information Criterion) (Akaike, 1973 ; Akaike 1978 a et b) pour choisir le meilleur modèle GAM (celui qui présente l'AIC le plus faible). L'AIC est une déviance pénalisée par ajout d'un terme fonction du nombre de paramètres :

AIC = - 2 * (maximum de la log-vraisemblance) + 2 * (nombre de paramètres)

7.3. Rappels de statistique bayésienne

L'approche bayésienne est aujourd'hui encore au centre d'un ensemble de débats et polémiques *philosophico-statistiques*. Mais, outre l'intérêt d'inclure des connaissances antérieures à l'analyse au cours de l'analyse, elle se prête avec souplesse à un ensemble de calculs, d'inférence et de créations d'échantillons. La littérature, volumineuse à présent, dans les ouvrages généraux (Saporta, 1990 ; Gourieroux, 1989) ou spécialisés (Robert, 1992) s'enrichit tous les jours de nouveaux développements (Société française de statistique, 1998).

7.3.1. Théorème de Bayes

7.3.1.1. Première formule de Bayes

A et B sont des événements

$P(A) \neq 0$

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

7.3.1.2. Deuxième formule de Bayes

Cette deuxième formulation est déduite de la première grâce à l'utilisation du théorème des probabilités totales.

- Pour deux événements

A et B sont des événements

$P(A) \neq 0$

$\overline{B}$ désigne le complémentaire de B

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\overline{B})P(\overline{B})}$$

- Cas général (plusieurs événements)

Soit $\{B_i\}$, i de 1 à n (n pouvant être infini), un système complet d'événements. Alors :

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)}$$, n pouvant être infini.

Remarque :

{B_i}est un système complet d'événements si :

$$\forall i \neq j, B_i \cap B_j = \varnothing \quad \text{et} \quad \bigcup B_i = \Omega$$

7.3.2. Application du théorème de Bayes aux variables aléatoires

7.3.2.1. Variables discrètes

X et Y sont deux variables aléatoires prenant un ensemble de valeurs respectives x_1, x_2, …, x_n et y_1, y_2, …, y_p, n et p pouvant être infinis.

$$P(X = x_i | Y = y_j) = \frac{P(Y = y_j | X = x_i) P(X = x_i)}{\sum_{k=1}^{n} P(Y = y_j | X = x_k) P(X = x_k)}$$

7.3.2.2. Variables continues

X et Y sont deux variables aléatoires prenant un ensemble de valeurs respectives x et y, réelles.

Soient f(x) et g(x), les densités de probabilité respectives de X et Y (densités marginales)

Soient f(x|y) et g(y|x), respectivement, les densités de probabilité de X conditionnellement à Y et de Y conditionnellement à X

Le théorème de Bayes s'exprime de la façon suivante :

$$f(x|y) = \frac{g(y|x) f(x)}{\int g(y|x) f(x) dx}$$

$$g(y|x) = \frac{f(x|y) g(y)}{\int f(x|y) g(y) dy}$$

Remarque :

La densité de la loi jointe de (Y,X) est :

$$h(x,y) = f(x|y)g(y) = g(y|x)f(x)$$

La densité de loi marginale de x est :

$$f(x) = \int h(x,y)\, dy$$
$$= \int f(x|y) g(y) dy$$

La densité de loi marginale de y est :

$$g(y) = \int h(x,y)\, dx$$
$$= \int g(y|x) f(x) dx$$

7.3.3. Inférence bayésienne

Soit Y, une variable aléatoire, y une réalisation de Y et Θ, le paramètre de la distribution de probabilité de Y. Le paramètre Θ est supposé inconnu mais distribué selon une loi de probabilité dont la densité est f(θ). Cette loi est dite loi *a priori*.

Soit g(y|θ), la densité de probabilité de Y conditionnellement à Θ.

Si f(θ|y) est la densité de probabilité de Θ conditionnellement à Y, le théorème de Bayes s'écrit en adaptant l'expression du 7.3.2.2. :

$$f(\theta|y) = \frac{g(y|\theta) f(\theta)}{\int g(y|\theta) f(\theta) d\theta}$$

f(θ|y) est la densité de la loi *a posteriori* de Θ

Remarque :

La densité de la loi jointe de (Θ,Y) est

$$h(\theta,y) = g(y|\theta) f(\theta)$$

La densité de loi marginale (loi prédictive) de Y est

$$g(y) = \int h(\theta, y)\, d\theta$$
$$= \int g(y|\theta) f(\theta) d\theta$$

Appliquée à vraisemblance, l'expression du théorème de Bayes devient :

$$f(\theta|y) = \frac{l(\theta|y) f(\theta)}{\int l(\theta|y) f(\theta) d\theta}$$

puisque $l(\theta|y) = g(y|\theta)$

Remarque : si la variance de la loi *a priori* est infinie (loi *a priori* non informative) ou si l'effectif de l'échantillon est infini, toute l'information provient de l'échantillon.

À partir de la formule précédente (expression de $f(\theta|y)$), il est possible de calculer la densité de probabilité marginale *a posteriori* de Θ (ou les densités marginales *a posteriori* des composantes de θ), l'espérance (conditionnelle) *a posteriori* et les matrices de covariances (conditionnelles) *a posteriori* de Θ par intégration dans le cas continu et par sommation dans le cas discret.
L'espérance *a posteriori* :

$$E(\theta|y) = \int \theta \, f(\theta|y) d\theta$$

La matrice de covariance *a posteriori* :

$$\mathrm{cov}(\theta|y) = \int (\theta - E(\theta|y))(\theta - E(\theta|y))' f(\theta|y) d\theta$$

Remarque : les intégrations ci-dessus ne sont analytiquement possibles que pour des modèles simples comme le modèle linéaire général (normal). Pour d'autres modèles (dont les plus importants en pratique), il n'existe pas de loi *a priori* conjuguée qui permette de calculer ces intégrales de façon analytique. Aussi il faut avoir recours à des intégrations numériques ou des procédures de type Monte Carlo (voir le paragraphe 7.4. : Échantillonnage de Gibbs). De plus, comme le paramètre à estimer peut avoir une dimension élevée, les calculs ne sont pas faciles. Un ensemble de méthodes a été proposée comme l'intégration de Gauss-Hermite ou l'intégration de type Monte Carlo (voir plus bas « Échantillonnage de Gibbs »).

7.3.4. Famille naturelle conjuguée, loi *a priori*, loi *a posteriori*

7.3.4.1. Choix de la loi *a priori*

La loi de distribution *a priori* est choisie souvent au sein d'un ensemble de lois classiques telles que la loi normale, une loi gamma, etc. Pour le paramètre, il convient de choisir une loi *a priori* conjuguée ; c'est une loi de densité qui est de forme semblable à la celle de la vraisemblance $l(y_1, y_2, \ldots, y_n ; \theta)$.
Dans ces conditions, la loi *a posteriori* associée a la même forme que la loi *a priori*.

7.3.4.2. Notion de naturelle conjuguée

Une famille naturelle conjuguée à un processus d'échantillonnage est un ensemble de lois assez riche et souple pour représenter l'information *a priori* sur les paramètres du modèle. Elle doit, d'autre part, se combiner facilement avec la fonction de vraisemblance dans la formule de Bayes afin de donner, *a posteriori*, une loi sur les paramètres du modèle qui appartienne à la même famille (Simar, 1998 b).

Le processus de Poisson, par exemple, peut se prêter à ce calcul (Simar, 1998 b) :

Soient n intervalles de temps donnés u_i, i de 1 à n.
Soit x_i, le nombre d'événements qui se sont produits pendant l'intervalle de temps u_i,.
La variable aléatoire correspondante X_i suit une loi de Poisson :

$$X_i \mid \lambda, u_i \sim P(\lambda u_i)$$

La vraisemblance s'écrit :

$$L(\lambda, ,x_1, x_2, \ldots x_n, u_1, u_2, \ldots, u_n) \propto \lambda^s e^{-\lambda t}$$

Avec $s = \sum_{i=1}^{n} x_i$ et $t = \sum_{i=1}^{n} u_i$

Si le problème est inversé :

Soit k un nombre d'événements donné.
Soit t le temps au bout duquel k événements sont observés.
Il est, alors, possible de montrer que t suit une loi Gamma :

$$t \mid \lambda, k \sim \Gamma(\lambda, k)$$

De plus, dans l'expression de la vraisemblance ci-dessus le noyau est un noyau d'une densité gamma
Donc la famille naturelle conjuguée de l'échantillonnage de Poisson est la famille des densités Gamma.

Comme pour le processus de Poisson, il est possible de montrer, de façon générale, que si la loi *a priori* est une loi gamma, alors la loi *a posteriori* est une loi gamma (Simar, 1978 b).

En effet, la densité *a priori* s'écrit :

$$p(\lambda) \propto \lambda^{s_0-1} e^{-\lambda t_0}$$
$$\lambda \sim \Gamma(t_0, s_0)$$

La combinaison de la vraisemblance vue plus haut et le noyau de la densité *a priori* donne l'expression suivante :

$$p(\lambda|s,t) \propto \lambda^{s+s_0-1} e^{-\lambda(t+t_0)}$$

Ceci signifie que :

$$\lambda|s,t \sim \Gamma(t_0+t, s_0+s)$$

Donc la densité *a posteriori* appartient à la famille des densités gamma.

7.4. Échantillonnage de Gibbs

7.4.1. Problématique

Comme il a été dit plus haut (paragraphe 7.3. : Rappels de statistiques bayésiennes), les intégrations nécessaires au calcul de la densité de la loi *a posteriori* de Θ, (f(θ|y)), de l'espérance *a posteriori* (E(θ|x)) et de la matrice de covariance *a p osteriori* (cov(θ|x)) ne sont pas toujours accessibles au calcul analytique ou numérique[12].
En effet les approximations analytiques consistent à remplacer la loi *a posteriori* par un équivalent asymptotique normal ou des approximations plus précises (approximation de Laplace). Ces méthodes

---

[12] La densité de la loi *a posteriori* de Θ s'écrit : $f(\theta|y) = \frac{g(y|\theta) f(\theta)}{\int g(y|\theta) f(\theta) d\theta}$ ou $f(\theta|y) = \frac{l(\theta|y) f(\theta)}{\int l(\theta|y) f(\theta) d\theta}$

si l'on utilise la vraisemblance l ;

L'espérance *a posteriori* s'écrit : $E(\theta|y) = \int \theta\, f(\theta|y) d\theta$;

La matrice de covariance *a posteriori* s'écrit : $\mathrm{cov}(\theta|y) = \int (\theta - E(\theta|y))(\theta - E(\theta|y))' f(\theta|y) d\theta$.

Pour établir ces fonctions, il faut d'abord calculer les intégrales qui apparaissent dans leurs expressions.

peuvent être efficaces mais le fait qu'elles soient fondées sur une approche asymptotique restreint leur application à des cas particuliers (Kass, 1989).

Quant aux intégrations numériques, elles se basent sur les méthodes telles que les développements de la méthode classique de Simpson (approximation de la fonction initiale par des fonctions polynomiales). Ces méthodes deviennent instables lorsque le paramètre est de dimension supérieure à 3 ou exigent une régularité suffisante de la fonction initiale (Robert, 1998).
Il est cependant possible d'utiliser des méthodes non analytiques comme les procédures de type Monte Carlo et plus particulièrement la technique d'échantillonnage de Gibbs (*Gibbs sampling*).
D'autre part, les calculs font intervenir des inversions de matrices de dimensions importantes ; là également, cette dernière méthode peut surseoir aux méthodes classiques (Harville, 1999).

L'échantillonnage de Gibbs fait partie des méthodes de Monte Carlo par chaîne de Markov (méthodes MCMC). Il se base sur deux ensembles de notions, les processus markoviens et les méthodes de Monte Carlo. Il est donc utile de préciser ces notions.

7.4.2. Notion de chaîne de Markov

7.4.2.1. Processus aléatoire

Le processus aléatoire est une extension du processus déterministe en ce qu'il rajoute la notion de variable aléatoire à la définition de celui-ci. Le processus est dit déterministe si son passé et son avenir (son évolution future) sont définis de façon unique par son état présent. L'ensemble des états du processus s'appelle « l'espace des phases » et est défini par un système d'équations différentielles (Arnold, 1988). Il est à noter que, malgré la notion de déterminisme qui les caractérise, certains de ces processus peuvent adopter un comportement très exotique comme ceux des systèmes chaotiques (May, 1976 ; Murray, 1989).

Le processus aléatoire est défini de la façon suivante :

Ω est l'univers des possibles (ensemble des expériences ω), T est l'ensemble des instants (ensemble des temps t).
Un processus aléatoire est une application de $\Omega \times T$ dans R, ensemble des réels.
Ainsi si Y est un processus aléatoire :

$$(\omega, t) \xrightarrow{Y} Y(\omega, t)$$

avec $(\omega,t) \in \Omega \times T$ et $Y(\omega,t) \in R$

Un processus aléatoire peut être considéré, plus simplement comme une suite de variables aléatoires indexées par le temps, Y_t, t = 1, 2, ….

Deux propriétés intéressantes peuvent caractériser un processus aléatoire : la stationnarité et l'ergodicité.

Un processus aléatoire Y_t est **stationnaire** au sens strict si la loi de probabilité de Y_t est la même, quel que soit t. Un processus aléatoire Y_t est stationnaire au sens large si $E(Y_t)$ = cte, $var(Y_t)$ = cte quel que soit t et si $cov(Y_t,Y_s) = \varphi(|t-s|)$.
Un processus aléatoire Y_t est dit **ergodique** si toutes ses caractéristiques peuvent être déterminées à partir d'une de ses trajectoires.

7.4.2.2. Processus de Markov

7.4.2.2.1. Définition

Un processus aléatoire, Y_t, est une **chaîne de Markov** (ou **processus markovien**) si Y_t peut prendre successivement un ensemble de valeurs appartenant à $E = \{E_1, E_2, \ldots, E_N\}$ que l'on appelle l'ensemble des états possibles du processus. L'ensemble E peut être de cardinal infini.
De plus, le processus est « muni » d'un ensemble de probabilités de transition p_{ij}, $(i,j) \in [1 ;N]^2$. La mesure p_{ij} est la probabilité que le processus se trouve dans l'état E_j au stade n, sachant qu'il se trouvait dans l'état E_i, au stade n-1.

Un processus aléatoire, Y_t, est une **chaîne de Markov du premier ordre** si la probabilité de Y_t, conditionnellement à son passé (Y_{t-1}, Y_{t-2}, …, Y_{t-k}, …) ne dépend que de Y_{t-1} (Lindsey, 1997). Ce qui peut s'écrire :

$$P(Y_t \mid Y_{t-1}, Y_{t-2}, \ldots, Y_{t-k}, \ldots) = P(Y_t \mid Y_{t-1})$$

Ce qui est équivalent à l'écriture suivante :

$$Y_t \sim K(Y_{t-1}, Y_t)$$

K étant une densité conditionnelle appelée noyau de transition (de densité de probabilité).

7.4.2.2.2. Exemple

Un processus autorégressif du premier ordre AR(1), par exemple, est un processus markovien. En effet :

$$Y_t = \varphi\, Y_{t-1} + \varepsilon_t$$

Remarque : dans ce cas particulier, le processus est appelé processus markovien linéaire.

Lorsque φ est égal à 1, Y_t est la marche aléatoire.

7.4.3. Méthodes de Monte Carlo

Les méthodes de Monte Carlo s'adressent à deux ensemble de problèmes : l'intégration et l'échantillonnage (Tanner, 1996 ; Gilks, 1996).

7.4.3.1. Intégration

L'une des intégrales classiques auxquelles aboutit l'analyse bayésienne est celle qui exprime la loi marginale de Y :

$$g(y) = \int g(y|x) f(x) dx$$

L'expression h(x,y) = g(y|x)f(x) est la loi de distribution jointe de X et Y.

Or, le produit g(y|x)f(x) n'est pas toujours intégrable par les méthodes analytiques. Il est possible alors d'avoir recours aux méthodes de Monte Carlo (Geweke, 1989).

7.4.3.1.1. Cas général

L'intégrale g(y) peut être approchée par l'expression :

$$\hat{g}_n(y) = \frac{1}{n} \sum_{i=1}^{n} g(y|x_i)$$

avec (x_1, …, x_i, …, x_n), échantillon tiré de la distribution de densité de probabilité f(x).

Quand $n \to \infty$:

$$\hat{g}_n(y) \xrightarrow{p.s.} g(y)$$

Ceci signifie l'expression approchée de g(y) converge presque sûrement vers elle[13].

L'erreur standard estimée de $\hat{g}_n(y)$ est :

$$\frac{1}{\sqrt{n}} \sqrt{\frac{\sum_{i=1}^{n} \left[f(y|x_i) - \hat{g}_n(y) \right]^2}{n-1}}$$

Il est, cependant, parfois difficile d'extraire un échantillon de la distribution f(x). Dans ce cas, il est possible, comme il sera vu dans le paragraphe suivant, d'avoir recours à une approximation de f(x) par une fonction facile à échantillonner.

7.4.3.1.2. Méthode de Monte Carlo de fonction d'importance I(x)

La fonction I(x), utilisée comme approximation, doit être une fonction de densité suffisamment proche de f(x). I(x) est appelée fonction d'importance (Fahrmeir, 1996).

La procédure utilisée par cette méthode est la suivante :

1°) (x_1, …x_j, …, x_n) est extrait de la distribution de densité I(x)

2°) Calcul de $\hat{g}(y) = \frac{1}{w} \sum_{i=1}^{n} w_i f(y|x_i)$ avec $w_i = \frac{g(x_i)}{I(x_i)}$ et $w = \sum_{i=1}^{n} w_i$

Cette méthode pondère $f(y|x_i)$ et donne ainsi plus d'importance aux données où I(x)<f(x).

