Prise en compte de la surdispersion par des modèles à mélange de Poisson

Sebastien Marque

UNIVERSITE VICTOR SEGALEN - BORDEAUX 2

Année 2003 Thèse n° xxx

THESE

pour le

DOCTORAT DE L'UNIVERSITE DE BORDEAUX 2

Mention : Sciences Biologiques et Médicales

Option : Epidémiologie et Intervention en Santé Publique

présentée et soutenue publiquement

le 03 Décembre 2003

par **Sébastien Marque**

Né le 11 décembre 1975 à Mont de Marsan

Prise en compte de la surdispersion par des modèles à mélange de Poisson

Membres du Jury

Monsieur le Professeur	Patrick BROCHARD	Président
Monsieur le Professeur	Avner BAR-HEN	Rapporteur
Monsieur le Professeur	Jean MACCARIO	Rapporteur
Madame le Docteur	Isabelle BALDY	Examinateur
Monsieur le Docteur	Daniel COMMENGES	Directeur de thèse

Résumé

Cette thèse propose une approche opérationnelle permettant de traiter des données environnementales surdispersées. Cette surdispersion, qui peut avoir pour origine une mauvaise spécification du modèle ou un recueil de données incomplet, entraîne un biais important dans l'estimation des paramètres.

Ce travail propose donc une approche basée sur la régression Arcsinus Stricte comme alternative à la régression Binomiale-Négative. Le second aspect est abordé en présentant un modèle hiérarchique encore méconnu en épidémiologie et une extension possible aux corrélations spatiales, qui permet de compléter l'information disponible dans les études écologiques. Chacun de ces deux aspects seront détaillés d'un point de vue théorique et par des études de simulation.

Enfin, nous préciserons les caractéristiques de la mortalité cardiovasculaire chez les personnes âgées par une analyse démographique complète. Nous détaillerons ensuite les facteurs de risque usuels de cette cause de décès ainsi que l'effet des éléments minéraux de l'eau de boisson, et principalement le calcium et le magnésium.

Mots-clé : Surdispersion, données groupées, modèles hiérarchiques, mélange de Poisson, Binomiale-Négative, données agrégées, études semi-écologiques.

Abstract

This thesis proposes a pragmatic approach in the frame of environmental data which are overdispersed. This overdispersion, which could be explained by a mispecification of the model or by an incomplete data collection, leads to an important bias in the estimation of parameters.

This research work proposes an approachbased on Strict Arcsine regression as an alternative as Binomial-Negative regression. The second point which is proposing concerns the presentation of an hierarchical model, relatively recent in epidemiology, and an potential extension to spatial correlations, which could extend information available in ecological studies by individual factors. Both of these aspects will be detailled theorically and by simulation studies.

In conclusion, we'll try to precise the different caracteristics of cardiovascular mortality among elderly by demographic study. In a second part, we'll detaill risk factors of this cause of death, and specially the effect of calcium and magnesium contained in drinking water.

Keywords : Overdispersion, grouped data, hierachical models, Poisson mixture, Negative-Binomial, Aggregated data, ecological study.

Table des matières

Table des figures

Liste des tableaux

Introduction

Position du problème

Le point de départ de ce travail concernait l'étude de l'effet des éléments minéraux de l'eau de boisson sur la mortalité cardiovasculaire chez les personnes âgées, et ses implications en terme de modélisation statistique. Nous nous sommes donc placés dans le cadre d'études environnementales, avec la volonté d'étudier les estimations des paramètres individuels et collectifs (propre à un sous-groupe de la population). Wakefield [109] a très récemment publié un article sur la réanalyse d'une étude portant sur des données environnementales (association entre la mortalité cardiovasculaire et le magnésium dans l'eau de boisson) et a mis en évidence de nombreux problèmes inhérents à ce type d'étude. Notre cadre de recherche est donc celui des études environnementales, s'intéressant particulièrement aux variations géographiques.

Dans ce type de schémas d'étude, différents niveaux de recueil sont envisageables. En effet, il est possible de collecter des données individuelles (consommation de tabac, hypertension) mais également des données "collectives", qui sont identiques chez les individus d'un même groupe.

Morgenstern [75] a montré que ces données peuvent être classées en deux catégories qui sont les données groupées et les données agrégées. Il est très important de faire la distinction entre ces deux catégories de variables car elles ne représentent pas le même niveau de collecte ni les mêmes implications en terme d'interprétation (et d'inférence) épidémiologique ou clinique. Les données "agrégées" sont des données recueillies au niveau de l'individu, qui ont été regroupées au sein d'une variable composite, contruite à partir de plusieurs personnes (composant un groupe ou une strate). C'est le cas, par exemple, de la proportion de fumeurs dans une unité géographique. Au contraire, les données groupées sont des variables qui existent et sont recueillies à un niveau supérieur à celui de l'individu, comme les zones géographiques, et qui n'ont pas d'interprétation (inférentielle) individuelle. C'est notamment le cas des variables environnementales, telles que celles traitant de la pollution ou de caractéristiques géographique (ensoleillement par exemple). Les situations où ce recueil peut apparaitre en épidémiologie sont nombreuses, notamment lors d'études concernant la pollution ou sur des pratiques de consommation (mesure de pollution par quartier, zone géographique ou enquêtes alimentaires par exemple).

Il est classique de traiter ces données qui sont recueillies à un niveau supérieur à l'individu par des modèles linéaires généralisés. En effet, cette classe de modèle permet de modéliser une variable aléatoire (par exemple, processus de dénombrement caractérisant le nombre d'événements observés dans une unité statistique) par différentes variables explicatives. Cependant, les variables explicatives incluses dans ces modèles doivent être recueillies au même niveau que la variable à expliquer. Il existe de nombreux problèmes en terme d'interprétation épidémiologique

que l'on retrouve sous le vocable de "biais écologiques" [32, 76, 87] lorsque ces analyses mêlent des données issues de ces deux sources. Ces biais apparaissent principalement lorsque les caractéristiques individuelles sont estimées à partir d'un niveau "groupe". Cependant, malgré la présence de ces biais, ces schémas d'étude écologiques restent très intéressants, principalement en raison d'une meilleure disponibilité de ces données écologiques (en terme de coût et de quantité). Une très bonne revue des principaux problèmes liés aux études multi-niveaux (également retrouvées sous le terme de semi-agrégées ou semi-écologiques) a été publiée récemment par Blakely *et al* [7].

De plus, lors de l'étude d'associations dans le cadre environnemental, les risques relatifs considérés sont très faibles (proches de 1) et leur interprétation est rendue difficile par l'importance que prennent alors l'ensemble des biais et le non respect des hypothèses sous-jacentes aux modèles utilisés. Une des hypothèses inhérentes à ces modèles de régression concerne l'indépendance nécessaire entre les unités statistiques (zones géographiques par exemple). Lorsque les individus d'une même unité géographique sont corrélés entre eux, cette notion d'indépendance est violée et cela peut conduire à faire varier le niveau de surdispersion intra- et inter- unité statistique.

La surdispersion peut être définie comme une variance observée des données supérieure à la variance théorique issue du modèle utilisé. De manière similaire, la sous-dispersion peut se définir comme une variance observée inférieure à celle induite parle modèle. Toutefois, cette dernière situation est très rarement rencontrée dans les problématiques épidémiologiques et l'explication de telles situations est relativement complexe à définir. Cependant, il existe

d'autres définitions de cette surdispersion. Dans le cadre de la modélisation de données de dénombrement, la surdispersion est classiquement définie par une variance supérieure à la moyenne (variance induite par l'utilisation d'un modèle de Poisson). Dans la suite de ce travail, seule la surdispersion, qui peut ainsi être définie comme un manque d'information prise en compte dans le modèle [32], sera considérée. Ce défaut de prise en compte peut être le résultat d'un manque d'information dû à une mauvaise connaissance éthiologique du processus qui nous intéresse (une mauvaise qualité de recueil ou un manque de variables retenues dans le protocole) ou il peut résulter d'un modèle mal spécifié (l'inadéquation du modèle peut être induite, par exemple, par la non prise en compte de l'hétérogénéïté spatiale).

Nous avons décidé d'aborder cette problématique en classant les différentes techniques qui ont été développées pour prendre en compte cette surdispersion en deux approches distinctes (distinction inhérente à la source de la surdispersion définie ci-dessus).

La première approche est basée sur le développement de modèles permettant un meilleur ajustement statistique qui conduit à une meilleure inférence. En effet, la classe des modèles linéaires généralisés spécifie que la variance est une fonction de la moyenne et il est possible de "travailler" sur une formulation différente de la variance (notamment en "relachant" les contraintes inhérentes aux hypothèses de ces modèles) afin d'obtenir un modèle plus adapté aux données. Cependant, cette approche, purement statistique, n'apporte pas de nouveaux éléments d'interprétation aux épidémiologistes ou cliniciens, puisqu'elle considère la surdispersion comme le résultat unique dun mauvaise spécification du modèle.

La seconde approche se place du côté de ces utilisateurs, en augmentant l'information dispo-

nible, ce qui se traduit par un meilleur ajustement aux données et une diminution de la surdispersion. Cet aspect nécessite l'introduction de nouvelles variables explicatives dans le modèle. Cette approche conduit cependant à d'autres problèmes méthodologiques liés à la combinaison de variables [75, 87] issues de multiples recueils dans des modèles "classiques". En effet, une grande majorité des problèmes de santé publique sont le résultat de combinaison de multiples facteurs de risques qui sont à la fois individuels et agrégés (environnementaux, génétique,...).

Plan de la thèse

Dans le cadre de ce travail, nous nous sommes intéressés à identifier la "meilleure" modélisation permettant de prendre en compte la surdispersion dans des modèles présentant une structure hiérarchique dans le recueil de données.

Dans le premier chapitre, les différentes techniques qui sont classiquement utilisées pour analyser les données de dénombrement et leurs méthodes d'estimation associées seront présentées. Nous nous attacherons à détailler les différentes techniques d'estimation des modèles linéaires généralisés mixtes, notamment par maximum de vraisemblance et par quasi-vraisemblance.

Dans le deuxième chapitre, nous présenterons un nouveau modèle, basé sur la loi Arc-sinus Stricte, comme une alternative à la loi Binomiale-Négative, permettant de prendre en compte la surdispersion dans des modèles exponentiels. En effet, les récents travaux de Koko-

nendji [50, 53] ont montré que la loi Binomiale-Négative n'était pas toujours la meilleure approche. Nous présenterons également une modification de l'extension du modèle hiérarchique agrégé, proposée par Guthrie, qui permet une meilleure interprétation épidémiologique. Chacune de ces deux parties fera l'objet d'une étude par simulation permettant d'évaluer et de bien appéhender les propriétés de ces méthodes.

Le chapitre 3 présentera la modélisation de la mortalité cardiovasculaire chez les personnes âgées et notamment l'impact du calcium et du magnésium contenus dans l'eau de boisson sur cette cause de décès. Dans ce chapitre, l'aspect démographique de cette cause de mortalité, ainsi que les différents facteurs de risque associés seront évalués. Puis, dans une dernière partie, nous présenterons les résultats issus de la modélisation par l'approche déterminée dans les chapitres précédents.

Une conclusion présentera l'originalité statistique de ce travail et commentera les résultats épidémiologiques sur la mortalité cardiovasculaire des personnes âgées.

Chapitre 1

Introduction théorique

Frome a montré que les modèles linéaires généralisés [28–30] sont particulièrement bien adaptés à l'analyse des données issues des processus de dénombrement. Cependant, pour prendre en compte certains éléments lors de cette analyse (surdispersion, structure spatiale), ces modèles ont été agrémentés d'une composante aléatoire. Les Modèles Linéaires Généralisés Mixtes (GLMM) sont ainsi apparus à la fin des années 1980.

La première partie présentera la théorie des Modèles Linéaires Généralisés Mixtes ainsi que les techniques d'estimations qui leur sont associées. Nous décrirons l'introduction des effets aléatoires dans ces modèles et notamment le cas de la loi Binomiale-Négative.

La seconde partie de ce chapitre sera consacrée à la présentation d'un modèle beaucoup plus récent permettant de prendre en compte simultanément des données individuelles et groupées. Ce modèle hiérarchique, à deux niveaux, est modélisé par un GLMM avec une technique d'estimation par Quasi-Vraisemblance. Cette technique, basée sur les Equations d'Estimations Généralisées (GEE), permet de modéliser les deux niveaux de recueil de données.

1.1 Modèles Linéaires Généralisés Mixtes

Les modèles linéaires généralisés (GLM) sont les modèles usuels qui permettent de modéliser les données de dénombrement (“count data”). Ces données de dénombrement sont par exemple le nombre de décès observés dans une région, durant une période de temps. Dans cette classe de modèles, les unités d’agrégation sont classiquement considérées comme unité statistique. Par exemple, dans notre application l’unité statistique que nous considérons est l’agrégation géographique définie par la commune de résidence. L’objectif est ainsi de modéliser le nombre d’événements observés, dans chacune des unités, à l’aide de variables explicatives.
Cependant, afin d’affiner l’ajustement de cette modélisation, il est possible de décomposer cette unité statistique en sous-unité par stratification. Cette stratification peut être définie par les modalités d’une variable (sexe par exemple) ou par la combinaison de modalités de plusieurs variables (sexe et classes d’âge par exemple). Il est ainsi possible de modéliser le nombre d’événements observés sur chacune des sous-populations considérées.

1.1.1 Modèles Linéaires Généralisés

Par la suite, l’indice k représentera la commune ($k \in \{1,...,K\}$) et l’indice s désignera la strate au sein de la commune ($s \in \{1,...,S\}$). Les variables observées seront donc doublement indicées par k et s.

Présentation

Les modèles linéaires généralisés (GLM) ont été introduits par Nelder et Wedderburn [79] en 1972 comme une extension des modèles linéaires classiques et ont été formalisés par McCullagh & Nelder [72]. Cette classe de modèles regroupent de nombreuses lois usuelles, telles que les lois Binomiale, Normale, Gamma ou de Poisson.

Nous nous situons dans le cadre de Familles Exponentielles Naturelles (FEN) et la fonction liant le prédicteur et le paramètre est donc exponentielle (nous nous plaçons dans le cas particuliers des modèles log-linéaires). On peut ainsi écrire que $\mathbb{E}[Y_{ks}] = \mu_{ks} = exp\{X_{ks}\beta\}$, où X_{ks} est la matrice des observations sur la strate k,s et β est le vecteur des paramètres à estimer.

Ces modèles sont basés sur la densité suivante :

$$f(Y,\theta_{ks}) = \exp\left\{\frac{Y\theta_{ks} - b(\theta_{ks})}{a(\phi)} + c(y,\phi)\right\} \tag{1.1}$$

où θ_{ks} est le paramètre canonique et ϕ (resp. $a(\phi)$) le paramètre de surdispersion (resp. la fonction de variance). Les fonctions $b(\theta_{ks})$ et $c(y,\phi)$ représente les fonctions spécifiques de chaque loi.

L'expression des deux premiers moments du modèle est aisément déduite de l'équation 1.1 :

$\mathbb{E}[Y_{ks}] = b'(\theta_{ks})$ et Var$[Y_{ks}] = a(\phi).b''(\theta_{ks})$.

Le tableau 1.1 présente des exemples de lois issues de l'équation 1.1 en spécifiant quelles en sont les fonctions associées (a, b, c). Ces différentes fonctions permettent de définir les paramètres θ_{ks} et ϕ qui caractérisent les lois d'une même famille.

TAB. 1.1 – *Paramètres de certaines lois issues de Familles Exponentielles Naturelles*

Lois	θ	b(θ)	a(ϕ)
Binomiale	$\ln(\frac{p}{1-p})$	$n\ln(1+\exp(\theta))$	1
Poisson	ln(λ)	exp(θ)	1
Normale	μ	$\theta^2/2$	σ^2
Gamma	$(a\lambda)^{-1}$	$ln(\theta)$	a^{-1}
Exponentielle	λ^{-1}	$ln(\theta)$	-1

Certaines lois issues de FEN n'ont qu'un seul paramètre à estimer. En effet, le paramètre de surdispersion ($a(\phi)$) est connu (et souvent fixé à l'identité) et cette contrainte ne permet donc pas de prendre en compte un coefficient de surdispersion. C'est notamment le cas des lois Binomiale, Exponentielle et de Poisson et c'est principalement dans ces modèles qu'apparaissent les problèmes de surdispersion. En effet, la variance observée est supérieure à la variance théorique induite par le modèle sous-jacent puisque le paramètre de surdispersion est fixé à une constante. La loi Binomiale-Négative est très souvent considérée comme la référence dans les cas de surdispersion dans la loi de Poisson car son paramètre de surdispersion est libre. Nous reviendrons sur ce cas particulier dans le chapitre suivant.

Estimation

Classiquement, l'estimation des paramètres de ces modèles est réalisée par la méthode du Maximum de Vraisemblance car il est ainsi possible d'obtenir des estimateurs sans biais, sous des conditions peu contraignantes. L'écriture de la log-vraisemblance du modèle, déduite de l'équation 1.1, s'écrit :

$$l = \sum_{k,s} l_{ks} = \sum_{k,s} [Y\theta_{ks} - b(\theta_{ks})]/a_{ks}(\phi) + c(Y,\phi)$$

L'expression analytique de la première dérivée de l par rapport au paramètre β_j peut s'écrire :

$$\frac{dl}{d\beta_j} = \sum_{k,s} X_{ks}^j \frac{g'(\mu_{ks})(Y - \mu_{ks})}{g'(\mu_{ks})^2 V[Y]}$$

Nous utilisons la notation $\sum_{k,s} \equiv \sum_{k=1}^{K} \sum_{s=1}^{S}$.

Cependant, l'équation annulant cette dérivée a rarement de solution analytique, et les solutions sont donc obtenues à l'aide d'un algorithme itératif [1] de type Newton-Raphson ou des scores de Fisher ("Fisher-scoring"). Les estimations des variances des paramètres sont obtenues en utilisant la dérivée seconde de cette log-vraisemblance.

La surdispersion

Une correction de la variance est possible en présence de surdispersion avérée [10, 101]. En effet, à partir de la statistique du χ^2 de Pearson définie par $X^2 = \sum_{k,s} \frac{(y_{ks} - \mu_{ks})^2}{\mu_{ks}}$ et suivant une loi χ^2_{n-p}, il est possible de "redresser" la variance des estimateurs en utilisant la proposition suivante :

Variance observée = ϕ Variance théorique

où ϕ est le paramètre de surdispersion. Il est possible de réaliser une estimation de cette surdispersion en définissant : $\phi = X^2/(n-p)$. Cette technique a l'avantage d'être applicable quelque soit le modèle exponentiel considéré et cette correction est opérationnelle facilement et rapidement. Cependant, elle a l'inconvénient majeur de ne pas autoriser d'inférence sur le paramètre de surdispersion (estimation empirique). En effet, il est difficile de calculer une estimation de la variance de l'estimateur de cette surdispersion et, par conséquence, de déterminer si elle est

statistiquement significative.

Exemple d'un GLM : la régression de Poisson

La régression de Poisson est classiquement utilisée pour modéliser des données de dénombrement ($Y_{ks} \sim P(\mu_{ks})$). La loi de Poisson permet également d'approximer certaines autres lois sous certaines hypothèses de convergence, telle que la loi Binomiale par exemple (*Cf.* tableau 1.1). La distribution sur laquelle le modèle s'appuie peut être formulée ainsi :

$$P[Y_{ks} = y] = \frac{e^{-\mu_{ks}} \mu_{ks}^{y}}{y!} \tag{1.2}$$

Les premiers moments de cette distribution sont $\mathbb{E}[Y_{ks}] = \mu_{ks} = \exp\{X_{ks}\beta\}$ et $\text{Var}[Y_{ks}] = \mu_{ks}$. On remarquera ainsi que l'espérance est identique à la variance (hypothèse implicite de la loi de Poisson). Dans la pratique, il est couramment observé une variance supérieure à l'espérance, ce qui traduit la présence d'une surdispersion des données par rapport au modèle utilisé.

L'estimation des paramètres de ce modèle de Poisson est communément réalisée par la méthode du maximum de vraisemblance, à partir de l'expression ci-dessous :

$$l = \sum_{k=1}^{K} l_{ks} = \sum_{k=1}^{K} \sum_{s=1}^{S} (\bar{y}_{ks} \ln(\mu_{ks}) - \mu_{ks} - \ln(\bar{y}_{ks}!))$$

La validation de l'ajustement du modèle de Poisson peut être réalisée par le test défini par Hosmer & Lemeshow [45], basé sur la loi du χ^2.

1.1.2 Modèles Linéaires Généralisés Mixtes (GLMM)

Dans cette partie, nous nous placerons dans le cadre restreint des GLMM définis par une distribution conditionnelle de Poisson. En effet, ce cadre est classiquement utilisé dans les études environnementales, qui motivent ce travail. Cependant, nous essaierons de généraliser cette présentation aux différents cas de FEN. Une très bonne description de ces modèles GLMM a été développée par McCulloch [73].

Présentation des GLMM

Cette classe de modèle a été définie, dans le cadre poissonnien, par Lawless [56] en 1987 et se présente comme une extension des GLM par adjonction d'une structure d'effets aléatoires [3, 46]. Le principe sous-jacent est similaire à celui qui a conduit à l'extension des modèles linéaires aux modèles linéaires mixtes dans les années 1970.
Cette extension est basée sur la spécification d'une distribution conditionnelle aux effets aléatoires, distribution qui reste une FEN, telle que décrite dans le paragraphe précédent. L'effet aléatoire (unidimensionnel) est considéré comme réprésentant l'hétérogénéïté induisant la surdispersion. Dans notre cas, nous nous intéressons à la surdispersion engendrée par une corrélation propre aux unités d'agrégation, telles que les communes. Il suit que les effets aléatoires (h_k) portent donc sur ces mêmes unités. On peut ainsi utiliser la notation suivante : $Y|h_k \sim FEN$.

Il est par conséquent possible de spécifier les deux premiers moments de ces modèles marginaux par : $\mathbb{E}[Y_{ks}] = h_k exp(X_{ks}\beta) = \mu_{ks}$ et $V[Y_{ks}] = \mathbb{E}[V[Y_{ks}|h_k]] + V[\mathbb{E}[Y_{ks}|h_k]]$

Le modèle ainsi caractérisé est un modèle conditionnel et dépendant de la loi de l'effet aléatoire h_k. Dans le cadre de données environnementales, il est classique de considérer que le modèle conditionnel suit une loi de Poisson. D'après les remarques de Cox [17], les estimations des paramètres ne sont pas biaisées par la présence d'une surdispersion, faible ou modérée, et ainsi nous considèrerons que $\mathbb{E}[h_k] = 1$ et $V[h_k] = \sigma^2$.

Nous nous retrouvons donc dans la situation d'un modèle avec deux paramètres : un vecteur de paramètres d'intérêt (β) et un paramètre de dispersion (σ^2) qui peut être considéré comme un paramètre de nuisance.

Remarque:

Lorsque l'on fixe $h_k = 1$, il s'ensuit que $\mu_{ks} = \exp\{X_{ks}\beta\}$, ce qui entraine l'égalité entre les deux premiers moments de la distribution ($\mathbb{E}[Y_{ks}] = V[Y_{ks}] = \mu_{ks}$). Il est possible de retrouver le modèle de Poisson décrit précédemment lorsque l'effet aléatoire est fixé à l'identité.

Problèmes d'estimation et solutions

Afin de réaliser les estimations des différents paramètres, il est nécessaire de déterminer la distribution marginale (Y_{ks}) du modèle en déconditionnant la distribution de ($Y_{ks}|h_k$) par rapport aux effets aléatoires (h_k). Cependant, cette technique pose de nombreux problèmes car il existe rarement de formes analytiques explicites pour les vraisemblances marginales.
Une solution classique consiste à déterminer une distribution de l'effet aléatoire qui conduise à

une distribution marginale ayant une forme connu et qui permette ainsi de mettre en œuvre les techniques d'estimations classiques par maximum de vraisemblance.
Lorsque la log-vraisemblance n'a pas de forme analytique, les estimations des paramètres sont obtenues en utilisant des approximations telles que la méthode du 1er ou 2nd ordre basé sur les séries de Taylor [5] ou des intégrations numériques, notamment par Quadrature de Gauss [105]. Cette dernière méthode est basée sur l'approximation des intégrales par des sommations pondérées sur des points prédéfinis. Ces méthodes d'estimation ont l'inconvénient majeur de nécessiter des calculs très couteux en temps, ce qui empêche une intégration sur plusieurs effets aléatoires si on considère de nombreux points de quadrature.

Modèles Linéaires Hiérarchiques (HLM)

Comme nous avons pu le voir dans les sections précédentes, les estimations des paramètres dans le cas général sont loin d'être aisées. Pour remédier à ce problème, de nombreux auteurs se sont récemment intéressés à l'apport de l'approche bayésienne dans cette problématique. L'aspect de la surdispersion liée à une agrégation des données a fait l'objet d'une classe particulière de modèles, dénommée Modèles Hiérarchiques Linéaires (HLM).
Ces modèles ont été spécifiquement développés pour éviter les inconvénients de l'analyse au niveau supérieur (problème de surdispersion) ou de celle au niveau inférieur (corrélations et non-indépendance entre individus). En effet, cette classe de modèles permet d'utiliser un prédicteur individuel pour considérer les données individuelles et un prédicteur log-linéaire au niveau du groupe pour les variables définies à ce niveau.

Une des méthodes les plus courantes permettant l'estimation des paramètres de modèles HLM nécessite une procédure en deux étapes [12]. Lors de la première étape, l'estimation concerne la relation entre individus pour chaque niveau d'agrégation. Les paramètres de régression sont donc estimés séparément dans chaque groupe. Dans une seconde étape, les résultats du premier passage sont considérés comme les événements de l'analyse entre les groupes.
Ces modèles ont l'avantage d'être opérationnels à partir de logiciels tels que S-plus ou Winbugs. Cependant, l'estimation des paramètres de ces modèles étant basés sur l'approche bayésienne par MCMC, l'interprétation des résultats doit prendre en compte cette situation, notamment concernant la validité des résultats de convergence.

Exemple classique de GLMM : le modèle Binomial-Négatif

Le modèle de régression Binomial-Négatif a été défini par Dean [21] en 1989 afin de répondre aux problèmes de surdispersion liés aux modèles de Poisson. Ce modèle illustre la méthode générale exposée ci-dessus car la loi Binomiale-Négative est le résultat d'un mélange d'une loi de Poisson (sur les données) et d'une loi Gamma (loi de l'effet aléatoire). En effet, lorsque $Y_{ks}|h_k \sim P(\mu_{ks})$ et que $h_k \sim Gamma(1,\alpha)$ ($\mathbb{E}[h_k] = \alpha^{-1}$ et $V[h_k] = \alpha^{-2}$) alors la distribution marginale $Y_{ks} \sim BN(\mu_{ks},\alpha^2)$. Cette distribution Binomiale-Négative est définie par la distribution ci-dessous :

$$P[Y_{ks} = y_{ks}] = \frac{\Gamma(y_{ks} + \alpha^{-1})}{y_{ks}!\Gamma(\alpha^{-1})} \left(\frac{\alpha\mu_{ks}}{1 + \alpha\mu_{ks}}\right)^{y_{ks}} \left(\frac{1}{1 + \alpha\mu_{ks}}\right)^{\alpha^{-1}} \tag{1.3}$$

Les deux premiers moments de cette loi sont $\mathbb{E}[Y_{ks}] = \mu_{ks}$ et $V[Y_{ks}] = \mu_{ks}(1 + \alpha\mu_{ks})$, où

$\mu_{ks} = \exp\{X_{ks}\beta\}$. On observe que la fonction variance est de forme quadratique, ce qui permet de prendre en compte une autre forme, plus flexible, de surdispersion.

Le modèle BN a de nombreux avantages, notamment son appartenance à la Famille Exponentielle ainsi qu'une forme analytique simple de sa log-vraisemblance marginale qui s'écrit sous la forme :

$$l(\beta,\alpha) = \sum_{k,s} \left(\sum_{j=0}^{y_{ks}-1} ln(1+\alpha j) + y_{ks} ln(\mu_{ks}) - (y_{ks}+\alpha^{-1}) ln(1+\alpha\mu_{ks}) \right) \tag{1.4}$$

Les estimations des paramètres β et α sont déterminées par la méthode du maximum de vraisemblance, mis en oeuvre par un algorithme itératif de type Newton-Raphson. Cette procédure a été décrite dans le chapitre précédent. Pour la maximisation de la vraisemblance, il est nécessaire d'obtenir une forme analytique simple de la dérivée de la log-vraisemblance. Les formes explicites de ces dérivées sont présentées ci-dessous :

$$\begin{aligned} \frac{\partial l}{\partial \beta_p} &= \sum_{k,s} \frac{X_{ks}^p (y_{ks}-\mu_{ks})}{1+\alpha\mu_{ks}} \\ \frac{\partial l}{\partial \alpha} &= \sum_{k,s} \left\{ \sum_{j=0}^{y_{ks}} \left(\frac{j}{1+\alpha j} \right) + \alpha^{-2} ln(1+\alpha\mu_{ks}) - \frac{(y_{ks}-\alpha^{-1})\mu_{ks}}{1+\alpha\mu_{ks}} \right\} \end{aligned}$$

A partir de ces dérivées premières, il est possible de déterminer la hessienne (dérivées seconde par rapport aux paramètres) nécessaire à l'estimation du maximum de vraisemblance, notam-

ment pour l'estimation des variances des paramètres.

$$\begin{aligned}
\frac{\partial^2 l}{\partial\beta_j\partial\beta_q} &= -\sum_{k,s}\frac{X_{ks}^j X_{ks}^q \mu_{ks}(1+\alpha y_{ks})}{(1+\alpha\mu_{ks})^2} \\
\frac{\partial^2 l}{\partial\beta_j\partial\alpha} &= -\sum_{k,s}\frac{X_{ks}^j \mu_{ks}(y_{ks}-\mu_{ks})}{(1+\alpha\mu_{ks})^2} \\
\frac{\partial^2 l}{\partial\alpha^2} &= -\sum_{k,s}\left\{\sum_{j=0}^{y_{ks}}\left(\frac{j}{1+\alpha j}\right)^2 + 2\alpha^{-3}ln(1+\alpha\mu_{ks}) - \frac{2\alpha^{-2}\mu_{ks}}{1+\alpha\mu_{ks}} - \frac{\mu_{ks}^2(y_{ks}-\alpha^{-1})}{(1+\alpha\mu_{ks})^2}\right\}
\end{aligned}$$

Il est à noter que la loi BN fait partie des Familles Exponentielles Naturelles à deux paramètres (dont un de dispersion), également appelées Familles Exponentielles de Dispersion (FED). Plus de détails sur cette famille peuvent être trouvés dans l'article de Jorgensen [48].

Il existe d'autres mélanges de Poisson dans la littérature, le plus connu étant le modèle Poisson-Inverse Gaussien, qui est un mélange obtenu lorsque l'effet aléatoire conjugué suit une loi Inverse-Gaussienne. Cet autre modèle, très bien décrit dans la littérature par Willmot [112] en 1987, est une alternative à la loi Binomiale-Négative. Il faut également noter que cette distribution conduit à une forme cubique de la variance, ce qui pourrait être plus adapté à certaines situations rencontrées dans le cadre de la surdispersion [53]. Enfin, l'autre intérêt de ce mélange différent réside dans la possibilité qu'il offre de prendre en compte la sous dispersion. En effet, le paramètre de dispersion peut exprimer la sous dispersion (lorsqu'il est négatif) aussi bien que la sur-dispersion (lorsqu'il est positif). Cependant, ainsi que nous l'avons précisé en introduction, ce cas de sous dispersion est rare dans les applications épidémiologiques.

1.1.3 Tests de surdispersion

Les problèmes liés à la surdispersion tiennent à sa quantification et à sa signification statistique. En effet, selon le modèle que l'on utilise, la distribution peut ajuster plus ou moins bien la surdispersion présente. Classiquement, le modèle de référence dans des processus de dénombrement est le modèle de Poisson. Cependant, il est nécessaire d'avoir un test statistique permettant de rejeter l'utilisation de ce modèle au risque de la présence d'une surdispersion.

Les tests les plus courants sont le test du rapport de vraisemblance et le test basé sur la déviance [11, 21]. Ce dernier est déterminé par :

$$D = \sum_{k,s} d_{ks}(\mu_{ks},y_{ks}) = 2\sum_{k,s}[l_{ks}(y_{ks},y_{ks}) - l_{ks}(\mu_{ks},y_{ks})] = -2\int_{y_{ks}}^{\mu_{ks}} \frac{y_{ks}-t}{v(t)}dt$$

Turnbull [107] a montré que sous l'hypothèse nulle, $D \sim \chi^2_{n-p}$.

TAB. 1.2 – *Log-vraisemblances et déviances de certaines Familles Exponentielles Naturelles*

Lois	Déviance	Log L
Binomiale	$2\sum[y_{ks}ln(\frac{y_{ks}}{\mu_{ks}}) + (1-y_{ks})ln(\frac{1-y_{ks}}{1-\mu_{ks}})]$	$\sum[y_{ks}ln(1-p) + (n_{ks}-y_{ks})ln(p)]$
Poisson	$2\sum[y_{ks}ln(\frac{y_{ks}}{\mu_{ks}}) - (y_{ks}-\mu_{ks})]$	$\sum y_{ks}ln(\mu_{ks}) - \mu_{ks}$
Normale	$\sum(y_{ks}-\mu_{ks})^2$	$-2^{-1}\sum[\frac{(y_{ks}-\mu_{ks})^2}{\phi} + ln(\phi) + ln(2\pi)]$
Gamma	$2\sum[-ln(\frac{y_{ks}}{\mu_{ks}}) + \frac{y_{ks}-\mu_{ks}}{\mu_{ks}}]$	$\sum \phi^{-1}ln(\frac{y_{ks}}{\mu_{ks}}) - \frac{y_{ks}}{\mu_{ks}} - ln(y_{ks}) - ln(\Gamma(\phi^{-1}))$

Plus récemment, Dean [22] a proposé une statistique de test permettant d'évaluer la surdispersion liée à un modèle de Poisson qui est basé sur le test du score :

H_0 : "Modèle de Poisson correct" – H_1 : "Surdispersion"

La statistique de test est définie par :

$$T_1 = \frac{\sum_{k,s}[(Y_{ks} - \mu_{ks})^2 - Y_{ks}]}{\sqrt{2\sum_{k,s}\mu_{ks}^2}}$$

Sous l'hypothèse nulle, cette statistique suit une loi normale centrée réduite. L'idée de ce test repose sur le fait que sous H_0, $\bar{Y}_{ks}$ et $n^{-1}\sum_{k,s}(Y_{ks} - \mu_{ks})^2$ (le numérateur de la statistique évalue la différence entre ces deux quantités) estiment la même quantité alors que sous l'hypothèse alternative, le second terme est d'autant plus grand que la surdispersion est importante.

Dean [22] a également déterminé une statistique de test permettant d'évaluer la surdispersion basée sur l'équation partielle du score standardisé. Cette statistique s'écrit :

$$T_2 = \frac{1}{\sqrt{2n}}\sum \frac{\{(Y_{ks} - \mu_{ks})^2 - Y_{ks}\}}{\mu_{ks}}$$

Il est possible d'évaluer la surdispersion car $T_2 \sim \chi_p^2$. De nombreuses extensions ont été développées à partir de ces statistiques du score, notamment par Dean [22] et Smith [101].

1.2 Estimation par Quasi-Vraisemblance (QL)

L'intérêt majeur de l'estimation par QL est d'étendre les techniques d'estimations classiques à des modèles ayant des fonctions variances n'appartenant pas à des familles exponentielles ainsi qu'à des modèles dont la distribution n'est pas totalement spécifiée [44]. En effet, cette technique d'estimation ne nécessite que la spécification des deux premiers moments de la dis-

tribution ainsi que de la fonction de lien entre ces deux moments [72, 111]. Cette approche est également appelée “approche marginale”.

Cette technique est basée sur les équations du quasi-score également appelées équations d’estimations généralisées (GEE) et qui ont été très bien décrites dans la littérature [13, 65, 113].
On peut définir le logarithme de la quasi-vraisemblance par :

$$QL(Y_{ks},\mu_{ks}) = \sum_{k,s} \int_{Y_{ks}}^{\mu_{ks}} \frac{Y_{ks} - t}{a(\phi)v(t)} dt$$

où $a(\phi)v(t)$ est la fonction variance, telle qu’elle a été décrite dans les sections précédentes.

L’expression de cette quasi-vraisemblance est ensuite utilisée comme une vraisemblance afin de réaliser l’estimation des paramètres. La maximisation de la quasi-vraisemblance consiste à rechercher les valeurs qui annulent $QL(y_{ks},\mu_{ks})$ et qui peuvent être obtenues par la résolution du système suivant :

$$U(\beta) = \sum_{k,s} D_{ks} V_{ks}^{-1} (Y_{ks} - \mu_{ks}) = 0$$

où D_{ks} représente la dérivée de μ_{ks} par rapport aux paramètres à estimer et V représente la matrice de variance-covariance associée au modèle (matrice de travail ou “working matrix”). Cette matrice peut être spécifiée de différentes manières, sans que cela n’altére les propriétés des estimateurs. $U(\beta)$ est appelée fonction de quasi-score.

Cette équation peut être résolue à l’aide d’un algorithme itératif de Newton-Raphson. La procédure

d'estimation s'arrête lorsque la convergence est atteinte, le critère de convergence étant défini par $\sum_p |\hat{\beta}_p^{i+1} - \hat{\beta}_p^i| < C_{cv1}$ ou $\sum_{k,s} U_{ks} < C_{cv2}$. Il est classique de considérer que ces critères soient inférieurs à 10^{-4} ou 10^{-6}.

L'estimateur de la matrice variance-covariance, qui fournit les estimateurs des variances associées aux estimateurs, est obtenu par l'estimateur Sandwich défini ci-dessous :

$$K(\sum_{k=1}^{K} D_{ks}^T V_{ks}^{-1} D_{ks})^{-1} \left(\sum_{k=1}^{K} D_{ks}^T V_{ks}^{-1} (y_{ks} - \mu_{ks})^2 V_{ks}^{-1} D_{ks}^T \right) (\sum_{k=1}^{K} D_{ks}^T V_{ks}^{-1} D_{ks})^{-1}$$

De nombreux travaux ont montré que les estimateurs restaient robustes et consistents même si la matrice de variance-covariance est mal spécifiée [74, 83] mais satisfaisant des conditions très générales. Cette approche a donc l'avantage de fournir des estimateurs asymptotiquement sans biais.

1.3 Modélisation des données agrégées

Nous avons vu dans le chapitre précédent, une méthode permettant de prendre en compte la surdispersion par une modélisation statistique plus adaptée. Comme nous l'avons précisé dans le chapitre introductif, cette surdispersion peut être le résultat de nombreux facteurs, notamment le manque d'information. Afin de prendre en compte le maximum d'information possible, une méthode permettant d'utiliser l'information issue de niveaux de recueil différents a été proposée par Prentice et Sheppard en 1995 [82] et a été développée dans quelques travaux ultérieurs par Anderson en 1998 [2] et Guthrie en 2002 [37]. Nous présenterons tout d'abord le modèle initial

et ensuite les extensions qui y ont été apportées.

1.3.1 Modèle Hiérarchique Agrégé

Soit un modèle pour données individuelles défini par :

$$P[Y_{iks} = 1] = p_{iks} = p_{ks0} exp\{Z_{iks}\alpha\} \tag{1.0}$$

où p_{ks0} est le "baseline", k est l'indice de la commune ($k \in \{1,K\}$), s est l'indice de la strate ($s \in \{1,S\}$) et i est l'indice de l'individu dans la commune k ($i \in \{1,...,n_{ks}\}$). $P[Y_{iks} = 1]$ représente la probabilité de connaitre l'événement pour l'individu i. Si la population (k) est structurée selon le sexe et l'âge ($strate\ s$), p_{iks} peut être considérée comme la fonction de risque d'un modèle de Cox. Ce modèle peut ainsi être considéré comme une approximation d'un modèle à risques proportionnels, et les estimations des paramètres s'interprétent comme des risques relatifs (RR).

Présentation du modèle

Prentice et Sheppard [82, 96, 97] proposent un modèle permettant de traiter l'agrégation obtenue à partir de ce modèle initial. En effet, par sommation de (1.0) sur les n_{ks} individus de chaque strate de chaque commune, on obtient :

$$p_{ks} = p_{ks0} \sum_{i=1} (exp\{Z_{iks}\alpha\}) \tag{1.1}$$

En définissant y_{iks} comme l'indicateur de survenue de la mortalité chez l'individu i de la cohorte k dans la strate s, on peut définir le taux de survenue de l'événement par $\bar{y}_{ks} = n_{ks}^{-1} \sum_{i=1}^{n_{ks}} y_{iks} = n_{ks}^{-1} y_{ks}$. Un effet aléatoire peut être introduit dans la caractérisation du terme de baseline. On peut donc spécifier $p_{ks0} = e^{\gamma_s} h_k$ et noter que l'effet aléatoire (h_k) porte uniquement sur la commune (et non sur la strate), ce qui est cohérent avec les hypothèses évoquées en introduction, sur le fondement de la surdispersion. D'après l'équation (1.1), on obtient :

$$P[Y_{ks} = \bar{y}_{ks}] = e^{\gamma_s} h_k \sum_i (exp\{Z_{iks}\alpha\}) \tag{1.2}$$

où Y_{ks} représente la variable aléatoire associée au nombre d'événements observés dans la strate s de la commune k.

Estimation par Quasi-Vraisemblance

Prentice et Sheppard [82] propose une approche par quasi-vraisemblance (QL) pour estimer les paramètres du modèle 1.2 (*Cf.* section 1.2). Ainsi que dans l'approche par GLMM, nous pouvons contraindre sans perte de généralité que $\mathbb{E}[h_k] = 1$. L'effet aléatoire n'est ainsi spécifié que par ses deux premiers moments, tels que $\mathbb{E}[h_k] = 1$ et $V[h_k] = \sigma^2$.
Notons, $n_k = \sum_{s=1}^{S} n_{ks}$ et $\bar{Y}_k = S^{-1} \sum_{s=1}^{S} \bar{Y}_{ks}$. L'effet aléatoire étant déjà indépendant des strates, nous appliquerons cette même hypothèse au baseline ($\gamma_s = \gamma$). Il est donc possible de reformuler l'expression du modèle comme suit:

$$\bar{Y}_k = h_k \varepsilon_k \, (exp\{X_{ks}\beta\}) \tag{1.3}$$

où, $X_{iks} = (1, Z_{iks})$ et $\beta^T = \left(\gamma, \alpha^T\right)$ et $\varepsilon_k[f(.)] = n_k^{-1} \sum_{s=1}^{S} \sum_{i=1}^{n_{ks}} f(.)$.

Les deux premiers moments du modèle peuvent être spécifiés par :

$$\mathbb{E}[\bar{Y_k}] = \mu_k = \mathbb{E}[\mathbb{E}\{\bar{Y_k}|h_k\}] = \varepsilon_k \left(exp\{X_{ks}\beta\}\right) \tag{1.4}$$

$$V[\bar{Y_k}] = \mathbb{E}[V\left(\bar{Y_k}|h_k\right)] + V[\mathbb{E}\left(\bar{Y_k}|h_k\right)] = \sigma^2\left(\mu_k^2 - \phi_k n_k^{-1}\right) + (\mu_k - \phi_k)n_k^{-1} \tag{1.5}$$

où $\phi_k = \varepsilon_k\{e^{2X_{ks}\beta}\}$.

Les équations du score sont déduites de (1.4) et sont définies par :

$$\sum_{k=1}^{K} D_k^T V_k^{-1}(\bar{y_k} - \mu_k) = 0 \tag{1.6}$$

$$I(\beta) = \sum_{k=1}^{K} D_k^T V_k^{-1} D_k \tag{1.7}$$

où $D_k = \partial\mu_k/\partial\alpha^T = \varepsilon_k \left(X_{iks} exp\{X_{iks}\beta\}\right)$, V_k est une matrice de travail quelconque et $I(\beta)$ représente la matrice d'information de Fisher, dépendant des paramètres β.

L'estimation des variances des estimateurs est fournie par un estimateur sandwich. Ce dernier a l'avantage d'être robuste quant aux hypothèses initiales. Ainsi, l'estimation des variances est donnée par :

$$\hat{V} = I^{-1}(\beta)\left[\sum_{k=1}^{K} D_k^T V_k^{-1}(\bar{Y_k} - \mu_k)(\bar{Y_k} - \mu_k)^T V_k^{-1} D_k\right] I^{-1}(\beta)$$

et l'estimateur de la variance σ^2 par :

$$\hat{\sigma^2} = K^{-1}\sum_{k=1}^{K} \max\{[(\bar{Y_k} - \mu_k)^2 - (\mu_k - \phi_k)n_k^{-1}][\mu_k^2 - \phi_k n_k^{-1}]^{-1}, 0\}$$

Liang et Zeger [63] ont montré que les estimateurs des paramètres et de leurs variances sont

consistants, lorsque $K \rightsquigarrow \infty$ et lorsque certaines conditions de convergence sont respectés (notamment si la matrice de travail est constante).

Extension aux échantillons

Le modèle décrit précédemment est basé sur l'hypothèse d'un recueil exhaustif de toutes les variables d'intérêt. Cependant, l'un des principaux avantages de ce modèle est son adaptabilité à des données issues d'un sous-échantillon. En effet, dans le modèle proposé ci-dessus, nous avons considéré que les données des variables explicatives et de la variable réponse étaient issues d'un même échantillon. Cependant, il pourrait être intéressant de pouvoir modéliser une variable réponse recueillie sur l'ensemble de la population et de l'expliquer par des variables issues d'un sous-échantillon de cette même population. L'intérêt de l'approche de Prentice et Sheppard est justement de pouvoir combiner ces deux sources de données.
Si nous nous replaçons dans notre application, il est ainsi possible d'expliquer la mortalité cardiovasculaire (taux de mortalité recueilli, sur tous les décès de la commune, grâce aux données administratives) par des facteurs de risque individuels issus d'un échantillon représentatif de cette population. Ainsi, les données individuelles sont recueillies sur les m_k individus constituant l'échantillon sélectionné sur les n_k individus de la population de la zone k.

Il est possible de réaliser un tel mélange entre les sources de données à la condition de

corriger le modèle précédemment établi par le modèle suivant:

$$\sum_{k=1}^{K} \tilde{D}_k^T \tilde{V}_k^{-1} (\hat{y_k} - \tilde{\mu}_k) = 0 \tag{1.8}$$

$$I(\beta) = \sum_{k=1}^{K} \tilde{D}_k^T \tilde{V}_k^{-1} \tilde{D}_k \tag{1.9}$$

où $\tilde{D}_k = \varepsilon_k^m (X_{ks} exp\{X_{ks}\beta\})$, $\tilde{\phi}_k = \varepsilon_k^m (exp\{2X_{ks}\beta\})$ et $\tilde{\mu}_k = \varepsilon_k^m (exp\{X_{ks}\beta\})$. Il est également nécessaire de redéfinir la fonction ε_k par $\varepsilon_k^m[f(k)] = m_k^{-1} \sum_{s=1}^{S} \sum_{i=1}^{m_k} f(k)$.

On peut noter que $\mathbb{E}[\tilde{D}_k] = D_k$ et $\mathbb{E}[\tilde{\mu_k}] = \mu_k$. L'estimation de σ^2 peut être biaisée par la sélection de l'échantillon, cependant ce biais diminue lorsque la taille de l'échantillon (m_k) augmente. Lorsque $m_k << n_k$, Prentice & Sheppard propose une estimation de la correction de ce biais [82].

Avantages de cette méthode

Les principaux avantages de cette récente approche concerne la possibilité de

(i) prendre en compte des données recueillies dans un petit échantillon ;

(ii) les mettre en relation avec des données globales issues de la population générale.

De plus, nous pouvons voir qu'à partir de données individuelles, ce modèle fournit un modèle qui est caractérisé au niveau de l'agrégation. Il est donc possible d'incorporer dans ce modèle, des variables explicatives qui sont issues de recueils mixtes, c'est-à-dire issues de différents niveaux de l'agrégation (comme des données individuelles et des données groupées au niveau de l'unité géographique par exemple).

1.3.2 Extension aux corrélations spatiales

En partant de l'équation 1.1, Guthrie et Sheppard [37] proposent une méthode d'estimation différente des GEE, détaillée dans le chapitre précédent. En effet, l'approche considérée pour prendre en compte les corrélations spatiales est basée sur les modèles hiérarchiques bayésiens (HLM) définie dans le chapitre 1.1.2, par une spécification particulière de h_k.

Modèle hiérarchique

Soit Y_{ks} le nombre de décès observés dans la strate s de la commune k et Y_k le nombre de ceux observés sur la commune k ($Y_k = \sum_{s=1}^{S} Y_{ks}$). On peut modéliser Y_k par une somme de loi de Bernoulli (loi de Y_{iks}) et cette sommation convergence vers une loi de Poisson.

$$Y_k \sim P(\mu_k)$$

où, $\mu_k = h_k e^{\gamma} \sum_s \sum_i^{n_{ks}} exp\{Z_{iks}\alpha\} = E_k h_k \varepsilon_k \left(exp\{Z_k\alpha\}\right)$ avec E_{ks} le nombre de décès attendus dans la commune k ($E_k = \sum_s n_{ks} e^{\gamma_s}$).

Dans ce cadre, il est classique de caractériser l'effet aléatoire traduisant une corrélation spatiale par $h_k = \exp(u_k + v_k)$, ainsi que décrit par Besag *et al* [6] en 1991. Cette formulation est basée sur un modèle de poisson conditionnel autoregressif (u_k) auquel on a rajouté une composante permettant de prendre en compte la surdispersion (v_k). On peut donc noter que l'effet aléatoire est composé de deux variables distinctes, portant sur les mêmes unités. La

première, u_k, réprésente l'hétérogénéïté spatiale (à travers une structure de corrélations spatiales autorégressive) alors que la seconde variable, v_k, traduit la structure de variation non expliquée par le modèle ou par les corrélations spatiales.
Dans le cadre de l'estimation bayésienne classique [6, 110] d'un tel modèle, on spécifie que $u_k \sim N(\bar{u}_{-k}, w_k^{-1}\sigma_u^2)$ et que $v_k \sim N(0, \sigma_v^2)$.

Afin de se retrouver avec des unités identiques, il convient de sommer l'équation précédente sur l'ensemble des strates de chaque commune. Si on considère que le baseline est indépendant des strates ($e^{\gamma_s} = e^{\gamma}, \forall s$), on peut donc la reformuler comme suit :
$\mu_k = E_k exp\{u_k + v_k\}\varepsilon_k \, (exp\{Z_k\alpha\})$.

L'estimation de ces paramètres se fait à l'aide d'un algorithme adapté au paradigme bayésien. La méthode proposée par les auteurs est basée sur un algorithme de Monte-Carlo par Chaînes de Markov (MCMC) avec une composante Metropolis-Hasting [41]. Cependant, Langford *et al* [55] ont montré que dans cette structure de modélisation spatiale, l'estimation des paramètres était possible par la technique du QL (*Cf.* section 1.2), ce qui permet une mise en œuvre affranchie des problèmes liés au paradigme bayésien.

Enfin, la validité de l'ajustement statistique peut être quantifiée par la statistique de test décrite par Spiegelhater [102] en 1998 qui est basée sur la déviance. Ce critère, nommé *Deviance Information Criterion* ou DIC, définit un critère de pénalisation de la déviance afin de déterminer un niveau de qualité d'ajustement du modèle et, par simplification, peut être rap-

proché du critère d'Akaike classique (AIC).

Extension aux échantillons

Ainsi que présenté au chapitre précédent, l'intérêt de cette approche est d'étendre cette technique au cas ou le recueil des données individuelles n'est pas réalisé sur l'ensemble de la population, mais uniquement sur un échantillon issu de celle-ci, alors que le recueil de la réponse est, elle, globale à la population. Conformément aux modifications présentées dans le chapitre précédent, on peut réécrire le modèle précédent par :

$$\mathbb{E}[Y_k] = \mu_k = E_k exp\{u_k + v_k\}\varepsilon_k^m \left(exp\{Z_k^m \alpha\}\right)$$

où $\varepsilon_k^m[f(.)] = m_k^{-1} \sum_{i=1}^{m_k} f(.)$ et Z_k^m représente le sous-ensemble de Z_k ramené aux m_k individus sélectionnés dans l'échantillon.

Une atténuation de l'effet étudié a été montrée [36] par des études de simulations lorsque la taille de l'échantillon devenait faible ($m_k << n_k$). De plus, cette approche semble effacer le risque de biais écologique lié à ce type d'études. Enfin, Guthrie *et al* [37] propose de corriger le biais introduit en sélectionnant l'échantillon, par la même technique que Prentice.

Autres approches pour les corrélations spatiales

D'autres approches basées sur les GEE ont été proposées pour prendre en compte la corrélation spatiale entre les communes. Par exemple, Anderson [2] a proposé en 1998 une méthode basée

sur des clusters avec une structure de corrélations entre clusters, déterminée a-priori.

Cette approche permet de relacher les hypothèses d'indépendance implicite au modèle de Prentice et Sheppard. L'idée est de regrouper les K communes de l'étude dans R clusters ($R < K$) afin d'obtenir :

$$P[Y_{ikr} = 1] = p_{ikr} = p_{kr0} exp\{Z_{ikr}\alpha\}$$

$$p_{kr} = p_{kr0}\varepsilon_k(exp\{Z_{ikr}\alpha\}) = h_{kr}e^{\gamma}\varepsilon_k(exp\{Z_{ikr}\alpha\}) = h_{kr}\varepsilon_k\,(exp\{X_{iks}\beta\})$$

où r est l'indice du cluster ($r \in \{1,R\}$), p_{kr0} est le "baseline", $X = (\mathbb{1}Z)$,$\beta = (\gamma\alpha)$, $\mathbb{E}[h_{kr}] = 1$,$V[h_{kr}] = \sigma^2$ et corr$[h_{jr},h_{ks}] = \mathbb{1}_{\{r=s\}}\rho_{rjk}$. La fonction ε_k est définie comme dans les modèles précédents.

De la même façon que dans le modèle initial, il est possible de déterminer les deux premiers moments de la variable $\bar{Y}_{rk}$, qui représente le taux de mortalité dans la commune k appartenant au cluster r. En effet,

$\mathbb{E}[Y_{rk}] = \mu_{rk} = \varepsilon_k(exp\{X_{ikr}\beta\})$

$V[Y_{rk}] = \sigma^2\left(\mu_{kr}^2 - \phi_k n_{kr}^{-1}\right) + (\mu_{kr} - \phi_k)n_{kr}^{-1}$

cov$(Y_{rj},Y_{sk}) = \rho_{rjk}\sigma^2\mu_{rj}\mu_{rk}\mathbb{1}_{[r=s;j\neq k]}$

Les équations du score associées sont :

$$\sum_{r=1}^{R} D_r^T V_r^{-1}(\bar{y_r} - \mu_r) = 0 \tag{1.10}$$

$$I(\beta) = R^{-1}\sum_{r=1}^{R} D_r^T V_r^{-1} D_r \tag{1.11}$$

où $D_r = \partial\mu_r/\partial\beta^T$ et V_r est la matrice de variance-covariance de Y_{rk}.

Le modèle spatial inclut une structure de covariance différente, cov(μ_{rj},μ_{rk}):

$$\begin{cases} \rho\sigma^2\mu_{rj}\mu_{rk} \text{ si } j \sim k \\ \\ \rho_{rjk}\sigma^2\mu_{rj}\mu_{rk} \text{ si } j \nsim k \\ \\ \sigma^2\mu_{rk}^2 + n_{rk}^{-1}\left(\mu_{rk} - (1+\sigma^2)\phi_{rk}\right) \text{ si } j = k \end{cases}$$

où $j \sim k$ définit une proximité spatiale entre j et k. Cependant, cette approche a l'inconvénient de ne considérer que les contiguïtés entre les communes et certaines études de simulations ont conclu que cette approche pouvait aboutir à une sous-estimation des paramètres de variance et des corrélations [36]. De plus, contrairement au modèle initial pour données agrégées, cette approche ne permet pas de stratifier la population sur des facteurs de risque établis, tels que l'âge et le sexe.

Il existe de nombreux autres travaux considérant la modélisation dans la cadre des corrélations spatiales [19]. Ces approches sont basées sur différentes méthodes. Par exemple, une de ces approches concerne la prise en compte dans les modèles de variables décrivant la localisation des unités géographiques, notamment par leur longitude et lattitude [81,86,88]. Cette approche peut être appliquée en se basant sur la comparaison des Rapports Standardisés de Mortalité (SMR). En effet, Cook *et al* [16] ont proposé une modélisation de ces SMR en fonction de diverses variables explicatives et en modélisant l'erreur aléatoire en fonction de la distance entre le centre de ces zones.

Chapitre 2

Présentation de nouvelles méthodologies

Comme nous avons pu le voir dans les sections précédentes, il existe plusieurs modèles de régression permettant de tenir compte de la surdispersion dans des modèles de dénombrements (modèles de Poisson). Dans la section suivante 2.1, nous présenterons la régression Arcsinus Stricte (AS) comme une approche complémentaire aux modèles de régression classiques que sont les modèles de Poisson ou Binomial-Négatif. La section 2.2 aura pour objectif de présenter une approche plus explicative du modèle proposé par Guthrie *et al* [37] en introduisant une autre paramétrisation des effets aléatoires. Enfin, nous tenterons d'évaluer les avantages et inconvénients de ces différentes techniques dans le chapitre **??** par des études de simulations et par Bootstrap.

Les simulations ont été réalisées à partir de jeux de données simulées en C++. Les estimations des paramètres à partir de ces données ont été réalisées à partir des programmes que nous avons développés. Afin de valider nos programmes, nous avons comparé nos résultats concernant la loi BN aux résultats fournis par la procédure Genmod de SAS. De plus, les résultats de

notre programme sur le modèle de Prentice et Sheppard ont été comparé aux résultats fournis par le programme S-plus obtenus par Guthrie et Sheppard.

2.1 Une approche complémentaire : le modèle Arcsinus Strict

Introduite en 1990 par Letac [60] et très récemment étudiée par Kokonendji *et al* [53] en 2003, la distribution Arcsinus Stricte fait partie de la Famille Exponentielle à Dispersion (FED), telle qu'elle a pu être décrite par Jorgensen [49], avec une variance de forme cubique $(V[Y] = \mu(\alpha^2\mu^2 + 1))$. Dans ce chapitre, nous présenterons un nouveau modèle de régression basé sur cette loi et nous décrirons comment il peut s'inscrire dans un aspect de complémentarité par rapport aux autres modèles. Cette partie fait l'objet d'une article soumis à *Communication in Statistics: Theory and Methods* dont le manuscrit est présenté dans l'annexe C.

2.1.1 Présentation du modèle

En nous basant sur les travaux de Kokonendji et Koudhar [53], il est possible de caractériser les différentes formes de variances en fonction de l'expression du prédicteur log-linéaire (correspondant à l'espérance dans les familles exponentielles), comme précisé dans le graphique 2.1.