---

[13] La notation $X_n \xrightarrow{p.s.} X$ traduit la notion de convergence presque sure (p.s. ; a.s. pour *almost surely*, en anglais) :

Une suite de variables aléatoires réelles, $(X_n)_{n \in N}$, est dite converger presque sûrement vers la variable aléatoire X, quand $n \to \infty$ si

$$Xn(\omega) \to X(\omega), \forall\ \omega \in N^c \text{ avec } P(N) = 0.$$

L'erreur standard est dans ce cas (Geweke 1989) :

$$\frac{1}{w}\sqrt{\sum_{i=1}^{n} w_i^2 \left[f\left(y|x_i\right) - \hat{g}_n\left(y\right)\right]^2}$$

7.4.3.2. Échantillonnage

7.4.3.2.1. Cas général

Les méthodes de Monte Carlo permettent d'extraire un échantillon (y_1, …y_j, …, y_n) de la distribution marginale $g(y) = \int g(y|x) f(x) dx$ de la façon suivante :

1.a) Extraction d'une valeur x_1 de la distribution f(x)
1.b) Extraction d'une valeur y_1 de la distribution $g(y|x_1)$
…….
i.a) Extraction d'une valeur x_i de la distribution f(x)
i.b) Extraction d'une valeur y_i de la distribution $g(y|x_i)$
…….
n.a) Extraction d'une valeur x_n de la distribution f(x)
n.a) Extraction d'une valeur y_n de la distribution $g(y|x_n)$

Cette méthode aboutit à l'élaboration de n couples (x_i,y_i), i = 1, 2, …, n. Les n couples sont un échantillon extrait de la distribution jointe de X et Y, h(x,y) = g(y|x)f(x). L'échantillon (y_1, …, y_i, …, y_n) est extrait de la distribution de densité marginale g(y).

7.4.3.2.2. Méthode de pondération et échantillonnage par la méthode d'acceptation-rejet

Comme il a été dit plus haut, lorsqu'il est difficile d'extraire un échantillon de la distribution f(x), cette dernière peut être remplacée par I(x), densité proche de f(x) et dont on peut extraire facilement un échantillon (Gilks, 1991 ; Gilks 1992).
En effet, l'intégrale précédente peut s'écrire :

$$g(y) = \int g(y|x) \frac{f(x)}{I(x)} I(x) dx$$

La procédure utilisée reproduit la procédure générale en remplaçant f(x) par I(x) puis en pondérant y :

1.a) Extraction d'une valeur x_1 de la distribution I(x)

1.b) Extraction d'une valeur y'_1 de la distribution $g(y|x_1)$

1.c) Calcul de $y_1 = \frac{w_1}{w} y'_1$

…….

i.a) Extraction d'une valeur x_i de la distribution I(x)

i.b) Extraction d'une valeur y'_i de la distribution $g(y|x_2)$

i.c) Calcul de $y_i = \frac{w_i}{w} y'_i$

…….

n.a) Extraction d'une valeur x_n de la distribution I(x)

n.a) Extraction d'une valeur y'_n de la distribution $g(y|x_1)$

n.c) Calcul de $y_n = \frac{w_n}{w} y'_n$

Dans le cas particulier où il existe M, constante positive telle que $\frac{g(x)}{I(x)} \leq M, \forall x$, il est possible d'améliorer la qualité de l'échantillonnage tiré de la densité g(y) (Tanner, 1996).

La méthode s'appelle algorithme d'acceptation-rejet et se présente de la façon suivante :

1.a) Extraction d'une valeur x_1 de la distribution I(x)

1.b) Extraction d'une valeur u de la distribution uniforme sur [0 ;1], u_1 indépendant de x_1

1.c) Si $u_1 \leq \frac{g(x_1)}{M.I(x_1)}$, x_1 est retenu, sinon la procédure 1.a), 1.b), 1.c) est répétée

1.b) Extraction d'une valeur y_1 de la distribution $g(y|x_1)$

……..

i.a) Extraction d'une valeur x_i de la distribution I(x)

i.b) Extraction d'une valeur u de la distribution uniforme sur [0 ;1], u_i indépendant de x_i

i.c) Si $u_i \leq \frac{g(x_i)}{M.I(x_i)}$, x_i est retenu, sinon la procédure i.a), i.b), i.c) est répétée

i.b) Extraction d'une valeur y_i de la distribution $g(y|x_i)$

……..

n.a) Extraction d'une valeur x_n de la distribution I(x)

n.b) Extraction d'une valeur u de la distribution uniforme sur [0 ;1], u_n indépendant de x_n

n.c) Si $u_n \leq \frac{g(x_n)}{M.I(x_n)}$, x_n est retenu, sinon la procédure n.a), n.b), n.c) est répétée

n.b) Extraction d'une valeur y_n de la distribution $g(y|x_n)$

7.4.4. Échantillonnage de Gibbs

7.4.4.1. Techniques de Monte Carlo par chaînes de Markov

Les techniques de Monte Carlo par chaînes de Markov (Markov Chain Monte Carlo : MCMC) ont pour but, comme les méthodes de Monte Carlo de façon générale, de générer des échantillons à partir d'une distribution de probabilité jointe afin d'estimer les espérances des distributions marginales à partir des moyennes des échantillons obtenus (Hastings, 1970 ; Tanner, 1996 ; Gilks, 1996). Mais dans les MCMC, les différentes valeurs x_i ne sont pas générées de façon parallèle (i.e. indépendamment) ; elles sont générées de façon *chaînée* c'est-à-dire déduites les unes des autres. Il existe deux MCMC remarquables : la méthode de Metropolis-Hastings (*Metropolis algorithm*) (Metropolis, 1953 ; Knorr-Held, 2000) et l'échantillonnage de Gibbs. Ces méthodes peuvent être utilisées lorsqu'il est difficile de simuler la loi g(y|x) par des algorithmes habituels comme la méthode d'acceptation-rejet vue plus haut. Très appréciées des chercheurs en intelligence artificielle parce qu'elles peuvent approcher et simuler les états successifs d'un réseau de neurones (Bellazzi, 1997), ces méthodes sont largement utilisées dans tous les domaines couverts par l'approche bayésienne (Smith, 1993).

7.4.4.2. Échantillonnage de Gibbs

7.4.4.2.1. Préliminaires et principe

Soit $E = \{E_1, E_2, \ldots, E_N\}$, l'ensemble des états possibles d'un processus de Markov c'est-à-dire l'ensemble des états dans lesquels le processus peut se trouver (N peut être infini).

Soit F_t, l'état dans lequel se trouve le processus après t transitions.

F_t peut être égal à l'un quelconques des états E_i avec une probabilité non nulle, ce qui revient à dire que le processus peut se trouver dans n'importe lequel des états E_i avec une probabilité non nulle et peut s'écrire :

$$P(F_t = E_i) > 0$$

Ceci est aussi équivalent à supposer que p_{ij}, la probabilité de transition de F_t de l'état E_i à l'état E_j, est supérieure à 0.

N'importe quel état pouvant être atteint au terme de t transitions, il est intéressant de savoir quelle est la probabilité pour cet état d'être atteint.

La propriété d'ergodicité (voir 7.4.2.1. : Processus aléatoire) permet d'approcher cette probabilité lorsque t est grand. En effet, la probabilité pour que l'état F_t soit égal à l'état E_i , $P(F_t = E_i)$, est égale à la fréquence de passage du processus par l'état E_i au cours d'une trajectoire quelconque. Il est ainsi possible d'approcher F_t en lançant le processus à partir de n'importe quel état, en le laissant évoluer selon un grand nombre d'itérations et de calculer la proportion de passage du processus par l'état E_i.

Il est ainsi possible d'approcher les $P(F_t = Ei)$, quelque soit i et de déterminer l'état le plus probable.

Si le processus Y_t est constitué d'un ensemble de p variables aléatoires :

$$Y_t = \{Y_{1,t}, Y_{2,t}, \ldots, Y_{p,t}\}$$

Soit $y_{j,t}$, la valeur prise par la variable Yj,t.

L'état de Y_t est désigné par F_t.

Ici $F_t = (Y_{1,t} = y_{1,t}, Y_{2,t} = y_{2,t}, \ldots, Y_{p,t} = y_{p,t}\}$

La probabilité de transition de F_t à F_{t+1} est représentée par un ensemble de p probabilités conditionnelles partielles, $P(Y_{j,t+1} \mid F_t)$ avec :

$$P(Y_{j,t+1} \mid F_t) = P(Y_j \mid Y_j, j \neq i)$$

Ceci signifie que la probabilité de transition ne dépend pas de l'ordre de l'itération mais des probabilités conditionnelles des différentes variables entre elles.

Les probabilité conditionnelles $P(Y_j \mid Y_j, j \neq i)$ sont obtenues à partir de la distribution jointe $P(Y_1, Y_2, \ldots, Y_p)$.

Ainsi, pour trouver les valeurs *a posteriori* des Y_j, les plus probables, il suffit de faire évoluer le processus selon un grand nombre d'itérations à partir de n'importe quel état et d'opérer comme précédemment.

L'échantillonnage de Gibbs est entièrement inspiré de ces notions.

Remarque : quand l'ensemble des valeurs prises par Y (espace des états) est fini, la distribution conditionnelle permettant d'obtenir Yt peut-être écrite comme une matrice de transition de probabilité. Quand l'ensemble des valeurs prises par Y n'est pas fini, la distribution conditionnelle peut-être vue comme un noyau de transition de probabilité.

7.4.4.2.2. Définition

Cette méthode est une généralisation de l'algorithme de Metropolis-Hastings (Metropolis, 1953). Elle a été introduite par Hastings (1970), utilisée par Geman et Geman (1984) dans le cadre de l'analyse de processus spatiaux mettant en jeu de nombreuses variables (analyse et reconstruction d'images) puis appliquée aux statistiques bayésiennes par Gelfand et Smith (Gelfand, 1990 b), Gelfand *et al.* (Gelfand, 1990 a), Carlin *et al.* (1992) et Gelfand *et al.* (1992). Il s'agit d'un algorithme markovien permettant de fabriquer des échantillons extraits d'une distribution conjointe donnée ; cette méthode procède par échantillonnage itératif à partir des densités conditionnelles déduites de la probabilité conjointe (Casella, 1992 ; Gelman, 1995 ; Tanner, 1996).
Dans le cas présent (statistiques bayésiennes), la densité conjointe est une densité conjointe *a posteriori*.

Remarque : il faut, bien sur, supposer que l'ensemble des distributions conditionnelles utilisées pour créer l'échantillon déduit de la probabilité conjointe déterminent celle-ci de façon unique (Besag, 1974 ; Besag, 1995 ; Hobert, 1997).

7.4.4.2.3. Mise en œuvre

Soit un ensemble de variables aléatoires Θ_1, Θ_2, …,Θ_i, …, Θ_p et un ensemble de données y (Dellaportas, 1993). La densité de probabilité conjointe *a posteriori* s'écrit $f(\theta_1, \theta_2, \ldots,\theta_i, \ldots, \theta_p|y)$. La densité conditionnelle de chacun des éléments Θ_i conditionnellement aux autres (et à y) s'écrit :

$$f(\theta_i|\theta_1, \theta_2,\ldots, \theta_{i-1}, \theta_{i+1}, \ldots,\theta_p, y),\ i = 1, 2, \ldots p.$$

L'échantillonnage de Gibbs attribue au p-uplet $(\theta_1, \theta_2, \ldots,\theta_i, \ldots, \theta_p)$, de façon itérative, un ensemble de valeurs extraites des distributions précédentes à partir de la connaissance de leurs densités conditionnelles.

1°) La procédure attribue tout d'abord au (p-1)-uplet $(\theta_2, \ldots, \theta_i, \ldots, \theta_p)$ des valeurs dites de début : $\theta_2^{(0)}, \ldots, \theta_i^{(0)}, \ldots, \theta_p^{(0)}$.

2°) Puis il génère le premier échantillon :

$\theta_1^{(1)}$ est généré à partir de la distribution de densité conditionnelle $f(\theta_1 \mid \theta_2^{(0)}, \theta_3^{(0)}, \ldots, \theta_p^{(0)}, y)$

$\theta_2^{(1)}$ est généré à partir de la distribution de densité conditionnelle $f(\theta_2 \mid \theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_p^{(0)}, y)$

…….

$\theta_p^{(1)}$ est généré à partir de la distribution de densité conditionnelle $f(\theta_p \mid \theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_{p-1}^{(1)}, y)$

3°) Puis il génère le deuxième échantillon :

$\theta_1^{(2)}$ est généré à partir de la distribution de densité conditionnelle $f(\theta_1 \mid \theta_2^{(1)}, \theta_3^{(1)}, \ldots, \theta_p^{(1)}, y)$

$\theta_2^{(2)}$ est généré à partir de la distribution de densité conditionnelle $f(\theta_2 \mid \theta_1^{(2)}, \theta_3^{(1)}, \ldots, \theta_p^{(1)}, y)$

…….

$\theta_p^{(2)}$ est généré à partir de la distribution de densité conditionnelle $f(\theta_p \mid \theta_1^{(2)}, \theta_2^{(2)}, \ldots, \theta_{p-1}^{(2)}, y)$

…….

k°) Puis il génère le k-ème échantillon :

$\theta_1^{(k)}$ est généré à partir de la distribution de densité conditionnelle $f(\theta_1 \mid \theta_2^{(k-1)}, \theta_3^{(k-1)}, \ldots, \theta_p^{(k-1)}, y)$

$\theta_2^{(k)}$ est généré à partir de la distribution de densité conditionnelle $f(\theta_2 \mid \theta_1^{(k)}, \theta_3^{(k-1)}, \ldots, \theta_p^{(k-1)}, y)$

…….

$\theta_p^{(k)}$ est généré à partir de la distribution de densité conditionnelle $f(\theta_p \mid \theta_1^{(k)}, \theta_2^{(k)}, \ldots, \theta_{p-1}^{(k)}, y)$

Si le vecteur $(\theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_p^{(i)})$ est noté $\theta^{(i)}$, la suite $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(i)}, \ldots, \theta^{(k)}$ est une réalisation d'une chaîne de Markov. La probabilité de transition de $\theta^{(i)}$ à $\theta^{(i+1)}$ s'écrit (noyau de transition ou *kernel*) :

$K(\theta^{(i)}, \theta^{(i+1)}) = f(\theta_1 \mid \theta_2^{(i)}, \theta_3^{(i)}, \ldots, \theta_p^{(i)}, y)\, f(\theta_2 \mid \theta_1^{(i)}, \theta_3^{(i)}, \ldots, \theta_p^{(i)}, y) \ldots\ldots f(\theta_p \mid \theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_{p-1}^{(i)}, y)$

Geman et Geman (1984) ont montré que, sous des conditions peu contraignantes, quand $k \to \infty$, le vecteur $(\Theta_1^{(k)}, \Theta_2^{(k)}, \ldots, \Theta_p^{(k)}) \to (\Theta_1, \Theta_2, \ldots, \Theta_p)$.

Ainsi, pour k suffisamment grand, $(\theta_1^{(k)}, \theta_2^{(k)}, \ldots, \theta_p^{(k)})$ peut être considéré comme une observation extraite de la distribution de densité conjointe $f(\theta_1, \theta_2, \ldots, \theta_p)$ et pour i = 1, 2, …, p, $\theta_i^{(k)}$ peut être considéré approximativement comme une observation extraite de la distribution de densité marginale $f(\theta_i)$.

D'autre part :

$$\frac{1}{k}\sum_{i=1}^{k} f\left(\theta^{(i)}\right) \xrightarrow{\text{p.s.}} \int f(\theta) f(\theta|y) d\theta \text{ quand } k \to \infty$$

Et :

$$\sqrt{k}\left(\frac{1}{k}\sum_{i=1}^{k} f\left(\theta^{(i)}\right) - \int f(\theta) f(\theta|y) d\theta\right) \xrightarrow{\text{faiblement}} N\left(0, \sigma_f^2\right)$$

La procédure ci-dessus (de 1°) à k°)) est répétée un grand nombre de fois et génère N valeurs $\theta_i^{(k)}$ pour i = 1, 2, …, p :

$$\theta_{i,1}, \theta_{i,2}, \ldots, \theta_{i,j}, \ldots, \theta_{i,N}$$

Ces N valeurs constituent un échantillon de taille N, extrait de la distribution de densité marginale $f(\theta_i)$. Si N est suffisamment grand, il est possible d'estimer à partir de cet échantillon, les densités conjointes et marginales des θ_i ainsi que leurs moments par des méthodes empiriques ou des méthodes utilisant des outils de lissage de la densité.

La densité marginale de θ_i, par exemple peut être obtenue de façon empirique, en calculant la moyenne des densités conditionnelles :

$$f(\theta_i) = \frac{1}{N}\sum_{j=1, j\neq i}^{N} f\left(\theta_i | \theta_j^{(k)}\right)$$

Des outils de lissage de la densité tels les estimateurs calculés par la méthode du noyau peuvent également être utilisés pour estimer les densités marginales.

Remarque : ces méthodes (MCMC, en général et échantillonnage de Gibbs, en particulier) ressemblent à première vue au « bootstrap » mais des différences de fond existent entre les principes de ces deux ensembles de techniques : dans l'approche classique (fréquentiste) du « bootstrap », les paramètres sont considérés comme des constantes et les données comme réalisation de variables aléatoires. Dans l'approche bayésienne l'ensemble des paramètres et des données sont issus de variables aléatoires. D'autre part le « bootstrap » échantillonne les variables correspondant aux données afin d'estimer des intervalles de confiance. Dans les MCMC, ce sont les paramètres qui sont échantillonnés, conditionnellement aux données (Clayton, 1991).