Cette représentation visuelle nous permet de mettre en évidence une différence importante entre les différents modèles classiquement utilisés et le modèle Arcsinus Strict. Cette représentation permet également de valider le fait que la loi AS permet d'obtenir une distribu-

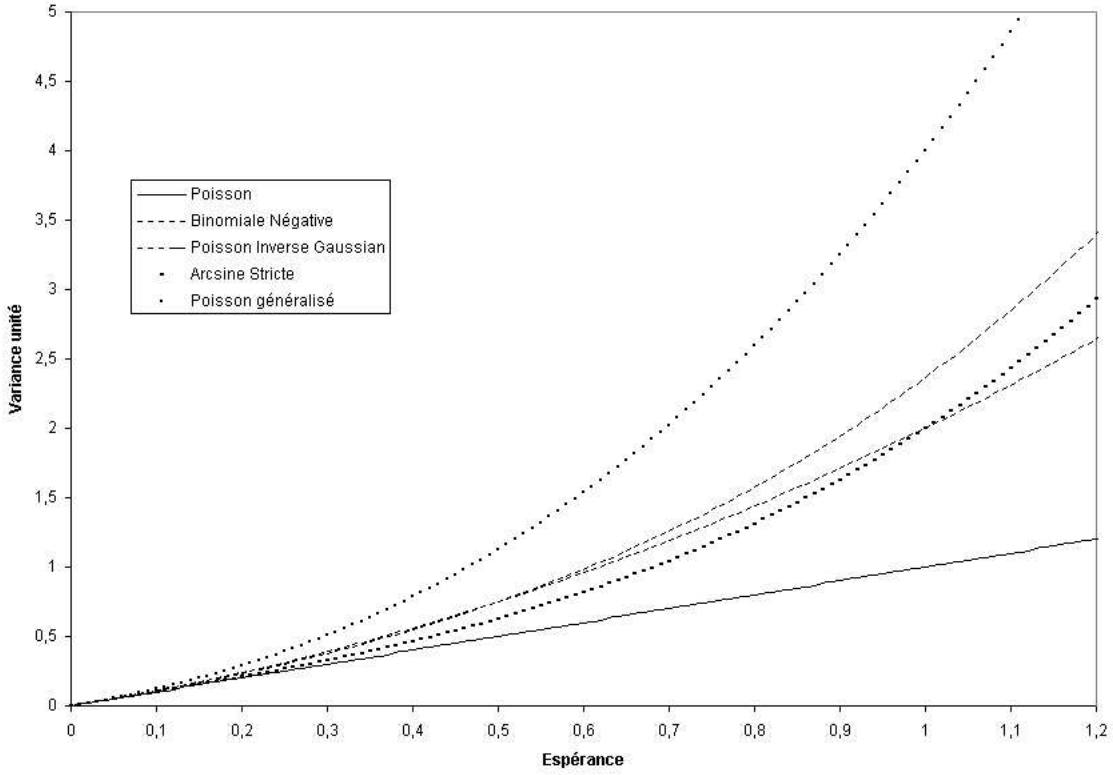

FIG. 2.1 – *Evolution des variances unitaires de certains modèles exponentiels en fonction de l'espérance*

tion de variance qui se situe autour de la loi Binomiale-Négative et Poisson-Inverse Gaussienne.

Définition du modèle

La distribution du modèle AS peut s'écrire sous la forme suivante:

$$P[Y_{ks} = y|\mu,\alpha] = \frac{A(y;\alpha)}{y!}\left(\frac{\mu_{ks}^2\alpha^2}{1+\mu_{ks}^2\alpha^2}\right)^{\frac{y}{2}} \exp\left\{-\alpha \arcsin\sqrt{\frac{\mu_{ks}^2\alpha^2}{1+\mu_{ks}^2\alpha^2}}\right\} \qquad (2.1)$$

où $y = 0,1,\cdots,\alpha > 0$ et $A(y;\alpha)$ est défini par :

$$A(y;\alpha) = \begin{cases} \prod_{k=0}^{z-1}(\alpha^2 + 4k^2) & \text{si } y = 2z \ ; \ A(0;\alpha) = 1 \\ \alpha \prod_{k=0}^{z-1}(\alpha^2 + (2k+1)^2) & \text{si } y = 2z+1 \ ; \ A(1;\alpha) = \alpha \end{cases} \quad (2.2)$$

A partir de l'équation 2.1, nous pouvons définir la log-vaisemblance (l) comme suit:

$$l = l(y|x;\beta,\alpha) = \sum_{k,s} l_{ks}$$

où

$$l_{ks}(\beta,\alpha) = -ln(y_{ks}!) + ln\{A(y_{ks};\alpha)\} + \frac{y_{ks}}{2} ln\left(\frac{\mu_{ks}^2\alpha^2}{1+\mu_{ks}^2\alpha^2}\right) - \left(\alpha \arcsin\sqrt{\frac{\mu_{ks}^2\alpha^2}{1+\mu_{ks}^2\alpha^2}}\right) \quad (2.3)$$

Estimation des paramètres

Nous utiliserons la méthode du maximum de vraisemblance (MLE) à partir du modèle 2.3 afin de réaliser l'estimation des paramètres. Il est possible d'utiliser d'autres techniques comme la méthode des moments ou la vraisemblance profilée afin de déterminer les estimations des paramètres. Nous préfèrerons la méthode du maximum de vraisemblance car ses estimateurs ont les meilleures propriétés asymptotiques.

L'estimation des paramètres par MLE est relativement facile à obtenir puisque la log-vraisemblance a une dérivée analytique de forme simple. Il est donc possible d'obtenir les solutions de cette régression en utilisant un algorithme itératif classique de Newton-Raphson. Pour ce faire, il est nécessaire de préciser les dérivées premières et secondes de la log-vraisemblance du modèle

par rapport aux paramètres à estimer. Ces dérivées sont présentées ci-dessous:

$$\begin{aligned}
\frac{\partial l}{\partial \beta_j} &= \sum_{k,s} \frac{X_{ks}^j (y_{ks} - \alpha^2 \mu_{ks})}{1 + \alpha^2 \mu_{ks}^2} \\
\frac{\partial l}{\partial \alpha} &= \sum_{k,s} \left[\frac{\partial ln A(y_{ks},\alpha)}{\partial \alpha} - \arcsin\left(\sqrt{\frac{\mu_{ks}^2 \alpha^2}{1 + \mu_{ks}^2 \alpha^2}} \right) + \frac{y_{ks} - \alpha^2 \mu_{ks}}{\alpha(1 + \alpha^2 \mu_{ks}^2)} \right]
\end{aligned}$$

et l'on peut définir

$$\partial(\ln A(y,\alpha))/\partial\alpha = \begin{cases} 2\alpha \sum_{k=0}^{z-1} (\alpha^2 + 4k^2)^{-1} & \text{si } y = 2z \\ \alpha^{-1} + 2\alpha \sum_{k=0}^{z-1} [\alpha^2 + (2k+1)^2]^{-1} & \text{si } y = 2z+1 \\ \partial lnA(1,\alpha)/\partial\alpha = \alpha^{-1} \text{ et } \partial lnA(0,\alpha)/\partial\alpha = 0 & \end{cases}$$

Nous pouvons également définir l'inverse de la matrice d'information de Fisher par la dérivée seconde de l par rapport à β et α. Les différentes composantes de cette matrice sont précisées ci-dessous:

$$\begin{aligned}
\frac{\partial^2 l}{\partial \beta_j \partial \beta_q} &= -\sum_{k,s} \frac{X_{ks}^j X_{ks}^q \alpha^2 \mu_{ks} (1 + 2\mu_{ks} y_{ks} - \mu_{ks}^2 \alpha^2)}{(1 + \mu_{ks}^2 \alpha^2)^2} \\
\frac{\partial^2 l}{\partial \beta_j \partial \alpha} &= -\sum_{k,s} \frac{2\alpha \mu_{ks} X_{ks}^p (1 + y_{ks} \mu_{ks})}{(1 + \mu_{ks}^2 \alpha^2)^2} \\
\frac{\partial^2 l}{\partial \alpha^2} &= -\sum_{k,s} \left[\frac{2\alpha^2 \mu_{ks} + y_{ks}(1 + 3\alpha^2 \mu_{ks}^2)}{\alpha^2 (1 + \mu_{ks}^2 \alpha^2)^2} - \frac{\partial^2 ln A(y_{ks},\alpha)}{\partial \alpha^2} \right]
\end{aligned}$$

où

$$\partial^2 lnA(y,\alpha)/\partial\alpha^2 = \begin{cases} 2\sum_{k=0}^{z-1} \frac{(4k^2-\alpha^2)}{(\alpha^2+4k^2)^2} & \text{si } y = 2z \\ -\alpha^{-2} + 2\sum_{k=0}^{z-1} \frac{\left[(2k+1)^2-\alpha^2\right]}{\left[\alpha^2+(2k+1)^2\right]^2} & \text{si } y = 2z+1 \\ \partial^2 lnA(1,\alpha)/\partial\alpha^2 = -\alpha^{-2} & \end{cases}$$

et $\mu_{ks} = \exp(X_{ks}^T\beta)$, $\mathbf{X}_i = (X_i^1,...,X_i^d)^T$, $\beta = (\beta_1,...,\beta_d)^T$ et $\alpha > 0$, comme défini dans la section précédente.

Il est possible de tester l'hypothèse que $\alpha = 0$, ce qui correspond à conclure que le modèle de Poisson serait le modèle le plus adapté à nos données. Plusieurs approches sont envisageables pour tester cette hypothèse, notamment le test du rapport de vraisemblance qui est défini par :

$$T = 2\left\{l(y|x;\hat{\beta},\hat{\alpha}) - l(y|x;\hat{\beta}_0,0)\right\}$$

où β_0 est la valeur de l'estimation de β sous l'hypothèse nulle.

2.1.2 Implémentation du modèle

Afin de permettre une utilisation aisée de ce modèle de régression, nous avons développé deux modèles utilisant des plate-formes différentes. Ce modèle est donc opérationnel à partir d'une macro SAS ainsi que par un programme réalisé en C++. Les sources du programme SAS sont disponibles en annexes D.

L'estimation des paramètres à partir de la macro SAS est basée sur la procédure GENMOD de la version 8 (et ultérieures). La variance de ce modèle généralisé a été implémentée selon

les spécifications définies ci-dessus. La convergence de l'estimation a été évaluée à partir de la différence de déviance ($D = 2[l(y|x;\hat{\mu},\hat{\alpha}) - l(y|x;y,0)]$) entre deux étapes de la procédure. La déviance du modèle Arscinus Strict peut être définie par:

$$D_{SA} = \sum yln\left[\frac{y^2(1+\mu^2\alpha^2)}{\mu^2(1+y^2\alpha^2)}\right] - 2\alpha^{-1}\left[\arcsin\left(\sqrt{\frac{y^2\alpha^2}{1+y^2\alpha^2}}\right) - \arcsin\left(\sqrt{\frac{\mu^2\alpha^2}{1+\mu^2\alpha^2}}\right)\right]$$

Cette macro fournit donc les estimations des paramètres et les variances qui y sont associées, ainsi que l'estimation du paramètre de surdispersion (α). Elle fournit également la valeur de la log-vraisemblance qui permet la comparaison de ce modèle avec d'autres modèles. Le principal problème lors de l'étape de calcul sous SAS vient de la cyclicité de la fonction A, présente dans la détermination de la vraisemblance et de ses dérivées.

Dans le programme C++, les estimations des paramètres s'obtiennent par l'annulation des dérivées de la log-vraisemblance du modèle. Cette résolution est obtenue à l'aide d'un algorithme itératif de Newton-Raphson. Le critère de convergence retenu est le maximum entre la modification de la valeur des paramètres ($|\beta^{p+1} - \beta^p|$) ou la valeur de la dérivée au point ($\frac{\partial l}{\partial \alpha}$). Dans toutes nos simulations et applications, ce critère de convergence devait être inférieur à 10^{-6}. Les avantages de ce programme par rapport à la macro SAS réside dans son estimation de la variance associée au paramètre de dispersion $\hat{\sigma}(\alpha)$, ce qui est impossible dans la macro SAS. ce programme fournit également la valeur de la log-vraisemblance au point de convergence.

2.1.3 Evaluation par simulations des modèles paramétriques : Poisson, Binomial-Negative et Arcsinus Strict

Nous avons validé notre modèle Arcsinus Strict à l'aide d'une étude par simulations. L'objectif de cette étude était de vérifier la cohérence des estimations par les deux programmes que nous proposons ainsi que de valider notre modèle de régression. Pour ce faire, nous avons simulé plusieurs jeux de données suivant une loi Arcsinus Stricte et Binomiale-Négative, en modifiant les paramètres initiaux. A l'issue de cette procédure de génération des données, nous avons effectué les analyses de régression sur ces jeux de données à l'aide des programmes proposés.

Afn de valider ces programmes dans nos conditions d'applications, nous avons choisi de fixer des paramètres initiaux proches de ceux que nous nous attendons à rencontrer dans l'étude de la mortalité cardiovasculaire chez les personnes âgées. Afin de ne pas influencer l'estimations des paramètres, notamment dans des cas de non-convergence, les paramètres initiaux étaient tirés au hasard sur l'étendue des valeurs plausibles. Par exemple, nous avons simulé des données à partir d'un modèle ayant pour paramètre β_0^0 compris entre -3 et -4. Dans ce cas là, la valeur initiale du paramètre estimé était tirée au hasard à partir d'une loi uniforme définie sur [-3;-4].

Nous avons réalisé de nombreuses simulations et nous ne présenterons dans cette partie que les deux simulations illustrant le mieux notre modèle. Le tableau 2.1 présente les résultats des modèles de Poisson, BN et AS effectués sur un jeu de données simulé à partir d'un loi AS. Le tableau 2.2 présente les résultats de ces mêmes modèles effectués sur un jeu de données simulé

à partir d'une loi BN. Les paramètres des lois simulées sont précisés dans le titre de chaque tableau.

TAB. 2.1 – *Simulations d'une distribution AS avec* $K = 100$, $S = 12$, $\beta_0 = -3$, $\beta_1 = -0{,}13$, $\beta_2 = -0{,}4$ *et* $\alpha = 0{,}25$. *Estimations des paramètres et des variances des paramètres par Poisson, BN and AS.*

	Poisson		Binomial-Négatif		ArcSinus Strict	
	$\hat{\beta}$	$\hat{\sigma}(\hat{\beta})$	$\hat{\beta}$	$\hat{\sigma}(\hat{\beta})$	$\hat{\beta}$	$\hat{\sigma}(\hat{\beta})$
Intercept	-3,06	0,109	-3,07	0,456	-3,09	0,662
X1	-0,12	0,077	-0,13	0,323	-0,13	0,468
X2	-0,42	0,078	-0,40	0,320	-0,38	0,456
α	-		1,24		0,20	
Log-vraisemblance	-4724,18		-4371,35		-4195,06	
Intercept	-2.96	0.331	-2.99	0.303	-2.98	0.332
X1	-0.09	0.02	-0.11	0.123	-0.12	0.202
X2	-0.44	0.079	-0.38	0.491	-0.39	0.101
α	-		1,33		0,27	
Log-vraisemblance	-5123.87		-4953.20		-4941.83	

Les deux tableaux ci-dessus présentent deux séries de deux simulations réalisées afin de valider la méthode Arcsinus proposée. Ces simulations ont été réalisées sur des jeux de 1000 répliques.

La première conclusion que l'on peut faire concerne la proximité des résultats entre les modèles BN et AS. Les estimations des paramètres dans ces modèles sont très proches même si les variances sont plus faibles pour le modèle AS par rapport à BN. Il est important de noter qu'il est impossible de comparer les estimations du paramètre de dispersion issues des modèles AS et BN car ils ne représentent pas la même quantité. Cependant, il serait intéressant de construire une statistique permettant de tester le degré de surdispersion et, éventuellement, de l'utiliser

TAB. 2.2 – *Simulations d'une distribution BN avec $K = 100$, $S = 12$, $\beta_0 = -4$, $\beta_1 = -0{,}4$, $\beta_2 = 0{,}2$ et $\alpha = 0{,}6$. Estimations des paramètres et des variances des paramètres par Poisson, BN and AS.*

	Poisson		Binomial-Négatif		ArcSinus Strict	
	$\hat{\beta}$	$\hat{\sigma}(\hat{\beta})$	$\hat{\beta}$	$\hat{\sigma}(\hat{\beta})$	$\hat{\beta}$	$\hat{\sigma}(\hat{\beta})$
Intercept	-3,99	0,019	-3,99	0,045	-3,99	0,048
X1	-0,40	0,013	-0,39	0,029	-0,40	0,031
X2	0,18	0,013	0,18	0,029	0,17	0,030
α	-		0,538		0,078	
Log-vraisemblance	-5136,21		-5076,75		-5088,71	
Intercept	-4.02	0.442	-3.98	0.487	-4.04	0.432
X1	-0.39	0.045	-0.41	0.243	-0.40	0.282
X2	0.18	0.009	0.19	0.042	0.18	0.099
α	-		0,613		0,094	
Log-vraisemblance	-5925.78		-5806.31		-5785.29	

pour déterminer le modèle le plus adéquat. Nous avons également mis en évidence de faibles différences entre les modèles de Poisson d'une part et BN et AS d'autre part. En revanche, même si la différence entre les estimateurs étaient faible, la variance des estimations des paramètres était nettement plus faible dans le modèle de Poisson.

Le principal avantage de cette approche est de fournir une alternative à la régression Binomiale-Négative (BN) afin de prendre en compte la surdispersion dans des données de dénombrement. Cependant, ces deux approches sont complémentaires et nous conseillons d'utiliser les deux modèles en pratique. En présence de surdispersion, une approche pourrait être de réaliser une régression selon ces deux modèles et d'utiliser la log-vraisemblance pour sélectionner le modèle le plus adapté. En effet, les simulations précédentes ont permis de montrer que la log-vraisemblance semblait être un bon critère de choix pour déterminer le modèle le plus adapté.

On peut également noter qu'il n'y a pas de valeur pour le coefficient de dispersion dans le cas

du modèle de Poisson, conformément à la présentation de la section 1.1.1. De plus, on peut constater que la différence des log-vraisemblances entre le modèle de Poisson et du meilleur modèle est d'autant plus grande que la surdispersion (traduite dans ce cas par α) est élevée. On peut donc en conclure que la modélisation "naïve" (modèle de Poisson) peut nettement dégrader la qualité de l'ajustement par rapport à un modèle plus souple, même s'il est mal spécifié. Enfin, la régression BN est la seule disponible dans les logiciels actuels, nous proposons deux programmes permettant l'estimation des paramètres du modèle Arcsinus Strict. En effet, les simulations menées dans ce chapitre montrent que les méthodes proposées dans le chapitre 2 sont opérationnelles et valides. Une autre manière de tester la qualité de l'ajustement aurait pu être de travailler sur les taux de recouvrement des estimations des paramètres lors de la modélisation par les deux méthodes différentes. Cette méthode n'a pu être mise en oeuvre pour des contraintes de temps mais sera prochainement explorée.

En ce qui concerne le modèle Arcsinus Strict, les résultats des deux programmes sont identiques. De plus, les simulations montrent que les estimations des paramètres fournies sont très proches des paramètres fixés, ce qui tend à valider notre modèle. De plus, ces études ont confirmé que la log-vraisemblance semble être un bon indicateur de la qualité de l'ajustement. En effet, cet indicateur a permis de sélectionner dans nos simulations le modèle "correct" malgré la proximité des estimations des paramètres. Nous pouvons donc proposer aux utilisateurs de réaliser les deux modèles (BN et AS) et de se baser sur ce critère pour retenir celui qui est le plus adéquat aux données. Enfin, conformément aux remarques de Cox [17] concernant les estimations des paramètres, ces simulations semblent montrer que la surdispersion n'implique

qu'une mauvaise estimation de la variance sans altérer l'estimation des paramètres.

2.2 Adaptation du modèle spatial

Le modèle proposé par Guthrie *et al* [37] et présenté en 1.3.2 permet de prendre en compte les corrélations spatiales dans un modèle hiérarchique pour données agrégées. Cependant, ce modèle est basé sur une prise en compte de ces corrélations par deux variables aléatoires, une pour l'hétérogénéité spatiale et l'autre pour la corrélation entre les variables.
Nous pensons que ce modèle, malgré une validité statistique certaine, n'offre pas une interprétation épidémiologique aisée pour les utilisateurs. En effet, les interprétations des estimations de ces paramètres restent très compliquées et obscures en terme d'association.
Afin de répondre à un souci de meilleure intréprétation et compréhension de cette ubiquité des effets aléatoires, nous proposons un modèle sur le même schéma mais avec une distribution des effets aléatoires différente.

2.2.1 Présentation du modèle

D'après le modèle défini à la section 1.3.2, on peut écrire que $\mu_k = \exp(X\beta + u_k)$. Ainsi que Leroux [59] l'a proposé, il est alors possible de poser que $u_k \sim N(O,D)$ où D représente la matrice de variance-covariance associée aux effets aléatoires. Nous pouvons déterminer D par :

$$\sigma^2 D = \lambda R + (1-\lambda) I$$

où R est la matrice de proximité spatiale et $\lambda \in [0,1]$ représente un paramètre de distribution de l'hétérogénéité spatiale. En effet, lorsque λ=0, la matrice de variance-covariance est $D = \sigma^{-2} I$ et donc le modèle ne comprend qu'une corrélation entre variables. Au contraire, lorsque λ=1, $D = \sigma^{-2} R$ et donc la matrice de variance-covariance représente uniquement une corrélation spatiale. Le paramètre σ^2 représente l'hétérogénéité totale du modèle. La proximité spatiale [57, 62] entre les zones géographiques est prise en compte à travers la matrice R comme:

$$R_{ij} = \begin{cases} n_i & \text{si } i = j \\ -\mathbb{1}\{i \sim j\} & \text{si } i \neq j \end{cases} \tag{2.-2}$$

Il est ainsi possible de préciser le niveau de proximité entre deux zones géographiques. Par exemple, dans le cas d'études portant sur de grandes zones , deux zones peuvent être considérées comme très proches même si elles ne sont pas contigües. On peut donc définir quel est le critère de proximité géographique entre les différentes zones.

Cette approche conduit à une interprétation épidémiologique des paramètres en deux étapes. Tout d'abord, une interprétation de σ^2 comme la surdispersion classique puis vient l'interprétation de λ comme un indicateur (quasiment un pourcentage) de l'origine de la surdispersion.

Donc, ce modèle est capable de quantifier la surdispersion, mais également de permettre de quantifier la part de chaque "cause" de cette surdispersion (corrélation spatiale ou manque d'information).

2.2.2 Implémentation du modèle

Ce modèle est estimable à l'aide d'un algorithme de Newton-Raphson complété par une extension de Monte-Carlo (NRMC) ou à l'aide d'un algorithme de Monte-Carlo à Chaînes de Markov (échantillonneur de Gibbs ou Metropolis par exemple). Cette dernière solution a l'avantage d'être opérationelle à partir de logiciels existants, tels que WinBugs.

2.2.3 Modèles hiérarchiques agrégés

La méthode des Modèles Hiérarchiques Agrégés a été décrite et évaluée, à notre connaissance, que par 4 auteurs (Prentice, Sheppard, Anderson et Guthrie), malgré l'intérêt de cette approche. De plus, ce modèle a toujours été évalué sur le même jeu de données et à partir des mêmes simulations, ce qui conduit à affaiblir la reproductibilité de ces estimations. De plus, ces simulations et applications portaient sur un faible nombre de zones géographiques ($K = 21$) et seulement dans un cas précis de problème nutritionnel.
Cela nous conduit à vouloir évaluer cette méthode à l'aide de nouvelles simulations, à partir de paramètres différents de ceux déjà introduits, avant de la mettre oeuvre sur notre application. Cette phase d'évaluation sera conduite à partir de techniques de Bootstrap afin d'obtenir des résultats robustes.

Rappel sur la technique du Bootstrap

L'approche par bootstrap est une méthode de calcul permettant de déterminer des estimateurs optimaux lorsque nous sommes en présence de facteurs de nuisance [26], tels que des

échantillons de faibles tailles ou un modèle de régression qui ne seraient pas bien adapté aux données.

Dans notre cas, l'utilisation du bootstrap devrait nous permettre d'obtenir des estimateurs des variances correctes. En effet, Cox [17] a montré qu'en présence de surdispersion, seules les estimations des variances étaient affectées et non pas les estimateurs des paramètres lorsque l'on effectuait une régression de Poisson classique.

Le bootstrap a pour principe de générer des échantillons aléatoirement à partir des données initiales, d'effectuer les estimations par des techniques classiques et de calculer les estimateurs des deux premiers moments de façon non-paramétriques à partir des résultats obtenus sur chaque échantillon. Plus de détails concernant la technique du bootstrap peuvent être retrouvés dans la littérature et notamment dans le papier d'Efron [26].

Le principal avantage de cette technique réside dans la robustesse des estimations fournies. En effet, comme cela est décrit ci-dessus, les estimations sont basées sur des approches non-paramétriques à partir de chaque estimation intermédiaire (sur chaque échantillon bootstrapé) permettant de s'affranchir aussi d'une éventuelle mauvaise spécification du modèle statistique sous-jacent.

Présentation des résultats

Dans la simulation présentée dans le tableau 2.3, nous avons voulu évaluer le Modèle Hiérarchique Agrégé (MHA) par rapport au modèle classique que représente BN. Cette simulation a été réalisée sur un jeu de données simulé à partir de deux variables globales (X1, X2) et de deux variables individuelles (X3, X4). Il faut noter que le modèle BN ne prend en compte

que les variables globales (X1, X2). De plus, afin de valider, sur des échantillons plus grands, les simulations effectuées par Prentice et Sheppard [82, 96], nous avons également comparé ces deux modèles à un modèle MHA par bootstrap.

TAB. 2.3 – *Simulations avec $K = 100$, $S = 12$, $\beta_0 = -4{,}00$, $\beta_1 = -0{,}01$, $\beta_2 = -0{,}10$, $\beta_3 = -0{,}20$, $\beta_4 = 0{,}10$ et $\sigma^2 = 1{,}2$. Estimations des paramètres et des variances des paramètres par un modèle BN et MHA (normal et bootstrapé).*

	Binomial-Négatif		Modèle MHA		Modèle MHA Bootstrap	
	$\hat{\beta}$	$\hat{\sigma}(\hat{\beta})$	$\hat{\beta}$	$\hat{\sigma}(\hat{\beta})$	$\hat{\beta}$	$\hat{\sigma}(\hat{\beta})$
Intercept	-4,00	0,213	-4,00	0,169	-3,99	0,143
X1	-0,01	0,003	-0,02	0,002	-0,01	0,003
X2	-0,09	0,113	-0,13	0,091	-0,11	0,121
X3	-	-	-0,17	0,031	-0,21	0,045
X4	-	-	0,12	0,097	0,11	0,091
σ^2	α=0,17		σ^2=1,25		σ^2=1,21	

On peut constater que les estimations dans le cadre du modèle hiérarchique agrégé (MHA) sont très proches de celles obtenues par un modèle identique auquel on a rajouté une estimation par bootstrap. On peut donc en conclure que les estimations issues du MHA sont non biaisées, tant en ce qui concerne les estimateurs des paramètres que des estimateurs des variances des paramètres. En revanche, on peut remarquer une légère différence entre les estimations issues d'un modèle BN par rapport à celles issues du modèle MHA.

Dans la simulation présentée dans le tableau 2.4, nous avons souhaité expliciter les résultats obtenus dans l'étude précédente, en ce qui concerne la distance entre BN et MHA. Nous avons donc confronté le modèle MHA avec deux techniques classiques utilisées en présence de recueil mixte. Nous avons donc simulé un jeu de données à partir d'un modèle comprenant deux

variables explicatives globales (X1 et X2) et de deux variables individuelles.

Les méthodes évaluées étaient le Modèle Hiérarchique Agrégé, le modèle BN ne considérant que les variables globales (X1, X2) et un modèle BN qui prenait en compte ces mêmes variables ainsi que les variables individuelles (X3, X4) recodées en moyenne par strate.

TAB. 2.4 – *Simulations avec $K = 100$, $S = 12$, $\beta_1 = -0.2$, $\beta_2 = 0.03$, $\beta_3 = 0.2$, $\beta_4 = 0.005$ et $\sigma^2 = 1.5$. Estimations des paramètres et des variances des paramètres par BN (avec et sans variable individuelle) et par MHA.*

	Binomial-Négatif1		Binomial-Négatif2		MHA	
	$\hat{\beta}$	$\hat{\sigma}(\hat{\beta})$	$\hat{\beta}$	$\hat{\sigma}(\hat{\beta})$	$\hat{\beta}$	$\hat{\sigma}(\hat{\beta})$
Intercept						
X1	-0,197	0,051	-0,260	0,067	-0,201	0,032
X2	0,030	0,023	0,041	0,052	0,031	0,021
X3	-0,154	0,123	-	-	-0,186	0,078
X4	0,001	0,009	-	-	0,004	0,008
Dispersion	α=0,23		α=1,52		σ^2=1,61	

A partir des résultats présentés dans le tableau 2.4, il est possible de valider le modèle proposé par Prentice et Sheppard, sur des données proches de celles que nous aurons dans notre application. On observe que les estimations des paramètres sont très proches dans les trois modèles évalués. En revanche, les estimateurs des variances de ces paramètres sont très différents.

On peut également remarquer qu'une fois que l'on introduit les variables individuelles dans le modèle BN (en moyennant les observations par unité géographique), ce modèle est très proche du MHA. Cependant, les résultats sont fortement biaisés lorsque ces variables ne sont pas considérées.

Enfin, le paramètre de dispersion n'est pas comparable entre les deux modèles car il ne représente pas la même quantité statistique, mais la surdispersion est bien détectée dans tous les cas.

En ce qui concerne le modèle hiérarchique agrégé (MHA), la première conclusion concerne la validation de ce modèle sur de grands échantillons. En effet, les simulations effectuées sur ce modèle résultaient uniquement d'une étude concernant 21 unités d'aggrégation (K=21) et il nous semblait donc important de confirmer ces résultats sur des tailles d'échantillons plus grandes (K=100 dans notre exemple).

De plus, la comparaison des résultats obtenus par le modèle "classique" et ceux obtenus par le modèle bootstrapé étaient très proches, ce qui a permis de renforcer la validation du modèle présenté. Enfin, on peut remarquer une grande proximité entre le modèle BN et le modèle MHA. Cela permet de confirmer que l'introduction d'une composante individuelle ne modifie pas nécessairement l'estimation des paramètres déjà présents. En effet, l'objectif de l'introduction de nouvelles variables est d'améliorer l'ajustement et de définir des facteurs de risque supplémentaires mais cela n'implique pas forcément une modification des effets des autres variables explicatives.

Conclusion du modèle spatial

Le principal intérêt de cette approche réside dans la meilleure interprétation épidémiologique qu'elle fournit des estimations de la surdispersion. En effet, l'approche classique des HLM fournit deux estimations de la surdispersion, une pour chacune des causes de surdispersion. Il est donc possible de juger de l'importance de la surdispersion ainsi que de son origine principale, mais cette paramétrisation permet d'affiner la complémentarité des sources qui y sont

associées. Il est en effet possible de quantifier la surdispersion (paramètre σ^2) mais également la part respective de la corrélation spatiale et du manque d'information dans cette surdispersion (paramètre λ).