RÉFÉRENCES

« Quand on proclama que la Bibliothèque comprenait tous les livres, la première réaction fut un bonheur extravagant. Tous les hommes se sentirent maîtres d'un trésor intact et secret. Il n'y avait pas de problème personnel ou mondial dont l'éloquente solution n'existât quelque part : dans quelque hexagone. L'univers se trouvait justifié, l'univers avait brusquement conquis les dimensions illimitées de l'espérance. »[14]

[14] Jorge Luis Borges. La biliothèque de Babel. In : Jorge Luis Borges, Fictions. Traduction : P Verdevoye, Ibarra et Roger Caillois. Paris, Gallimard 1965.

Les références ci-après ont été classées par ordre alphabétique. Il peut être utile de pouvoir retrouver l'endroit du texte où est inséré le renvoi à une référence donnée. Aussi, à la suite de cette dernière, figurent :

- **Entre parenthèses, l'écriture du renvoi, telle qu'elle se présente dans le texte ;**
- **Précédée d'une flèche, la numérotation du (ou des) paragraphe(s) où apparaît le renvoi.**

Aareleid T, Linsalu M, Rahu M, Baburin A. Lung cancer in Estonia in 1968-87: time trends and public health implications. European J Cancer Prevention 1994;3:419-25. (Aareleid, 1994). ➔ 4.4.1.

Ahsan H, Neugut AI, Bruce JN. Trends in incidence of primery malignant brain tumors in USA, 1981-1990. Int J Epidemiol 1995;24:1078-85. (Ahsan, 1995). ➔ 2.2.2.2.3.2.(a)(α)

Aitkin M, Anderson D, Francis Brian, Hinde J. Statistical modelling in GLIM. JB Copas, AP Dawid, GK Eagleson, DA Pierce, BW Silverman. eds. Clarendon Press - Oxford, Oxford 1989. (Aitkin, 1989). ➔ 2.2.2.2.3.2.(b)(α) note 6

Aitio A, Day NE, Heseltine E, Kaldor J, Miller AB, Parkin DM, Riboli E. Cancer: causes, occurrence and control. L Tomatis ed. Lyon, IARC Scientific Publications, 1990, 352 p. (Aitio, 1990). ➔ 1. ; 2.2.2.2.2.1.

Akaike H. Fitting autoregressive models for prediction. Annals Institute Statist Math 1969;21:243-7. (Akaike, 1969). ➔ 2.2.1.1.2.

Akaike H. Information theory and an extension of the maximum likelihood principle, in Second International Symposium on Information Theory. (eds B.N. Petrov and F. Czáki). Akademiai Kiadó, Budapest 1973, pp 267-81. (Akaike, 1973). ➔ 7.2.4.

Akaike H. On entropy maximisation principle. In: Applications of Statistics. Krisnaiah eds. North Holland 1977;27-41. (Akaike, 1977). ➔ 2.2.1.1.2.

Akaike H. A new look at the Bayes procedure. Biometrika 1978;65:53-9. (Akaike, 1978 a). ➔ 7.2.4.

Akaike H. On the likelihood of a time series model. Statistician 1978;27:215-235. (Akaike, 1978 b). ➔ 7.2.4.

Albert I, Jais JP. Gibbs sampler for the logistic model in the analysis of longitudinal binary data. Statist Med 1998;17:2905-21. (Albert, 1998). ➔ 3.2.2.1.3.

Alexander FE, Stockton D, Harvey C. Appendix I: Materials and statistical methodology. In Scottish Executive Health Department. Cancer scenarios: an aid to planning cancer services in Scotland in the next decade. R. Black, D. Stockton eds. The Scottish Executive, Edinburgh: 2001;342-349. (Alexander, 2001). ➔ 4.3.

Andreasen AH, Andersen KW, Madsen M, Mouridsen H, Olesen KP, Lynge E. Regional trends in breast cancer incidence and mortality in Denmark prior to mammographic screening. Br J Cancer 1994;70:133-7. (Andreasen, 1994). ➔ 2.2.2.2.3.2.(b)(ε)

Aragonés N, Pollán M, López-Abente et al. Time trend and age-period-cohort effects on gastric cancer incidence in Zaragoza and Navarre, Spain. J Epidemiol Community Health 1997;51:412-7. (Aragonés, 1997). ➔ 2.2.2.2.3.2.(b)(η)

Arnold V. Équations différentielles ordinaires. 4th ed. Moscou, Éditions Mir 1984. Traduction française, 1988. 334 p. (Arnold, 1988). ➔ 7.4.2.1.

Australian Institute of Health and Welfare (AIHW) and Australasian Association of Cancer Registries (AACR). Cancer in Australia 1991-1994 (with projections to 1999). Camberra: 1998. AIHW (Cancer Series N° 7). 90 p. (Australian Institute oh Health, 1998). ➔ 4.3.

Bashir SA, Estève J. Projecting cancer incidence and mortality using Bayesian age-period-cohort models. J Epidemiol Biostatist 2001;6:287-96. ➔ 2.2.2.2.3.2.(b)(δ) ; 3.2.1. ; 6.

Beard CM, Spencer RJ, Weiland LH, O'Fallon WM, Melton LJ-III. Trends in colorectal cancer over a half century in Rochester, Minnesota, 1940 to 1989. Ann-Epidemiol 1995;5(3):210-4. (Beard, 1995). ➔ 4.3.1.

Bellazzi R, Magni P, De Nicolao G. Dynamic probabilistic networks for modelling and identifying dynamic systems: a MCMC approach. Intell Data Analysis 1997;1:245-262. (Bellazzi, 1997). ➔ 7.4.4.1.

Benhamiche AM, Faivre J, Menegoz F, Grosclaude P. Les cancers digestifs en France à l'aube de l'an 2000. Hépato-Gastro 1997;4 (Suppl):8-12. (Benhamiche, 1997). ➔ 4.3.4.

Benhamiche AM. Cancer du côlon : épidémiologie descriptive et groupes à risque élevé. Gastroenterol Clin Biol 1998; 22(Suppl):S3-S11. (Benhamiche, 1998). ➔ 4.3.4.

Bergé P, Pomeau Y, Vidal Y. L'ordre dans le chaos ; vers une approche déterministe de la turbulence. Paris, Hermann 1988, 352 p. (Bergé, 1988). ➔ 1.

Berman IR, Ulcickas M, Yood SM, Grant RL. Cancer trends unique to Georgia. J Med Assoc Ga 1993 ;82:29-33. (Berman, 1993). ➔ 4.3.1.

Berzuini C, Clayton D. Bayesian analysis of survival on multiple time scales. Statist Med 1994;13:823-38. (Berzuini, 1994). ➔ 2.2.2.2.3.2.(b)(δ) ; 2.2.2.2.3.2.(b)(η) ; 3.2.1. ; 6.

Besag JE. Spatial interaction and the statistical analysis of lattice systems. J R Statist Soc Series B, 1974;36:192-236. (Besag, 1974). ➔ 7.4.4.2.2.

Besag JE, Green PJ, Higdon D, Mengersen KL. Bayesian computation and stochastic systems. Statist Sci 1995;10:3-66. (Besag, 1995). ➔ 7.4.4.2.2.

Bornefalk A, Persson I, Bergström R. Trends in breast cancer mortality among Swedish women 1953-92: analysis by age, period and birth cohort. British J cancer 1995;72:493-7. (Bornefalk, 1995). ➔ 2.2.2.2.3.2.(b)(γ)

Boutron-Ruault M. Alimentation et cancérogenèse colorectale : éléments pour une prévention primaire. Gastroenterol Clin Biol 1998;22 (Suppl 3):S12-20. (Boutron-Ruault, 1998). ➔ 4.3.4.

Bowerman RJ. Alaska Native cancer epidemiology in the Arctic. Public Health 1998;112:7-13. (Bowerman, 1998). ➔ 4.3.1.

Box GEP, Jenkins GM. Time series analysis: forecasting and control. San Francisco, Hoden-Day 1970. (Box, 1970). ➔ 2.2.1.1.2.

Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. J Am Stat Ass 1993;88:9-25. (Breslow, 1993). ➔ 2.2.2.2.3.2.(b)(δ) ; 2.2.2.2.3.2.(b)(η) ; 3.2.1. ; 4.1.4. ; 4.2.4. ; 6.

Breslow NE. Extra-Poisson variation in log-linear models. Appl Statist 1984;33:38-44. (Breslow, 1984). ➔ 2.2.2.1.2. ; 6.

Bresson G, Pirotte A. Économétrie des séries temporelles. Théorie et applications 1st ed. Paris Presses universitaires de France, 1995. 58 p. (Bresson, 1995). ➔ 2.1.1. ; 2.2.1.2.

Brockwell PJ, Davis RA. Times series. Theory and methods. 1st ed. New York : Springer-Verlag, 1987. 519 p. (Brockwell, 1987). ➔ 2.2.1.1.2.

Brookmeyer R, Day NE. Two-stage models for the analysis of cancer screening data. Biometrics 1987;43:657-69. (Brookmeyer, 1987). ➔ 4.1.4.

Capocaccia R, De Angelis R, Frova L, Gatta G, Sant M, Micheli A et al. Estimation and projections of colorectal cancer trend in Italy. Int J Epidemiol 1997;26:924-32. (Capocaccia, 1997). ➔ 4.3.4.

Carlin BP, Gelfand AE, Smith AFM. Hierarchical Bayesian analysis of changepoint problems. Appl Statist 1992;42:389-405. (Carlin, 1992). ➔ 7.4.4.2.2.

Casella G, George EI. Explaining the Gibbs sampler. The Am Statistician 1992;46:167-74. (Casella, 1992). ➔ 7.4.4.2.2.

Cherruault Y, Loridan P. Modélisation et méthodes mathématiques en biomédecine. Paris, Masson 1977, 144 p. (Cherruault, 1977). ➔ 2.3.

Ciatto S, Cecchini S, Iossa A , Grazzini G, Bonardi R, Zappa M et al. Trends in cervical cancer incidence in the district of Florence. Eur J Cancer 1995;31A(3):354-5. (Ciatto, 1995). ➔ 4.2.1. ; 4.2.4.

Clayton DG. A Monte Carlo method for Bayesian inference in frailty models. Biometrics 1991;47:467-85. (Clayton, 1991). ➔ 3.2.2.1.3. ; 7.4.4.2.3.

Clayton D, Schifflers E. Models for temporal variation in cancer rates. I. Age-period and age-cohort models. Stat Med 1987;6:449-67. (Clayton, 1987 a). ➔ 2.2.2.2.3.2. (b)(γ) ; 4.3.

Clayton D, Schifflers E. Models for temporal variation in cancer rates. II. Age-period-cohort models. Stat Med 1987;6:469-81. (Clayton, 1987 b). ➔ 2.2.2.2.3.2. (b)(γ) ; 4.3.

Cleveland WS, Devlin SJ. Locally-weighted regression, An approach to regression analysis by local fitting. J Am Statist Assoc 1988;83:597-610. (Cleveland, 1988). ➔ 7.2.4.

Coleman MP, Estève J, Damiecki P, Arslan A, Renard H. Trends in cancer incidence and mortality. Lyon: IARC Scientific Publications. 1993 (Vol. 121). (Coleman, 1993). ➔ 2.2.2.2.3.2.(a)(α) ; 4.1.1. ; 4.2.1. ; 4.3.1. ; 4.3.4. ; 4.4.1.

Coleman MP, Lutz JM. Tendances évolutive du cancer et prévision des besoins : vers une meilleurs utilisation des registres du cancer. Rev Epidemiol Sante Publique 1996;44:S2-S6. (Coleman, 1996). ➔ 1. ; 3.1.2.

Comité national des registres. Rapport d'activité, février 1986-1989. INSERM, DGS. INSERM, Paris: 1989. 48 p. (Comité national des registres, 1989). ➔ 3.1.2.

Comité national des registres. Rapport d'activité, 1996-1999. DGS, INSERM, InVS, Paris: 2000. 176 (Comité national des registres, 2000). ➔ 3.1.2.

Coutrot B, Droesbeke JJ. Les méthodes de prévision. 2nd ed. Paris : Presses universitaires de France, 1990. 128 p. (Que sais-je ?). (Coutrot, 1990). ➔ 2.2.1.1.2.

Cox DR. Some remarks on overdispersion. Biometrika 1983;70:269-74. (Cox, 1983). ➔ 6.

Day NE, Walter SD. Simplified models of screening for chronic disease: estimation procedures from mass screening programmes. Biometrics 1984;40 :1-14. (Day, 1984). ➔ 4.1.4.

Decarli A, La Vecchia C. Age, period and cohort models : review of knowledge and implementation in GLIM. Rivistica di Statistica Applicata 1987;20:397-410. (Decarli, 1987). ➔ 2.2.2.2.3.2.(b)(α)

Dellaportas P, Smith AFM. Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. Appl Statist 1993;42:443-59. (Dellaportas, 1993). ➔ 7.4.4.2.3.

Dockery DW, Pope CA, Xu X, Spengler JD, Ware JH, Fay ME Ferris BG, Speizer FE. An association between air pollution and mortality in six U.S. cities. N Engl J Med 1993;329:1753-9. (Dockery, 1993). ➔ 4.3.

Dos-Santos-Silva I, Swerdlow AJ. Sex differences in time trends of colorectal cancer in England and Wales: the possible effect of female hormonal factors. Br J Cancer 1996; 73:692-7. (Dos-Santos-Silva, 1996). ➔ 4.3.4.

Droesbeke JJ, Fichet B, Tassi P. Modélisation ARCH. Théorie statistique et applications dans le domaine de la finance. 1st ed. Bruxelles : Éditions de l'université de Bruxelles, 1994. 242 p. (Droesbeke, 1994). ➔ 2.2.1.2.

Dubrow R, Bernstein J, Holford TR. Age-period-cohort modelling of large-bowel-cancer incidence by anatomic sub-site in Connecticut. Int J Cancer 1993;53:907-13. (Dubrow, 1993). ➔ 2.2.2.2.3.2.(b)(η)

Dubrow R, Johansen, C, Skov T, Holford TR. Age-period-cohort modelling of large-bowel-cancer incidence by anatomic sub-site in Denmark. Int J Cancer 1994;58:324-29. (Dubrow, 1994). ➔ 2.2.2.2.3.2.(b)(η)

Ducimetière P, Montaville B, Schaffer P et al. Recherche et politiques de santé ; l'apport des registres de morbidité. Inserm La documentation française, Paris, 1992. 229 p. (Ducimetière, 1992). ➔ 3.1.2.

Dyba T, Hakulinen T, Comparison of different approaches to incidence prediction based on simple interpolation techniques. Statist Med 2000;19:1741-52. (Dyba, 2000). ➔ 2.2.2.2.1.2.

Dyba T, Hakulinen T, Päivärinta L. A simple non-linear model in incidence prediction. Stat Med 1997;16:2297-309. (Dyba, 1997). ➔ 2.2.2.2.1.2. ; 2.2.2.2.2.3.(b)(β) ; 4.1.1.

Eilstein D, Quénel P, Hédelin G, Kleinpeter J, Arveiler D, Schaffer P. Pollution atmosphérique et infarctus du myocarde. Strasbourg, 1984-1989. Rev Epidemiol Sante Publique 2001;49(1):13-25. (Eilstein, 2001). ➔ 7.2.4.

Elsaleh H, Joseph D, Grieu F, Zeps N, Spry N, Iacopetta B. Association of tumour site and sex with survival benefit from adjuvant chemotherapy in colorectal cancer. Lancet 2000;35:1745-50. (Elsaleh, 2000). ➔ 4.3.4.

Engeland A, Haldorsen T, Tretli S et al. Prediction of cancer incidence in the Nordic cancer registries [Report]. A collaborative study of the five Nordic cancer registries. APMIS Suppl 1993;101. (Engeland, 1993). ➔ 4.1.1. ; 4.1.4. ; 4.2.1. ; 4.2.4. ; 4.3.4.

Estève J, Benhamou E, Raymond L. Méthodes statistiques en épidémiologie descriptive. Paris : Les éditions INSERM, 1993. 307 p. (Esteve, 1993). ➔ 2.2.2.1.2.

Evstifeeva TV, MacFarlane GJ, Robertson C. Trends in cancer mortality in central European countries. The effect of age, birth cohort and time-period. European J Public Health 1997;7:169-76. (Evstifeeva, 1997). ➔ 2.2.2.2.3.2.(b)(α)

Ewertz M, Carstensen B. Trends in breast cancer incidence and mortality in Denmark, 1943-1982. Int J Cancer 1988;41:46-51. (Ewertz, 1988). ➔ 2.2.2.2.3.2.(b)(γ) ; 4.1.1.

Fahrmeir L, Tutz G. Multivariate statistical modelling based on generalized linear models. 2nd rev ed. New York : Springer – Verlag, 1996. 426 p. Springer series in statistics. (Fahrmeir, 1996). ➔ 2.2.2.1.1. ; 7.2.4. ; 7.4.3.1.2.

Faivre J, Boutron MC, Senesse P et al. Environmental and familial risk factors in relation to the colorectal adenoma-carcinoma sequence: results of a case-control study in Burgundy (France). Eur J Cancer Prev 1997;6:127-31. (Faivre, 1997). ➔ 4.3.4.

Feuer EJ, Wun LM. How much of the recent rise in breast cancer incidence can be explained by increases in mammography utilization ? A dynamic population model approach. Am J Epidemiol 1992;12:1423-36. (Feuer, 1992). ➔ 4.1.1. ; 4.1.4. ; 6.

Francis B, Green M, Payne C, editors. The GLIM System Release 4 Manual. Oxford: Clarendon Press, 1994. (Francis, 1994). ➔ 2.2.2.2.3.2.(b)(α) note 6

Garfinkel L, Boring CC, Heath CW. Changing trends. An overview of breast cancer incidence and mortality. Cancer 1994;74:222-7. (Garfinkel, 1994, a). ➔ 4.1.1.

Garfinkel L, Mushinski M. Cancer incidence, mortality and survival: trends in four leading sites. Stat Bull Metrop Insur Co 1994;75:19-27. (Garfinkel, 1994, b). ➔ 4.3.1.

Gelfand AE, Hills SE, Racine-Poon A, Smith AFM. Illustration of Bayesian inference in normal data models using Gibbs sampling. J Am Statist Ass 1990; 85:972-85. (Gelfand, 1990 a). ➔ 7.4.4.2.2.

Gelfand AE, Smith AFM, Lee TM. Bayesian analysis of constrained parameter and truncated data problems. J Am Statist Ass 1992;87:523-32. (Gelfand, 1992). ➔ 7.4.4.2.2.

Gelfand AE, Smith AFM. Sampling based approaches to calculating marginal densities. J Am Statist Ass 1990;85:398-409. (Gelfan, 1990 b). ➔ 7.4.4.2.2.

Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. London: Chapman and Hall, 1995. (Gelman, 1995). ➔ 7.4.4.2.2.

Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans Pattn Anal Mach Intell 1984;6:721-41. (Geman, 1984). ➔ 7.4.4.2.2. ; 7.4.4.2.3.

Geweke J. Bayesian inference in econometric models using Monte Carlo integration. Econometrica 1989;57:1317-39 (Geweke, 1989). ➔ 7.4.3.1. ; 7.4.3.1.2.

Gibson L, Spiegelhalter DJ, Camilleri-Ferrante C, Day NE. Trends in invasive cervical cancer incidence in East Anglia from 1971 to 1993. J Med Screening 1997;4:44-8. (Gibson, 1997). ➔ 4.2.1.

Gilks WR, Richardson S, Spiegelhalter D. Markov chain Monte-Carlo in practice. London: Chapman and Hall, 1996. (Gilks, 1996). ➔ 7.4.3. ; 7.4.4.1.

Gilks WR, Wild P. Adaptative rejection sampling for Gibbs sampling. Technical support number UR-90-01., Medical Research Council Biostatistics Unit, 5 February, 1991. (Gilks, 1991). ➔ 7.4.3.2.2.

Gilks WR, Wild P. Adaptative rejection sampling for Gibbs sampling. Appl Statist 1992;41:337-48. (Gilks, 1992). ➔ 7.4.3.2.2.

Giraud R, Chaix N. Économétrie. 2nd ed. Paris : Presses universitaires de France, 1994. 353 p. (Giraud, 1994). ➔ 2.1.1.

Glass L, Mackey MC. From clocks to Chaos ; The rhythms of life. Princeton, Princeton University Press 1988, 248 p. (Glass, 1988). ➔ 1.

Glymour C, Madigan D, Pregibon D, Smyth P. Statistical inference and data mining. Communications of the ACM 1996;39:35-41. (Glymour, 1996). ➔ 3.2.2.1.

Gourieroux C, Monfort A. Cours de séries temporelles. 1st ed. Paris : Economica, 1983. 380 p. (Gourieroux, 1983). ➔ 2.1.1.

Gourieroux C, Monfort A. Séries temporelles et modèles dynamiques. 1st ed. Paris : Economica, 1990. 780 p. (Gourieroux, 1990). ➔ 2.1.1. ; 2.1.1.1.1. ; 2.2.1.1.1. ; 2.2.1.1.2.

Gourieroux C, Monfort A. Statistique et modèles économétriques. Vol 1 Notions générales, estimation, prévision, algorithmes. 1st ed. Paris : Economica, 1989. 565 p. (Gourieroux, 1989). ➔ 7.3.

Gourieroux C. Modèles ARCH et applications financières. 1st ed. Paris : Economica, 1992. 288 p. (Gourieroux, 1992). ➔ 2.2.1.2.

Gregor A, Milroy R. Lung cancer. In Scottish Executive Health Department. Cancer scenarios: an aid to planning cancer services in Scotland in the next decade. R. Black, D. Stockton eds. The Scottish Executive, Edinburgh: 2001;64-81. (Gregor, 2001). ➔ 4.3.

Guégan D. Séries chronologiques non linéaires à temps discret. 1st ed. Paris : Economica, 1994. 301 p. (Guégan, 1994). ➔ 2.2.1.2.

Hakulinen T, Dyba T. Precision of incidence predictions based on Poisson distributed observations. Stat Med 1994;13:1513-23. (Hakulinen, 1994). ➔ 2.2.2.2.1.2.

Hakulinen T, Hakama M. Predictions of epidemiology and the evaluation of cancer control measures and the setting of policy priorities. Soc Sci Med 1991;33:1379-83. (Hakulinen, 1991). ➔ 1. ; 2.2.2.2.2.1. ; 2.2.2.2.2.2.

Hakulinen T, Pukkala E. Future incidence of lung cancer: forecasts based on hypothetical changes in the smoking habits of males. Int J Epidemiol 1981;10:233-40. (Hakulinen, 1981). ➔ 2.2.2.2.2.2. ; 4.3.

Hartigan JA. Linear Bayesian models. JR Statist Soc B 1969;31:446-54. (Hartigan, 1969). ➔ 6.

Harville DA. Use of Gibbs sampler to invert large, possibly sparse, positive definite matrices. Linear Algebra Applic 1999;289:203-24. (Harville, 1999). ➔ 7.4.1.

Hastie TJ, Tibshirani RJ. Generalized Additive Models. 1st ed. London: Chapman & Hall 1990. 336 p. (Hastie, 1990). ➔ 7.2.4.

Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 1970;57:97-109. (Hastings, 1970). ➔ 7.4.4.1.

Hertzman C, Frank J, Evans RG. L'hétérogénéité de l'état de santé et les déterminants de santé des populations. In: Être ou ne pas être en bonne santé ; biologie et déterminants sociaux de la maladie. RG Evans, ML Barer, TR Marmor ed. Les Presses de l'Université de Montréal, Montréal / John Libbey Eurotext, 1996, 77-101. (Hertzman, 1996). ➔ 1.

Heuer C. Modelling of time trends and interactions in vital rates using restricted regression splines. Biometrics 1997;53:161-77. (Heuer, 1997). ➔ 7.2.4.

Heuer C, Blettner. Trendanalyse der Morbus Hodgkin Mortalität in der Bundesrepublik Deutschland 1935-1989: Ein Vergleich von Log-linearen Modellen mit deskriptiven Stanardmethoden. Soz Präventivmed 1994;39:217-26. (Heuer, 1994). ➔ 2.2.2.2.3.2.(b)(γ)

Hobert JP, Robert CP, Goutis C. Connectedness conditions for the convergence of the Gibbs sampler. Statist Probability Lett 1997;33:235-40. (Hobert, 1997). ➔ 7.4.4.2.2.

Holford TR, Roush GC, McKay LA. Trends in female breast cancer in Connecticut and the United States. J Clin Epidemiol 1991;44:29-39. (Holford, 1991). ➔ 2.2.2.2.3.2.(b)(β) ; 4.1.1.

Holford TR. The estimation of age period and cohort effects for vital rates. Biometrics 1983;39:311-24. (Holford, 1983). ➔ 2.2.2.2.3.2.(b)(β)

Hristova L, Dimova I, Iltcheva M. Projected cancer incidence rates in Bulgaria, 1968-2017. Int J Epidemiol 1997;26:469-75. (Hristova, 1997). ➔ 4.1.1. ; 4.1.4. ; 4.2.1. ; 4.2.4.

Janssen-Heitjnen MLG, Nab HW, Van Reek J, Van der Heijden, Schipper R, Coebergh JWW. Striking changes in smoking behaviour and lung cancer incidence by histological type in South-east Netherland, 1991-1991. European J Cancer 1995;31A(6):949-52. (Janssen-Heitjnen, 1995). ➔ 4.3.

Jensen OM, Parkin DM, MacLennan R, Muir CS, Skeet RG. Enregistrement des cancers. Princimes et méthodes. IARC Publications scientifiques N° 95, Lyon, 1996. 216 p. (Jensen, 1996). ➔ 3.1.2.

Jolley D, Giles GG. Visualizing age-period-cohort trend surfaces: a synoptic approach. Int J Epidemiol 1992;21:178-82. (Jolley, 1992). ➔ 5.3.

Jordan P, Brubacher D, Tsugane S, Tsubono Y, Gey KF, Moser U. Modelling of mortality data from a multi-centre study in Japan by means of Poison regression with error in variables. Int J Epidemiol 1997;26:501-7. (Jordan, 1997). ➔ 3.2.2.1.3.

Kass RE, Steffey D. Approximate Bayesian inference in conditionnally independant hierarchical models (parametrical empirical Bayes models). J Am Statistic Ass 1989;87:717-26. (Kass, 1989). ➔ 7.4.1.

Kelsall JE, Samet JM, Zeger SL, Xu J. Air pollution and mortality in Philadelphia, 1974-1988. Am J Epidemiol 1997;146:750-62. (Kelsall, 1997). ➔ 7.2.4.

Kesley JL, Horn-Ross PL. Breast cancer: magnitude of the problem and descriptive epidemiology. Epidemiol Rev 1993;15:7-16. (Kesley, 1993). ➔ 4.1.1.

Kessler LG, Feuer EJ, Brown ML. Projections of the breast cancer burden to US women:1990-2000. Prev Med 1991;20:170-82. (Kessler, 1991). ➔ 2.2.2.2.2.2. ; 4.1.4. ; 6.

Knorr-Held L. Bayesian modelling of inseparable space-time variation in disease risk. Statist Med 2000;19:2555-67. (Knorr-Held, 2000). ➔ 6. ; 7.4.4.4.1.

Koyama Y, Kotake K. Overview of colorectal cancer in Japan: report from the Registry of the Japanese Society for Cancer of the Colon and Rectum. Dis Colon Rectum 1997;40(Suppl):S2-9. (Koyama, 1997). ➔ 4.3.1.

Krause A, Olson M. The basics of S and S-Plus. J Chambers, W Eddy, W Härdle, S Sheather, L Tierney eds. Springer Verlag, New York, 1997. (Krause, 1997). ➔ 5.3.

Laplace PS. Essai philosophique sur les probabilités. 1776. Christian Bourgois, Paris, 1986. (Laplace, 1776). ➔ 1.

Launois G. Épidémiologie du cancer colo-rectal. In Tumeurs colorectales. J. Faivre, M. Gignoux eds. Doin, Paris 1995;1-12. (Launois, 1995). ➔ 4.3.4.

Lee WC, Lin RS. Analysis of cancer rates using excess risk age-period-cohort models. Int J Epidemiol 1995 ;24:671-7. (Lee, 1995). ➔ 2.2.2.2.3.2.(b)(ε)

Lee WC, Lin RS. Autoregressive age period cohort models. Statist Med 1996;15:273:81. (Lee, 1996). ➔ 2.2.2.2.3.2.(b)(η)

Levi F, La Vecchia C, Randimbison L, Te VC. Incidence, mortality and survival from invasive cervical cancer in Vaud, Switzerland, 1974-1991. Ann Oncol 1994;5:747-52. (Levi, 1994). ➔ 4.2.1.

Lichtman SM, Mandel F, Hoexter B et al. Prospective analysis of colorectal carcinoma. Determination of an age-site and stage relationship and the correlation of DNA index with clinicopathologic parameters. Dis Colon Rectum;1994:37: 1286-90. (Lichtman, 1994). ➔ 4.3.4.

Lindsey JK. Applying generalized linear models. G Casella, S Fienberg, I Olkin eds. Springer-Verlag, New York, 1997. 256 p. (Lindsey, 1997). ➔ 2.2.2.1.1. ; 7.2.3. ; 7.4.2.2.1.

Liu JS, Wong WH, Kong A. Covariance structure of the Gibbs sampler with application to the comparisons of estimators and augmentation schemes. Biometrika 1994;81:27-40. (Liu, 1994). ➔ 6.

Loffeld R, Putten A, Balk A. Changes in the localization of colorectal cancer: implications for clinical practice. J Gastroenterol Hepatol 1996;11:47-50. (Loffeld, 1996). ➔ 4.3.4.

MacLennan R, Muir C, Steinitz R, Winkler A. Cancer registration and its techniques. W Davis ed. IARC Scientific Publications N° 21, Lyon, 1978. 235p. (MacLennan, 1978). ➔ 3.1.2. ; 6.

Maheswaran R, Strachan DP, Elliott P, Shipley MJ. Trends in stroke mortality in Greater London and south east England—evidence for a cohort effect? J Epidemiol Community Health 1997;51:121-6. (Maheswaran, 1997). ➔ 2.2.2.2.3.2.(b)(γ)

Manton KG. Health forecasting and models of aging. In: Forecasting the health of elderly populations. KG Manton, BH Singer, RM Suzman eds. Springer Verlag, New York 1993;80-106. (Manton, 1993 a). ➔ 2.3.

Manton KG, Singer BH, Stallard E. Cancer forecasting: cohort models of disease progression and mortality. In: Forecasting the health of elderly populations. KG Manton, BH Singer, RM Suzman eds. Springer Verlag, New York 1993;80-106. (Manton, 1993 b). ➔ 2.3.

Martinez ME, Grodstein F, Giovannucci E et al. A prospective study of reproductive factors, oral contraceptive use, and risk of colorectal cancer. Cancer Epidemiol Biomarkers Prev 1997; 6:1-5. ➔ 4.3.4.

McCullagh P, Nelder JA. Generalized linear models. 2nd ed. London : Chapman and Hall, 1989. (McCullagh, 1989). ➔ 2.2.2.1.1. ; 7.2.3.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equations of state calculations by fast computing machines. J Chem Phys 1953;21:1087-91. (Metropolis, 1953). ➔ 7.4.4.1. ; 7.4.4.2.2.

McNally RJQ, Alexander FE, Staines A, Cartwright RA. A comparison of three methods of analysis for age-period-cohort models with application to incidence data on non-Hodgkin's lymphoma. Int J Epidemiol 1997;26:32-46. (McNally, 1997). ➔ 2.2.2.2.3.2.(b)(α)

Massé R. Culture et santé publique. G Morin ed. Montréal, 1995, 499 p. (Massé, 1995). ➔ 1.

MathSoft, Data Analysis Products Division. S-Plus 4, User's guide. Seattle, Washington: MathSoft, Inc, 1997, 620 p. (Mathsoft, 1997 a). ➔ 5.3.

MathSoft, Data Analysis Products Division. S-Plus 4, Programmer's guide. Seattle, Washington: MathSoft, Inc, 1997, 582 p. (Mathsoft, 1997 b). ➔ 5.3.

MathSoft, Data Analysis Products Division. S-Plus 4, Guide to statistics. Seattle, Washington: MathSoft, Inc, 1997, 877 p. (Mathsoft, 1997 c). ➔ 5.3.

MathSoft, Data Analysis Products Division. S-Plus 4, Getting started with S-Plus 4.5. Seattle, Washington: MathSoft, Inc, 1997, 877 p. (Mathsoft, 1998). ➔ 5.3.

May RM. Simple mathematical models with very complicated dynamics. Nature 1976;261:459-67. (May, 1976). ➔ 2.2.1.2. ; 7.4.2.1.

Mezzetti M, Robertson C. A hierarchical Bayesian approach to age-specific back-calculation of cancer incidence rates. Stat Med 1999;18:919-33. (Mezzetti, 1999). ➔ 6.

Miller BA, Feuer EJ, Hankey BF. Recent incidence trends for breast cancer in women and the relevance of early detection: an update. Cancer J Clin 1993;43:27-41. (Miller, 1993). ➔ 4.1.4.

Miller BA, Feuer EJ, Hankey BF. The increasing incidence of breast cancer since 1982: relevance of early detection. Cancer Causes Control 1991;2:67-74. (Miller, 1991). ➔ 4.1.4.

Monfort A. Cours de probabilités. 1st ed. Paris : Économica, 1980. 210 p. (Monfort, 1980). ➔ 7.1.

Mouchart M. L'inférence bayésienne : principes généraux. In Société française de statistique Méthodes bayésiennes en statistique. Notes de cours Journées d'études en statistiques. CIRM 5-9 oct 1998. Chap 3, p. 42-4. (Mouchart, 1998). ➔ 3.2.1.

Murray JD. Mathematical biology. Berlin, Springer Verlag 1989. (Murray, 1989). ➔ 7.4.2.1.

Negri E, La Vecchia C, Levi F, Randriamiharisoa A, Decarli A, Boyle P. The application of age, period and cohort models to predict Swiss cancer mortality. J Cancer Res Clin Oncol 1990;116:207-214. (Negri, 1990). ➔ 2.2.2.2.3.2.(b)(α)

Nelder JA, Wedderburn RWM. Generalized linear models. J R Statist Soc A 1972;135:370-84. (Nelder, 1972). ➔ 2.2.2.1.1. ; 7.2.3.

Nelson RL, Persky V, Turyk M.Time trends in distal colorectal cancer subsite location related to age and how it affects choice of screening modality. J Surg Oncol 1998; 69: 235-8. (Nelson, 1998). ➔ 4.3.4.

Obrand DI, Gordon PH. Continued change in the distribution of colorectal carcinoma. Br J Surg;1998: 85: 246-8. (Obrand, 1998). ➔ 4.3.4.

O'Callaghan FJK, Osmond C, Martyn CN. Trends in epilepsy mortality in England and Wales and the United States, 1950-1994. Am J Epidemiol 2000;151:182-9. (O'Callaghan, 2000). ➔ 2.2.2.2.3.2.(b)(α)

Osmond C, Gardner MJ. Age-period and cohort models applied to cancer mortality rates. Stat Med 1982;1:245-59. (Osmond, 1982 a). ➔ 2.2.2.2.3.2.(b)(α)

Osmond C, Gardner MJ, Acheson ED. Analysis of trends in cancer mortality in England and Wales during 1951-80 separating changes associated with period of birth and period of death. Br Med J 1982;284:1005-8. (Osmond, 1982 b). ➔ 2.2.2.2.3.2.(b)(α)

Parkin DM, Whelan SL, Ferlay J, Raymond L, Young J. Cancer incidence in five continents. Vol VII. IARC Scientific Publications No. 143, Lyon 1997. (Parkin, 1997). ➔ 4.1.1.