Le second intérêt qui en découle est de pouvoir identifier les causes de la surdispersion et d'y remédier par un design approprié lors d'études ultérieures. Par exemple, si on détermine que le facteur principal de survenue de la surdispersion est un manque d'information, il sera nécessaire d'envisager un schéma d'étude (design) incluant plus de variables explicatives et un recueil plus complet.

Enfin, il existe de nombreux travaux sur la prise en compte de la corrélation spatiale dans la littérature statistique et de nombreux modèles ont été proposés. Sans revenir sur cette bibliographie, il est important de rappeler qu'une prise en compte aisée de ces corrélations peut être faite par l'introduction de variables décrivant la lattitude et la longitude des zones géographiques concernées [88]. Cette approche a l'avantage d'être facile de mise en oeuvre dans les modèles classiques. Cependant, cette approche néglige la notion de structure hiérarchique cachée (ou inhérente) aux données.

Chapitre 3

Application à la mortalité cardiovasculaire chez les personnes âgées

En France, comme dans tous les pays développés, la mortalité cardiovasculaire est un réel problème de santé publique puisque c'est la première cause de mortalité chez les personnes âgées de plus de 65 ans. Malgré le déclin de l'incidence de cette cause de décès depuis la fin des années 60, l'augmentation importante de la population âgée dans ces pays entraîne une augmentation du nombre de cas absolus observés.

La lutte contre la mortalité cardiovasculaire passe par une lutte contre les facteurs de risque qui en sont à la cause. Afin de lutter plus efficacement contre cette pathologie, il est donc important de comprendre et d'identifier au mieux les nouveaux facteurs de risque qui pourraient intervenir dans le plan de réduction des risques.

Dans ce chapitre, nous nous attacherons d'abord à mieux décrire cette cause de mortalité, tant sous ses aspects étiologiques que démographiques. Ensuite, nous essaierons de mieux caractériser l'effet de nouveaux facteurs de risque que sont les éléments minéraux de l'eau de boisson.

Une partie de cette application a déjà fait l'objet d'un article publié dans le journal *European*

Journal of Epidemiology et dont une copie est présentée en annexe B.

3.1 Epidémiologie de la mortalité cardiovasculaire chez les personnes âgées

La mortalité cardiovasculaire regroupe sous une même dénomination deux causes de décès qu'il convient de séparer. La première cause est d'origine cérébrovasculaire alors que la seconde est d'origine cardiaque ou non cérébrovasculaire. La distinction est importante car même si les deux causes ont une origine commune (atteinte du coeur ou des vaisseaux), les facteurs de risques et les conséquences peuvent être différents. Par la suite, nous essaierons de distinguer autant que possible ces deux sous-causes de décès.

3.1.1 Facteurs de risques usuels

Une des difficultés qui apparait lorsque l'on s'intéresse à la mortalité cardiovasculaire vient de la nature multifactorielle du risque. Depuis des décennies, de nombreuses recherches ont été menées afin de déterminer les facteurs de risques qui peuvent être associés à cette pathologie. Même si, désormais, beaucoup de ces facteurs sont clairement identifés, certaines causes entrainant cette mortalité restent encore plus ou moins inconnues.

Nous proposons de présenter les facteurs de risque classiques de la mortalité cardiovasculaire, en essayant de les regrouper par thème. Ceux-ci, au nombre de trois, sont les facteurs démographiques, les facteurs comportementaux et les facteurs individuels. Bien entendu, cette classification tripartite reste arbitraire et certains facteurs pourraient être classés dans une autre

des catégories. Nous avons choisi cette représentation pour distinguer les facteurs pour lesquels l'individu peut agir directement (tabac par exemple) de ceux dont la modification passe par une prise en compte de facteurs associés (nutrition notamment).

Facteurs démographiques

Ces facteurs de risque ont la particularité de ne pouvoir être modifiés par l'individu. Ces facteurs de risque sont principalement l'âge et le sexe. En effet, les femmes sont beaucoup moins touchées que les hommes par la mortalité cardiovasculaire, quelle que soit la sous-cause considérée. Il est également prouvé que l'accroissement de l'âge est très liée à l'augmentation de la fréquence de cette cause de décès.

Facteurs comportementaux (vin, alcool, tabac)

Le principal facteur de risque comportemental est la consommation de tabac. Cette consommation a été associée dans de très nombreuses études à la mortalité cardiovasculaire [64, 80], pour les deux sous-causes [98]. Il semblerait que la consommation de tabac ait un effet dose-réponse sur l'infarctus du myocarde [94] et que les effets soient irréversibles sur la santé à partir d'une certaine dose (évaluée à 20 paquets-années). Il a également été montré que l'arrêt du tabac conduit à une réduction du risque mais ne le ramène pas à celui d'un non-fumeur [8]. Cet impact semblerait plus important chez les femmes que chez les hommes [84]. Enfin, le tabac a été associé à une potentialisation de l'effet délétère d'autres facteurs de risque sur cette cause de mortalité. Les hypothèses biologiques concernant l'impact du tabac sur l'aggravation de la

mortalité cardiovasculaire portent principalement sur la modification de la répartition du cholestérol (augmentation de la fraction LDL) ainsi que sur la modification du flux sanguin.
La consommation d'alcool est le second facteur de risque comportemental, puisqu'elle joue un rôle important sur cette pathologie. En effet, les études qui ont été menées sur l'effet de la consommation d'alcool concluent toutes qu'une consommation modérée d'alcool réduit le risque de survenue de la mortalité cardiovasculaire [77, 106] alors qu'une consommation importante est délétère sur cette même cause de décès [66]. Il semblerait que cet effet soit plus marqué chez les personnes âgées [52].

Facteurs individuels (HTA, diabète, obésité, cholestérol)

Il existe de nombreux autres facteurs de risque de la mortalité cardiovasculaire identifiés. Il s'agit principalement de facteurs de risque liés à un état de santé altéré et qui entrainerait une augmentation du risque de mortalité par cette cause de décès.
L'obésité ou le surpoids a été identifié comme un facteur de risque important de la MCV [15, 51, 89, 103, 104]. Ces complications cardiovacsulaires semblent dues aux effet de l'obésité sur les éléments responsables du développement des plaques d'athérome (arthérosclérose), notamment par l'élévation du taux de cholestérol et de glucose. Ce surpoids peut être évalué par l'indice de masse corporelle (BMI en anglais) ou indice de Quetelet.
L'hypertension artérielle (HTA) constitue également l'un de ces principaux facteurs de risque et son action sur la mortalité a été très bien décrit, tant sur l'aspect cérébrovasculaire des personnes âgées [61] que sur la mortalité cardiovasculaire non dissociée [27, 70].

Enfin, il est important de considérer le diabète comme un autre facteur de risque à ne pas négliger dans la mortalité cardiovasculaire et notamment cérébrovasculaire [54]. En effet, il a été démontré que la baisse du taux de MCV, durant ces deux dernières décennies, est plus faible dans la population des diabètiques par rapport à la baisse observée dans la population générale [25, 35].

Il convient de noter que ces facteurs de risque sont très liés entre eux, puisqu'il a été montré que l'hypertension artérielle et le diabète sont associés à la présence d'un surpoids (ou d'obésité) chez l'individu.

3.1.2 Le rôle des minéraux dans l'eau de boisson

Depuis près 40 ans, de nombreuses études épidémiologiques ont suggéré une relation inverse entre le nombre de décès imputables aux maladies cardio-vasculaires et la dureté de l'eau [18, 24, 93, 95]. Cette dureté de l'eau est principalement déterminée par les concentrations de calcium et de magnésium dans l'eau et ces deux minéraux, ensemble ou séparément, peuvent jouer un rôle sur la mortalité cardio-vasculaire.

Le calcium est le cation le plus abondant dans le corps humain et le magnésium est le deuxième cation intracellulaire derrière le potassium. Le calcium et le magnésium sont des activateurs enzymatiques antagonistes. Le calcium est indispensable à la coagulation, à l'influx nerveux et à la contraction musculaire, notamment cardiaque [34, 70]. Le magnésium intervient dans le transfert et la libération d'énergie, il participe aussi à la physiologie cardiaque [58].

Dans les pays développés, les produits laitiers représentent la plus importante source de cal-

cium. Cependant, il est admis qu'avec l'âge l'absorption intestinale du calcium diminue, surtout chez les femmes [42, 114]. De plus, les personnes âgées développent parfois une intolérance au lactose [4, 38], ce qui conduit à une réduction de l'apport calcique. L'eau de boisson constitue donc un apport supplémentaire de calcium sous forme ionisée pour les personnes âgées. Quant au magnésium, ses principales sources proviennent de l'alimentation solide. La quantité de magnésium consommée quotidiennement dans les pays développés, dont la France, est insuffisante par rapport à celle qui est recommandée (environ 500mg/jour) [23]. De plus, le magnésium contenu dans l'eau de boisson étant sous forme ionisée, il serait plus biodisponible que celui apporté par les aliments solides [31, 39]. Ainsi, l'eau de boisson constituerait une source non négligeable des différents minéraux (calcium, magnésium).

Des études cliniques ont montré que le niveau de magnésium dans le myocarde était plus bas chez les personnes décédées par maladie cardio-vasculaire que chez celles décédées d'une autre cause [14]. En revanche, dans ces même études la concentration de calcium dans le myocarde était significativement plus haute pour les décès par cardiopathie ischémique. La relation entre ces éléments et la mortalité cardio-vasculaire n'est pas si évidente, car les modifications de ces éléments pourraient également apparaître lors de la mort cellulaire.

Les résultats d'études nutritionnelles laissent suggérer que des supplémentations en magnésium pourraient diminuer l'incidence de certaines maladies cardiovasculaires [47, 99, 100]. Toutefois, il n'a pas été trouvé de relation entre une supplémentation de calcium et la mortalité cardiovasculaire [85, 108]. En revanche, chez les personnes âgées, la déficience en Calcium pourrait contribuer au développement de l'hypertension artérielle [71].

Les différentes études écologiques menées sur les éléments de l'eau de boisson ont souvent re-

trouvé un effet protecteur du calcium et du magnésium sur la mortalité cardio-vasculaire [58]. Mais ces travaux reposaient surtout sur l'étude des corrélations [93, 95] entre les concentrations de calcium et de magnésium dans l'eau et les taux de mortalité dans différentes zones géographiques. Les facteurs de risque individuels, mis à part le sexe et l'âge, n'étaient pas pris en compte.

Rylander [92] a trouvé que les décès de cardiopathies ischémiques étaient inversement associés aux concentrations de calcium et de magnésium dans l'eau, pour les hommes et les femmes. D'après cet auteur, le calcium aurait aussi un effet protecteur sur les décès par maladies cérébro-vasculaires pour les hommes. Mais cette relation n'était pas retrouvée par Yang [115–117], pour qui ce serait le magnésium de l'eau de boisson qui jouerait un rôle protecteur significatif sur le risque de maladie cérébro-vasculaire. Rubenowitz [90, 91] a trouvé des associations différentes suivant le sexe chez des personnes âgées de 50 à 70 ans. Pour les hommes, il existait une relation inverse entre le magnésium de l'eau de boisson et le nombre de décès par infarctus du myocarde et l'effet du calcium n'était pas significatif [90]. En revanche, des effets inverses étaient retrouvés chez les femmes [91].

Enfin, il est communément admis que l'absorption du calcium est très liée à la présence de vitamine C et D [9]. La présence de cette dernière vitamine dans l'organisme est principalement le résultat de sa synthèse par l'organisme, processus très dépendant de l'exposition au soleil. On peut donc en conclure que les disparités régionales (axe Nord-Sud) peuvent être très importantes dans les études quantifiant les effets du calcium sur la mortalité.

3.2 Population : présentation de l'étude Paquid

Le programme PAQUID (Personnes âgées QUID) est une enquête épidémiologique prospective dont l'objectif est d'étudier le vieillissement cérébral et la dépendance des personnes âgées de plus de 65 ans, vivant à domicile lors de l'inclusion [20]. Cette enquête s'est attachée, entre autres, à mettre en évidence les facteurs de risque de détérioration intellectuelle. Le suivi de cette cohorte a débuté en 1988.

Pour réaliser cet objectif, une cohorte de 4134 personnes âgées en Gironde et en Dordogne, dont 3777 vivant à domicile, a été mise en place. Les sujets ont été tirés au sort sur les listes électorales de 37 communes de Gironde et 38 de Dordogne. Le plan de sondage a été stratifié sur trois critères : l'âge (en 3 classes, 65-74 ans, 75-81 ans et 85 ans et plus), le sexe et la taille des unités urbaines (en cinq classes).

Les critères d'inclusion dans cette cohorte étaient les suivants :

- être âgé d'au moins 65 ans au 31 décembre 1987,
- vivre à son domicile au moment du recueil initial des données,
- être inscrit sur les listes électorales.

3.2.1 Constitution de l'échantillon

La population-cible de l'étude est constituée par les personnes, âgées de 65 ans et plus en 1988, vivant dans les départements de Gironde et Dordogne. L'échantillon a été constitué en combinant les deux départements :

– en Gironde, 37 communes ont été choisies au hasard, après stratification en fonction de la taille de l'unité urbaine ; 4050 personnes de 65 ans et plus vivant à leur domicile ont été tirées au sort sur les listes électorales de ces communes, après stratification sur l'âge et le sexe. Parmi celles-ci 2792 ont accepté de participer à l'étude, soit un taux d'acceptation de 69%. Malgré les refus, la distribution par âge et sexe de l'échantillon étudié est représentative de celle de la population âgée de Gironde vivant à son domicile. Ces personnes ont été enquêtées à partir de janvier 1988 ;
– en Dordogne, la même procédure a conduit à la sélection de 1505 personnes vivant à leur domicile, dans 38 communes différentes. Parmi elles, 985 ont accepté de participer à l'enquête (66%) et ont été visitées à partir de 1989.

3.2.2 Recueil des informations

Après une prise de rendez-vous, le recueil des données est effectué au domicile de la personne âgée par des enquêtrices psychologues spécifiquement formées. Les variables recueillies initialement portaient notamment sur les caractéristiques socio-démographiques : âge, sexe, statut matrimonial, niveau d'études, ancienne profession, lieu de résidence des individus.
Tous les sujets ont fait ensuite l'objet d'un suivi pendant 10 ans, en alternant les visites au domicile et les enquêtes postales ou téléphoniques. Les suivis à domicile ont eu lieu à un (T1), trois (T3), cinq (T5), huit (T8) et dix (T10) ans en Gironde et trois, cinq, huit et dix ans en Dordogne avec les mêmes procédures que lors du bilan initial. Les événements enregistrés systématiquement sont en particulier le décès, l'entrée en institution, l'évolution de l'état fonc-

tionnel et la survenue d'une détérioration intellectuelle faisant craindre le début d'une démence. Les données du suivi 10 n'étant disponibles que depuis peu de temps, elles n'ont pu être utilisées pour les analyses de cette thèse.

Les différentes variables

Les variables d'ajustement que nous avons utilisées dans nos analyses ont toutes été recueillies lors des différentes visites intervenues dans le cadre du protocole PAQUID, tel qu'il est décrit ci-dessus. Notre analyse porte sur toutes les variables observées entre l'inclusion (T0) et la huitième année de suivi (T8).

Pour l'ensemble de nos analyses, l'âge sera recodé en classe de cinq années d'amplitude afin d'être cohérent avec les résultats et données disponibles par ailleurs (INSEE, INSERM).

L'hypertension artérielle et le diabète ont été définis par la prise de médicaments associés à ces pathologies.

Le recueil auprès du SC8

Afin de pallier le faible effectif de la cohorte PAQUID nécessaire à l'étude sur la mortalité cardiovasculaire, nous avons considéré la population vivant dans les 75 communes sélectionnées dans le programme PAQUID, entre le 01/01/1990 et 31/12/1996 et qui était âgée de plus de 65 ans. Le Service Commun 8 de l'INSERM (SC8), chargé de l'élaboration et de la diffusion de la "statistique nationale des causes de décès", nous a communiqué les enregistrements de tous les décès survenus durant les sept années de suivi sur les 75 communes considérées. Ces enregis-

trements sont composés de l'âge (en année révolu), le sexe, la commune de décès, la cause de décès (primaire et secondaire, codées sur 4 chiffres) et l'année de décès.

Nous avons également recueilli les effectifs dans chacune de ces communes, par classe d'âge de cinq ans et par sexe. Cette distribution nous a été fournie par l'INSEE. Nous avons donc pu déterminer le nombre de personnes-années comme la moyenne de la population entre les deux dates, pour chaque classe d'âge.

Les causes de décès

Les causes de décès ont été identifiées pour tous les décès observés par deux méthodes différentes. La première méthode était le certificat médical qui nous a été communiqué par le service commun de l'INSERM (SC8). La recherche de la cause du décès était basée sur le nom, prénom, date et lieu de naissance de toutes les personnes décédées entre l'inclusion et la huitième année de suivi.

La seconde méthode consistait à rechercher l'information auprès du médecin généraliste. L'objectif de cette nouvelle recherche visait à obtenir des résultats d'une source indépendante afin de confirmer le diagnostic. Le délai maximum entre la date du décès et le contact avec le médecin était de 3 années (entre les suivis à 5 et 8 ans). Les personnes ayant une cause de décès d'origine cardiovasculaire (code 390-459 de la Classification Internationale des Maladies) déclarées par une des deux parties et non confirmées par la seconde étaient exclues de l'étude.

Une étude de concordance a été menée par un retour a-posteriori sur les dossiers des patients. Ce retour a permis de recoder certaines causes qui avaient été mal codées lors de la saisie ini-

tiale. Les conclusions de ce travail ont fait l'objet d'une thèse de médecine [78].

3.2.3 Mesure de l'exposition

Le calcium et le magnésium contenus dans l'eau de boisson constituent nos principales variables d'exposition. Le problème de ce type de variable réside dans sa complexité à l'évaluer, puisque l'eau de boisson est composée simultanément de l'eau d'adduction (l'eau du robinet) et de l'eau minérale, dont la consommation est en constante augmentation.
Nous préciserons donc les différents recueils qui ont été nécessaires à la détermination de ces paramètres, ainsi que son introduction dans les modèles statistiques.

L'eau d'adduction

Les Directions Départementales des Affaires Sanitaires et Sociales (DDASS) de Gironde et de Dordogne ont fourni les renseignements concernant toutes les ressources en eau potable utilisées entre 1991 et 1994 par les 75 communes sélectionnées dans PAQUID. Ces renseignements concernaient les débits horaires et les périodes d'utilisation de chaque réserve. Une moyenne pondérée de chaque source d'approvisionnement a été réalisée pour chaque commune. Cette moyenne représente l'exposition de tous les individus résidant dans cette commune. Les techniques de mesure des composants de l'eau sont décrites dans une étude antérieure [1].
La commune était exclue de l'enquête si les valeurs manquaient pour les sources importantes

1. Jacqmin H, Commenges D. Aluminium-maladie d'Alzheimer : rapport de recherche. Universit´e Bordeaux2. D´epartement d'information m´edicale. Bordeaux, 1992.

(plus de 20% de l'approvisionnement en eau d'une commune). Nous n'avions donc pas de valeurs d'exposition pour sept communes qui ont été exclues de l'enquête.

La principale commune de l'étude, Bordeaux, a été décomposée en quatre zones distinctes d'approvisionnement. Cette commune avait donc quatre valeurs d'exposition différentes. Mais les données que nous avons recueillies à partir des services de diffusion des certificats de décès ne précisaient que la commune. Il a donc fallu que nous calculions une moyenne pondérée de l'exposition pour cette commune.

Lors de l'enquête PAQUID, le tirage au sort était équiprobable sur l'ensemble de la commune de Bordeaux, sans aucune stratification sur les quartiers. Nous pouvions donc supposer que les personnes incluses dans cette étude étaient aléatoirement distribuées sur toute la commune. Ainsi, la proportion de personnes dans chaque zone d'approvisionnement était identique dans PAQUID et dans la population totale. Sous cette hypothèse, nous avons calculé les valeurs d'exposition en effectuant une moyenne pondérée des quatre zones. Le poids de chaque zone était affecté selon le nombre de personnes dans l'étude PAQUID.

Cependant il n'a été possible d'avoir des données complètes concernant le calcium et le magnésium que pour 71 zones, correspondant à 68 communes. Les données concernant les concentrations d'aluminium, de silice, de sodium, de potassium, de fluor, de fer et de cuivre ainsi que le niveau de pH étaient également disponibles. Ces divers éléments ont donc pu être pris en compte en tant que facteurs d'ajustement.

L’eau de boisson

Lors du suivi à trois ans (T3), le questionnaire administré contenait une partie concernant la consommation d’eau des sujets. Les questions portaient sur le volume de leur consommation d’eau minérale (variable avec quatre modalités) ainsi que sur les marques des eaux minérales qu’ils buvaient. A partir de ces items, nous avons reconstruit une variable estimant la consommation individuelle d’eau minérale.

Cependant, pour des contraintes de coût, ce questionnaire n’a été proposé qu’aux individus résidant en Gironde. Les habitants de Dordogne ainsi que les personnes n’ayant pas participé au suivi à T3 n’ont donc pas de valeurs pour cette variable calculée. Nous avons ainsi retenu uniquement 1286 réponses exploitables à cet item.

Enfin, comme nous l’avons indiqué, cette variable a été codée en quatre modalités de fréquence de consommation (tout le temps, souvent, rarement et jamais). Il est donc difficile d’interpréter de telles variables dans un modèle statistique. Afin de pouvoir affecter une valeur de consommation à chaque individu, nous avons pondéré la valeur des minéraux de l’eau qu’il consommait en fonction de la réponse à sa fréquence de consommation. Nous avons également fait varier ce système de pondération afin de ne retenir que le codage le plus associé. Par exemple, la modalité “*tout le temps*” a été associée aux poids 0.9, 0.8, 0.7, cela permet de considérer qu’une personne qui ne boit que de l’eau minérale a des apports de minéraux autre que cette eau minérale équivalent à 0.1, 0.2, 0.3. Ces poids peuvent ainsi être associés à la part de l’eau minérale dans la consommation d’eau chez les personnes. Ce codage multiple nous a conduit à créer 27 variables de consommation effective d’eau.

Recodage

La consommation des éléments minéraux de l'eau de boisson est donc une variable continue. Cependant, l'effet des minéraux qui nous intéressent n'est pas forcément de forme linéaire. Au contraire, certaines hypothèses peuvent nous laisser penser que l'effet de ces variables pourrait être en forme de "U" (*U-shape*). En effet, il est possible qu'un excès ou une carence en Calcium ou en Magnésium entraîne des conséquences sur la mortalité cardiovasculaire et qu'une concentration intermédiaire soit le plus protecteur pour cette cause de décès.
Afin d'étudier au mieux ces éléments, nous considérerons cette consommation en continu mais également en trois classes. Comme il n'existe pas de références épidémiologiques ou cliniques, les bornes de ces classes seront définies par les terciles des variables afin de fournir des effectifs suffisants dans chacune des classes.

Comme nous l'avons précédemment indiqué, les variables d'exposition seront étudiées en continu (selon tous les codages), ainsi que découpées en trois classes. La multiplicité des tests inhérente à cette situation peut engendrer une distorsion des résultats obtenus. Afin de prendre en compte cette situation, nous avons tout d'abord sélectionné le meilleur codage de la variable par la technique développée par Hashemi et Commenges [40], basée sur un modèle de Cox. Ensuite, nous avons effectué les modèles de régression uniquement sur la variable qui avait été retenue.

3.3 Méthodes

Sans revenir sur les méthodes décrites dans les chapitres précédents, nous présenterons dans ce chapitre les aspects méthodoligiques qui ont permis de réaliser l'analyse démographique présentée dans la partie résultat. Les détails des autres méthodes peuvent être retrouvés dans le cahpitre 1 et 2.

3.3.1 Les hypothèses démographiques

La principale hypothèse que nous élaborerons dans ce travail concerne la migration de notre sous-population. En effet, les données que nous avons obtenues concernent les effectifs de notre sous-population ainsi que les décès qui s'y rapportent, mais sans que l'on puisse identifier les éventuelles migrations. Celles-ci peuvent être détaillées en deux points. Le premier concerne les migrations entre la France et les pays étrangers et le second vise les migrations entre l'Aquitaine et le reste du pays (ou les pays étrangers éventuels).

S'agissant des migrations internationales, on peut penser qu'elles sont faibles chez les personnes âgées de plus 65 ans. En effet, la très grande majorité de cette sous-population est à la retraite, ce qui est une situation relativement stable, et donc qu'il est peut probable que l'on assiste à un changement de situation géographique, mis à part les potentiels retours au pays. Pour les migrations concernant l'Aquitaine, cette hypothèse pourrait être plus discutable puisqu'il est fréquent que les personnes âgées quittent leur lieu de vie pour se rapprocher de leurs enfants ou d'infrastructures adaptées. Toutefois, un élément de réponse a été fourni par une étude menée sur la cohorte PAQUID [20] qui a montré que dans cette sous-population, plus de 75% des per-

sonnes résidaient dans leurs communes depuis plus de 40 ans.
On peut ainsi en conclure que la mobilité des personnes âgées de plus de 65 ans est un phénomène mineur, qui a peu de chance de modifier significativement notre étude et nous permet de considérer cette population comme fermée.

La seconde hypothèse qui est faite dans ce mémoire est une hypothèse implicite à l'utilisation des données de l'INSEE. En fait, c'est une hypothèse complexe qui est faite lors de l'estimation des effectifs dans les populations entre les recensements. En effet, il est nécessaire de disposer des effectifs de la population au 1er janvier de chaque année afin de calculer les taux de mortalité annuels. Lors des années des recensements, ces renseignements sont obtenus aisément, mais des difficultés se rencontrent lors des années intercensitaires. L'INSEE fait des estimations des effectifs par extrapolation durant ces années "sans information" en adoptant une hypothèse de linéarité de l'évolution. En effet, nous travaillions sur les données communales, et nous avons donc utilisé la méthode OMPHALE qui est un modèle beaucoup moins précis que ELP mais qui est le seul à fournir les estimations au niveau local.
Cependant, il convient de relativiser l'importance de cette hypothèse par les corrections qui sont apportées par l'INSEE à ces estimations. En effet, après chaque recensement, les estimations sont remodélisées par interpolation, en tenant compte des migrations et des changements intervenus entre les deux recensements.

Méthodes de calcul des paramètres démographiques

Nous avons pris en compte les décès par âge (en années révolues entières) et par année. Dans les analyses démographiques, ous avons donc estimé le nombre de décès dans le parallélogramme à base verticale à partir de la formule suivante :

$$\hat{Dc}_{Gx}^{1990-95} = 0{,}9 * Dc_{Gx}^{1990-91} + 0{,}7 * Dc_{Gx}^{1991-92} + 0{,}5 * Dc_{Gx}^{1992-93} + 0{,}3 * Dc_{Gx}^{1993-94} + 0{,}1 * Dc_{Gx}^{1994-95}$$

$$+0{,}1 * Dc_{Gx-1}^{1990-91} + 0{,}3 * Dc_{Gx-1}^{1991-92} + 0{,}5 * Dc_{Gx-1}^{1992-93} + 0{,}7 * Dc_{Gx-1}^{1993-94} + 0{,}9 * Dc_{Gx-1}^{1994-95}$$

où $\hat{Dc}_{Gx}^{\alpha-\beta}$ représente l'estimation du nombre de décès entre les années α et β. Cette formule a été estimée à l'aide du diagramme de Lexis, tel que définit ci-dessous :

Le quotient de mortalité, qui n'est pas présenté dans un souci de lisibilité, est donc calculé selon la formule suivante :

$$ {}_aq_c = \frac{d_c(x; x+a)}{s_x - d_{c'}(x; x+a)/2} \tag{3.1}$$

où ${}_aq_c$ représente le quotient de mortalité cardiovasculaire, $d_c(x; x+a)$ le nombre de décès d'origine cardiovasculaire et $d_{c'}(x; x+a)$ celui d'origine autre que cardiovasculaire. s_x représente le nombre de personnes à un âge compris entre x et $x+a$ au temps d'origine.

L'espérance de vie est calculée selon la formule suivante (2) :

$$e_x = 0{,}5 + \frac{\sum_{i=x+1}^{\omega} S_i}{S_x} \tag{3.2}$$

où e_x représente l'espérance de vie à l'âge x. Cette espérance est soumise à une légère imprécision car nous ignorons l'espérance de vie à 100 ans, qui est nécessaire au calcul. En conséquence, nous avons émis l'hypothèse que cette valeur était comprise entre 5 et 15 ans. Les

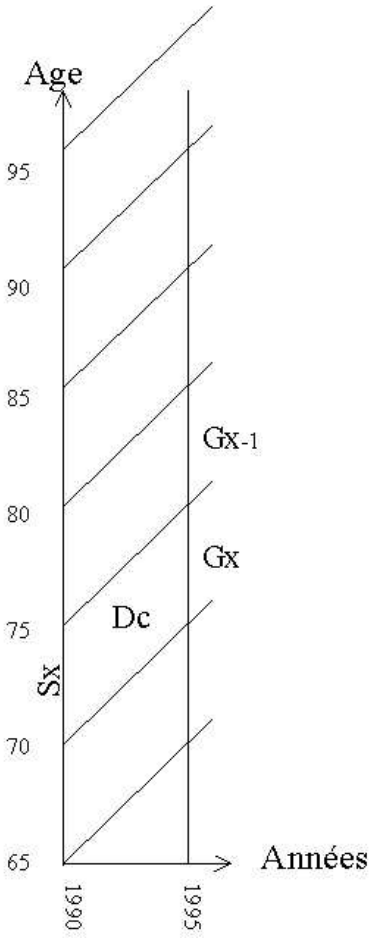

FIG. 3.1 – *Diagramme de Lexis représentant l'inclusion dans la cohorte*

résultats présentés sont ainsi les espérances calculées selon l'hypothèse retenue. On constate que les valeurs changent assez nettement selon l'hypothèse retenue mais que l'écart relatif entre les deux sexes reste faible. Cela signifie que malgré la surmortalité masculine, l'écart reste assez faible puisque les espérances de vie, dans le cas où l'unique cause de décès serait cardiovasculaire, seraient proches entre les deux sexes.