Poincaré H. Science et méthode. Flammarion, Paris 1903. (Poincaré, 1903). ➔ 1.

Pollock AM, Benster R, Vickers N. Why did treatment rates for colorectal cancer in south east England fall between 1982 and 1988? The effect of case ascertainment and registration bias. J Public Health Med 1995 ;17:419-28. (Pollock, 1995). ➔ 4.3.1.

Pope CA, Thun M, Namboodiri M et al. Particulate Air Pollution as a Predictor of Mortality in a Prospective Study of US Adults. American J Respir Critical Care Med 1995;151:669-74. (Pope, 1995). ➔ 4.3.

Powell J. Data sources and reporting. In: Cancer registration ; principles and methods. OM Jensen, DM Parkin, R MacLennan, CS Muir, RG Skeet eds. Lyon, IARC Scientific Publications N° 95, 1991;29-42. (Powell, 1991). ➔ 3.1.2. ; 6.

Price B. Analysis of current trends in United States mesothelioma incidence. Am J Epidemiol 1997;145 :2111-8. (Price, 1997). ➔ 2.2.2.2.3.2.(a)(β)

Prigogine, I. (Le leggi del caos. Laterza, Rome 1993). Version française revue et complétée par l'auteur : Les lois du chaos. Flammarion, 1994, 127 p. (Prigogine, 1994). ➔ 1.

Prorok PC. Mathematical models of breast cancer screening. In: Screening for breast cancer. Day NE, Miller AB, editors. Toronto: Hans Huber Publishers 1988:95-104. (Prorok, 1988). ➔ 4.1.4.

Raymond L, Menegoz F, Fioretta G. Recent trends in incidence of cervical cancer in several regions of South Western Europe. Rev Epidemiol Santé Publ 1995 ;43 :122-6. (Raymond, 1995). ➔ 4.2.1.

Réseau Francim. Le cancer en France : incidence et mortalité situation en 1995, évolution entre 1975 et 1995 [Rapport]. Ministère de l'emploi et de la solidarité. La Documentation Française, 1998, 182 p. (Réseau Francim, 1998). ➔ 4.1.1. ; 4.2.1. ; 4.2.4. ; 4.3.1. ; 4.4.1.

Richardson S, Gilks WR. A Bayesian approach to measurement error problems in Epidemiology using conditional independence models. Am J Epidemiol 1993;138:430-42. (Richardson, 1993). ➔ 6.

Robert C. L'analyse statistique bayésienne. 1st ed. Paris : Economica, 1992. 396 p. (Robert, 1992). ➔ 7.3.

Robert C. Méthodes de calcul en analyse bayésienne. In Société française de statistique Méthodes bayésiennes en statistique. Notes de cours Journées d'études en statistiques. CIRM 5-9 oct 1998. Chap 6. (Robert, 1998). ➔ 7.4.1.

Robertson C, Boyle P. Age, period and cohort models: The use of individual records. Statist Med 1986;5:527-38. (Robertson, 1986). ➔ 2.2.2.2.3.2.(b)(θ)

Robertson C, Boyle P. Age-period-cohort models of chronic disease rates I: modelling approach. Statist Med 1998;17:1305-23. (Robertson, 1998 a). ➔ 2.2.2.2.3.2.(b)(θ)

Robertson C, Boyle P. Age-period-cohort models of chronic disease rates II: graphical approaches. Statist Med 1998;17:1305-23. (Robertson, 1998 b). ➔ 5.3.

Rostgaard K, Vaeth M, Holst H, Madsen M, Lynge Elsebeth. Age-period-cohort modelling of breast cancer incidence in the Nordic countries. Statist Med 2001;20:47-61. (Rostgaard, 2000). ➔ 4.1.2.2.

Ruelle D. Déterminisme et prédictibilité. In: L'ordre du chaos. Paris, Pour la science 1989;52-63. (Ruelle, 1989). ➔ 1.

Saporta G. Probabilités, analyse des données et statistique. 1st ed. Paris : Éditions Technip, 1990. 493 p. (Saporta, 1990). ➔ 7.1. ; 7.3.

Schaffer P. Épidémiologie des cancers du col et du corps de l'utérus. Bull Acad Natle Med 1997;181(7):1347-64. (Schaffer, 1997). ➔ 4.2.1.

Schaffer P, Lavillaurex J. Le cancer dans le Bas-Rhin. Incidence des nouveaux cas de 1975 à 1977. Paris : Economica, Coll Medica, 1981. 147 p. (Schaffer, 1981). ➔ 3.1.2.

Schwartz J. Multinational trends in cancer mortality rates. Methological issues and results. In Trends in cancer mortality in industrial countries. DL Davis, D Hoel eds. Annals of the NY Academy of Sciences 1990;136-145. (Schwartz, 1990 a). ➔ 7.2.4.

Schwartz J. Multinational trends in multiples myeloma. In Trends in cancer mortality in industrial countries. DL Davis, D Hoel eds. Annals of the NY Academy of Sciences 1990;215-224. (Schwartz, 1990 b). ➔ 7.2.4.

Schwartz J. Non parametric smoothing in the analysis of air pollution and respiratory illness. The Canadian Journal of Statistics 1994;22(4):471-487. (Schwartz, 1994 a). ➔ 7.2.4.

Schwartz J. Air pollution and hospital admissions for the elderly in Birmingham, Alabama. Am J Epidemiol 1994;139:589-98. (Schwartz, 1994 b). ➔ 7.2.4.

Schwartz J, Spix C, Touloumi G et al. Methodological issues of air pollution and daily counts of deaths or hospital admissions. Journal of Epidemiology and Community Health 1996;50:S19-S36. (Schwartz, 1996). ➔ 7.2.4.

Schwartz J. Use of multistage model to predict time trends in smoking induced lung cancer. J Epidemiol Community Health 1992;46:311-5. (Schwartz, 1992). ➔ 4.3.

Sigurdsson K, Adalsteinsson S, Ragnardsson J. Trends in cervical and breast cancer in Iceland. A statistical evaluation of trends in incidence and mortality for the period 1955-1989, their relation to screening and prediction to the year 2000. Int J Cancer 1991;48:523-8. (Sigurdsson, 1991). ➔ 4.1.4. ; 4.2.1. ; 4.2.4.

Simar L. Le paradigme bayésien. In Société française de statistique. Méthodes bayésiennes en statistique. Notes de cours Journées d'études en statistiques. CIRM 5-9 oct 1998. Chap 2. (Simar, 1998 a). ➔ 7.1.

Simar L. Les modèles de base de l'analyse bayésienne. In Société française de statistique. Méthodes bayésiennes en statistique. Notes de cours Journées d'études en statistiques. CIRM 5-9 oct 1998. Chap 4. (Simar, 1998 b). ➔ 7.3.4.2.

Skeet RG. Quality and quality control. In: Cancer registration ; principles and methods. OM Jensen, DM Parkin, R MacLennan, CS Muir, RG Skeet eds. Lyon, IARC Scientific Publications N° 95, 1991;101-7. (Skeet, 1991). ➔ 3.1.2. ; 6.

Smith AFM, Roberts GO. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. J R Statist Soc Series B, 1993;55:3-23. (Smith, 1993). ➔ 7.4.4.1.

Société française de statistique. Méthodes bayésiennes en statistique. Notes de cours Journées d'études en statistiques. CIRM 5-9 oct 1998. (Société française de statistique, 1998). ➔ 7.3.

Sondik EJ. Breast cancer trends incidence, mortality and survival. Cancer 1994;74:995-9. (Sondick, 1994). ➔ 4.1.1.

Spiegelhalter D, Thomas A, Best N, Gilks W. BUGS 0.5 Bayesian inference using Gibbs sampling Manual (version ii). 1996. (Spiegelhalter, 1996 a). ➔ 3.2.2.1.1.

Spiegelhalter D, Thomas A, Best N, Gilks W. BUGS 0.5 Examples Volume 1 (version i). 1996. (Spiegelhalter, 1996 b). ➔ 3.2.2.1.1.

Spiegelhalter D, Thomas A, Best N, Gilks W. BUGS 0.5 Examples Volume 2 (version ii). 1996. (Spiegelhalter, 1996 c). ➔ 3.2.2.1.1.

Spiegelhalter D, Thomas A, Best N, Gilks W. Ice: non-parametric smoothing in an age-cohort model. In: BUGS 0.5 Examples Volume 2 (version ii), 1996:30-6. (Spiegelhalter, 1996 d). ➔ 3.2.2.1.2.

Steenland K, Nowlin S, Palu S. Cancer incidence in the National Health and Nutrition Survey I. Follow-up data: diabetes, cholesterol, pulse and physical activity. Cancer Epidemiol Biomarkers Prev 1995;4:807-11. (Steenland, 1995). ➔ 4.3.4.

Sugihara G, May RM. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. Nature 1990;344:734-41. (Sugihara, 1990). ➔ 2.2.1.2.

Sun D, Tsutakawa RK, Kim H, He Z. Spatio-temporal interaction with disease mapping. Stat Med 2000;19:2015-35. (Sun, 2000). ➔ 6.

Tanner MA. Tools for statistical inference. 3rd ed. New York : Springer Verlag, 1996. 202 p. (Tanner, 1996). ➔ 7.4.3. ; 7.4.3.2.2. ; 7.4.4.1. ; 7.4.4.2.2.

Tazi MA, Faivre J, Lejeune C, Benhamiche AM, Dassonville F. Performances du test Hemoccult dans le depistage des cancers et des adenomes colorectaux. Resultats de cinq campagnes de depistage en Saone-et-Loire. Gastroenterol Clin Biol 1999;23:475-80. (Tazi, 1999). ➔ 4.3.4.

Tominaga S, Kuroishi T. Epidemiology of breast cancer in Japan. Cancer Lett 1995;90:75-9. (Tominaga, 1995). ➔ 4.1.1.

Van Oomarssen GJ, Boer R, Habbema JDF. Modelling issues in cancer screening. Stat Meth Med Res 1995;4:33-54. (Van Oomarssen, 1995). ➔ 4.1.4.

Venables WN, Ripley BD. Modern applied statistics with S-PLUS. 2nd ed. New York: Springer 1997. 548 p. (Venables, 1997). ➔ 5.3. ; 7.2.4.

Vioque J, Navarro Gracia JF, Millas Ros J, Mateo de las Heras E. Evolución y predicción de la incidencia de cáncer de mama en Zaragoza, 1961-2000. Med Clin (Barc) 1993;101:12-7. (Vioque, 1993). ➔ 2.2.2.2.3.2.(b)(β) ; 4.1.1. ; 4.1.4.

Vizcaino AP, Moreno V, Bosch FX, Munoz N, Barros-Dioz XM, Borras J et al. International trends in the incidence of cervical cancer: II. Squamous cell carcinomas. Int J Cancer 2000;86:429-35. (Vizcaino, 2000). ➔ 4.2.1

Vizcaino AP, Moreno V, Bosch FX, Munoz N, Barros-Dioz XM, Parkin DM. International trends in the incidence of cervical cancer: I. Adenocarcinoma and adenosquamous cell carcinomas. Int J Cancer 1998;75:536-45. (Vizcaino, 1998). ➔ 4.2.1.

Wakai K, Suzuki S, Ohno Y, Kawamura T, Tamakoshi A, Rie A. Epidemiology of breast cancer in Japan. Int J Epidemiol 1995;24:285-91. (Wakai, 1995). ➔ 4.1.1.

Waterhouse J, Muir C, Correa P, Powell J. Cancer incidence in five continents. Lyon: IARC Scientific Publications N° 15. 1976 (Vol 3). (Waterhouse, 1976). ➔ 4.2.4.

Weidmann C, Schaffer P, Hédelin G, Arveux P, Chaplain G, Exbrayat C et al. L'incidence du cancer du col régresse régulièrement en France. Bull Epidemiol Hebdomadaire 1998;5:17-19. (Weidmann, 1998). ➔ 4.2.1. ; 4.2.4.

Wedderburn RWM. Quasi-likelihood functions, generalized liner models, and the Gauss-Newton method. Biometrika 1974;61:439-47. (Wedderburn, 1974). ➔ 6.

Wiklund K, Hakulinen T, Sparén P. Prediction of cancer mortality in the Nordic countries in 2005: effects of various interventions. European J Cancer Prevention 1992;1:247-58. (Wiklund, 1992). ➔ 2.2.2.2.1.2.

Wingo PA, Ries LA, Rosenberg HM, Miller DS, Edwards BK. Cancer incidence and mortality, 1973 1995: a report card for the U.S. Cancer 1998;82:1197-207. (Wingo, 1998). ➔ 4.3.1.

Wun LM, Feuer EJ, Miller BA. Are increases in mammographic screening still a valid explanation for trends in breast cancer incidence in the United States ? Cancer Causes Control 1995;6 :135-44. (Wun, 1995). ➔ 4.1.1. ; 4.1.4. ; 6.

Zaridze DG, Gurevicius R. Lung cancer in the USSR: patterns and trends. In: DG Zaridze, R Peto eds. Tobacco: a major international health hazard. Lyon: IARC Scientific Publications N° 74. 1986:87-101. (Zaridze, 1986). ➔ 4.3.

Zerr-Fuhrmann J. Epidémiologie descriptive du cancer du sein dans le département du Bas-Rhin [Thèse]. Strasbourg, France: Université Louis Pasteur, 1992. (Zerr-Fuhrmann, 1992). ➔ 4.1.1.

Zheng T, Holford TR, Ward BA, McKay L, Flannery J, Boyle P. Time trend in pancreatic cancer incidence in Connecticut, 1935-1990. Int J Cancer 1995;61:622-7. (Zheng, 1995). ➔ 2.2.2.2.3.2.(b)(β)

INDEX

« LE SAGE ET L'ASTRONOMIE. — Tant que tu considères les étoiles comme quelque chose qui est « au dessus de toi », il te manque le regard de celui qui cherche la connaissance. »[15]

[15] Friedrich Nietzsche. Par delà le bien et le mal. Traduction : Henri Albert, revue par Marc Sautet. Paris, Le Livre de Poche, Librairie Générale Française, 1991.

A

B

C

D

E

F

G

H

I

L

M

N

P

R

S

T

V

ANNEXES

ANNEXE 1. DONNÉES DE POPULATION

A1.1. Population bas-rhinoise

Les données de populations sont fournies par l'INSEE (méthode PRUDENT)

Tableau A1.1. Effectifs de la population féminine du Bas-Rhin, par tranches d'âge de 5 ans et par périodes de 5 années.

Âge\Période	1975-1979	1980-1984	1985-1989	1990-1994	1995-1999	2000-2004	2005-2009	2010-2014	2015-2019
00-04	154263	151661	155538	161197	158557	153463	148880	149276	151394
05-09	176241	155644	153178	157278	160394	159809	154744	150221	150682
10-14	187712	179120	157515	154693	160908	164909	164242	159029	154372
15-19	193951	205131	196926	168573	172352	178167	182382	181432	175658
20-24	197144	206724	214124	207851	181505	184547	190103	193999	192667
25-29	177644	189378	192561	200812	201214	176601	179499	184737	188373
30-34	133567	177398	191518	197936	199100	199506	175340	178244	183415
35-39	125169	132442	175520	191458	195307	198095	198485	174663	177580
40-44	129626	124355	131786	176792	188653	193919	196746	197149	173653
45-49	133063	127842	122267	131083	173163	186582	191928	194864	195328
50-54	131714	129462	125165	121491	128147	170683	184034	189545	192664
55-59	103375	128378	126948	123755	118505	126100	168031	181354	187079
60-64	90456	98835	123311	123487	119600	115650	123300	164493	177769
65-69	112993	85112	93926	118485	117602	115455	112150	119956	160425
70-74	100887	101407	76899	87360	110205	110802	109612	107232	115345
75-79	73473	82677	85211	66699	76981	98652	100567	100649	99515
80-84	40738	50555	59289	64432	51440	62159	80683	83971	85550
85-89	15608	20692	27290	34901	39602	32563	42103	55547	59609
90-94	3755	4950	7201	10821	14376	17100	14415	20596	27607
95-99	589	740	1036	1694	2531	3687	4715	4313	6561
00-99	2281968	2352503	2417209	2500798	2570142	2648449	2721959	2791270	2855246

Tableau A1.2. Effectifs de la population féminine du Bas-Rhin, par tranches d'âge de 4 ans et par périodes de 4 années (ces données sont utilisées pour l'analyse du cancer du sein).

Âge\Période	1975-1976	1977-1980	1981-1984	1985-1988	1989-1992	1993-1996	1997-1900	1901-1904	1905-1908
25-28	56273	118784	121858	122838	128380	129077	125960	108990	115781
29-32	40666	104604	118778	123254	126000	128541	128087	125005	108330
33-36	40155	80670	103546	117659	123508	126068	127097	127443	124394
37-40	40508	79908	80152	103386	117926	123600	124765	126693	127026
41-44	42691	81684	79302	79516	103396	117749	121938	123902	125852
45-48	42597	85153	80661	78301	78965	102688	115762	120883	122890
49-52	42285	84064	83146	79271	77654	78331	100963	114536	119654
53-56	39538	83683	82228	82060	78574	76846	76783	99654	113096
57-60	22125	66202	81445	80306	80600	77583	75316	75774	98378
61-64	35933	50328	63737	78948	78305	78814	75314	73706	74247
65-68	37106	70234	48020	61274	76218	75728	75948	73304	71950
69-72	33689	69001	64911	44421	57869	72655	72078	72948	70708
73-76	27582	58483	61261	58799	40953	53457	67219	67451	68743
77-80	19195	43097	47591	50652	50002	35655	47295	60068	60920
81-84	11074	25825	30773	35362	39265	39617	28379	38980	49941
85-88	5119	11925	15000	18655	22697	26295	27141	19862	28493
25-88	536536	1113645	1162409	1214702	1280312	1342704	1390045	1429199	1480403

Tableau A1.3. Effectifs de la population masculine du Bas-Rhin, par tranches d'âge de 5 ans et par périodes de 5 années.

Âge\Période	1975-1979	1980-1984	1985-1989	1990-1994	1995-1999	2000-2004	2005-2009	2010-2014	2015-2019
00-04	160215	159460	162663	168155	167258	161884	157071	157518	159778
05-09	183212	161575	159898	163675	168418	168837	163495	158762	159291
10-14	195440	186696	162666	161295	166234	171829	172212	166792	161987
15-19	196333	205094	196414	170106	173096	177895	183719	184025	178236
20-24	204633	207979	214124	206546	183003	185047	189845	195647	195700
25-29	197123	197878	196023	204936	200891	177598	179604	184002	189415
30-34	150581	192339	195836	197602	198702	195850	173240	175215	179451
35-39	138944	146208	187365	193989	190800	193623	190855	168947	170885
40-44	140108	135310	142419	185313	187469	186415	189290	186701	165438
45-49	137345	134851	130492	140424	178574	182539	181775	184805	182473
50-54	111832	129935	128594	127090	133557	171532	175505	175112	178340
55-59	80686	103560	120177	121589	119006	126468	162570	166550	166647
60-64	68964	72391	93446	110659	111580	110387	117767	151732	155874
65-69	82028	58747	63592	83361	98228	100399	100197	107677	139481
70-74	67038	64291	47163	53211	70070	84276	87318	88251	95933
75-79	40450	46150	45770	35375	40978	55128	67601	71269	73254
80-84	18078	22673	27049	28986	22639	27688	38027	47764	51499
85-89	6507	7380	9394	12274	14498	11589	15229	21426	27722
90-94	1431	1603	2109	2937	4012	4807	3937	5697	8239
95-99	283	264	259	380	537	781	988	876	1346
00-99	2181231	2234384	2285453	2367903	2429550	2494572	2550245	2598768	2640989

A1.2. Populations standard

Tableau A1.4. Populations de référence par tranches d'âge de 5 ans.

Tranches d'âge	Population mondiale	Population européenne
0-4	12000	8000
5-9	10000	7000
10-14	9000	7000
15-19	9000	7000
20-24	8000	7000
25-29	8000	7000
30-34	6000	7000
35-39	6000	7000
40-44	6000	7000
45-49	6000	7000
50-54	5000	7000
55-59	4000	6000
60-64	4000	5000
65-69	3000	4000
70-74	2000	3000
75-79	1000	2000
80-84	500	1000
+85	500	1000
Total	100000	100000

Tableau A1.5. Populations de référence par tranches d'âge de 4 ans.
(ces données sont utilisées pour l'analyse du cancer du sein)

Tranches d'âge	Population mondiale	Population européenne
25-28	6320	5600
29-32	5280	5600
33-36	4800	5600
37-40	4800	5600
41-44	4800	5600
45-48	4760	5600
49-52	4240	5600
53-56	3600	5200
57-60	3200	4560
61-64	3080	3920
65-68	2480	3280
69-72	1840	2640
73-76	1200	2000
77-80	680	1360
81-84	380	760
85-88	172	348
25-88	51632	63268

Valeurs obtenues par interpolation.

ANNEXE 2. FICHIERS BUGS

A2.1. Écriture générale des fichiers

Le principe et les hypothèses servant de base à la modélisation ont été explicités au paragraphe 2.2.2.1.2.

Les fichiers nécessaires aux calculs sont :
Les fichiers « Données » : il contiennent les données disponibles (effectifs de population, nombre de cas incidents) ;
Le fichier « Valeurs initiales » : il donne des valeurs arbitraires initiales aux paramètres ;
Le fichier « Programme » : il construit le modèle et énonce les hypothèses ;
Le fichier « Commandes » : il contient les instructions de routine (nombre d'itérations, etc.) ;
Fichier « Résultats » : il affiche les valeurs des paramètres estimés.

A2.1.1. Fichier « Données »

Il y a deux fichiers de données :

A2.1.1.1. Fichier « Données de population »

Nom générique du fichier : Nomfichpopu.**dat**

Les trois dernières colonnes correspondent, respectivement, à la classe d'âge **a**, à la période **p** et à la #cohorte **c**. La première colonne contient l'effectif de la fraction de la population correspondant à #l'âge **a**, à la période **p** et à la cohorte **c**.

…… … … …
…… … … …

A2.1.1.2. Fichier « Données d'incidence »

Nom générique du fichier : Nomfichcas.**dat**

#Ce fichier contient le nombre de cas correspondant à l'âge **a**, à la période **p** et à la cohorte **c**, dans le #même ordre que dans le fichier population.

…
…

A2.1.2. Fichier « Valeurs initiales »

Nom générique du fichier : Nomfichvalinit.**in**

list(taua=1, taup=1, tauc=1, alpha=c(0,….,0), beta =c(0,…,0),
gamma=c(0,…,0))

A2.1.3. Fichier « Programme »

Nom générique du fichier : Nomfichprog.**bug**

#Proj nb cas incid cancer d… chez … à l'horizon … à partir des périodes … à …

model APCnorm;
const
N = nombre_de_classes, I = nombre_de_classes_d_age, J = nombre_de_periodes, K = nombre_de_cohortes, M = nombre_de_periode_a_predire; # M projections to be made
var
age[N], period[N], cohort[N], cas[N-M*I], popn[N],

```
    mu[N-M*I], pred.mu[M*I], total, pred.rate[M*I],

    alpha[I], alphamean[I], alphaprec[I],
    taua, sigmaa, da, ra, tau.likea[I], Nneighsa[I],

    beta[J], betamean[J],  betaprec[J], taup, sigmap,

    gamma[K], gammamean[K], gammaprec[K], tauc, sigmac;

 data  popn, period, age, cohort  in "chemin\ Nomfichpopu.dat",
 cas in "chemin\ Nomfichcas.dat";

 inits in "chemin\ Nomfichvalinit.in";

{
  for (n in 1:N-M*I) {
    cas[n]        ~ dpois(mu[n]);
    log(mu[n])  <- log(popn[n]) + alpha[age[n]] + beta[period[n]]
                 + gamma[cohort[n]];
  }

  for (i in 1:M*I) {
    log(pred.mu[i])  <- log(popn[N-M*I+i]) + alpha[age[N-M*I+i]] +
              beta[period[N-M*I+i]] + gamma[cohort[N-M*I+i]];
    pred.rate[i] <- 100000*pred.mu[i]/popn[N-M*I+i];
  }

  total <- sum(pred.mu[]);

  betamean[1]    <- 0.0;
  betaprec[1]    <- taup*1.0E-6;
  betamean[2]    <- 0.0;
  betaprec[2]    <- taup*1.0E-6;
  for (j in 3:J){
    betamean[j]    <- 2*beta[j-1] - beta[j-2];
    betaprec[j]    <- taup;
  }
```

```
for (j in 1:J){
   beta[j]       ~ dnorm(betamean[j],betaprec[j]);
 }

 taup           ~ dgamma(1.0E-3,1.0E-3);
 sigmap         <- 1/sqrt(taup);

 alphamean[1]    <- 2*alpha[2] - alpha[3];
 Nneighsa[1]     <- 1;
 alphamean[2]    <- (2*alpha[1] + 4*alpha[3] - alpha[4])/5;
 Nneighsa[2]     <- 5;
 for (i in 3:(I-2)){
   alphamean[i]    <- (4*alpha[i-1] + 4*alpha[i+1]- alpha[i-2]
                    - alpha[i+2])/6;
   Nneighsa[i]     <- 6;
 }
 alphamean[I-1]  <- (2*alpha[I] + 4*alpha[I-2] - alpha[I-3])/5;
 Nneighsa[I-1]   <- 5;
 alphamean[I]    <- 2*alpha[I-1] - alpha[I-2];
 Nneighsa[I]     <- 1;
 for (i in 1:I){
   alphaprec[i]    <- Nneighsa[i] * taua;
 }

 for (i in 1:I){
   alpha[i]       ~ dnorm(alphamean[i],alphaprec[i]);
   tau.likea[i]   <- Nneighsa[i] * alpha[i] * (alpha[i]
                    - alphamean[i]);
 }

 da             <- 0.0001 + sum(tau.likea[])/2;
 ra             <- 0.0001 + I/2;
 taua           ~ dgamma(ra,da);
 sigmaa         <- 1/sqrt(taua);

 gammamean[1]    <- 0.0;
```

```
  gammaprec[1]    <- tauc*1.0E-6;
  gammamean[2]    <- 0.0;
  gammaprec[2]    <- tauc*1.0E-6;
  for (k in 3:K){
    gammamean[k]    <- 2*gamma[k-1] - gamma[k-2];
    gammaprec[k]    <- tauc
  }

  for (k in 1:K){
    gamma[k]        ~ dnorm(gammamean[k],gammaprec[k]);
  }

  tauc            ~ dgamma(1.0E-3,1.0E-3);
  sigmac          <- 1/sqrt(tauc);
}
```

A2.1.4. Fichier « Commandes »

Nom générique du fichier : Nomfichcommand.**cmd**

```
compile("chemin\Nomfichprog.bug")
update(nombre_iterations_tour_de_chauffe)
monitor(mu)
monitor(pred.mu)
update(nombre_iterations_pour_calculs)
stats(mu)
stats(pred.mu)
q()
```

A2.1.5. Fichier « Résultats »

Nom générique du fichier : Nomfichresult.**log**

Welcome to BUGS on 10 th Jun 2001 at 16:37:1
BUGS : Copyright (c) 1992 .. 1995 MRC Biostatistics Unit.
All rights reserved.
Version 0.600 for 32 Bit PC.
For general release: please see documentation for disclaimer.
The support of the Economic and Social Research Council (UK)
is gratefully acknowledged.
Bugs>compile("chemin\ Nomfichprog.bug ")
Proj nb cas incid cancer d… chez … à l'horizon … à partir des périodes … à …

```
model APCnorm;
const
    N = nombre_de_classes, I = nombre_de_classes_d_age, J = nombre_de_periodes, K =
nombre_de_cohortes, M = nombre_de_periode_a_predire; # M projections to be made
var
    age[N], period[N], cohort[N], cas[N-M*I], popn[N],
    mu[N-M*I], pred.mu[M*I], total, pred.rate[M*I],

    alpha[I], alphamean[I], alphaprec[I],
    taua, sigmaa, da, ra, tau.likea[I], Nneighsa[I],

    beta[J], betamean[J], betaprec[J], taup, sigmap,

    gamma[K], gammamean[K], gammaprec[K], tauc, sigmac;

 data popn, period, age, cohort in "chemin\ Nomfichpopu.dat",
 cas in "chemin\ Nomfichcas.dat";

 inits in "chemin\ Nomfichvalinit.in";
```

```
{
  for (n in 1:N-M*I) {
    cas[n]        ~ dpois(mu[n]);
    log(mu[n])  <- log(popn[n]) + alpha[age[n]] + beta[period[n]]
                  + gamma[cohort[n]];
  }

  for (i in 1:M*I) {
    log(pred.mu[i])  <- log(popn[N-M*I+i]) + alpha[age[N-M*I+i]] +
                beta[period[N-M*I+i]] + gamma[cohort[N-M*I+i]];
    pred.rate[i] <- 100000*pred.mu[i]/popn[N-M*I+i];
  }

  total <- sum(pred.mu[]);

  betamean[1]    <- 0.0;
  betaprec[1]    <- taup*1.0E-6;
  betamean[2]    <- 0.0;
  betaprec[2]    <- taup*1.0E-6;
  for (j in 3:J){
    betamean[j]    <- 2*beta[j-1] - beta[j-2];
    betaprec[j]    <- taup;
  }

 for (j in 1:J){
    beta[j]        ~ dnorm(betamean[j],betaprec[j]);
  }

  taup            ~ dgamma(1.0E-3,1.0E-3);
  sigmap          <- 1/sqrt(taup);

  alphamean[1]    <- 2*alpha[2] - alpha[3];
  Nneighsa[1]     <- 1;
  alphamean[2]    <- (2*alpha[1] + 4*alpha[3] - alpha[4])/5;
  Nneighsa[2]     <- 5;
```

```
for (i in 3:(I-2)){
  alphamean[i]    <- (4*alpha[i-1] + 4*alpha[i+1]- alpha[i-2]
                   - alpha[i+2])/6;
  Nneighsa[i]     <- 6;
}
alphamean[I-1]  <- (2*alpha[I] + 4*alpha[I-2] - alpha[I-3])/5;
Nneighsa[I-1]   <- 5;
alphamean[I]    <- 2*alpha[I-1] - alpha[I-2];
Nneighsa[I]     <- 1;
for (i in 1:I){
  alphaprec[i]    <- Nneighsa[i] * taua;
}

for (i in 1:I){
  alpha[i]        ~ dnorm(alphamean[i],alphaprec[i]);
  tau.likea[i]   <- Nneighsa[i] * alpha[i] * (alpha[i]
                   - alphamean[i]);
}

da             <- 0.0001 + sum(tau.likea[])/2;
ra             <- 0.0001 + I/2;
taua            ~ dgamma(ra,da);
sigmaa         <- 1/sqrt(taua);

gammamean[1]    <- 0.0;
gammaprec[1]    <- tauc*1.0E-6;
gammamean[2]    <- 0.0;
gammaprec[2]    <- tauc*1.0E-6;
for (k in 3:K){
  gammamean[k]    <- 2*gamma[k-1] - gamma[k-2];
  gammaprec[k]    <- tauc
}

for (k in 1:K){
  gamma[k]        ~ dnorm(gammamean[k],gammaprec[k]);
}
```

```
  tauc          ~ dgamma(1.0E-3,1.0E-3);
  sigmac        <- 1/sqrt(tauc);
}
```

```
Parsing model declarations.
Loading data value file(s).
Loading initial value file(s).
Parsing model specification.
Checking model graph for directed cycles.
Possible directed cycle or undirected link in model
Generating code.
Generating sampling distributions.
Checking model specification.
Choosing update methods.
compilation took  00:00:00
Bugs>update(4000)     time for    4000   updates was  00:00:14
Bugs>monitor(mu)
Bugs>monitor(pred.mu)
Bugs>update(5000)     time for    5000   updates was  00:00:21
Bugs>stats(mu)

             mean       sd      2.5% :  97.5%  CI    median      sample
[1]            ……..      ……..       ……..        ……..      ……..      ……..
[2]            ……..      ……..       ……..        ……..      ……..      ……..
       ……..
[n]            ……..      ……..       ……..        ……..      ……..      ……..

Bugs>stats(pred.mu)
             mean       sd      2.5% :  97.5%  CI    median      sample
[1]            ……..      ……..       ……..        ……..      ……..      ……..
[2]            ……..      ……..       ……..        ……..      ……..      ……..
……..
[p]            ……..      ……..       ……..        ……..      ……..      ……..
Bugs>q()
```

A2.2. Exemple de fichiers BUGS : le cancer du côlon

Fichiers « Données de population » (pofl55.dat) et « Données d'incidence » (cofl55.dat)

pofl55.dat				cofl55.dat
177644	1	1	13	3
133567	1	2	12	2
125169	1	3	11	7
129626	1	4	10	11
133063	1	5	9	26
131714	1	6	8	36
103375	1	7	7	50
90456	1	8	6	53
112993	1	9	5	86
100887	1	10	4	81
73473	1	11	3	90
40738	1	12	2	45
15608	1	13	1	22
189378	2	1	14	1
177398	2	2	13	2
132442	2	3	12	5
124355	2	4	11	7
127842	2	5	10	22
129462	2	6	9	30
128378	2	7	8	54
98835	2	8	7	64
85112	2	9	6	78
101407	2	10	5	101
82677	2	11	4	129
50555	2	12	3	74
20692	2	13	2	27
192561	3	1	15	1
191518	3	2	14	6

175520	3	3	13	7
131786	3	4	12	16
122267	3	5	11	15
125165	3	6	10	48
126948	3	7	9	56
123311	3	8	8	84
93926	3	9	7	80
76899	3	10	6	112
85211	3	11	5	126
59289	3	12	4	97
27290	3	13	3	52
200812	4	1	16	0
197936	4	2	15	4
191458	4	3	14	9
176792	4	4	13	15
131083	4	5	12	24
121491	4	6	11	42
123755	4	7	10	66
123487	4	8	9	79
118485	4	9	8	107
87360	4	10	7	110
66699	4	11	6	114
64432	4	12	5	121
34901	4	13	4	80
201214	5	1	17	
199100	5	2	16	
195307	5	3	15	
188653	5	4	14	
173163	5	5	13	
128147	5	6	12	
118505	5	7	11	
119600	5	8	10	
117602	5	9	9	
110205	5	10	8	
76981	5	11	7	
51440	5	12	6	
39602	5	13	5	

```
176601    6    1   18
199506    6    2   17
198095    6    3   16
193919    6    4   15
186582    6    5   14
170683    6    6   13
126100    6    7   12
115650    6    8   11
115455    6    9   10
110802    6   10    9
 98652    6   11    8
 62159    6   12    7
 32563    6   13    6
179499    7    1   19
175340    7    2   18
198485    7    3   17
196746    7    4   16
191928    7    5   15
184034    7    6   14
168031    7    7   13
123300    7    8   12
112150    7    9   11
109612    7   10   10
100567    7   11    9
 80683    7   12    8
 42103    7   13    7
```

Fichier « Valeurs initiales » : cof55.in

```
list(taua=1, taup=1, tauc=1, alpha=c(0,0,0,0,0,0,0,0,0,0,0,0,0),
beta =c(0,0,0,0,0,0,0),
gamma=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0))
```