Evolutions projectives

Perspectives générales

Les projections concernant la mortalité des personnes de plus de 65 ans dépendent de deux

paramètres : les effectifs de la sous-population et les taux de décès que l'on doit y appliquer. Ainsi, nous devons projeter les deux composantes afin de déterminer la projection du nombre de décès d'origine cardiovasculaire. En effet, le nombre de décès d'origine cardiovasculaire pour une année x, est calculé selon la formule suivante :

$$Dc^x = \sum_{i=65}^{\omega} Dc_i^x = \sum_{i=65}^{\omega} S_i^x \times m_i^x$$

où S_i^x (resp. m_i^x) est le nombre de personnes (resp. taux de mortalité) au 1er janvier x, entre l'âge i et $i+4$ et Dc_i^x est le nombre de décès estimé pour l'année x entre les âges i et $i+4$.

La première étape de ces projections est de définir les hypothèses de projection qui seront nécessaires à l'estimation des S_i^x et m_i^x. En ce qui concerne les projections de la structure de la population des plus de 65 ans, nous pouvons réutiliser les résultats de l'INSEE dans ces projections, qui définissent un effectif ainsi qu'une distribution par classe d'âge.
S'agissant de la projection des taux de mortalité, nous avons élaboré trois hypothèses d'évolution que l'on pourrait caractériser d'hypothèses maximale, moyenne et minimale. Il est très probable que la mortalité cardiovasculaire ne va pas ré-augmenter dans le futur, donc l'hypothèse maximale serait une stabilisation du taux de mortalité à son niveau de 2000. L'hypothèse moyenne serait une baisse identique à celle observée sur ces 20 dernières années, baisse estimée par une identification en séries chronologiques. Enfin, l'hypothèse minimale serait une accélération de la baisse de cette cause de mortalité, en supposant que la vitesse de diminution soit doublée.
L'INSEE a réalisé des projections de population sur une période de 50 ans. Il est peut-être présomptueux de garder un tel recul pour effectuer nos projections quand on connaît la sensibilité des projections des taux de mortalité aux hypothèses. Aussi, il nous parait plus raisonnable de projeter les taux avec un recul à 10, 15 et 20 ans. L'estimation pourrait être faite à partir des données collectées sur les 30 dernières années, même si la similitude des codages n'est pas parfaite (comme nous l'avons discuté précédemment) avec les outils classiques de série chronologiques. La technique des moyennes mobiles pourrait ainsi être utilisée sur les taux ainsi que sur les logarithmes des taux, ce qui permettrait de tenir compte du domaine de définition ([0,1])

ainsi que d'éprouver la sensibilité des résultats au modèle.

Résultats attendus

Nous avons considéré la population des personnes âgées de plus de 65 ans comme une population fermée. Cette hypothèse est très "pratique" dans le cadre des projections mais sa validité est incertaine dans le futur. En effet, les régions françaises investissent pour attirer cette population que les médias nomment les "jeunes-vieux", en raison d'un fort pouvoir d'achat et d'une espérance de vie relativement longue. Le résultat de cette tentative d'attraction peut être un déséquilibre de la structure de cette sous-population par une augmentation massive des 65-75 ans ce qui aurait pour conséquence d'accroitre le nombre total de personnes âgées mais de diminuer la part des plus de 75 ans. Cet éventuel biais peut être modulé par le fait que les départs en retraite se déroulent majoritairement avant 65 ans et que les migrations potentielles devraient se dérouler au début de la cessation d'activité professionnelle.

3.4 Les résultats

Afin d'étudier de manière optimale la mortalité cardiovasculaire, nous avons souhaité prendre en compte des variables individuelles correspondant à des facteurs de risque important ou à des variables d'exposition composite (consommation de calcium ou magnésium). C'est dans cette situation que le modèle hiérarchique pour données agrégées (MHA), présenté dans le chapitre 1.3, sera un élément important de ces analyses. De plus, le modèle Arcsinus Strict (AS), complémentaire à la régression Binomiale-Négative (BN) sera également évalué sur ces données réelles.

Comme nous avons pu le présenter dans les sections précédentes, la mortalité cardiovasculaire est une thématique très large qu'il est possible d'aborder selon plusieurs aspects. L'objectif de notre travail est de déterminer l'effet du calcium et du magnésium sur la survenue de cette cause de mortalité. Cependant, étant donnée la multiplicité des sources d'exposition (eau d'ad-

duction et de consommation), des modèles et des causes à prendre en compte (cardiovasculaire global ou décomposés en sous-causes), nous ne pourrons pas détailler l'ensemble des analyses réalisées. Seules les analyses les plus intéressantes en terme d'interprétation épidémiologique ou introduisant un élément de discussion sur nos méthodes seront présentées.

Toutes les analyses qui ont été réalisées ont été ajustées sur l'âge (recodé en classes d'âge de 5 ans) et sur le sexe. Cependant, dans un souci de clarté de présentation et vu le faible intérêt des résultats (l'effet de l'âge et du sexe est très bien connu sur la mortalité cardiovacsulaire), les résultats concernant ces variables ne seront pas présentés dans les tableaux.

Ce chapitre débutera par une analyse démographique de la population d'étude, puis par une description de notre jeu de données, notamment par une présentation de l'incidence de la mortalité cardiovasculaire, ainsi que des différents facteurs de risque qui y sont associés. Ensuite, nous présenterons les résultats du modèle de Poisson par sous-cause ainsi que les résultats issus des modèles AS, MHA et MHA avec corrélations spatiales.

3.4.1 Aspects démographiques de la mortalité cardiovasculaire chez les personnes âgées

La mortalite cardiovasculaire a fait l'objet de beaucoup d'études d'un point de vue démographique dans la population française. Cette partie a pour objectif de préciser ces aspects dans une sous-population particulière que représentent les personnes âgées de plus de 65 ans dans le sud-ouest de la France.
Ce travail a fait l'objet d'un mémoire pour l'obtention du Diplôme d'Etudes Approfondies d'Analyse Démographique (Université Bordeaux IV) et fait l'objet d'un article prochainement soumis au journal *Population*.

Description

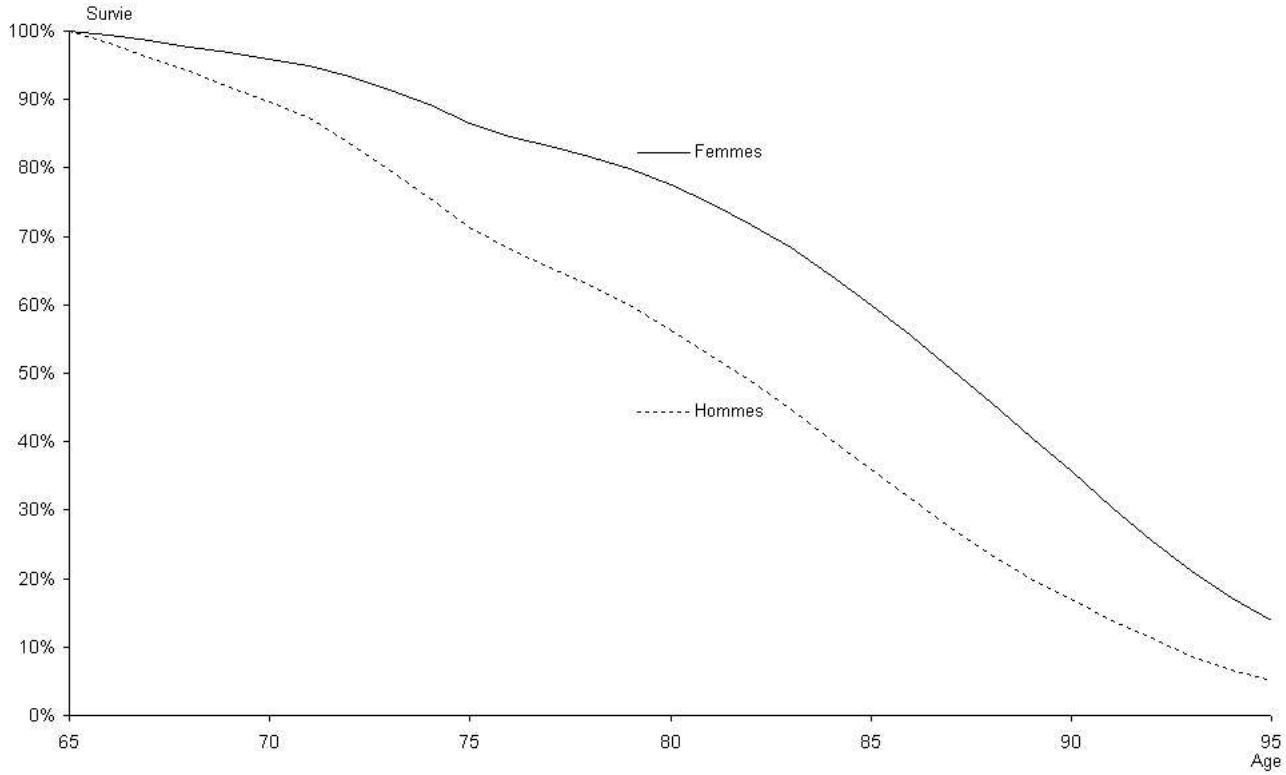

FIG. 3.2 – *Représentation de la fonction de survie (sous l'hypothèse d'une cause de décès uniquement cardio-vasculaire) en fonction du sexe, chez les personnes de plus de 65 ans dans les 75 communes*

La figure 3.2 représente la fonction de survie selon le sexe. On peut observer que la mortalité cardiovasculaire est plus élevée chez les hommes puisque la probabilité d'être toujours en vie à un âge x est plus forte chez les femmes, quel que soit l'âge considéré. La différence entre les deux sexes est très importante puisque l'écart entre les médianes de survie est d'environ 5 ans entre les hommes et les femmes (respectivement 82 et 87 ans pour les survivants à 65 ans).
La mortalité différentielle dépendante du sexe est essentielle car elle conditionne la structure de la population ainsi que l'effet de sélection décrit précédemment. La surmortalité masculine d'origine cardiovasculaire en fonction de l'âge est décrite dans la figure 3.3.
La tendance, mise en évidence dans la figure précédente est confirmée, avec une diminution de la surmortalité masculine qui passe de plus de 2,5 à 1 entre 65 et 95 ans. La surmortalité cardio-

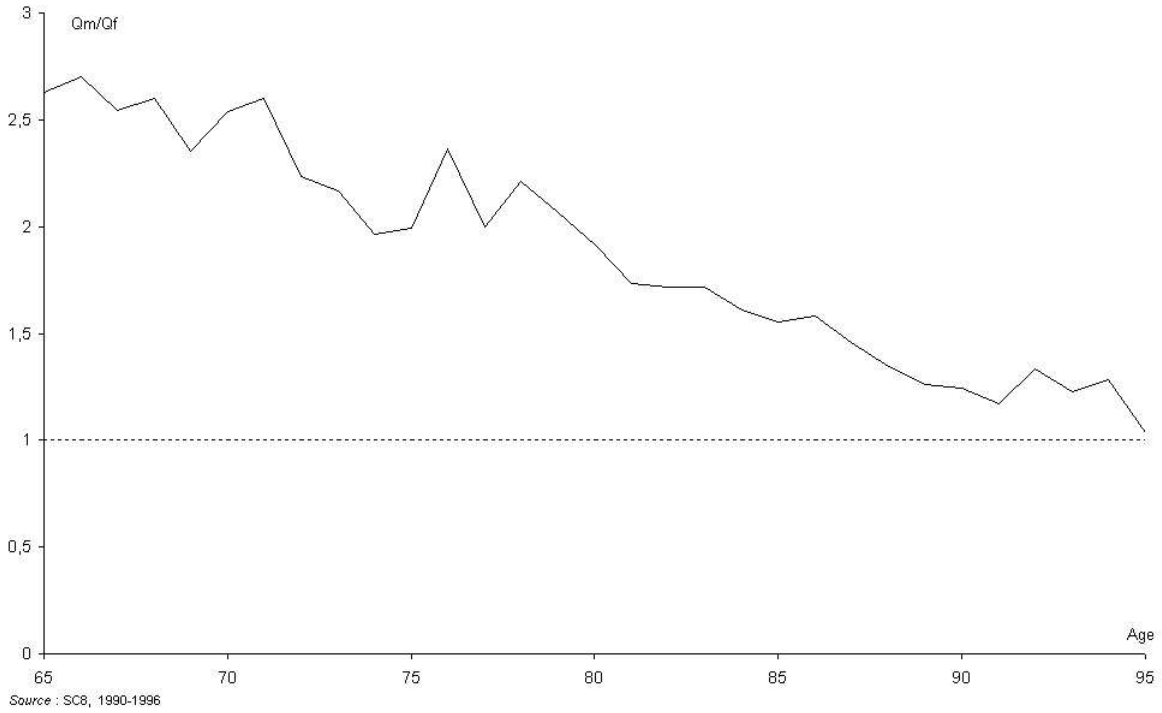

FIG. 3.3 – *La surmortalité d'origine cardiovasculaire des hommes par rapport aux femmes, calcul basé sur les quotients de mortalité*

vasculaire masculine est identique à la surmortalité masculine générale qui est classiquement mise en évidence [67].

Même si les causes de cette surmortalité sont encore mal connues, certains facteurs ont pu être identifiés, tels que la différence de consommation de tabac et d'alcool, ainsi qu'une exposition professionnelle plus importante de la population masculine. Ce "rapprochement" est également le résultat d'une sélection de la population, puisqu'à ces âges avancés seuls les hommes les plus résistants sont encore vivants.

La table 3.1 présente les tables de mortalité d'origine cardiovasculaire dans chacune des deux sous-populations que constituent les hommes et les femmes âgés de plus de 65 ans résidant dans les 75 communes de Gironde et Dordogne. L'événement étudié est donc le décès d'origine cardiovasculaire, dans la cohorte constituée respectivement des hommes et des femmes âgés de plus de 65 ans au 1er janvier 1990. Toutefois, l'étude portant exclusivement sur une cause par-

TAB. 3.1 – *Table de mortalité cardiovasculaire de la population résidant dans les 75 communes sélectionnées en Gironde et Dordogne.*

Age x	Femmes		Hommes	
en a.r.*	S_x	D_x	S_x	D_x
65-69	100000	1534	100000	4146
70-74	98466	2752	95854	4995
75-79	95715	6906	90859	11036
80-84	88809	11991	79823	13674
85-89	76818	17772	66149	16000
90-94	59046	17302	50149	14481
95-99	41744	14091	35669	10095
100+	27653	27653	25574	25574
e_{65}†	*24,4 - 27,2*		*22,7 - 24,7*	

* a.r. : années révolues
† espérance de vie à 65 ans

ticulière, les décès d'origine différente de l'appareil circulatoire doivent être considérés comme des événements perturbateurs, puisqu'ils empêchent d'observer l'événement d'intérêt.

Evolutions projectives

Nous avons effectué un essai de projection sur notre population d'étude. Toutefois la collecte des données des taux de mortalité cardiovasculaire n'ayant pu être réalisée qu'au niveau national, nous avons utilisé cette sous-population comme base de projection. Les effectifs en 2010, 2015 et 2020 ont été estimés par l'INSEE et fournis par classes d'âge de 5 ans entre 65 et 90 ans. La classe d'âge 90 ans et plus n'était pas détaillée.

Nous avons réalisé la projection des taux de mortalité cardiovasculaire avec comme horizon les années 2010, 2015 et 2020, à partir des données obtenues entre 1965 et 1995. Nous avons ainsi mené une étude en séries chronologiques (ARMA - moyenne mobiles) et une régression linéaire sur les séries de données. Ces deux techniques donnaient des résultats très similaires nous permettant de conserver la technique de régression linéaire pour un souci de simplification des calculs.

La série des données a tout d'abord été projetée sans transformation. Mais les résultats obtenus étant parfois "négatifs", nous en avons conclu à une mauvaise spécification du modèle linéaire,

nous conduisant à utiliser un modèle log-linéaire. Les taux ont donc été transformés par la fonction de logarithme népérien. Les résultats obtenus pour les projections des taux sont présentés dans le tableau 3.2.

TAB. 3.2 – *Projections des taux de mortalité et estimations du nombre de décès cardiovasculaires en 2010, 2015 et 2020, selon les hypothèses maximale et moyenne*

Age x	En 2010		En 2015		En 2020	
en a.r.‡	Effectif	Taux	Effectif	Taux	Effectif	Taux
	Millier	(/100000)	Millier	(/100000)	Millier	(/100000)
65-69	2497	221,41	3551	186,79	3736	157,59
70-74	2386	437,03	2343	370,55	3355	314,19
75-79	2249	941,34	2144	817,25	2129	709,53
80-84	1743	2015,25	1837	1794,98	1781	1598,79
85-89	1107	4117,08	1191	3782,73	1283	3475,52
90-94	407	9481,52	662	9234,02	815	8992,98
Décès*	156418		187106		208963	
Décès†	194101		227575		252962	

* Estimation selon l'hypothèse moyenne (taux calculés et présentés dans le tableau)
† Estimation selon l'hypothèse maximale (taux identiques à ceux de 1997)
‡ a.r. : années révolues

Le tableau 3.2 présente les estimations des taux de mortalité d'origine cardiovasculaire en 2010, 2015 et 2020 sous l'hypothèse moyenne qui est une baisse du taux de décès identique à celle connue lors de ces trente dernières années. Le nombre de décès que nous pourrions observer, si cette hypothèse est vérifiée, est présenté dans ce même tableau (décès*). La seconde estimation du nombre de décès est réalisée sous l'hypothèse maximale qui était que les taux de mortalité restent constants à ceux observés en France en 1997.

Dès lors, on peut constater qu'avec les deux conditions différentes, le nombre de décès d'origine cardiovasculaire augmente assez nettement entre 2010 et 2020 (entre 30% et 35% selon les hypothèses). Ce chiffre peut être rapproché des 153 472 décès observés en 1997 et imputés à l'appareil circulatoire.

Incidence de la baisse des causes cardiovasculaires

Afin de mesurer les conséquences qu'aurait la suppression de la mortalité cardiovasculaire sur la situation démographique du pays, nous avons calculé une table de mortalité en l'absence de cette cause nous permettant de déduire une augmentation de l'espérance de vie.
L'événement étudié, dans les deux tables, est le décès des personnes présentes dans la cohorte des personnes âgées de plus de 65 ans au 1er janvier 1990 et résidant dans les 75 communes de PAQUID. Cette population est fermée puisque nous avons montré que les migrations pouvaient être négligeables dans cette classe d'âge. Néanmoins, une différence existe entre les deux tables. En effet, dans la table générale nous avons effectivement une population fermée et par corrollaire, le calcul de la table de mortalité se réduit à un rapport de proportionnalité afin de définir un effectif différent à l'origine. En revanche, dans la table en l'absence de mortalité cardiovasculaire, cette cause de mortalité est considérée comme un événement perturbateur puisqu'elle empêche d'observer le décès en l'absence de cause de décès. Ce point est essentiel dans la calcul des quotients de mortalité.

TAB. 3.3 – *Table de mortalité de la population résidant dans les 75 communes sélectionnées en Gironde et Dordogne*

Age x	En l'absence de MCV		Toute cause de décès	
en a.r.	S_x	D_x	S_x	D_x
65-69	100000	2739	100000	8917
70-74	97261	3853	91083	10768
75-79	93409	9083	80315	18074
80-84	84326	13735	62241	21143
85-89	70591	19123	41098	20760
90-94	51468	18422	20338	13420
95-99	33046	13603	6918	5070
100+	19443	19443	1848	1848
e_{65}	22,0		15,6	

Le tableau 3.3 représente la table de mortalité de la population résidant dans les 75 communes représentatives de la population d'Aquitaine ainsi que la table de mortalité en l'absence de décès d'origine cardiovasculaire. Les quotients permettant de calculer la table "en l'absence de mor-

talité cardiovasculaire" ont été estimés à partir de la formule3.1. On constate que la suppression de la mortalité d'origine cardiovasculaire entraînerait une augmentation de l'espérance de vie à 65 ans d'environ 6 ans, sous l'hypothèse que cette cause soit indépendante des autres causes de décès. Ce constat représentant près de 30% d'augmentation est un facteur important dans la politique générale de la santé publique.

3.4.2 Description de la mortalité cardiovasculaire sur la cohorte Paquid

La table 3.4 représente l'incidence de la mortalité cardiovasculaire, détaillée selon que la cause soit d'origine cérébrovasculaire ou pas. Cette indicidence est calculée à partir de la population générale des 68 communes sélectionnées à partir de l'échantillonnage de la cohorte Paquid, en stratifiant sur le sexe.

TAB. 3.4 – *Incidence de la mortalité cardiovasculaire dans la population générale des 68 communes par sous-cause et par sexe, pour 1000 PA.*

	CérébroVasculaire (N=3819)		Cardiovasculaire (N=10492)		Total (N=14311)	
	Homme	Femme	Homme	Femme	Homme	Femme
Age						
65-69	0,96	0,45	4,36	1,37	5,31	1,82
70-74	2,82	1,46	9,91	3,91	12,73	5,37
75-79	3,83	2,32	12,36	5,77	16,19	7,69
80-84	8,80	6,08	25,68	14,75	34,48	20,83
85-89	16,18	13,81	46,91	33,14	63,10	46,95
90-94	26,26	25,90	79,66	64,69	105,92	90,59
Plus de 95	37,76	35,62	100,70	109,94	138,46	145,57

On observe que la principale cause de décès d'origine cardiovasculaire est la cause "non cérébrovasculaire", puisqu'elle représente environ 75% de tous les décès observés, et ce quel que soit l'âge considéré. On peut également noter que les taux de mortalité sont plus élevés chez les hommes par rapport aux femmes. Les effectifs exprimés dans les colonnes de titre représentent les nombres de décès observés qui ont conduits à l'estimation de ces taux de mor-

talité.

La table 3.5 représente l'incidence de la mortalité cardiovasculaire dans la population générale des 68 communes que nous avons considérées (identique au tableau ci-dessus), ainsi que les taux relevés dans la région Aquitaine et en France. Ces deux derniers taux sont ceux de l'année 1994, année qui correspond au milieu du suivi de la cohorte, ce qui permettra de comparer les taux avec ceux issus de la population générale des 68 communes. Ces taux sont présentés par sexe.

TAB. 3.5 – *Incidence de la mortalité cardiovasculaire dans la population générale (68 communes de Paquid, Aquitaine et France), pour 1000 PA.*

	Population "Paquid"		Aquitaine		France	
	Homme	Femme	Homme	Femme	Homme	Femme
Age						
65-69	5,31	1,82	5,84	2,01	5,95	2,31
70-74	12,73	5,37	11,23	5,44	10,36	4,63
75-79	16,19	7,69	18,61	9,81	17,94	10,00
80-84	34,48	20,83	33,78	21,97	33,31	22,19
85-89	63,10	46,95	61,45	46,78	57,56	45,15
90-94	105,92	90,59	87,37	79,87	89,03	78,56
Plus de 95	138,46	145,57	147,00	151,89	128,83	127,95

source : INED, Base de données Vallin&Mesle, 1994

La première observation que l'on peut faire concerne la proximité de ces taux de mortalité entre les trois populations représentées. En effet, on retrouve des taux sensiblement identiques, quel que soit le sexe considéré. Ce résultat est important car il pourra nous permettre d'étendre nos conclusions à l'ensemble de la population (aquitaine ou française) puisqu'il confirme, sur l'événement d'intérêt de notre analyse, la représentativité de la cohorte d'étude.
On peut également noter que la différence entre les hommes et les femmes, en faveur du sexe féminin, s'atténue avec l'âge. Cette constation confirme ce que l'on avait pu observer sur la figure 3.3, concernant une décroissance de la surmortalité masculine.

Le tableau 3.6 présente les principales caractéristiques des individus issus de la cohorte Pa-

quid et de la population générale des 68 communes sélectionnées de cette étude. Cette table contient également la répartition des différents facteurs de risques individuels que nous utiliserons dans nos analyses et qui ne sont disponibles que sur la population issue de la cohorte Paquid.

TAB. 3.6 – *Distribution des différents facteurs de risques individuels de la mortalité cardiovasculaire dans PAQUID et dans la population générale des 68 communes.*

	Paquid		Population générale	
	N (%)	Moy (et)	N (%)	Moy (et)
Age moyen au décès	83,26	(7,21)	83,50	(7,82)
Homme	1458	(41,53%)	32 928	(40,64%)
Décès CV	441	(36,15%)	14 311	(39,74%)
Effectif (PA)	22 363		516 082	
Consommation				
Ex fumeurs	959	(27,31%)	-	
Fumeurs	339	(9,66%)	-	
Vin 25cl/js	1424	(40,56%)	-	
Vin >25cl/js	543	(15,47%)	-	
HTA	1263	(39,44%)	-	
Diabète	276	(8,60%)	-	
IMC	24,55	(3,99)	-	

On peut observer que les caractéristiques de la population générale et de Paquid sont très proches en ce qui concerne l'incidence du décès d'originie cardiovasculaire, ainsi que de son âge moyen de survenue. Ces deux populations sont également comparables en ce qui concerne la répartition par sexe, puisque les différences observées ne sont pas significatives. Enfin, on peut noter que l'introduction des informations issues de la population générale permet d'augmenter significativement la quantité d'information disponible, puisque les cas recensés de décès augmentent de 441 à 14311.

En ce qui concerne les facteurs de risque individuels, 959 personnes étaient d'anciens fumeurs (27% de l'échantillon) et un peu moins de 10% fumaient toujours (339 personnes). Un peu moins de 40% des individus étaient sous médicament anti hypertenseurs (soit 1263 personnes) et 9% (276 personnes) étaient diabétiques. Enfin, la valeur moyenne de l'Indice de Masse Corporelle (IMC) était de 25.

Le tableau 3.7 présente les éléments descriptifs du calcium et du magnésium recueillis. Ainsi que nous l'avons détaillé dans le chapitre précédent, la consommation de calcium et de magnésium à travers l'eau minérale a été considérée par de multiples estimations. Les éléments présentés dans toutes nos analyses sont issus du recodage ayant permis de détecter la "meilleure" association (*Cf.* chapitre précédent).

TAB. 3.7 – *Descriptif des facteurs de risques liés à l'eau de boisson dans PAQUID.*

	N	Min	Moyenne	Max	1er tercile	Médiane	2d Tercile
Adduction*							
Calcium	68	8,90	72,60	146,20	52,58	79,44	93,71
Magnésium	68	1,15	9,13	34,00	3,98	6,97	10,75
Consommation†							
Calcium	1526	4,16	104,72	514,12	55,40	77,4	99,69
Magnésium	1526	1,58	20,98	101,96	7,08	12,80	22,02

* N=68 communes
† N'=1568 individus

Les données concernant l'eau d'adduction (l'eau du robinet) ont été recueillies sur 68 communes, alors que les données de consommation individuelle l'ont été sur 1526 individus de Gironde.

La première observation concerne la différence entre les deux manières de considérer ces éléments minéraux. La prise en compte de l'eau minérale augmente la moyenne de ces deux éléments, ainsi que l'étendue de consommation. Par exemple, la quantité maximale de calcium absorbée passe de 146 à 514 mg.

Cette différence se retrouve dans l'estimation des terciles puisque ceux-ci ont des valeurs différentes selon que l'on considère l'eau d'adduction et l'eau de consommation. Cependant, lors du recodage des éléments selon les terciles, il est nécessaire de retenir des bornes identiques pour les deux valeurs. En effet, si l'on souhaite comparer les résultats des modèles, nos classes doivent être identiques. En conséquence, nous retiendrons les valeurs des terciles définis par l'eau d'adduction.

3.4.3 Résultat des modèles de Poisson par sous-causes

Dans cette partie, nous présenterons les résultats qui sont extraits de l'article publié dans le journal *Eur J Epidemiol*. Ces tableaux présentent les résultats d'un modèle de Poisson avec prise en compte de la surdispersion (par le χ^2 de Pearson) sur les deux sous-causes de mortalité cardiovasculaire. Ces résultats sont à considérer comme l'approche naïve de cette application. Dans cette section, les éléments minéraux de l'eau de boisson seront considérés uniquement en trois classes, définies dans la section précédente.

TAB. 3.8 – *Modèles de Poisson sur la mortalité cardiovasculaire (non-cérébrovasculaire), ajusté sur le sexe et l'âge.*

	Homme		Femme		Ensemble	
	RR	IC 95%	RR	IC 95%	RR	IC 95%
Calcium (mg/l)						
$[\ 9;\ 53[$	1,00	-	1,00	-	1,00	-
$[\ 53;\ 94[$	0,94	0,86 ; 1,03	0,96	0,88 ; 1,05	0,95	0,88 ; 1,01
$[\ 94; 146]$	0,90*	0,81 ; 0,99	0,91*	0,83 ; 0,99	0,90†	0,84 ; 0,97
Magnesium (mg/l)						
$[\ 1;\ 4[$	1,00	-	1,00	-	1,00	-
$[\ 4; 11[$	0,94	0,85 ; 1,04	0,90*	0,82 ; 0,99	0,92*	0,86 ; 0,99
$[11; 34]$	1,02	0,90 ; 1,15	0,90	0,79 ; 1,02	0,96	0,87 ; 1,05
Ruralité	1,02	0,99 ; 1,05	1,02	0,99 ; 1,06	1,02	1,00 ; 1,05
Echelle §		0,98		1,10		1,11

* p≤0,05 † p≤0,01

§ permet de prendre en compte la surdispersion (ϕ)

Le tableau 3.8 représente les résultats d'un modèle de Poisson, avec prise en compte de la surdispersion par la statistique de Pearson, réalisé sur la mortalité cardiovasculaire (non-cérébrovasculaire). Ces résultats suggèrent un effet protecteur du calcium contenu dans l'eau d'adduction, avec un potentiel effet dose-réponse, et cet effet semble être identique dans les deux sexes. L'effet protecteur est significatif lorsque la concentration de calcium est supérieur à 94 mg/l.

Le magnésium semble avoir un effet proctecteur pour des doses modérées (RR=0,9 et légèrement significatif) mais ne semble pas le conserver lorsque la concentration devient plus importante.

TAB. 3.9 – *Modèles de Poisson sur la mortalité cérébrovasculaire, ajusté sur le sexe et l'âge.*

	Homme		Femme		Ensemble	
	RR	IC 95%	RR	IC 95%	RR	IC 95%
Calcium (mg/l)						
$[\ 9;\ 53[$	1,00	-	1,00	-	1,00	-
$[\ 53;\ 94[$	0,94	0,79 ; 1,11	0,90	0,79 ; 1,03	0,91	0,82 ; 1,01
$[\ 94; 146]$	0,89	0,74 ; 1,07	0,84*	0,74 ; 0,97	0,86†	0,77 ; 0,96
Magnesium (mg/l)						
$[\ 1;\ 4[$	1,00	-	1,00	-	1,00	-
$[\ 4; 11[$	0,74†	0,63 ; 0,88	0,79†	0,69 ; 0,91	0,77†	0,69 ; 0,86
$[11; 34]$	0,91	0,72 ; 1,14	0,93	0,78 ; 1,11	0,92	0,80 ; 1,06
Ruralité	1,07	1,00 ; 1,14	1,10	1,05 ; 1,15	1,09	1,05 ; 1,13
Echelle §		0,99		1,01		1,00

* p≤0,05 † p≤0,01

§ permet de prendre en compte la surdispersion (ϕ)

Le tableau 3.9 représente les résultats du même modèle de Poisson que précédemment mais en considérant la mortalité cérébrovasculaire. Lorsque l'on s'intéresse à cette cause spécifique, les résultats concernant l'effet du calcium semble similaires à ceux obtenus ci-dessus. Ce minéral semble avoir un effet protecteur pour des concentrations supérieures à 94mg/l. Cet effet est retrouvé de manière identique dans les deux sexes, même s'il apparait un peu plus marqué chez les femmes.

Le magnésium, quant à lui, exhibe plutôt un effet en "U", puisqu'il semble être significatif pour des concentrations comprises entre 4 et 11 mg/l (RR=0,8 et statistiquement significatif), alors que son effet est moins marqué lorsque la concentration augmente.

3.4.4 Résultat du modèle Arcsinus

Nous avons conclu dans le chapitre précédent que la régression Arcsinus Stricte était une alternative et une approche complémentaire à la régression Binomiale-Négative. Nous avons également présenté la log-vraisemblance comme le critère de choix entre ces deux modèles. Ainsi, le tableau 3.10 présente les résultats de ces deux modèles sur les données de notre application.

TAB. 3.10 – *Modèles Poisson, Binomial-Négatif et Arcsinus Strict sur la population générale des 68 communes de Gironde et Dordogne, entre 1990 et 1997. Résultats concernant l'eau d'adduction considérée en continu.*

	Poisson		Binomial-Négatif		Arcsinus Strict	
	RR	IC 95%	RR	IC 95%	RR	IC 95%
Calcium	0,999*	0,998 ; 1,000	0,999	0,998 ; 1,001	0,999*	0,998 ; 1,000
Magnésium	1,002	0,998 ; 1,006	1,006*	1,000 ; 1,012	1,005*	1,001 ; 1,009
Ruralité	1,151*	1,092 ; 1,214	1,115†	1,031 ; 1,206	1,107*	1,030 ; 1,191
α	-		0,111		0,062	
Log-Likelihood	-2254,58		-2061,38		-2015,69	

* p≤0,05 † p≤0,001
‡ vs Homme

Ainsi que nous l'avons développé dans les simulations (*cf.* chapitre **??**), nous pouvons déterminer le meilleur modèle à partir de la valeur de la log-vraisemblance, la plus grande étant la meilleure. La table 3.10 montre que le modèle AS est le meilleur modèle, avec une log-vraisemblance (l_{AS}) égale à -2016, comparée à l_{BN}=-2061. Ainsi, même si les estimations des paramètres sont très proches, nous pouvons conclure à un meilleur ajustement par le modèle AS par rapport au modèle BN. Si on compare ces résultats avec ceux du modèle de Poisson, on peut constater que même si les estimations des paramètres ne sont pas modifiée, les estimations de la variance sont nettement plus faibles dans le modèle sans surdispersion (Poisson). On peut également observer que le paramètre de surdispersion n'est pas très élévé (α=0,06 dans AS). De plus, comme nous l'avons explicité dans la présentation du modèle, on ne peut pas comparer les valeurs de ce paramètre dans les deux modèles.

Ce modèle met en évidence un léger effet protecteur du calcium pris en continu. Cet effet semble faible car il représente la différence de risque pour une différence de concentration d'un milligramme de Calcium, ce qui est une très faible différence par rapport à l'amplitude de cette variable. Si l'on ramène ce risque pour une variation 100mg (ce qui est plus facilement interprétable), le risque relatif est de 0,92, ce qui est conforme à ce que l'on a pu mettre en évidence par ailleurs, même si son effet n'est pas significatif. L'effet du magnésium semble délétère (RR=1,03 pour une variation de 10 unités) mais sa significativité statistique est très faible.

TAB. 3.11 – *Modèles Poisson, Binomial-Négatif et Arcsinus Strict sur la population générale des 68 communes, concernant l'eau d'adduction considérée en terciles.*

	Poisson		Binomial-Négatif		Arcsinus Strict	
	RR	IC 95%	RR	IC 95%	RR	IC 95%
Calcium (mg/l)						
$[\ 9;\ 53[$	1,000	-	1,000	-	1,000	-
$[\ 53;\ 94[$	0,963	0,921 ; 1,007	0,972	0,864 ; 1,094	0,970	0,893 ; 1,053
$[\ 94; 146]$	0,911†	0,864 ; 0,949	0,937	0,849 ; 1,061	0,944	0,868 ; 1,026
Magnesium (mg/l)						
$[\ 1;\ 4[$	1,000	-	1,000	-	1,000	-
$[\ 4; 11[$	0,905†	0,852 ; 0,961	0,951	0,869 ; 1,042	0,939	0,865 ; 1,020
$[11; 34]$	0,932	0,866 ; 1,003	1,011	0,911 ; 1,122	1,021	0,931 ; 1,119
Ruralité	1,088*	1,023 ; 1,157	1,083	0,997 ; 1,178	1,074	0,997 ; 1,157
α		-		0,109		0,060
Log-Likelihood		-2217,23		-2058,65		-2019,05

* p≤0,05 † p≤0,001
‡ vs Homme

Les résultats des modèles réalisés en considérant le calcium et le magnésium en trois classes sont présentés dans le tableau 3.11. Ces résultats confirme un potentiel effet dose-réponse protecteur du calcium, même si cet effet n'est pas significatif après avoir sélectionné le modèle adéquat. De même, le magnésium semble confirmer son effet en "cloche" inversée, effet non significatif également.

On peut enfin observer que les paramètres de dispersion sont très proches des estimations obtenues lorsque ces variables étaient considérées en continu (0,111 vs 0,109 pour BN et 0,062 vs 0,060 pour AS). Il est également important de noter que le mauvais choix du modèle pourrait conduire à des conclusions faussement significatives, puisque l'effet du calcium est statistiquement significatif si l'on utilise un modèle de Poisson classique. Enfin, on peut conclure que même si l'explication épidémiologique est plus pertinente en catégorisant les variables, l'ajustement statistique n'est que très peu modifié selon le codage adopté, même s'il est possible de suspecter un effet seuil.

3.4.5 Résultat du modèle pour données agrégées

A la suite des simulations (*cf.* section 2.2.3), nous avons pu observer que les estimations des paramètres issues de BN et de MHA sont très proches et que seules les estimations des variances sont biaisées. Nous effectuerons donc les deux modèles (BN et MHA) sur chaque jeu de données afin de déterminer le modèle le plus adapté.
Le tableau 3.12 présente les résultats lorsque les minéraux sont considérés comme des variables continues issues uniquement de l'eau d'adduction.

TAB. 3.12 – *Modèles Binomial-Négatif et de MHA sur les données de l'eau d'adduction. Les minéraux sont considérés en continu.*

	Binomial-Négatif			MHA		
	RR	IC 95%	RR / 100mg	RR	IC 95%	RR / 100mg
Calcium	0,999	0,998 ; 1,001	0,905	1,000	0,998 ; 1,001	0,980
Magnésium	1,006*	1,000 ; 1,012	1,062†	1,001	0,990 ; 1,012	1,014†
Ruralité	1,115*	1,030 ; 1,191		1,234*	1,108 ; 1,397	
Diabète	-	-		1,168	0,532 ; 2,568	
HTA	-	-		0,614	0,344 ; 1,098	
Consommation de tabac						
Non Fumeurs	-	-		1,000	-	
Anciens fumeurs	-	-		1,238*	1,015 ; 1,511	
Fumeurs actuels	-	-		1,718*	1,389 ; 2,128	
Dispersion		α=0,111			σ^2=1,057	

* p≤0,05 † RR pour 10mg

On peut observer dans le tableau précédent que les estimateurs des paramètres du Calcium et du Magnésium sont très proches de 1. Il convient de préciser que ces estimations sont données pour une variation d'une unité, ce qui est très faible lorsque l'on considère le Calcium qui a une étendue de plus de 100 unités. Ce même risque relatif est égal à 0,97 pour une variation de 100mg/l, ce qui est plus facilement interprétable d'un point de vue épidémiologique.
La seconde conclusion que l'on peut extraire de ce tableau est que même si ces deux éléments paraissent avoir un effet protecteur sur la mortalité cardiovasculaire, ils ne sont pas statistiquement significatifs. Enfin, on peut noter que ces données présentent une légère surdispersion.

Le tableau 3.13 représente les résultats des mêmes modèles que le précédent à la différence que les éléments minéraux sont déterminés par rapport à l'eau consommée. Cette variable étant une variable individuelle, nous avons introduit la moyenne de ces consommations par strate et commune dans le modèle BN.

TAB. 3.13 – *Modèles Binomial-Négatif et de MHA sur les données de l'eau de consommation. Les minéraux sont considérés en continu.*

	Binomial-Négatif			MHA		
	RR	IC 95%	RR / 100mg	RR	IC 95%	RR / 100mg
Calcium	0,998	0,994 ; 1,002	0,82	0,998*	0,996 ; 1,000	0,82
Magnésium	1,022*	1,004 ; 1,040	1,25†	1,011	0,993 ; 1,029	1,12†
Ruralité	1,103*	1,031 ; 1,180		1,122	0,941 ; 1,338	
Diabète	-	-		1,022	0,342 ; 3,056	
HTA	-	-		0,571	0,326 ; 1,002	
Consommation de tabac						
Non Fumeurs	-	-		1,000	-	
Anciens fumeurs	-	-		1,570*	1,209 ; 2,041	
Fumeurs actuels	-	-		1,319*	1,025 ; 1,695	
Dispersion		α=0,096			σ^2=1,048	

* p≤0,05 † RR pour 10mg

On peut observer que l'effet du calcium paraît plus important que dans l'analyse précédente. Le risque relatif associé à une variation de consommation de 100 mg est d'environ 0,80, ce qui représentent un effet protecteur non négligeable. De plus, cet effet semble être légèrement significatif ($p = 0{,}06$).

En revanche, l'effet du magnésium réellement consommé est différent de celui mis en évidence lors de l'analyse sur les données concernant l'eau d'adduction, puisque ce minéral semble avoir un effet délétère (RR=1,25 pour une variation 10mg). Cependant, cet effet n'est pas du tout significatif et la différence entre les deux analyses peut s'expliquer par la forte variance qui est associée à ces estimations.

Le tableau 3.14 présente les résultats des modèles décrits dans les tableaux précédents, mais considère que les éléments minéraux contenus dans l'eau d'adduction sont catégorisés en trois classes, afin de mettre en évidence un éventuel effet "en U".

TAB. 3.14 – *Modèles de MHA et Binomial-Négatif sur les données de l'eau d'adduction, considérées en trois classes.*

	Binomial-Négatif		MHA	
	RR	IC 95%	RR	IC 95%
Calcium (mg/l)				
$[\ 9;\ 53[$	1,000	-	1,000	-
$[\ 53;\ 94[$	0,972	0,864 ; 1,094	0,975	0,880 ; 1,081
$[\ 94; 146]$	0,949	0,849 ; 1,061	0,930	0,806 ; 1,072
Magnesium (mg/l)				
$[\ 1;\ 4[$	1,000	-	1,000	-
$[\ 4; 11[$	0,951	0,869 ; 1,042	0,847*	0,750 ; 0,956
$[11; 34]$	1,011	0,911 ; 1,122	0,877	0,757 ; 1,016
Ruralité	1,083	0,997 ; 1,178	1,151*	1,082 ; 1,224
Diabète	-	-	1,174	0,474 ; 2,903
HTA	-	-	0,617	0,349 ; 1,093
Consommation de tabac				
Non Fumeurs	-	-	1,000	-
Anciens fumeurs	-	-	1,366*	1,158 ; 1,612
Fumeurs actuels	-	-	1,665*	1,344 ; 2,049
Dispersion	α=0,105		σ^2=1,091	
* p$\leq$0,05			† RR pour 10mg	

Les résultats du tableau précédents confirment les résultats déjà obtenus à partir des tableaux 3.8, 3.9 et 3.11 concernant les effets du calcium et du magnésium contenus dans l'eau d'adduction. En effet, ces deux modèles confirment l'effet dose-réponse observé sur le calcium contenu dans l'eau d'adduction. L'effet du magnésium paraît, quant à lui, assez différent selon que l'on considère un modèle BN ou MHA. Cependant, ces effets ne sont pas statistiquement significatifs.

3.4.6 Résultat du modèle avec corrélations spatiales

Dans les tableaux suivants, nous présenterons les résultats des analyses avec prise en compte des corrélations spatiales. Ces analyses comprennent une analyse avec un modèle BN prenant en compte l'association spatiale par la longitude et la lattitude du centre de la commune. Nous présenterons également les résultats d'un modèle MHA avec corrélations spatiales, telle que Guthrie l'a définit ainsi qu'avec notre adaptation de cette structure de corrélation. La proximité entre deux communes sera définie par une distance entre les centres de ces zones inférieures à 50 kilomètres.

TAB. 3.15 – *Modèles MHA et Binomial-Négatif adaptés au design spatial sur les données de l'eau d'adduction*

	BN+Li+li			Spatial		
	RR	IC 95%	RR/100mg	RR	IC 95%	RR/100mg
Calcium	0,998	0,992 ; 1,004	0,82	0,998	0,990 ; 1,005	0,79
Magnésium	1,081	1,056 ; 1,107*†	1,06	1,000	0,998 ; 1,002	1,00†
Longitude	1,006	0,914 ; 1,107		-	-	
Lattitude	0,992	0,942 ; 1,046		-	-	
Ruralité	1,019	0,953 ; 1,094		1,142*	1,082 ; 1,216	
Diabète	-	-	-	1,134	0,592 ; 2,172	
HTA	-	-	-	0,660	0,429 ; 1,016	
Dispersion		α=0,093			σ^2=1,012	

* p$\leq$0,05 † RR pour 10mg

On constate que les résultats issus de ces deux modèles sont très proches, tant en ce qui concerne l'estimation des paramètres que dans l'estimation de la surdispersion. On peut également remarquer que l'effet de la longitude et de la lattitude n'est pas significatif dans notre modèle. On peut enfin noter que l'effet du calcium est atténué par rapport à celui observé dans le tableau 3.12 alors que celui du magnésium semble plus "amorti".

Afin de valider la sensibilité de notre définition de proximité, nous avons réalisé ces mêmes analyses en prenant comme critère de proximité une distance de 25 et 75 kilomètres entre les centres des communes. La modification des estimateurs et de leurs variances étant inférieure à

15%, nous pouvons considérer que les résultats obtenus sont très peu sensibles à la définition du critère de proximité.

3.5 Conclusion de l'application

Les différentes analyses réalisées semblent montrer une relation dose-réponse non significative entre le calcium contenu dans l'eau de boisson et la mortalité cardiovasculaire. En accord avec les résultats obtenus dans notre première analyse [68], une concentration de calcium supérieure à 94 mg/l serait associée à un RR=0,9 par rapport à une concentration inférieure à 53 mg. Ces résultats semblent également mettre en évidence un effet en "U" de la concentration de magnésium sur cette cause de décès.
Cependant, il est important de noter que ces résultats perdent toute significativité statistique lorsque l'on considère des modèles permettant de prendre en compte une éventuelle surdispersion. De plus, toutes ces analyses ont pris en compte l'âge et le sexe comme facteurs d'ajustement de cette analyse, même si les estimations des risques relatifs ne sont pas présentés dans les tableaux. Ces estimations étaient conformes à celles obtenues dans les publications précédentes.
Enfin, il semblait que l'effet de ces deux minéraux était identique dans les sous-causes de mortalité cardiovasculaire que nous avons considéré.

La principale originalité de notre application concernait la population cible de l'étude. En effet, l'effet du calcium et du magnésium (séparément) a rarement été étudié sur la mortalité des personnes âgées. De plus, nous avons pu déterminer la part de chacune des deux sous-causes de mortalité.

La principale faiblesse de la majorité des études similaires concerne la définition de la variable d'exposition. En effet, il est très difficile de déterminer, dans les études épidémiologiques, une exposition reflétant les multiples sources d'apports en calcium et magnésium dans l'eau d'adduction. Afin de prendre en compte cette incertitude de mesure, nous avons considéré des valeurs moyennes au cours du temps (valeurs relevées et fournies par la DDASS) pour chacune des sources d'approvisionnement des villes.
Nous nous sommes intéressés aux effets des éléments minéraux de l'eau de boisson sur la mortalité cardiovasculaire. Il était donc important de prendre en compte la consommation minérale

dans nos analyses. L'une des principales faiblesses de notre mresure de l'exposition vient de sa détermination. En effet, nous n'avions pas de mesure précise pour l'ensemble de la population, mais seulement pour la population suivie à 3 ans (T3) en Gironde. De plus, ce recueil était relativement imprécis puisqu'il ne permettait pas de quantifier l'apport en minéraux issu de cette consommation. Cependant, la méthode de pondération que nous avons utilisée pour déterminer cette exposition semble valide et la méthode permettant de déterminer le meilleur codage a fait l'objet d'une publication [40]. Enfin, les analyses de sensibilité réalisées sur les différents codage nous permettent de conclure à une bonne détermination de notre définition de l'exposition. Afin de modéliser au mieux cette association, nous avons considéré de multiples codages, en terciles et en continu. Cependant, dans un souci de comparabilité, le recodage en trois classes des minéraux s'est fait avec les mêmes seuils pour l'eau d'adduction et l'eau de consommation. La différence entre les résultats de ces deux variables, considérées en trois classes, peut s'expliquer en partie par ce recodage. En effet, on peut penser que les valeurs des minéraux étant très différentes selon leur origine (les concentrations sont beaucoup plus importantes dans l'eau embouteillée par rapport à l'eau d'adduction), la catégorisation de ces variables peut entrainer une distorsion dans les effets observés.

Les causes de décès ont été fournies par le SC8 qui est un service spécialement dédié à l'analyse des causes de décès. Cependant, il peut apparaitre un certain biais d'information dans la classification de ces causes. En effet, les certificats de décès sont souvent remplis de manière parcellaire et l'extraction d'informations à partir de ceux-ci peut être délicate.

Nous avons donc mené une étude de concordance entre les causes de décès fournies par le SC8 et les causes recueillies dans la cohorte PAQUID. Cette dernière procédure est basée sur la collecte des informations auprès des médecins généralistes (voir publications antérieures [20, 43] pour plus d'informations sur le recueil). Cette concordance semblait très bonne et le peu de cas discordants relevés ne pouvaient être caractérisé comme une erreur de l'une des deux sources (manque d'information ou mauvais codage). Il apparait donc que les certificats de mortalité d'origine cardiovacsulaire sont des données plutôt fiables.

Il est cependant nécessaire de nuancer cette remarque en précisant que cette concordance était

beaucoup moins importante lorsque l'on considérait la mortalité par sous-cause. En effet, même si le SC8 reporte très bien le décès d'origine cardiovasculaire, les informations concernant la sous-cause (cérébrovasculaire par exemple) semblent beaucoup moins fiables. Il serait donc nécessaire de mener des études plus précises sur les éventuelles mauvaises classifications par sous-causes de mortalité.

Ces analyses ont également été ajustées sur le statut urbain ou rural de la commune de résidence. Cet ajustement a été considéré car il peut être considérée comme une variable synthétique reflétant des variables difficilement mesurables, telles que le niveau socio-économique de la zone géographique ou le niveau de pollution. Cependant, cette variable était rarement significative dans nos analyses et son impact semble donc limité.

En ce qui concerne les différents facteurs d'ajustement introduits dans les modèles, on peut constater que le diabète ne semblait pas associé à la mortalité cardiovasculaire. On peut également noter que l'hypertension artérielle est associée à une réduction du risque de mortalité cardiovasculaire dans nos analyses (mais son effet semble non significatif). Cet effet peut avoir deux causes distinctes. La première concerne le mode de recueil de cette variable. En effet, l'hypertension artérielle (HTA) est définie par la médication à visée anti-hypertensive, ce qui peut donc entrainer un biais de mesure, les personnes présentant une HTA étant déjà sous traitement. La seconde explication de cette situation tient à la sous-population à laquelle nous nous intéressons. En effet, il est possible que les personnes présentant une HTA et qui étaient à risque de décéder par MCV n'ont peut être pas survécu jusqu'à 65 ans (âge d'inclusion dans notre étude) et nous avons donc uniquement des personnes très résistantes dans cette catégorie (présentant une hypertension), ce qui entrainerait un biais de sélection.

Seule, la consommation de tabac semble être associée à une augmentation de cette cause de décès. On retrouve ainsi, dans les trois analyses réalisées, un effet plus important chez les fumeurs actuels par rapport aux non fumeurs mais également par rapport aux anciens fumeurs. Les risques ainsi mis en évidence sont similaires à ceux rapportés dans les études antérieures. En ce qui concerne la consommation de vin, aucune association n'a été trouvée avec la mortalité cardiovasculaire.

La non-significativité des paramètres de positionnement spatial que sont la longitude et la latitude peut s'expliquer par la faible distance qui sépare les communes dans notre échantillon. En effet, celui-ci est composé des seules communes de Gironde et de Dordogne, ce qui représente un éloignement maximum inférieur à 250km. Il serait donc nécessaire, afin d'augmenter la variabilité de ces paramètres, de constituer un échantillon comprenant des variations géographiques beaucoup plus importantes, telles que différentes régions françaises ou européennes.

Enfin, la partie concernant la démographie devrait permettre de mieux définir la politique de santé publique en matière de prise en charge de la mortalité cardiovascualire. En effet, les projections montrent que, en dépit d'une baisse continuelle du taux de mortalité cardiovasculaire chez les personnes âgées, le nombre absolu de décès imputables à cette cause devrait considérablement augmenter pour atteindre un chiffre compris entre 209000 et 253000 décès par an. Cette situation est le résultat de l'accroissement du nombre de personnes âgées et de l'allongement de la survie de ces personnes.
De plus, ces projections montrent que si la mortalité cardiovasculaire était supprimée, l'espérance de vie à 65 ans serait augmentée de sept années. Bien qu'il soit impossible de se placer dans une telle situation, cette hypothèse permet de comparer l'impact de chacune des causes de décès sur l'allongement de la vie. C'est sur ces comparaisons que peuvent s'appuyer les choix en matière de politique de santé publique et ces résultats confirment l'intérêt de lutter contre cette cause de décès, même chez les personnes âgées, qui n'est pas une cible actuellement privilégiée par les actions de sensibilisation.

Chapitre 4

Discussion

Ce travail de recherche avait pour objectif de déterminer une approche permettant d'analyser de manière optimale des données de mortalité sur de très grands échantillons. En effet, les professionnels de santé publique ont de plus en plus souvent accès à des sources d'information contenant de très nombreuses informations et une des problématiques émergentes est de comprendre et mettre en œuvre des outils permettant de dégager l'information la plus fiable de ces bases de données. Si l'on s'intéresse aux problèmes qui concernent plus spécifiquement le statisticien, la surdispersion est l'un des problèmes principaux qui nuit à la fiabilité des résultats dégagés lors de la modélisation.
Afin d'apporter une réflexion complète autour de ces problèmes, nous nous sommes attachés à cerner les causes de cette surdispersion et nous proposons des solutions opérationnelles afin d'améliorer les conclusions dégagées lors de ces analyses.

4.1 Conclusion sur les résultats épidémiologiques

Comme nous avons pu l'appréhender dans notre application, la mortalité cardiovasculaire semble associée aux minéraux contenus dans l'eau de boisson, notamment le calcium et le magnésium. Cependant, cette association n'était pas statistiquement significative à partir de nos données. Ce manque de significativité peut s'expliquer notamment par la très faible dispersion

géographique de notre échantillon, qui était limité au Sud-Ouest de la France.
En effet, on peut envisager de mener une étude basée sur un schéma similaire en recueillant les données de mortalité de plusieurs communes auprès du SC8 et de les mettre en relation avec des données individuelles qui auraient pu être recueillies dans d'autres échantillons représentatifs constitués pour étudier cette cause de mortalité. Cette étude complémentaire fournirait une plus grande variabilité dans les variables explicatives, mais permettrait également de définir une variable d'exposition plus précisément.

4.2 Conclusion sur les modèles statistiques

Dans ce travail, nous avons notamment proposé une approche complémentaire à la régression BN. En effet, dans certains cas, ce modèle de régression n'est pas bien adapté aux données surdispersées et nous avons montré que la régression Arcsinus Stricte (AS) pouvait être une bonne alternative à cette méthode. Cette méthode étant proposée à l'aide d'une macro SAS, elle est opérationnelle et d'une mise en oeuvre facile [69].
Comme nous l'avons vu dans le chapitre 2.1, ce modèle est basé sur une loi AS, qui est une loi appartenant à la Famille Exponentielle à Dispersion. Intuitivement, on peut penser que cette distribution est un mélange de Poisson. Cependant, la loi conguguée n'a pas encore été identifiée [53] et une des perspectives de ce travail est de déterminer la loi suivie par l'effet aléatoire afin que la distribution marginale du modèle suive une loi AS.

Ce travail nous a aussi permis d'évaluer et de mettre en pratique une méthode originale qui permet de modéliser des données issues de recueil hiérarchique. Il apparait que ce modèle est tout à fait adapté aux problématiques des épidémiologistes travaillant sur de données de mortalité. En effet, il offre la possibilité de mêler des données recueillies au niveau de la population (notamment auprès de sources administratives) et des données issues d'un échantillonnage de cette population pour collecter des variables d'ajustement ou d'exposition (cohorte d'étude).

De plus, cette méthode permet des économies substantielles dans la réalisation d'études de grandes échelles. L'intérêt majeur des études écologiques concerne les très grandes tailles des effectifs considérés, qui permettent de mettre en évidence des risques très faibles mais qui ont l'inconvénient d'être très suceptibles aux différents biais présents (biais écologique, de mesure, d'information,...) [32, 33, 75, 76]. L'approche des modèles hiérarchiques agrégés permet de ne conserver que les avantages de cette approche (effectifs très importants) en les combinant avec les avantages des études basées sur les échantillons (variables individuelles).

Même si notre application n'a pas permis de mettre en évidence de modification de l'effet des variables d'exposition en utilisant ce modèle par rapport à d'autres approches, il a cependant permis d'évaluer sa faisabilité et de valider les résultats fournis par rapport aux méthodes "classiques". En accord avec les remarques de Cox [17], ce travail a également montré qu'en présence d'une légère surdispersion, seules les estimations des variances étaient biaisées et que les estimations des paramètres n'étaient que très peu altérées. Enfin, il est envisageable que ce modèle, introduisant une corrélation spatiale, n'ai pu mettre en évidence de différence des estimations par le simple fait du manque de variabilité qu'il existait entre les communes représentées dans Paquid. Il serait important de réaliser ces mêmes analyses sur des communes présentant un éloignement spatial beaucoup plus important.
Cependant, ce modèle, même s'il est bien implémenté dans les logiciels statistiques usuels, reste d'une mise en oeuvre relativement complexe et la modélisation rencontre souvent des problèmes de convergence ou de validité des résultats obtenus. Les modèles concernant l'eau de consommation et le découpage en tercile n'ont, ainsi, pu converger dans notre modélisation, ce qui explique l'absence de résultats dans l'application.

Enfin, il existe quelques travaux étudiant la relation entre les modèles de dénombrement et les modèles de survie. En effet, il est possible d'étudier la mortalité avec l'une ou l'autre de ces approches selon la question et les moyens disponibles, ainsi que selon la force de l'association considérée. Cependant, même si on peut intuitivement considérer que ces approches sont proches, les relations statistiques entre ces deux modèles ne sont pas nettement établies.

Le Modèle Hiérarchique Agrégé donne un élément de réponse à cette question. La base de ce modèle est basé sur une modèle individuel qui peut être rapproché d'un modèle à risque proportionnel de Cox (si l'on suppose que le taux de mortalité observé est identique à une probabilité de connaître l'événement). Ce modèle individuel est ainsi approché par un modèle conditionnel de Poisson, par sommation sur les individus d'une même zone géographique. Cela amène donc un élément de discussion nouveau sur les relations qui unissent ces deux classes de modèles.

Perspectives

En accord avec les conclusions détaillées ci-dessus, les perspectives de ce travail peuvent être orientées selon deux axes. Le premier concerne l'étude plus approfondie de la relation entre le calcium et la mortalité cardiovasculaire, notamment par la constitution d'un schéma d'étude incluant plusieurs régions françaises avec une méthode statistique basée sur le MHA. Cet échantillon permettrait d'augmenter la variabilité de nombreuses variables disponible, telle que la concentration des minéraux dans l'eau d'adduction ainsi que les variations géographiques. Ce prolongement pourrait également inclure une composante alimentaire de l'apport de calcium et de magnésium au niveau individuel.

Le second axe qu'il serait souhaitable d'approfondir concerne les modèles statistiques. En effet, nous avons montré que le modèle Binomial-Négatif n'était pas le modèle le plus adapté dans toutes les situations. La régression AS permet une approche complémentaire lorsque la surdispersion est faible ou modérée. Il convient tout d'abord de compléter la caractérisation de cette distribution, notamment par la définition de la loi conjuguée à la loi de Poisson qui permettrait d'obtenir une telle distribution. De plus, en présence d'une surdispersion plus importante, ces deux modèles ne semblent plus adapté et il conviendrait d'adopter un modèle à trois paramètres tel que Binomial-Négatif Généralisé [50]. Cependant, même si cette distribution est bien définie, il est nécessaire de détailler les différents paramètres de régression qui y sont associés ainsi que la méthode d'estimation la plus adéquate (un travail est actuellement en cours pour déterminer une approche en deux étapes, basée sur la vraisemblance profilée). Enfin, il serait intéressant

d’étudier plus en détail les performances des tests d’hypothèses concernant le paramètre de surdispersion dans le modèle AS et, éventuellement, de développer une combinaison de tests qui permettrait de valider le modèle adéquat et de réaliser une inférence sur le paramètre de surdispersion. Ce travail fait l’objet d’une formalisation actuelle en vue d’un travail de recherche futur.