Fichier « Programme » : cof55.bugs

```
#Proj nb cas incid cancer colon chez les femmes à l'horizon 05-09 à
partir des périodes 75-79 à 90-94

model APCnorm;
const
      N = 91, I = 13, J = 7, K = 19, M = 3; # 5 projections to be
made
var
      age[N], period[N], cohort[N], col[N-M*I], popn[N],
      mu[N-M*I], pred.mu[M*I], total, pred.rate[M*I],

      alpha[I], alphamean[I], alphaprec[I],
      taua, sigmaa, da, ra, tau.likea[I], Nneighsa[I],

      beta[J], betamean[J],  betaprec[J], taup, sigmap,

      gamma[K], gammamean[K], gammaprec[K], tauc, sigmac;

  data  popn, period, age, cohort  in
"d:\dan\cancolo\inca\bayes\bugs\pofl55.dat",
  col in "d:\dan\cancolo\inca\bayes\bugs\cofl55.dat";

  inits in "d:\dan\cancolo\inca\bayes\bugs\cof55.in";

{
   for (n in 1:N-M*I) {
      col[n]        ~ dpois(mu[n]);
      log(mu[n])  <- log(popn[n]) + alpha[age[n]] + beta[period[n]]
                     + gamma[cohort[n]];
   }

   for (i in 1:M*I) {
      log(pred.mu[i])  <- log(popn[N-M*I+i]) + alpha[age[N-M*I+i]] +
                  beta[period[N-M*I+i]] + gamma[cohort[N-M*I+i]];
      pred.rate[i] <- 100000*pred.mu[i]/popn[N-M*I+i];
```

```
 }

 total <- sum(pred.mu[]);

 betamean[1]    <- 0.0;
 betaprec[1]    <- taup*1.0E-6;
 betamean[2]    <- 0.0;
 betaprec[2]    <- taup*1.0E-6;
 for (j in 3:J){
    betamean[j]    <- 2*beta[j-1] - beta[j-2];
    betaprec[j]    <- taup;
 }

for (j in 1:J){
    beta[j]        ~ dnorm(betamean[j],betaprec[j]);
 }

 taup            ~ dgamma(1.0E-3,1.0E-3);
 sigmap         <- 1/sqrt(taup);

 alphamean[1]    <- 2*alpha[2] - alpha[3];
 Nneighsa[1]     <- 1;
 alphamean[2]    <- (2*alpha[1] + 4*alpha[3] - alpha[4])/5;
 Nneighsa[2]     <- 5;
 for (i in 3:(I-2)){
    alphamean[i]    <- (4*alpha[i-1] + 4*alpha[i+1]- alpha[i-2]
                          - alpha[i+2])/6;
    Nneighsa[i]     <- 6;
 }
 alphamean[I-1]  <- (2*alpha[I] + 4*alpha[I-2] - alpha[I-3])/5;
 Nneighsa[I-1]   <- 5;
 alphamean[I]    <- 2*alpha[I-1] - alpha[I-2];
 Nneighsa[I]     <- 1;
 for (i in 1:I){
    alphaprec[i]    <- Nneighsa[i] * taua;
 }
```

```
   for (i in 1:I){
      alpha[i]        ~ dnorm(alphamean[i],alphaprec[i]);
      tau.likea[i]   <- Nneighsa[i] * alpha[i] * (alpha[i]
                           - alphamean[i]);
   }

   da             <- 0.0001 + sum(tau.likea[])/2;
   ra             <- 0.0001 + I/2;
   taua            ~ dgamma(ra,da);
   sigmaa         <- 1/sqrt(taua);

   gammamean[1]    <- 0.0;
   gammaprec[1]    <- tauc*1.0E-6;
   gammamean[2]    <- 0.0;
   gammaprec[2]    <- tauc*1.0E-6;
   for (k in 3:K){
      gammamean[k]    <- 2*gamma[k-1] - gamma[k-2];
      gammaprec[k]    <- tauc
   }

   for (k in 1:K){
      gamma[k]        ~ dnorm(gammamean[k],gammaprec[k]);
   }

   tauc            ~ dgamma(1.0E-3,1.0E-3);
   sigmac         <- 1/sqrt(tauc);
}
```

Fichier « Commandes » : cof55.cmd

```
compile("d:\dan\cancolo\inca\bayes\bugs\cof55.bug")
update(500)
monitor(mu)
monitor(pred.mu)
update(2000)
stats(mu)
stats(pred.mu)
```

```
q()
```

Fichier résultats : cof55.log

```
Welcome to BUGS on 21 st Dec 1998  at 16:20:51
BUGS : Copyright (c) 1992 .. 1995 MRC Biostatistics Unit.
All rights reserved.
Version 0.600 for 32 Bit PC.
For general release: please see documentation for disclaimer.
The support of the Economic and Social Research Council (UK)
is gratefully acknowledged.
Bugs>compile("d:\dan\cancolo\inca\bayes\bugs\cof55.bug")
#Proj nb cas incid cancer colon chez les femmes à l'horizon 05-09 à
partir des périodes 75-79 à 90-94

model APCnorm;
const
      N = 91, I = 13, J = 7, K = 19, M = 3; # 5 projections to be
made
var
      age[N], period[N], cohort[N], col[N-M*I], popn[N],
      mu[N-M*I], pred.mu[M*I], total, pred.rate[M*I],

      alpha[I], alphamean[I], alphaprec[I],
      taua, sigmaa, da, ra, tau.likea[I], Nneighsa[I],

      beta[J], betamean[J],  betaprec[J], taup, sigmap,

      gamma[K], gammamean[K], gammaprec[K], tauc, sigmac;

  data  popn, period, age, cohort  in
"d:\dan\cancolo\inca\bayes\bugs\pofl55.dat",
  col in "d:\dan\cancolo\inca\bayes\bugs\cofl55.dat";

  inits in "d:\dan\cancolo\inca\bayes\bugs\cof55.in";
```

```
{
   for (n in 1:N-M*I) {
      col[n]        ~ dpois(mu[n]);
      log(mu[n])  <- log(popn[n]) + alpha[age[n]] + beta[period[n]]
                     + gamma[cohort[n]];
   }

   for (i in 1:M*I) {
      log(pred.mu[i])  <- log(popn[N-M*I+i]) + alpha[age[N-M*I+i]] +
                   beta[period[N-M*I+i]] + gamma[cohort[N-M*I+i]];
      pred.rate[i] <- 100000*pred.mu[i]/popn[N-M*I+i];
   }

   total <- sum(pred.mu[]);

   betamean[1]    <- 0.0;
   betaprec[1]    <- taup*1.0E-6;
   betamean[2]    <- 0.0;
   betaprec[2]    <- taup*1.0E-6;
   for (j in 3:J){
      betamean[j]    <- 2*beta[j-1] - beta[j-2];
      betaprec[j]    <- taup;
   }

  for (j in 1:J){
      beta[j]        ~ dnorm(betamean[j],betaprec[j]);
   }

   taup            ~ dgamma(1.0E-3,1.0E-3);
   sigmap         <- 1/sqrt(taup);

   alphamean[1]    <- 2*alpha[2] - alpha[3];
   Nneighsa[1]     <- 1;
   alphamean[2]    <- (2*alpha[1] + 4*alpha[3] - alpha[4])/5;
   Nneighsa[2]     <- 5;
   for (i in 3:(I-2)){
      alphamean[i]    <- (4*alpha[i-1] + 4*alpha[i+1]- alpha[i-2]
```

```
                        - alpha[i+2])/6;
      Nneighsa[i]     <- 6;
   }
   alphamean[I-1]  <- (2*alpha[I] + 4*alpha[I-2] - alpha[I-3])/5;
   Nneighsa[I-1]   <- 5;
   alphamean[I]    <- 2*alpha[I-1] - alpha[I-2];
   Nneighsa[I]     <- 1;
   for (i in 1:I){
      alphaprec[i]    <- Nneighsa[i] * taua;
   }

   for (i in 1:I){
      alpha[i]        ~ dnorm(alphamean[i],alphaprec[i]);
      tau.likea[i]   <- Nneighsa[i] * alpha[i] * (alpha[i]
                           - alphamean[i]);
   }

   da             <- 0.0001 + sum(tau.likea[])/2;
   ra             <- 0.0001 + I/2;
   taua            ~ dgamma(ra,da);
   sigmaa         <- 1/sqrt(taua);

   gammamean[1]    <- 0.0;
   gammaprec[1]    <- tauc*1.0E-6;
   gammamean[2]    <- 0.0;
   gammaprec[2]    <- tauc*1.0E-6;
   for (k in 3:K){
      gammamean[k]    <- 2*gamma[k-1] - gamma[k-2];
      gammaprec[k]    <- tauc
   }

   for (k in 1:K){
      gamma[k]        ~ dnorm(gammamean[k],gammaprec[k]);
   }
   tauc            ~ dgamma(1.0E-3,1.0E-3);
   sigmac         <- 1/sqrt(tauc);
}
```

```
Parsing model declarations.
Loading data value file(s).
Loading initial value file(s).
Parsing model specification.
Checking model graph for directed cycles.
Possible directed cycle or undirected link in model
Generating code.
Generating sampling distributions.
Checking model specification.
Choosing update methods.
compilation took  00:00:00
Bugs>update(500)      time for    500    updates was  00:00:04
Bugs>monitor(mu)
Bugs>monitor(pred.mu)
Bugs>update(2000)      time for    2000   updates was  00:00:14
Bugs>stats(mu)
        mean        sd       2.5% :  97.5%  CI    median      sample
[1]    1.625E+0   4.123E-1   9.143E-1   2.550E+0   1.590E+0     2000
[2]    2.624E+0   4.682E-1   1.805E+0   3.645E+0   2.587E+0     2000
[3]    5.182E+0   6.790E-1   3.863E+0   6.607E+0   5.140E+0     2000
[4]    1.085E+1   1.125E+0   8.738E+0   1.308E+1   1.081E+1     2000
[5]    2.094E+1   1.763E+0   1.771E+1   2.465E+1   2.090E+1     2000
[6]    3.595E+1   2.604E+0   3.110E+1   4.116E+1   3.589E+1     2000
[7]    4.371E+1   2.916E+0   3.827E+1   4.971E+1   4.357E+1     2000
[8]    5.307E+1   3.389E+0   4.663E+1   5.997E+1   5.291E+1     2000
[9]    8.555E+1   4.968E+0   7.634E+1   9.514E+1   8.548E+1     2000
[10]   9.618E+1   5.225E+0   8.629E+1   1.067E+2   9.600E+1     2000
[11]   8.506E+1   4.887E+0   7.557E+1   9.436E+1   8.511E+1     2000
[12]   5.117E+1   3.679E+0   4.383E+1   5.838E+1   5.113E+1     2000
[13]   2.147E+1   2.361E+0   1.708E+1   2.621E+1   2.141E+1     2000
[14]   1.782E+0   4.091E-1   1.051E+0   2.699E+0   1.766E+0     2000
[15]   3.601E+0   5.463E-1   2.578E+0   4.751E+0   3.606E+0     2000
[16]   5.708E+0   6.270E-1   4.477E+0   6.966E+0   5.682E+0     2000
[17]   1.088E+1   9.418E-1   9.017E+0   1.269E+1   1.088E+1     2000
[18]   2.129E+1   1.523E+0   1.852E+1   2.442E+1   2.130E+1     2000
[19]   3.719E+1   2.277E+0   3.288E+1   4.175E+1   3.713E+1     2000
```

[20]	5.716E+1	3.108E+0	5.130E+1	6.348E+1	5.711E+1	2000
[21]	6.230E+1	3.129E+0	5.620E+1	6.855E+1	6.229E+1	2000
[22]	7.200E+1	3.522E+0	6.537E+1	7.921E+1	7.185E+1	2000
[23]	1.096E+2	4.888E+0	1.002E+2	1.194E+2	1.096E+2	2000
[24]	1.104E+2	4.977E+0	1.010E+2	1.208E+2	1.103E+2	2000
[25]	7.427E+1	3.646E+0	6.709E+1	8.158E+1	7.425E+1	2000
[26]	3.315E+1	2.593E+0	2.836E+1	3.828E+1	3.314E+1	2000
[27]	1.868E+0	4.377E-1	1.108E+0	2.886E+0	1.841E+0	2000
[28]	4.026E+0	5.999E-1	2.911E+0	5.324E+0	4.012E+0	2000
[29]	7.860E+0	8.510E-1	6.185E+0	9.571E+0	7.834E+0	2000
[30]	1.205E+1	1.026E+0	1.003E+1	1.413E+1	1.203E+1	2000
[31]	2.135E+1	1.458E+0	1.864E+1	2.437E+1	2.130E+1	2000
[32]	3.814E+1	2.258E+0	3.395E+1	4.249E+1	3.815E+1	2000
[33]	5.962E+1	3.179E+0	5.352E+1	6.609E+1	5.952E+1	2000
[34]	8.203E+1	3.858E+0	7.450E+1	8.972E+1	8.197E+1	2000
[35]	8.556E+1	3.926E+0	7.839E+1	9.383E+1	8.544E+1	2000
[36]	9.311E+1	4.111E+0	8.547E+1	1.011E+2	9.295E+1	2000
[37]	1.294E+2	5.579E+0	1.187E+2	1.414E+2	1.291E+2	2000
[38]	1.007E+2	4.487E+0	9.220E+1	1.095E+2	1.006E+2	2000
[39]	5.123E+1	3.422E+0	4.479E+1	5.843E+1	5.110E+1	2000
[40]	2.001E+0	5.422E-1	1.099E+0	3.267E+0	1.933E+0	2000
[41]	4.267E+0	7.776E-1	2.839E+0	5.883E+0	4.236E+0	2000
[42]	8.826E+0	1.226E+0	6.462E+0	1.139E+1	8.798E+0	2000
[43]	1.667E+1	1.881E+0	1.309E+1	2.065E+1	1.664E+1	2000
[44]	2.372E+1	2.081E+0	1.969E+1	2.805E+1	2.366E+1	2000
[45]	3.845E+1	2.767E+0	3.308E+1	4.424E+1	3.836E+1	2000
[46]	6.106E+1	3.800E+0	5.397E+1	6.848E+1	6.109E+1	2000
[47]	8.581E+1	4.864E+0	7.606E+1	9.516E+1	8.576E+1	2000
[48]	1.128E+2	6.223E+0	1.006E+2	1.253E+2	1.126E+2	2000
[49]	1.127E+2	5.942E+0	1.015E+2	1.244E+2	1.125E+2	2000
[50]	1.122E+2	5.823E+0	1.013E+2	1.248E+2	1.120E+2	2000
[51]	1.231E+2	6.312E+0	1.111E+2	1.354E+2	1.232E+2	2000
[52]	7.503E+1	5.358E+0	6.508E+1	8.576E+1	7.480E+1	2000

```
Bugs>stats(pred.mu)
        mean        sd       2.5% :  97.5%  CI    median      sample
[1]    2.085E+0   6.933E-1   9.606E-1   3.698E+0   1.999E+0     2000
[2]    4.439E+0   1.095E+0   2.439E+0   6.793E+0   4.408E+0     2000
[3]    9.301E+0   1.895E+0   5.695E+0   1.343E+1   9.304E+0     2000
[4]    1.843E+1   3.243E+0   1.244E+1   2.557E+1   1.832E+1     2000
[5]    3.249E+1   4.758E+0   2.362E+1   4.261E+1   3.235E+1     2000
[6]    4.221E+1   5.181E+0   3.286E+1   5.281E+1   4.192E+1     2000
[7]    6.099E+1   6.549E+0   4.900E+1   7.431E+1   6.067E+1     2000
[8]    8.764E+1   8.658E+0   7.105E+1   1.051E+2   8.722E+1     2000
[9]    1.173E+2   1.133E+1   9.680E+1   1.399E+2   1.170E+2     2000
[10]   1.492E+2   1.433E+1   1.225E+2   1.793E+2   1.487E+2     2000
[11]   1.387E+2   1.310E+1   1.155E+2   1.665E+2   1.378E+2     2000
[12]   1.094E+2   1.036E+1   9.102E+1   1.308E+2   1.091E+2     2000
[13]   9.619E+1   1.023E+1   7.742E+1   1.177E+2   9.558E+1     2000
[14]   1.953E+0   8.746E-1   6.396E-1   3.971E+0   1.839E+0     2000
[15]   4.681E+0   1.619E+0   1.951E+0   8.212E+0   4.541E+0     2000
[16]   9.873E+0   3.012E+0   4.726E+0   1.628E+1   9.798E+0     2000
[17]   1.980E+1   5.503E+0   1.065E+1   3.179E+1   1.960E+1     2000
[18]   3.665E+1   9.088E+0   2.192E+1   5.607E+1   3.612E+1     2000
[19]   5.885E+1   1.288E+1   3.712E+1   8.663E+1   5.778E+1     2000
[20]   6.812E+1   1.328E+1   4.527E+1   9.566E+1   6.700E+1     2000
[21]   8.916E+1   1.650E+1   5.992E+1   1.225E+2   8.801E+1     2000
[22]   1.225E+2   2.208E+1   8.569E+1   1.684E+2   1.205E+2     2000
[23]   1.586E+2   2.867E+1   1.112E+2   2.202E+2   1.562E+2     2000
[24]   1.880E+2   3.354E+1   1.319E+2   2.606E+2   1.857E+2     2000
[25]   1.427E+2   2.532E+1   1.002E+2   1.962E+2   1.409E+2     2000
[26]   8.875E+1   1.640E+1   6.046E+1   1.229E+2   8.729E+1     2000
[27]   2.329E+0   1.774E+0   4.096E-1   6.767E+0   1.915E+0     2000
[28]   4.553E+0   2.562E+0   1.162E+0   1.052E+1   4.113E+0     2000
[29]   1.083E+1   5.663E+0   3.310E+0   2.305E+1   9.951E+0     2000
[30]   2.183E+1   1.080E+1   7.605E+0   4.550E+1   2.033E+1     2000
[31]   4.090E+1   1.901E+1   1.582E+1   8.086E+1   3.798E+1     2000
[32]   6.888E+1   2.989E+1   2.921E+1   1.342E+2   6.388E+1     2000
[33]   9.842E+1   4.040E+1   4.410E+1   1.825E+2   9.215E+1     2000
[34]   1.032E+2   4.037E+1   4.789E+1   1.904E+2   9.679E+1     2000
[35]   1.297E+2   5.096E+1   6.087E+1   2.390E+2   1.214E+2     2000
```

```
[36]   1.728E+2   6.726E+1   8.381E+1   3.166E+2   1.617E+2     2000
[37]   2.101E+2   8.338E+1   1.022E+2   3.910E+2   1.964E+2     2000
[38]   2.028E+2   7.919E+1   9.757E+1   3.699E+2   1.902E+2     2000
[39]   1.281E+2   5.036E+1   6.026E+1   2.310E+2   1.212E+2     2000
Bugs>q()
```