Bibliographie

[1] M Aitkin. A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, 6(3):332–9, 1996.

[2] AB Anderson and RL Prentice. On the accommodation of disease rate correlations in aggregate data studies of disease risk factors. *Biometrics*, 54:1527–40, 1998.

[3] DA Anderson and JP Hinde. Random effects in generalized linear models and the EM algorithm. *Communication in Statistics: Theory and Methods*, 17(11):3847–56, 1988.

[4] I Aptel, A Cance-rouzaud, and H Grandjean. Association between calcium ingested from drinking water and femoral bone density in elderly women: evidence from the EPIDOS cohort. *J Bone Miner Res*, 14(5):829–33, 1999.

[5] SL Beal and LB Sheiner. Estimating population kinetics. *CRC Crit Rev Biomed Eng*, 8:195–222, 1982.

[6] J Besag, J York, and A Mollie. Bayesian image restoration with two applications in spatial statistics. *Ann Inst Stat Math*, 43(1):1–59, 1991.

[7] TA Blakely and AJ Woodward. Ecological effects in multi-level studies. *J Epidemiol Community Health*, 54:367–74, 2000.

[8] G Bolinder, L Alfredsson, A Englund, and U de Faire U. Smokeless tobacco use and increased cardiovascular mortality among swedish construction workers. *Am J Public Health*, 84(3):399–404, 1994.

[9] RM Bostick, LH Kushi, Y Wu, KA Meyer, TA Sellers, and AR Folsom. Relation of calcium vitamin D and dairy food intake to ischemic heart disease mortality among post-menopausal women. *Am J Epidemiol*, 149(2):151–61, 1999.

[10] NE Breslow. Extra-poisson variation in log-linear models. *Applied Statistics*, 33:38–44, 1984.

[11] NE Breslow. Tests of hypotheses in overdispersed poisson regression and other quasi-likelihood models. *JASA*, 85(410):565–71, 1990.

[12] AS Bryck and SW Raudenbush. *Hierarchical linear Models*. Newbury Park, 1992.

[13] JJ Chen and H Ahn. Fitting mixed Poisson regression models using quasi-likelihood methods. *Biometrical J*, 38:81–96, 1996.

[14] B Chipperfield and JR Chipperfield. Magnesium and potassium content of normal heart muscle in areas of hard and soft water. *Lancet*, i:121–2, 1976.

[15] PH Chyou, CM Burchfiel, K Yano, DS Sharp, BL Rodriguez, and JD Curb. Obesity alcohol consumption smoking and mortality. *Ann Epidemiol*, 7:311–7, 1997.

[16] DG Cook and SJ Pocock. Multiple regression in geographical mortality studies with allowance for spatially correlated errors. *Biometics*, 39(2):361–71, 1983.

[17] DR Cox. Some remarks on overdispersion. *Biometrika*, 70(1):269–74, 1983.

[18] MD Crawford, MJ Gardner, and JN Morris. Mortality and hardness of local water supplies. *Lancet*, 1(7547):827–31, 1968.

[19] N Cressie. *Statistics for spatial data*. John Wiley & Sons, 1993.

[20] JF Dartigues, M Gagnon, L Letenneur, D Commenges, P Barberger-Gateau, and S Auriacombe et al. Le programme de recherche PAQUID sur l'épidémiologie de la démence méthodes et résultats initiaux. *Rev Neurol*, 147(3):225–30, 1991.

[21] C Dean, JF Lawless, and GE Willmot. A mixed Poisson-Inverse Gaussian regression model. *Canad J Statist*, 17:171–81, 1989.

[22] CB Dean. Testing for overdispersion in Poisson and Binomial regression models. *JASA*, 87(418):451–7, 1992.

[23] G Debry. Les besoins nutritionnels des personnes âgées. *Néphrologie*, 11:307–11, 1990.

[24] AJ Dzik. Cerebrovascular disease mortality rates and water hardness in North Dakota. *S D J Med*, 42(4):5–7, 1989.

[25] LE Eberly, JD Cohen, R Prineas, and L Yang L for the Intervention Trial Research group. Impact of incident diabetes and incident nonfatal cardiovascular disease on 18-year mortality: the multiple risk factor intervention trial experience. *Diabetes Care*, 26(3):848–54, 2003.

[26] B Efron and RJ Tibshirani. An introduction to the bootstrap. *Monographs on Statistics and Applied Probability*, 1993.

[27] AR Folsom and RJ Prineas. Drinking water composition and blood pressure: a review of the epidemiology. *Am J Epidemiol*, 115(6):818–32, 1982.

[28] EL Frome. Poisson regression analysis. *American Statistician*, 35:262–3, 1981.

[29] EL Frome. The analysis of rates using Poisson regression models. *Biometrics*, 39:665–74, 1983.

[30] EL Frome and H Checkoway. Use of Poisson regression models in estimating incidence rates and ratios. *Am J Epidemiol*, 121(2):309–23, 1985.

[31] P Galan, P Preziosi, V Durlach, P Valeix, L Ribas, and D Bouzid et al. Dietary magnesium intake in a french adult population. *Magnes Res*, 10:321–8, 1997.

[32] S Greenland. Divergent biases in ecologic and individual-level studies. *Stat Med*, 11(9):1209–23, 1992.

[33] S Greenland and J Robins. Invited commentary: ecologic studies–biases misconceptions and counterexamples. *Am J Epidemiol*, 139(8):747–60, 1994.

[34] LE Griffith, GH Guyatt, RJ Cook, HC Bucher, and DJ Cook. The influence of dietary and nondietary calcium supplementation on blood pressure. *Am J Hypertens*, 12:84–92, 1999.

[35] K Gu, CC Cowie, and Harris MI. Diabetes and decline in heart disease mortality in US adults. *JAMA*, 281(14):1291–7, 1999.

[36] KA Guthrie and L Sheppard. Overcoming biases and misconceptions in ecological studies. *J R Statist Soc A*, 164(1):141–54, 2001.

[37] KA Guthrie, L Sheppard, and J Wakefield. A hierarchical aggregate data model with spatially correlated diseases rates. *Biometrics*, 58(4), 2002.

[38] GM Halpern, J Van de Water, AM Delabroise, CL Keen, and ME Gershwin. Comparative uptake of calcium from milk and a calcium-rich mineral water in lactose intolerant adults: implications for treatment of osteoporosis. *Am J Prev Med*, 7(6):379–83, 1991.

[39] BS Haring and W Van Delft. Changes in mineral composition of food as a result of cooking in hard and soft waters. *Arch Environ Health*, 36(1):33–5, 1981.

[40] R Hashemi and D Commenges. Correction of p-values after multiple tests in Cox regression. *Lifetime Data Anal*, 8(4):335–48, 2002.

[41] WK Hasting. Monte carlo sampling methods using markov chains and their applications. *Biometrica*, pages 97–109, 1970.

[42] RP Heaney, RR Recker, MR Stegman, and AJ Moy. Calcium absorption in women: relationship to calcium intake estrogen status and age. *J Bone Miner Res*, 4(4):469–75, 1989.

[43] C Helmer, P Baberger-Gateau, L Letenneur, and JF Dartigues. Subjective health and mortality in french elderly women and men. *J Gerontol Soc Sci*, 54B(2):S84–92, 1999.

[44] C Heyde. *Quasi-Likelihood and its applications*. Springer Verlag, 1997.

[45] DW Hosmer and S Lemeshow. *Applied Logistic Regression*. John Wiley & Sons, Inc., 1989.

[46] P Houggaard, ML Lee, and GA Whitmore. Analysis of overdispersed count data by mixtures of poisson variables and poisson processes. *Biometrics*, 53(4):1225–38, 1997.

[47] K Itoh, T Kawasaki, and M Nakamura. The effects of high oral magnesium supplementation on blood pressure, serum lipids and related variables in apparently healthy japanese subjects. *B J Nutr*, 78:737–50, 1997.

[48] B Jorgensen. Exponential dispersion models. *J R Stat Soc B*, 49:127–62, 1987.

[49] B Jorgensen. *The theory of dispersion models*. Chapman & Hall London, 1997.

[50] A Jourdan and CC Kokonendji. Surdispersion et modèle binomial négatif généralisé. *Rev Statistique Appliquée*, L(3):73–86, 2002.

[51] WB Kannel, PW Wilson, BH Nam, and RB D'Agostino RB. Risk stratification of obesity as a coronary risk factor. *Am J Cardiol*, 90(7):697–701, 2002.

[52] AL Klatsky, MA Armstrong, and GD Friedman. Alcohol and mortality. *Ann Intern Med*, 117(8):646–54, 1992.

[53] CC Kokonendji and M Koudhar. On strict arcsine distribution. *Communication in Statistics Theory and Methods*, 2003. Submitted.

[54] SP Laing, AJ Swerdlow, LM Carpenter, SD Slater, AD Burden, JL Botha, AD Morris, NR Waugh, W Gatling, EA Gale, CC Patterson, Z Qiao, and H Keen. Mortality from cerebrovascular disease in a cohort of 23 000 patients with insulin-treated diabetes. *Stroke*, 34(2):418–21, 2003.

[55] IH Langford, AH Leyland, J Rasbash, and H Goldstein. Multilevel modelling of the geographical distributions of diseases. *Appl Statist*, 48(2):253–68, 1999.

[56] JF Lawless. Negative binomial and mixed Poisson regression. *Canad J Statist*, 15(3):209–25, 1987.

[57] AB Lawson and FLR Williams. Spatial competing risk models in disease mapping. *Stat Med*, 19:2451–67, 2000.

[58] EM Van Leer, JC Seidell, and D Kromhout. Dietary calcium potassium magnesium and blood pressure in the Netherlands. *Int J Epidemiol*, 24(6):1117–23, 1995.

[59] BG Leroux. Modelling spatial disease rates using maximum likelihood. *Stat Med*, 19:2321–32, 2000.

[60] G Letac and M Mora. Natural real exponential families with cubic variance functions. *Ann. Statist.*, 18:1–37, 1990.

[61] S Lewington, R Clarke, N Qizilbash, R Peto, and R Collins; Prospective Studies Collaboration. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *Lancet*, 360(9349):1903–13, 2002.

[62] AH Leyland, IH Langford, J Rasbash, and H Goldstein. Multivariate spatial models for event data. *Stat Med*, 19:2469–78, 2000.

[63] KY Liang and SL Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

[64] KM Liaw and CJ Chen. Mortality attributable to cigarette smoking in Taiwan: a 12-year follow-up study. *Tob Control*, 7(2):141–8, 1998.

[65] WS Lu and RK Tsutakawa. Analysis of mortality rates via marginal extended quasi-likelihood. *Stat Med*, 15(13):1397–407, 1996.

[66] S Malyutina, M Bobak, S Kurilovitch, V Gafarov, G Simonova, Y Nikitin, and M Marmot. Relation between heavy and binge drinking and all-cause and cardiovascular mortality in Novosibirsk Russia: a prospective cohort study. *Lancet*, 360(9344):1448–54, 2002.

[67] S Marque. La mortalité cardiovasculaire chez les personnes: description et perspectives. Master's thesis, DEA Analyse Démographique - Université Bordeaux IV, 2000.

[68] S Marque, H Jacqmin-Gadda, J Dartigues, and D Commenges. Cardiovascular mortality and calcium and magnesium in drinking water: an ecological study in elderly people. *Eur J Epidemiol*, 18(4):305–9, 2003.

[69] S Marque and CC Kokonendji. A strict arcsine regression. *Communication in Statistics: Theory and Methods*, 2003. (Submitted).

[70] DA McCarron. Calcium metabolism and hypertension. *Kidney Int*, 35:717–36, 1989.

[71] DA McCarron. *Calcium-regulating Hormones; Roles in disease and aging*, chapter Epidemiological evidence and clinical trials of dietary calcium's effect on blood pressure. Contrib Nephrol Basel, Karger, 1991.

[72] P McCullagh and JA Nelder. *Generalized Linear Models*. Chapman & Hall London, 1993.

[73] CE McCulloch. Maximum likelihood alogrithm for generalized linear mixed models. *JASA*, 92:162–70, 1997.

[74] DF Moore and A Tsiatis. Robust estimation of the variance in moment methods for extra-binomial and extra-poisson variation. *Biometrics*, 47:383–401, 1991.

[75] H Morgenstern. Ecologic studies in epidemiology: concepts principles and methods. *Annu Rev Public Health*, 16:61–81, 1995.

[76] H Morgenstern and D Thomas. Principles of study design in environmental epidemiology. *Environ Health Perspect*, 101(S4):23–38, 1993.

[77] KJ Mukamal, KM Conigrave, MA Mittleman, CA Camargo, MJ Stampfer, WC Willett, and EB Rimm. Roles of drinking pattern and type of alcohol consumed in coronary heart disease in men. *N Engl J Med*, 348(2):109–18, 2003.

[78] F Najeme. *Thèse de médecine, Université Bordeaux2*, 2002.

[79] JA Nelder and RWM Wedderburn. Generalized linear models. *J R Stat Soc B*, 54(1):3–40, 1972.

[80] R Peto, AD Lopez, J Boreham, M Thun, and C Heath. Mortality from tobacco in developed countries: indirect estimation from national vital statistics. *Lancet*, 339(8804):1268–78, 1992.

[81] M Pollan and P Gustavsson. High-risk occupations for breast cancer in the swedish female working population. *Am J Public Health*, 89(6):875–81, 1999.

[82] RL Prentice and L Sheppard. Aggregate data studies of disease risk factors. *Biometrika*, 82(1):113–25, 1995.

[83] RL Prentice and LP Zhao. Estimating equations for parameters in mean and covariances of multivariate discrete and continuous responses. *Biometrics*, 47:825–39, 1991.

[84] E Prescott, H Scharling, M Osler, and P Schnohr. Importance of light smoking and inhalation habits on risk of myocardial infarction and all cause mortality. A 22 years follow up of 12 149 men and women in the Copenhagen city heart study. *J Epidemiol Community Health*, 56(9):300–6, 2002.

[85] A Reunanen, P Knekt, J Marniemo, J Maki, J Maatela, and A Aromaa. Serum calcium magnesium copper and zinc and risk of cardiovascular death. *Eur J Clin Nutr*, 50:431–7, 1996.

[86] S Richardson. A method for testing the significance of geographical correlations with application to industrial lung cancer in france. *Stat Med*, 9(5):515–28, 1990.

[87] S Richardson, C Guihenneuc-Jouyaux, and V Lasserre. Ecologic studies–biases misconceptions and counterexamples. *Am J Epidemiol*, 143(5):522–3, 1996.

[88] S Richardson, C Monfort, M Green, G Draper, and C Muirhead. Spatial variation of natural radiation and childhood leukaemia incidence in Great Britain. *Stat Med*, 14(21-22):2487–501, 1995.

[89] RG Rogers, RA Hummer, and PM Krueger. The effect of obesity on overall circulatory disease- and diabetes-specific mortality. *J Biosoc Sci*, 35(1):107–29, 2003.

[90] E Rubenowitz, G Axelsson, and R Rylander. Magnesium in drinking water and death from acute myocardial infarction. *Am J Epidemiol*, 143(5):456–62, 1996.

[91] E Rubenowitz, G Axelsson, and R Rylander. Magnesium and calcium in drinking water and death from acute myocardial infarction in women. *Epidemiology*, 10(1):31–6, 1999.

[92] R Rylander, H Bonevik, and E Rubenowitz. Magnesium and calcium in drinking water and cardiovascular mortality. *Scand J Work Environ Health*, 17(2):91–4, 1991.

[93] N Sakamoto, M Shimizu, I Wakabayashi, and K Sakamoto. Relationship between mortality rate of stomach cancer and cerebrovascular disease and concentrations of magnesium and calcium in well water in Hyogo prefecture. *Magnes Res*, 10(3):215–23, 1997.

[94] WH Sauer, JA Berlin, BL Strom, C Miles, JL Carson, and SE Kimmel. Cigarette yield and the risk of myocardial infarction in smokers. *Arch Intern Med*, 162(3):300–6, 2002.

[95] HA Schroeder. Relation between mortality from cardiovascular disease and treated water supplies. *JAMA*, 172(17):1902–8, 1960.

[96] L Sheppard and RL Prentice. On the reliability and precision of within- and between-population estimates of relative rate parameters. *Biometrics*, 51:853–63, 1995.

[97] L Sheppard, RL Prentice, and MA Rossing. Design considerations for estimation of exposure effects on disease risk using aggregate data studies. *Stat Med*, 15:1849–1858, 1996.

[98] CB Sherman. Health effects of cigarette smoking. *Clin Chest Med*, 12(4):643–58, 1991.

[99] RB Singh. effects of dieatry magnesium supplementation on prevention of coronary heart disease and sudden cardiac death. *Magnesium Trace Elem*, 9:143–51, 1990.

[100] RB Singh, SS Rastogi, S Ghosh, and MA Niaz. Dietary and serum magnesium levels in patients with acute myocardial infarction, coronary artery disease and noncardiac diag-

noses. *J Am College Nutr*, 13(2):139–43, 1994.

[101] PJ Smith and DF Heitjan. Testing and adjusting for departures from nominal dispersion in generalized linear models. *Appl Statist*, 42:31–41, 1993.

[102] D Spiegelhalter, N Best, and N Carlin. Bayesian deviance, the effective number of parameters and the comparison of arbitrarily complex models. *Research Report 98-0009(unpublished)*, 1998. Division of Biostatistics, Uni. of Minnesota.

[103] J Stevens, J Cai, KR Evenson, and R Thomas. Fitness and fatness as predictors of mortality from all causes and from cardiovascular disease in men and women in the lipid research clinics study. *Am J Epidemiol*, 156(9):832–41, 2002.

[104] J Stevens, J Cai, ER Pamuk, DF Williamson, MJ Thun, and JL Wood. The effect of age on the association between body-mass index and mortality. *N Engl J Med*, 338(1):1–7, 1998.

[105] AH Stroud. *Gaussian Quadrature Formulas*. Prentice-Hall, Englewood Cliffs, 1966.

[106] MJ Thun, R Peto, AD Lopez, JH Monaco, SJ Henley, and CW et al Heath. Alcohol consumption and mortality among middle-aged and elderly U.S. adults. *N Engl J Med*, 337(24):1705–14, 1997.

[107] BW Turnbull and L Weiss. A likelihood ratio statistic for testing goodness of fit with randomly censored data. *Biometrics*, 34(3):367–75, 1978.

[108] LPL Van Der Vijver, MAE Van Der Waal, KGC Weterings, JM Dekker, EG Schouten, and FJ Kok. Calcium intake and 28-year cardiovascular and coronary heart disease mortality in dutch civil servants. *Int J Epidemiol*, 21(1):36–9, 1992.

[109] J Wakefield. Sensitivity analyses for ecological regression. *(to appear)*, 2003.

[110] J Wakefield, NG Best, and L Waller. *Spatial Epidemiology*, chapter Bayesian approach to disease mapping, pages 104–27. Oxford University Press, 2000.

[111] RWM Wedderburn. Quasi-likelihood functions generalized linear models and the gauss-newton method. *Biometrika*, 61:439–47, 1974.

[112] GE Willmot. The poisson inverse gaussian distribution as an alternative to the negative-binomial. *Scand Actuar J*, pages 113–27, 1987.

[113] R Wolfinger and M O'Connell. Generalized linear mixed models: a pseudo-likelihood approach. *J Statist Comput Simul*, 48:233–43, 1993.

[114] RJ Wood, JC Fleet, K Cashman, ME Bruns, and HF Deluca. Intestinal calcium absorption in the aged rat: evidence of intestinal resistance to 1,25(OH)2 vitamin D. *Endocrinology*, 139(9):3843–8, 1998.

[115] CY Yang. Calcium and magnesium in drinking water and risk of death from cerebrovascular disease. *Stroke*, 29:411–4, 1998.

[116] CY Yang and HF Chiu. Calcium and magnesium in drinking water and risk of death from hypertension. *Am J Hypertens*, 12:894–9, 1999.

[117] CY Yang, JF Chiu, HF Chiu, TN Wang, CH Lee, and YC Ko. Relationship between water hardness and coronary mortality in Taiwan. *J Toxicol Environ Health*, 49(1):1–9, 1996.

Annexe A

Cartes des communes de Paquid

Dordogne

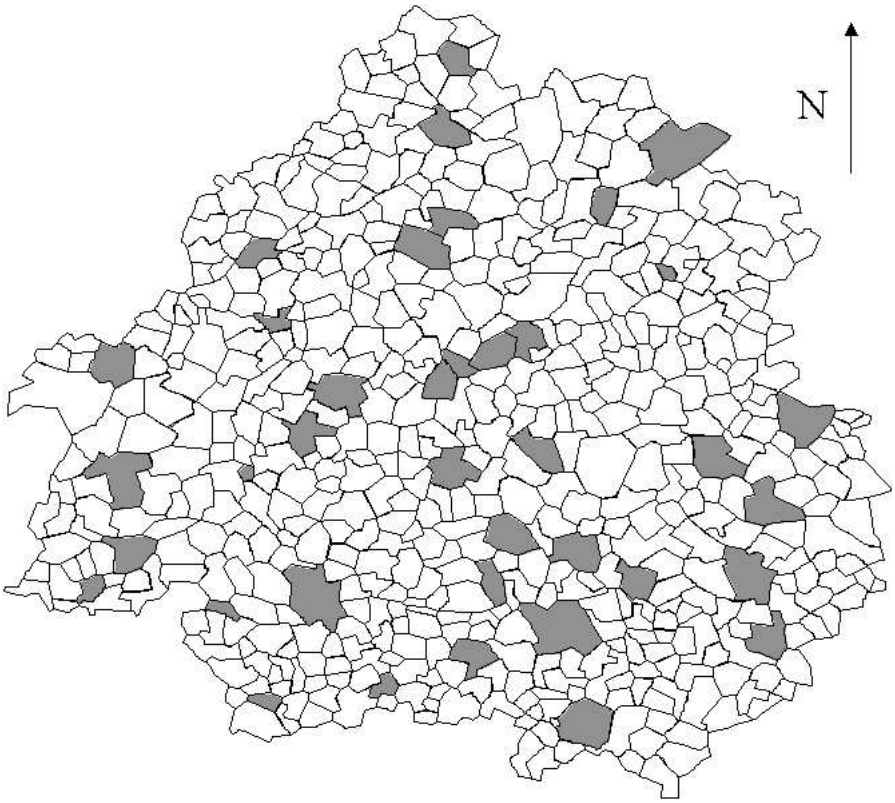

FIG. A.1 – *Les 37 communes de Dordogne sélectionnées dans PAQUID*

Gironde

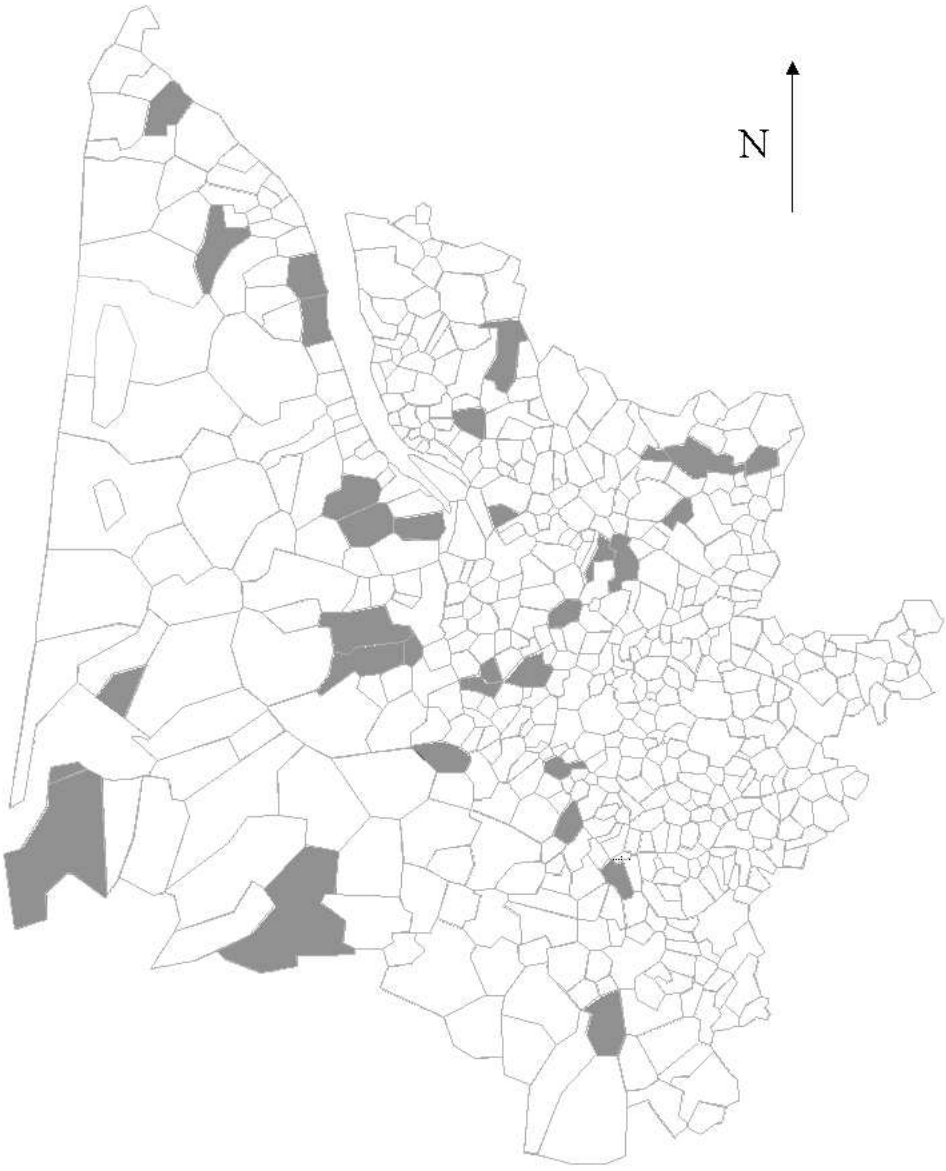

FIG. A.2 – *Les 38 communes de Gironde sélectionnées dans PAQUID*

Annexe B

Article paru dans EJE

Cardiovascular mortality and calcium and magnesium in drinking water: an ecological study in elderly people

Sébastien MARQUE[1], Hélène JACQMIN-GADDA[1], Jean-Francois DARTIGUES[1]

et Daniel COMMENGES[1]

[1]INSERM U330, 146 rue léeo Saignat, 33076 Bordeaux Cedex, France

Tel: +33(0)557571136

Fax: +33(0)556240081

Sebastien.Marque@isped.u-bordeaux2.fr

14th November 2002

Summary

Background

Previous studies found relations between cardiovascular mortality and minerals in drinking water, but the major works considered water hardness or neglected the differences between adults and elderly. Drinking water is an important source of calcium in the elderly particularly because of increased needs and decreased consumption of dairy products.

Methods

We collected informations about all deaths (14311) occuring in 69 parishes of the South-West of France during seven years (1990-1996). We obtained the causes of deaths from a special service of INSERM for each death, with age at death and sex. The exposure value was supplied by administrative source (DDASS) and by measurement surveys. We use an extra-Poisson variation model to take into account the heterogeneity of the population of these parishes.

Results

A significant relationship was observed between calcium and cardiovascular mortality with a RR=0.90 for non-cerebrovascular causes and RR=0.86 for cerebrovascular (when calcium is higher than the second tercile: 94 mg/l). We found a protective effect of magnesium concentrations between 4 and 11 mg/l with a RR=0.92 for non-cerebrovascular and RR=0.77 for cerebrovascular mortality, as compared to concentrations lower than 4 mg/l.

Conclusions

These findings strongly suggest a potential protective dose-effect relation between calcium in drinking water and cardiovascular causes. For magnesium, a U-shape effect is possible, especially for cerebrovascular mortality.

Key word: Calcium, Magnesium, Drinking Water, Cardiovascular Mortality, Elderly, Poisson Regression.

Introduction

Cardiovascular mortality represents the main cause of mortality in people over 65 years of age, with an incidence of 1.9 per 1000 person-years among men and 1.3 per 1000 among women [1] in 1990 in France. Previous epidemiologic studies have suggested an inverse relationship between the number of deaths by cardiovascular pathologies and water hardness [2], mainly determined by concentrations of calcium and magnesium.

Calcium and magnesium are antagonistic enzymatic activators. Calcium is essential to coagulation, nerve impulse and muscular contraction, in particular for the cardiac muscle. Magnesium is involved in the transfer and the release of energy and takes part in cardiac physiology. The results of epidemiologic studies have suggested a potential hypotensive action for calcium and an effect against thrombosis for magnesium [3–5].

In developed countries, dairy products represent the most significant source of calcium. However, it is recognized that the intestinal absorption of calcium decreases with aging [6], and particularly among women [7]. Moreover, elderly are sometimes confronted with a problem of lactose intolerance; for these subjects drinking water becomes the major source of calcium [8, 9]. The daily intake of magnesium in the developed countries is insufficient compared to recommended values [10]. Moreover, since the magnesium in drinking water is in ionized form, it might be more bioavailable than that provided by solid foods [11, 12]. This may also be the case for calcium which seems to be at least as bioavailable in water as in milk [13].

The various studies about the role of calcium and magnesium have often shown a protective effect of these two elements on cardiovascular mortality. However these works are essentially based on studies of correlations between the concentrations of these minerals in water and the death rates in the various geographical areas [14, 15]. Moreover, some of these studies have considered the hardness of water as a global factor rather than the separate effect of calcium and magnesium [16].

Recently, Rylander et al [17] found that deaths by ischaemic cardiopathies were inversely associated with the concentrations of calcium and magnesium in drinking water. According to the same authors, calcium may have a protective effect on cerebrovascular mortality in men. However, this relation was not found by Yang et al [18, 19] who suggested that magnesium in drinking water may have a protective effect. This protective effect of magnesium was also found in other studies on cardiovascular mortality [20]. Nevertheless, some doubts persist regarding these possible relationships because of contradictory [21] or non-significant [22, 23] results. Finally, the relationship between cardiovascular mortality and the mineral elements of drinking water has rarely been studied in elderly [24].