ANNEXE 3. ANALYSE DU CANCER DU SEIN INVASIF DANS S-PLUS

A3.1. Tableau de données

Notations

inv : nombre de cas incidents

pop : population

per : période

age : âge comme variable quantitative

coh : cohorte

agef : âge comme variable qualitative

lo(variable,fenêtre) : fonction de lissage non paramétrique

NA : valeur manquante

N° enr.	inv	pop	per	age	coh	agef
1	7.00	56273.00	1.00	1.00	16.00	1
2	8.00	40666.00	1.00	2.00	15.00	2
3	10.00	40155.00	1.00	3.00	14.00	3
4	31.00	40508.00	1.00	4.00	13.00	4
5	53.00	42691.00	1.00	5.00	12.00	5
6	53.00	42597.00	1.00	6.00	11.00	6
….						
95	125.00	39617.00	6.00	15.00	7.00	15
96	72.00	26295.00	6.00	16.00	6.00	16
97	NA	125960.00	7.00	1.00	22.00	1
98	NA	128087.00	7.00	2.00	21.00	2
……..						
143	NA	49941.00	9.00	15.00	10.00	15
144	NA	28493.00	9.00	16.00	9.00	16

A3.2. Modèle à un facteur

Introduction de agef (âge sous forme de variable qualitative)

```
inv.glm1_glm(inv~offset(log(pop/100000))+agef,family=poisson,data=se
inv,na=na.omit)

summary.glm(inv.glm1,cor=F)

> summary.glm(inv.glm1,cor=F)

Call: glm(formula = inv ~ offset(log(pop/100000)) + agef, family =
poisson, data = seinv, na.action = na.omit)
Deviance Residuals:
       Min       1Q     Median       3Q     Max
 -4.045301 -1.37414 -0.3709762 1.189252 4.63622

Coefficients: (1 not defined because of singularities)
                 Value  Std. Error    t value
(Intercept) 4.88508607 0.014185950 344.360877
      agef1 0.47328350 0.082162711   5.760320
      agef2 0.41597785 0.034458851  12.071727
      agef3 0.34492080 0.019213871  17.951656
      agef4 0.31694550 0.012047825  26.307280
      agef5 0.27035283 0.008435941  32.047738
      agef6 0.20248151 0.006625710  30.559971
      agef7 0.15743684 0.005394557  29.184389
      agef8 0.13138416 0.004585009  28.655160
      agef9 0.11554489 0.003922635  29.455936
     agef10 0.10045478 0.003446325  29.148377
     agef11 0.08755011 0.003122125  28.041834
     agef12 0.07651983 0.002945918  25.974865
     agef13 0.06218210 0.002995121  20.761131
```

```
     agef14 0.05086680 0.003234609  15.725794
     agef15 0.04073801 0.004110548   9.910604
     agef16         NA          NA         NA

(Dispersion Parameter for Poisson family taken to be 1 )

    Null Deviance: 5673.8 on 95 degrees of freedom

Residual Deviance: 318.3985 on 80 degrees of freedom

Number of Fisher Scoring Iterations: 3
>
```

A3.3. Modèles à deux facteurs

A3.3.1. Âge et période

A3.3.1.1. Introduction de agef+lo(per,.7)

```
inv.gam2_gam(inv~offset(log(pop/100000))+agef+lo(per,.7),family=pois
son,data=seinv,na=na.omit)
```

Le graphe de la courbe représentative de l'effet des variables explicatives (commande « plot.gam ») montre une relation en S avec per (forme d'une fonction logistique) mais proche d'une fonction logarithme pour les valeurs de per les plus élevées.

D'où :

A3.3.1.2. Introduction de agef+log(per)

```
inv.glm2_glm(inv~offset(log(pop/100000))+agef+log(per),family=poisso
n,data=seinv,na=na.omit)
```

Les graphes (commande « plot.glm ») montrent l'adéquation du modèle.

```
summary.glm(inv.glm2,cor=F)
```

```
> summary.glm(inv.glm2,cor=F)

Call: glm(formula = inv ~ offset(log(pop/100000)) + agef + log(per),
family = poisson, data = seinv, na.action
=
     na.omit)
Deviance Residuals:
       Min         1Q      Median        3Q      Max
 -2.589493 -0.6699477 -0.01086282 0.7680925 2.616129

Coefficients: (1 not defined because of singularities)
                 Value  Std. Error    t value
(Intercept) 4.53665042 0.030113227 150.653082
      agef1 0.46834537 0.082153239   5.700875
      agef2 0.41220415 0.034458694  11.962269
      agef3 0.34292887 0.019213297  17.848517
      agef4 0.31710961 0.012047282  26.322086
      agef5 0.27227539 0.008436775  32.272447
      agef6 0.20510057 0.006628265  30.943328
      agef7 0.15916332 0.005396459  29.494030
      agef8 0.13066885 0.004585439  28.496474
      agef9 0.11519681 0.003922676  29.366896
     agef10 0.10108879 0.003446489  29.330951
     agef11 0.08870832 0.003123211  28.402918
     agef12 0.07781559 0.002947602  26.399622
     agef13 0.06262915 0.002995347  20.908817
     agef14 0.04972038 0.003235567  15.366820
```

```
     agef15 0.03882226 0.004112872   9.439208
     agef16         NA          NA         NA
   log(per) 0.27421954 0.020372462  13.460304

(Dispersion Parameter for Poisson family taken to be 1 )

    Null Deviance: 5673.8 on 95 degrees of freedom

Residual Deviance: 128.427 on 79 degrees of freedom

Number of Fisher Scoring Iterations: 3
>
```

A3.3.2. Âge et cohorte

A3.3.2.1. Introduction de agef+lo(coh,.7)

```
inv.gam3_gam(inv~offset(log(pop/100000))+agef+lo(coh,.7),family=pois
son,data=seinv,na=na.omit)
```

Le graphe représentatif de l'effet (commande « plot.gam ») montre une relation strictement linéaire avec coh.

Donc :

A3.3.2.2. Introduction de agef+coh

```
inv.glm3_glm(inv~offset(log(pop/100000))+agef+coh,family=poisson,dat
a=seinv,na=na.omit)
```

```
Les graphes (commande « plot.glm » montre l'adéquation du modèle).
```

```
summary.glm(inv.glm3,cor=F)
```

```
> summary.glm(inv.glm3,cor=F)

Call: glm(formula = inv ~ offset(log(pop/100000)) + agef + coh,
family = poisson, data = seinv, na.action = na.
omit)
Deviance Residuals:
       Min         1Q    Median        3Q      Max
 -2.379869 -0.5861652 0.1398144 0.7196561 2.469883

Coefficients: (1 not defined because of singularities)
                 Value  Std. Error   t value
(Intercept) 3.81711652 0.074603414 51.165440
      agef1 0.51527140 0.082222368  6.266803
      agef2 0.45830607 0.034583393 13.252201
      agef3 0.38869387 0.019443110 19.991343
      agef4 0.36297396 0.012444193 29.168139
      agef5 0.31867648 0.009048115 35.220204
      agef6 0.25200554 0.007429347 33.920282
      agef7 0.20605604 0.006325176 32.577124
      agef8 0.17761611 0.005561812 31.934935
      agef9 0.16160195 0.005017057 32.210506
     agef10 0.14736155 0.004688212 31.432356
     agef11 0.13514669 0.004488511 30.109469
     agef12 0.12463954 0.004399482 28.330504
     agef13 0.10954059 0.004398336 24.905004
     agef14 0.09623964 0.004469806 21.531053
     agef15 0.08513546 0.005100507 16.691569
     agef16         NA          NA        NA
        coh 0.09311415 0.006322688 14.726989

(Dispersion Parameter for Poisson family taken to be 1 )

    Null Deviance: 5673.8 on 95 degrees of freedom

Residual Deviance: 98.5718 on 79 degrees of freedom
Number of Fisher Scoring Iterations: 3
>
```

A3.4. Modèle à trois facteurs

Introduction de agef+log(per)+coh

```
inv.glm_glm(inv~offset(log(pop/100000))+agef+log(per)+coh,family=poi
sson,data=seinv,na=na.omit)

summary.glm(inv.glm,cor=F)

> summary.glm(inv.glm,cor=F)

Call: glm(formula = inv ~ offset(log(pop/100000)) + agef + log(per)
+ coh, family = poisson, data = seinv, na.a
ction
      = na.omit)
Deviance Residuals:
       Min         1Q    Median        3Q      Max
 -2.645753 -0.6197745 0.1507556 0.6347573 2.495563

Coefficients: (1 not defined because of singularities)
                 Value Std. Error    t value
(Intercept)  3.4093737 0.19182559  17.773300
      agef1  0.5436422 0.08314706   6.538320
      agef2  0.4861005 0.03661910  13.274507
      agef3  0.4161821 0.02279588  18.256903
      agef4  0.3903851 0.01719421  22.704446
      agef5  0.3462903 0.01499101  23.099867
      agef6  0.2798713 0.01416659  19.755732
      agef7  0.2339734 0.01363995  17.153537
      agef8  0.2057255 0.01337929  15.376413
      agef9  0.1893592 0.01301975  14.544005
     agef10  0.1749850 0.01284336  13.624550
     agef11  0.1628328 0.01279690  12.724392
```

```
     agef12  0.1525419 0.01285405  11.867225
     agef13  0.1375522 0.01289888  10.663880
     agef14  0.1241207 0.01286986   9.644294
     agef15  0.1129426 0.01307257   8.639659
     agef16         NA         NA         NA
   log(per) -0.1815291 0.07840525  -2.315268
        coh  0.1487737 0.02490863   5.972779

(Dispersion Parameter for Poisson family taken to be 1 )

    Null Deviance: 5673.8 on 95 degrees of freedom

Residual Deviance: 93.26714 on 78 degrees of freedom

Number of Fisher Scoring Iterations: 3
>
```

A3.5. Comparaison des modèles à l'aide de l'AIC

```
AIC(inv.glm1)
AIC(inv.glm2)
AIC(inv.glm3)
AIC(inv.glm)

> AIC(inv.glm1)
glm(formula = inv ~ offset(log(pop/100000)) + agef, family =
poisson, data = seinv, na.action = na.omit)
Degrees of Freedom Total =  96
Degrees of Freedom Residual =  80
Residual Deviance =  318.3985
AIC= 350.3985
> AIC(inv.glm2)
glm(formula = inv ~ offset(log(pop/100000)) + agef + log(per),
family = poisson, data = seinv, na.action =
     na.omit)
Degrees of Freedom Total =  96
```

```
Degrees of Freedom Residual =  79
Residual Deviance =  128.427
AIC= 162.427
> AIC(inv.glm3)
glm(formula = inv ~ offset(log(pop/100000)) + agef + coh, family =
poisson, data = seinv, na.action = na.omit)
Degrees of Freedom Total =  96
Degrees of Freedom Residual =  79
Residual Deviance =  98.5718
AIC= 132.5718
> AIC(inv.glm)
glm(formula = inv ~ offset(log(pop/100000)) + agef + log(per) + coh,
family = poisson, data = seinv, na.action
      = na.omit)
Degrees of Freedom Total =  96
Degrees of Freedom Residual =  78
Residual Deviance =  93.2671
AIC= 129.2671
>
```

Donc la variance résiduelle et l'AIC montrent que le meilleur modèle est celui qui prend en compte agef + log(per) + coh

ANNEXE 4. JOSIAH WILLARD GIBBS

Les notes qui suivent sont inspirées d'un texte découvert sur un site Internet (www-groups.dcs.st-andrews.ac.uk/~history/Matematicians/Gibbs.html).

Josiah Willard Gibbs était professeur de littérature sacrée à l'université de Yale. Profession respectable mais n'ayant rien à voir *a priori* avec les mathématiques et encore moins avec les statistiques. Sans doute conscient de la nécessité historique d'inscrire un jour le patronyme dans une littérature tout aussi respectable mais statistique celle-là, il a décidé de mettre au monde le 11 février 1839 à New Haven, Connecticut, USA, un fils auquel il donna, comme cela se fait souvent en Amérique, non seulement le même nom mais les mêmes prénoms que lui. Il ne mit bien évidemment pas ce fils au monde lui-même mais confia ce soin à son épouse, mère par conséquent du deuxième du nom dont on dit qu'il tenait d'elle son physique… Entre le moment de sa naissance et celui de sa mort le 28 avril 1903 (à New Haven, Connecticut, USA, aussi), 65 années du dit fils ont été dédiées à la vie et plus de 35 ans à la science.

Josiah Willard Gibbs (II ou Jr, par conséquent) entre au Yale College à l'âge de 15 ans. Il en sort avec un prix d'excellence en latin et en mathématiques. Il entreprend des études d'ingénieur ainsi qu'une thèse dans laquelle il étudie la forme des engrenages. Son doctorat obtenu à 25 ans, il enseigne pendant trois ans le latin et la philosophie! Puis il embarque pour l'Europe où il étudie la physique pendant 4 ans (il avait hérité un petit pécule de ses parents, décédés quelques années auparavant) : Paris, Berlin puis Heidelberg où il rencontre Helmoltz. L'Europe faisait partie de l'histoire de la famille puisque celle-ci était originaire de Warwickshire en Angleterre et avait émigré à Boston en 1658. Il retourne à Yale en 1869 où, deux ans plus tard, il devient professeur de physique. Ses premières publications datent de 1873 (il avait 34 ans) : *Graphical methods in the thermodynamics of fluids* et *A method of geometrical representation of the thermodynamic poperties of substances by means of surfaces*. Trois ans plus tard il publie la première partie du travail qui l'a rendu célèbre : *On the equilibrium of heterogeneous substances*.

Outre ses recherches en thermodynamique, Gibbs a travaillé dans le domaine de l'analyse vectorielle, la physique des corps célestes, la théorie électromagnétique, la mécanique statistique (avec les conséquences que l'on sait sur la mécanique quantique).

Un ensemble d'ouvrages et d'articles ont été écrits sur J Willard Gibbs. Il est possible de consulter une liste de références sur le même site :

www-groups.dcs.st-andrews.ac.uk/~history/References/Gibbs.html

ARTICLES

RÉSUMÉ

Le cancer mobilise d'importantes ressources sanitaires et pèse fortement sur la mortalité de la population générale. Appréhender son évolution dans les années futures est donc essentiel pour la santé publique. L'objectif de ce travail est d'estimer l'incidence et le nombre de cas incidents à venir pour les cancers dans le Bas-Rhin et de tester la validité d'une méthode d'analyse fondée sur une approche bayésienne.

La méthode retenue se base sur un modèle âge-période-cohorte analysé grâce à un échantillonnage de Gibbs. Les nombres de cas incidents annuels, base de la prévision, sont extraits du registre des cancers du Bas-Rhin qui existe depuis 1975. Les effectifs de populations actuels et futurs sont estimés par l'INSEE. La méthode est explicitée puis appliquée à quatre des cancers les plus fréquents : cancer du sein invasif, du col de l'utérus (*in situ* et invasif), du poumon et cancer colo-rectal, dans les deux sexes selon le cas. Les prévisions, établies jusqu'en 2009 grâce à cette méthode, sont comparées aux résultats obtenus par d'autres techniques.

L'analyse met en évidence une croissance future de l'incidence des cancers invasifs du sein et des tumeurs *in situ* du col. L'incidence du cancer invasif du col diminue. Une augmentation de l'incidence du cancer du poumon et du côlon est prévue dans les deux sexes. Le cancer du rectum chez la femme augmente également ; chez l'homme, par contre, la prédiction va dans le sens d'une légère diminution.

TITLE

Prediction of the incidence of cancer in the Bas-Rhin. Bayesian method. Principles and applications.

ABSTRACT

Cancer is the cause of a large number of deaths. Predicting its evolution is important in Public Health. The objective of this work is to predict the trend in the incidence and in number of incident cases of cancer among population in the Bas-Rhin and to assess a Bayesian approach.

The method is based on an age-period-cohort model analysed by a Gibbs sampler. The number of incident cases was provided by the Bas-Rhin Tumour Registry. The population of the Department was estimated by the INSEE. Four cancers are studied : breast, cervix, lung and colorectal cancer. Trends are projected to 2009.

Incidence of invasive cancer of breast and in situ tumours of cervix are predicted to increase. Incidence of invasive cancer of cervix is predicted to decrease. Lung and colorectal cancers and rectum cancer in women increase in the two sexes but rectum cancer decrease in males.

MOTS-CLÉS

Épidémiologie, cancer, incidence, prévision, modèle âge-période-cohorte, statistique bayésienne, méthodes de Monte Carlo par chaînes de Markov, échantillonnage de Gibbs.

KEYWORDS

Epidemiology, cancer, incidence, prediction, age-period-cohort model, Bayesian statistics, Markov Chain Monte Carlo methods, Gibbs sampler.