The major aim of this study was to assess the relationship between cardiovascular mortality in the elderly and the concentration of calcium and magnesium in drinking water. In effect, the few number of studies among elderly encourage us to investigate more precisely this relationship [6, 7]. The study was based on the population of the 75 parishes from which the PAQUID sample was drawn. We have also studied separately cerebrovascular and other cardiovascular mortality. The water distribution network had already been studied for these parishes and mean concentrations of calcium and magnesium were already available. Deaths certificates were collected for a 7-years period and the analysis beared on 14,331 deaths and 777,493 persons-years.

Methods

Populations

Our sample was composed by the population living in the 75 parishes included in PAQUID, which is a prospective cohort in the South-West of France. These parishes were drawn at random among all parishes of Gironde and Dordogne according to a design in 5 strata based on the size of the parish. We

considered the population between 01/01/1990 and 31/12/1996, aged over 65 years and we collected all deaths occuring in this population. The number of person-years was calculated using the average of the population between the two dates, for sex and age (5-years classes). We obtained the distribution of the population of the parishes by age and sex from the "Institut National de la Statistique et des Etudes Economiques" (INSEE). The causes of deaths were provided by the SC8 (a specialized department of the "Institut Nationale de la Santé Et de la Recherche Médicale") from the death certificates. For each death, we recorded sex, age at death (one-year precision), parish of residence and main cause of death (coding on 4 numbers accorded to ICD-9). We observed 14,311 deaths and the total number of person-years was 777,493.

Measure of exposure

We divided each parish into distribution zones, supplied by one particular water source, on the basis of information given by the sanitary administration. Two measurement surveys were carried out in 1991 to measure pH and concentrations of calcium and magnesium in each zone. For each water supply, the mean of the determinations for the two surveys and of the routine measurements, collected by the sanitary administration between 1991 and 1994, were computed. For each distribution zone, we computed average values for calcium and magnesium concentration using the hourly flow or the relative contribution of each water source in the composition of drinking water in the zone. The details of the measures of the concentrations of the elements were described in previous publications [25]. Six parishes were excluded from the study because of recent changes in the distribution network. Finally, our sample considered 69 parishes for which data were available.

Statistical Analysis

We used a model of extra-Poisson variation regression in order to study the relationship between cardiovascular mortality and calcium and magnesium on the grouped data: the variable to be explained was the number of cardiovascular deaths in the parishes [26]. The heterogeneity within the parishes was

taken into account by the introduction of an over-dispersion parameter. Indeed, this heterogeneity could introduce an underestimation of the variance and thus lead to anti-conservative tests. The extra-Poisson variation model was performed using the GENMOD procedure of the SAS software with the 'dscale' option. This parameter was used in order to correct the estimators of variances without modify estimation of the parameters themselves. So, the scale option changes only the confidence interval limits. All these models were adjusted on the sex and age (by 5-years classes) of the subjects and on the rurality of the parishes.

Calcium and magnesium concentrations were regarded categorized in three classes in order to allow for a possible "U-shape" or for a dose-effect relation. The cut-off points were fixed at the tertiles values. The rural status of a parish was based on the five categories of the INSEE definition and we re-coded this variable into a dichotomic variable taking urban status as the reference class.

Results

When we considered global cardiovascular mortality (see table1), we observed a protective effect of higher calcium concentration (RR=0.90 - Confidence Interval of 95% CI95 [0.84 ; 0.96] - for concentration greater than 94 mg/l). We also found a protective effect of magnesium between 4 and 11 mg/l with estimated relative risk between 0.88 and 0.92 (respectively among women and men with , CI95 [0.81 ; 0.96] and , CI95 [0.84 ; 1.00]).

We then analysed separetely non-cerebrovascular (cardiovascular) mortality (10492 deaths) and cerebrovascular mortality (3819 deaths). Table 2 shows the results of the extra-variation Poisson models for cardiovascular mortality, excluding cerebrovascular origin. These models suggested a protective effect of a high concentration of calcium on cardiovascular mortality with a dose-response effect. This effect seemed

to be similar in both sexes, with a significant relative risk in the parishes having a water with a content higher than 94 mg/l (RR=0.90, CI95 [0.84 ; 0.97]). Concentration of magnesium between 4 and 11 mg/l seemed to be slightly protective on cardiovascular mortality (RR=0.92, CI95 [0.86 ; 0.99]).

Table 3 shows the results of the extra-variation Poisson models for cerebrovascular mortality. About calcium, we found roughly similar results on cerebrovascular mortality than on cardiovascular mortality. We observed a protective relationship for concentration greater than 94 mg/l. This effect seemed slightly clearer among women (RR=0.84 vs 0.89, with CI95 respectively [0.74 ; 0.97] and [0.74 ; 1.07]). However, we observed a more significant protective effect for magnesium for concentration between 4 and 11 mg/l. This effect seemed similar on both sexes and highly significant (RR=0.77, $p \leq 0.001$).

The rural status of parishes did not seem to be associated with cardiovascular mortality, and we found this result in all models we performed.

Discussion

These findings suggest a protective effect of calcium in drinking water with a dose-effect relationship. This effect is similar in both sexes with a RR=0.90 in cardiovascular mortality and RR=0.86 in cerebrovascular mortality.

Magnesium seems to be protective on cardiovascular mortality for concentrations between 4 and 11 mg/l (RR=0.90). The effect of magnesium seems to be higher on cerebrovascular mortality (RR=0.77).

Considering two different causes in CV mortality among the elderly was the major originality of this analysis. Few analysis were performed among this population and all of them considered only cardiovascular cause, without distinction between the sub-causes. Furthermore, we could estimate independently the

effect of Calcium and Magnesium on CVD, which has never been investigated among the elderly.

A weakness of our study may come from the exposure measurement. We measured concentration of these two minerals on several public taps in each parish to compare the reliability. This reliability of the measurement of calcium concentration is greater than 90%[1] [25]. Moreover, we used the average of several values to reduce the variability of these mineral concentrations. However the concentrations of the minerals were collected only at community level: we supposed that all persons living in a given parish had the same exposure, while in fact concentrations could be different between households. There are also other sources of calcium, e.g. nutrition and medication. Indeed, the main calcium intake comes from food, although its bioavailability may be lower than in drinking water. This hypothesis was made in a recent study by Bostick et al [27] who found a significant relation between ischemic mortality and calcium, but not with milk products (or vitamin D). Furthermore, the mineral composition of food (especially potatoes and vegetables) can be modified when they are cooked in water, and these changes depend on water hardness [28]. The concentration of minerals in the drinking water of a parish may be a good indicator of exposure to these minerals for subjects who live a long time in these parishes. At the beginning of the Paquid study, the average stay of subject in the same parish was about 40 years.

Causes of deaths were collected from death certificates, which are completed by general practitioners. They are certainly some misclassifications but it is very likely that misclassification did not depend on calcium concentration. Such non differential misclassification can only make the RRs closer to one.

Calcium and magnesium were introduced as categorized variables with three levels given by the tertiles. This had the advantage of providing a certain flexibility (possibility of observing U-shape or dose effect relation) and avoiding multiplicity of tests (as compared to trying several different codings or using a variable with a larger number of categories).

We also adjusted on rural place of residence which may be considered as a good indicator of several confounders such as dietary habits (wine or water consumption for exemple) or economical level of

[1] unpublished technical report INSERM U330, April 1993

household. Furthermore, rural place of residence might be a confounding factor, if both classification error and cardiovascular mortality depend on it ; However this variable was not significant.

As in the review by Sharett [29], our work supports the hypothesis of a protective effect of calcium in drinking water. Sharett reported the results of many epidemiologic studies of the 70's in Great-Britain which found strong correlations between calcium and cardiovascular mortality, but without any effect of magnesium. A possible protective effect of calcium on cardiovascular diseases has already been described in nutritional studies: diets rich in calcium and magnesium are associated with lower blood pressure [4], especially among women.

In conclusion, the present study has found a dose-effect relationship between calcium contained in drinking water and cardiovascular mortality and a U-shape effect of magnesium more pronounced for cerebrovascular mortality. These findings meet several criteria that suggest an interpretation in terms of causal effect. The statistical tests are highly significant ; there is a consistency between men and women ; people have been exposed to drinking water before developing the cardiovascular disease and dying, and there is no doubt about the direction of causal relationship if it exists: we cannot imagine that the calcium concentration is influenced by the rate of cardiovascular mortality. There is a biological plausibility and a biological gradient (monotone dose-effect relationship) for calcium. In view of this, we think that more precise studies should now be undertaken, in particular cohort or intervention studies. Cohort design would permit us to control the other sources of calcium intake and to investigate more precisely the cardiovascular causes of death. An intervention studyof Calcium supplementation would have the advantage of randomization, leading to more secure interpretation in term of causality.

References

1. Organisation Mondiale de la Santé. Epidémiologie et prévention des maladies cardio-vasculaires chez les personnes âgées. *Série de rapports techniques*, 1995;853.

2. Comstock GW. Water hardness and cardiovascular diseases. *Am J Epidemiol*, 1979;110(4):375–87.

3. McCarron DA. Calcium metabolism and hypertension. *Kidney Int*, 1989;35:717–36.

4. Van Leer E, Seidell J, and Krohout D. Dietary calcium, potassium magnesium and blood pressure in the Netherlands. *Int J Epidemiol*, 1995;24:1117–23.

5. Griffith LE, Guyatt GH, Cook RJ, Bucher HC, and Cook DJ. The influence of dietary and nondietary calcium supplementation on blood pressure. *Am J Hypertens*, 1999;12:84–92.

6. Wood RJ, Fleet JC, Cashman K, Bruns ME, and Deluca HF. Intestinal calcium absorption in the aged rat: evidence of intestinal resistance to 1,25(OH)2 vitamin D. *Endocrinology*, 1998;139(9):3843–8.

7. Heaney RP, Recker RR, Stegman MR, and Moy AJ. Calcium absorption in women: relationship to calcium intake, estrogen status and age. *J Bone Miner Res*, 1989;4(4):469–75.

8. Halpern GM, Van de Water J, Delabroise AM, Keen CL, and Gershwin ME. Comparative uptake of calcium from milk and a calcium-rich mineral water in lactose intolerant adults: implications for treatment of osteoporosis. *Am J Prev Med*, 1991;7(6):379–83.

9. Aptel I, Cance-rouzaud A, and Grandjean H. Association between calcium ingested from drinking water and femoral bone density in elderly women: evidence from the EPIDOS cohort. *J Bone Miner Res*, 1999;14(5):829–33.

10. Galan P, Preziosi P, Durlach V, Valeix P, Ribas L, and Bouzid D et al. Dietary magnesium intake in a french adult population. *Magnes Res*, 1997;10:321–8.

11. Van Dokkum W, De la Gueronniere V, Schaafsma G, Bouley C, Luten J, and Latge C. Bioavailability of calcium of fresh cheeses, enteral food and mineral water. A study with stable calcium isotopes in young adult women. *Br J Nutr*, 1996;75(6):893–903.

12. Heaney RP and Dowell MS. Absorbability of the calcium in a high-calcium mineral water. *Osteoporos Int*, 1994;4(6):323–4.

13. Couzy F, Kastenmayer P, Vigo M, Clough J, Munoz-Box R, and Barclay DV. Calcium bioavailability from calcium- and sulfate-rich mineral water, compared with milk, in young adult women. *Am J Clin Nutr*, 1995;62(6):1239–44.

14. Schroeder HA. Relation between mortality from cardiovascular disease and treated water supplies. *JAMA*, 1960;172(17):1902–8.

15. Sakamoto N, Shimizu M, Wakabayashi I, and Sakamoto K. Relationship between mortality rate of stomach cancer and cerebrovascular disease and concentrations of magnesium and calcium in well water in Hyogo prefecture. *Magnes Res*, 1997;10(3):215–23.

16. Punsar S and Karvonen MJ. Drinking water quality and sudden death: observations from West and East Finland. *Cardiology*, 1979;64:24–34.

17. Rylander R, Bonevik H, and Rubenowitz E. Magnesium and calcium in drinking water and cardiovascular mortality. *Scand J Work Environ Health*, 1991;17(2):91–4.

18. Yang CY. Calcium and magnesium in drinking water and risk of death from cerebrovascular disease. *Stroke*, 1998;29:411–4.

19. Yang CY and Chiu HF. Calcium and magnesium in drinking water and risk of death from hypertension. *Am J Hypertens*, 1999;12:894–9.

20. Purvis JR and Movahed A. Magnesium disorders and cardiovascular diseases. *Clin Cardiol*, 1992;15:556–68.

21. Rubenowitz E, Axelsson G, and Rylander R. Magnesium and calcium in drinking water and death from acute myocardial infarction in women. *Epidemiology*, 1999;10(1):31–6.

22. Reunanen A, Knekt P, Marniemo J, Maki J, Maatela J, and Aromaa A. Serum calcium, magnesium, copper and zinc and risk of cardiovascular death. *Eur J Clin Nutr*, 1996;50:431–7.

23. Maheswaran R, Morris S, Falconer S, Grossinho A, Perry I, Wakefield J, and Elliott P. Magnesium in drinking water supplies and mortality from acute myocardial infarction in North West England. *Heart*, 1999;82(4):455–60.

24. Folsom AR and Prineas RJ. Drinking water composition and blood pressure: a review of the epidemiology. *Am J Epidemiol*, 1982;115(6):818–32.

25. Jacqmin H, Commenges D, Letenneur L, Baberger-Gateau P, and Dartigues JF. Components of drinking water and risk of cognitive impairment in the elderly. *Am J Epidemiol*, 1994;139(1):48–57.

26. Fay MP and Feuer EJ. A semi-parametric estimate of extra-poisson variation for vital rates. *Stat Med*, 1997;16:2389–401.

27. Bostick RM, Kushi LH, Wu Y, Meyer KA, Sellers TA, and Folsom AR. Relation of calcium, vitamin D, and dairy food intake to ischemic heart disease mortality among postmenopausal women. *Am J Epidemiol*, 1999;149(2):151–61.

28. Haring BS and W Van Delft. Changes in mineral composition of food as a result of cooking in "hard" and "soft" waters. *Arch Environ Health*, 1981;36(1):33–5.

29. Sharett A. The role of chemical constituent of drinking water in cardiovascular diseases. *Am J Epidemiol*, 1979;93:256–66.

Table 1: *Poisson models on global cardiovascular mortality, adjusted on age (France, 1990-1997)*

	Men		Women		Both sexes	
	RR	95% CI	RR	95% CI	RR	95% CI
Calcium (mg/l)						
[9; 53[	1.00	-	1.00	-	1.00	-
[53; 94[	0.96	0.89 ; 1.05	0.95	0.88 ; 1.03	0.95	0.90 ; 1.01
[94;146]	0.91*	0.83 ; 0.99	0.90†	0.83 ; 0.97	0.90†	0.84 ; 0.96
Magnesium (mg/l)						
[1; 4[	1.00	-	1.00	-	1.00	-
[4;11[	0.92*	0.84 ; 1.00	0.88†	0.81 ; 0.96	0.90†	0.85 ; 0.96
[11;34]	0.98	0.88 ; 1.10	0.89	0.80 ; 0.99	0.93	0.86 ; 1.01
Urban §	1.00	0.99 ; 1.03	1.02	0.99 ; 1.05	1.01	0.99 ; 1.04
Scale §§		1.00		1.14		1.15

* $p \leq 0.05$
† $p \leq 0.01$
‡ $p \leq 0.001$
§ vs rural
§§ allows overdispersion relatively to a regular Poisson model

Table 2: *Poisson models on cardiovascular but non-cerebrovascular mortality, adjusted on age (France, 1990-1997)*

	Men		Women		Both sexes	
	RR	95% CI	RR	95% CI	RR	95% CI
Calcium (mg/l)						
[9; 53[	1.00	-	1.00	-	1.00	-
[53; 94[	0.94	0.86 ; 1.03	0.96	0.88 ; 1.05	0.95	0.88 ; 1.01
[94;146]	0.90*	0.81 ; 0.99	0.91*	0.83 ; 0.99	0.90†	0.84 ; 0.97
Magnesium (mg/l)						
[1; 4[	1.00	-	1.00	-	1.00	-
[4;11[	0.94	0.85 ; 1.04	0.90*	0.82 ; 0.99	0.92*	0.86 ; 0.99
[11;34]	1.02	0.90 ; 1.15	0.90	0.79 ; 1.02	0.96	0.87 ; 1.05
Urban §	1.02	0.99 ; 1.05	1.02	0.99 ; 1.06	1.02	1.00 ; 1.05
Scale §§		0.98		1.10		1.11

* $p \leq 0.05$
† $p \leq 0.01$
‡ $p \leq 0.001$
§ vs rural
§§ allows overdispersion relatively to a regular Poisson model

Table 3: *Poisson models on cerebrovascular mortality, adjusted on age (France, 1990-1997)*

	Men		Women		Both sexes	
	RR	95% CI	RR	95% CI	RR	95% CI
Calcium (mg/l)						
[9; 53[	1.00	-	1.00	-	1.00	-
[53; 94[	0.94	0.79 ; 1.11	0.90	0.79 ; 1.03	0.91	0.82 ; 1.01
[94;146]	0.89	0.74 ; 1.07	0.84*	0.74 ; 0.97	0.86†	0.77 ; 0.96
Magnesium (mg/l)						
[1; 4[	1.00	-	1.00	-	1.00	-
[4;11[	0.74‡	0.63 ; 0.88	0.79†	0.69 ; 0.91	0.77‡	0.69 ; 0.86
[11;34]	0.91	0.72 ; 1.14	0.93	0.78 ; 1.11	0.92	0.80 ; 1.06
Urban §	1.07	1.00 ; 1.14	1.10	1.05 ; 1.15	1.09	1.05 ; 1.13
Scale §§		0.99		1.01		1.00

* $p \leq 0.05$
† $p \leq 0.01$
‡ $p \leq 0.001$
§ vs rural
§§ allows overdispersion relatively to a regular Poisson model

Annexe C

Manuscript de l’article sur Arcsinus Strict

A Strict Arcsine Regression Model

Sébastien MARQUE and Célestin C. KOKONENDJI*
University of Bordeaux 2 and University of Pau

Abstract

The strict arcsine has been recently studied by Kokonendji and Khoudar in univariate problems involving counts. We propose a strict arcsine regression model for regression analysis of counts. The model can be derived from an attractive framework for incorporating random effect in Poisson regression models and in handling extra-Poisson variation. Comparison with negative binomial model is investigated by simulations and application on data concerning cardiovascular mortality among the elderly.

Key words and phrases. Extra-Poisson variation, maximum likelihood, overdispersion, Poisson mixture, regression analysis of counts, variance function.

2000 Mathematics Subject Classification: 62J02; 62J12.

Short running title: Strict arcsine regression.

1 Introduction

Overdispersion as compared to Poisson in regression analysis of count data was widely studied during the last two decades (since Cox, 1983). The main consequence of overdispersion is the under-estimation of variance and so the misspecification of significance level and confidence interval. The classical model proposed to deal with this phenomen is the negative binomial regression model, also called the Poisson-gamma model (e.g., Lawless, 1987;

*The corresponding author.

McCullagh and Nelder, 1989). It is based on the conditional distribution of a response variable Y such that, given the random effect ν, Y had a Poisson distribution with mean $\mu = \mu(\mathbf{x}; \boldsymbol{\beta})$ depending on a vector $\mathbf{x}$ of covariates and a vector $\boldsymbol{\beta}$ of unknown regression coefficients, and a specific distribution for the random effect ν with density $f(\nu), \nu > 0$. In the Poisson-gamma case, ν follows a gamma distribution which leads to obtention of the negative binomial distribution. This distribution is characterized by a quadratic form of variance:

$$Var(Y) = \mu(1 + a\mu),$$

where $a > 0$ can be considered as the overdispersion parameter such that, if $a < 1$ ($a = 1$ and $a > 1$, respectively) then the overdipersion will be said to be lower (middle and higher, respectively) with respect to the negative binomial model.

Jourdan and Kokonendji (2002) showed that other forms of variance could be more interesting in specific cases, and they focused their attention on a Lagrangian generalization of the negative binomial distribution, called the generalized negative binomial having a cubic variance function. But it would be more adequate for an higher-overdispersion with respect to the negative binomial model. It must be noted that, in the case of a very lower-overdispersion, many overdispersed regression models are competitive and also better for the support of data. In order to find an alternative distribution of negative binomial, some other mixtures were also studied, like the Poisson-inverse Gaussian distribution which is not easily applicable here (e.g., Dean et al., 1989; see also Kokonendji and Khoudar, 2002, for some recent references on the distribution).

The purpose of this paper is to present a regression model based on additive Exponential Dispersion Models (EDMs) (Jørgensen, 1997; Wei, 1998), with a cubic form of variance and which provides reasonable complementarity to the overdispersed regression models like the negative binomial. This model bears on the strict arcsine distribution which was presented by Letac and Mora (1990) and studied recently by Kokonendji and Koudhar (2002). Section 2 presents the model and its specification but also its implementation. In Section 3, we perform simulations in order to proove the validity of our model and we apply the latter, in section 4, to the dataset on cardiovascular mortality among the elderly in France. Section 5 concludes the paper by some remarks on regression and overdispersion.

2 The strict arcsine regression model

The classical way to introduce a flexible component of overdispersion (compared to the Poisson) is by random effect. Currently, Kokonendji and Khoudar (2002) are studying the potential distribution which can lead to the strict arcsine distribution as a Poisson mixture. Another way is by using the extra-Poisson variation into a Poisson model. Without further explanation, we here consider the strict arcsine as an overdispersed EDM compared to the Poisson model.

2.1 Definition

Let Y be a count response variable, and let $\mathbf{x}$ be an associated $d \times 1$ vector of explanatory variables. The strict arcsine regression model for Y on $\mathbf{x}$ is defined by

$$\Pr(Y = y|\mathbf{x}) = \frac{A(y;\alpha)}{y!}\left(\frac{\alpha^2\mu^2}{1+\alpha^2\mu^2}\right)^{y/2} \exp\left\{-\alpha \arcsin\sqrt{\frac{\alpha^2\mu^2}{1+\alpha^2\mu^2}}\right\}, \quad (1)$$

for $y = 0, 1, \cdots$, where $\mu = \mu(\mathbf{x})$ is a positive-valued function, $\alpha > 0$ represents a dispersion parameter, and $A(y;\alpha)$ is given explicitly by

$$A(y;\alpha) = \begin{cases} \prod_{k=0}^{z-1}(\alpha^2 + 4k^2) & \text{if } y = 2z; \text{and } A(0;\alpha) = 1 \\ \alpha\prod_{k=0}^{z-1}[\alpha^2 + (2k+1)^2] & \text{if } y = 2z+1; \text{and } A(1;\alpha) = \alpha. \end{cases} \quad (2)$$

For convenience we will write $Y \sim SA(\mu(\mathbf{x});\alpha)$ to denote this model. The model Y, given the covariates $\mathbf{x}$, is a kind of generalized (non)linear model (McCullagh and Nelder, 1989; Wei, 1998) which is characterized by their two first moments:

$$\mathrm{E}(Y) = \mu(\mathbf{x}) = \mu \quad \text{and} \quad Var(Y) = \mu(1+\alpha^2\mu^2). \quad (3)$$

We assume that $\mu(\mathbf{x}) = h(\mathbf{x}^\mathsf{T}\boldsymbol{\beta})$, where the vector $\boldsymbol{\beta}$ is an unknown regression coefficient and h is a known link function. Here we consider the common log-linear link for h; that is $\mu(\mathbf{x}) = \exp(\mathbf{x}^\mathsf{T}\boldsymbol{\beta})$. In terms of unknown

parameters in the model (1) with $\alpha > 0$, it is sometimes denoted as $\mu = \mu(\mathbf{x}) = \mu(\beta) > 0$.

Generally, the parameter α in (3) does not have the same interpretation in other models handling extra-Poisson variations, like a in the negative binomial with $Var(Y) = \mu(1 + a\mu)$. However, for the unit variance function $V(\mu)$ with $\alpha = 1$ in (3), Kokonendji and Khoudar (2002) compared the strict arcsine with $V_{\mathsf{SA}}(\mu) = \mu(1+\mu^2)$ to negative binomial with $V_{\mathsf{NB}}(\mu) = \mu(1+\mu)$ and to Poisson-inverse Gaussian with $V_{\mathsf{PIG}}(\mu) = \mu(1 + \mu^2/2) + (\mu^2/2)(2 + \mu^2)^{1/2}$, respectively, for showing their complementarity as overdispersed models with respect to the Poisson with $V_{\mathsf{P}}(\mu) = \mu$. For this paper, the indicator of goodness-of-fit will be the log-likelihood criterion.

We conclude this section by noting the following fact, which would be the cause of paired definition (2) of $A(.;.)$ in the strict arcsine model. From the ratio of recursive probabilities (1), the strict arcsine distribution $SA(\mu;\alpha)$ is unimodal or bimodal at the successive points y and $y + 1$ if

$$A(y + 1;\alpha)/A(y;\alpha) = (y + 1)\left(\frac{\alpha^2\mu^2}{1 + \alpha^2\mu^2}\right)^{-y/2}, \quad y = 0, 1, 2, \cdots$$

(Kokonendji and Khoudar, 2002). But, there will be no incidence in the next section.

2.2 Estimation of parameters

We chose to use the maximum likelihood method for estimating, simultaneously, the unknown parameters α and $\boldsymbol{\beta}$ in the model (1) with $\mu = \mu(\mathbf{x};\boldsymbol{\beta}) = \exp(\mathbf{x}^{\mathsf{T}}\boldsymbol{\beta})$ and from a random sample of observations. Alternatively, weighted least squares, or quasilikelihood, could be used for the regression analysis of count data.

According to our next example in medecine or epidemiological study, we can represent the sample as follows. If we consider geographical areas as statistical units, $Y = Y_{\mathsf{r}}$ represents the number of events occuring during a specific period on the r^{th} area. But, in many studies, we need to have more information in order to avoid confusion; bias, for exemple. The classical adjustement collected concerns age and sex. So, we could define several strata, based on age-sex distribution. This design allows us a greater precision on the collection of data. Then, we obtain $Y = Y_{\mathsf{rs}}$ which represents the number of events in the strata s, on the r^{th} region. Thus, we define μ_{rs} as the log-linear predictor associated with the r^{th} region and s^{th} strata. Hence, from

(1) and (2), the log-likelihood function from a random sample of observations $(\mathbf{y}, \mathbf{x}) = \{(y_{rs}, \mathbf{x}_{rs}), r = 1, \cdots, R \text{ and } s = 1, \cdots, S\}$ can be written as

$$\begin{aligned} l(\mathbf{y}|\mathbf{x}; \boldsymbol{\beta}, \alpha) &= \sum_{r,s} \ln \Pr(Y_{rs} = y_{rs}|\mathbf{x}_{rs}; \boldsymbol{\beta}, \alpha) \\ &= \sum_{r,s} \left\{ \ln\left(\frac{1}{y_{rs}!}\right) + \ln A(y_{rs}; \alpha) + \frac{y_{rs}}{2} \ln\left(\frac{\alpha^2\mu_{rs}^2}{1+\alpha^2\mu_{rs}^2}\right) \right\} \\ &\quad -\alpha \sum_{r,s} \arcsin\sqrt{\frac{\alpha^2\mu_{rs}^2}{1+\alpha^2\mu_{rs}^2}}, \end{aligned} \tag{4}$$

where, for convenience, we write $\mu_{rs} = \exp(\mathbf{x}_{rs}^{\mathsf{T}}\boldsymbol{\beta})$ with $\mathbf{x}_{rs} = (x_{rs}^1, \cdots, x_{rs}^d)^{\mathsf{T}}$ for $r = 1, \cdots, R$ and $s = 1, \cdots, S$. Then the first and second derivatives of $l(\mathbf{y}|\mathbf{x}; \boldsymbol{\beta}, \alpha)$ with respect to $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_d)^{\mathsf{T}}$ and α can be expressed as

$$\begin{aligned} \frac{\partial l}{\partial \beta_i} &= \sum_{r,s} \frac{x_{rs}^i(y_{rs} - \alpha^2\mu_{rs})}{1+\alpha^2\mu_{rs}^2}, \quad i = 1, \cdots, d, \\ \frac{\partial l}{\partial \alpha} &= \sum_{r,s} \left[\frac{\partial \ln A(y_{rs}; \alpha)}{\partial \alpha} + \frac{y_{rs} - \alpha^2\mu_{rs}}{\alpha(1+\alpha^2\mu_{rs}^2)} - \arcsin\sqrt{\frac{\alpha^2\mu_{rs}^2}{1+\alpha^2\mu_{rs}^2}} \right] \end{aligned}$$

with

$$\frac{\partial \ln A(y; \alpha)}{\partial \alpha} = \begin{cases} 2\alpha \sum_{k=0}^{z-1} (\alpha^2 + 4k^2)^{-1} & \text{if } y = 2z, \\ \alpha^{-1} + 2\alpha \sum_{k=0}^{z-1} [\alpha^2 + (2k+1)^2]^{-1} & \text{if } y = 2z+1, \end{cases}$$

and

$$\begin{aligned} \frac{\partial^2 l}{\partial \beta_i \partial \beta_j} &= -\sum_{r,s} \frac{x_{rs}^i x_{rs}^j \alpha^2 \mu_{rs}(1 + 2\mu_{rs}y_{rs} - \alpha^2\mu_{rs}^2)}{(1+\alpha^2\mu_{rs}^2)^2}, \quad i, j = 1, \cdots, d, \\ \frac{\partial^2 l}{\partial \beta_i \partial \alpha} &= -2\sum_{r,s} \frac{\alpha\mu_{rs}x_{rs}^i(1 + y_{rs}\mu_{rs})}{(1+\alpha^2\mu_{rs}^2)^2}, \quad i = 1, \cdots, d, \\ \frac{\partial^2 l}{\partial \alpha^2} &= -\sum_{r,s} \left[\frac{2\alpha^2\mu_{rs} + y_{rs}(1 + 3\alpha^2\mu_{rs}^2)}{\alpha^2(1+\alpha^2\mu_{rs}^2)^2} - \frac{\partial^2 \ln A(y_{rs}; \alpha)}{\partial \alpha^2} \right] \end{aligned} \tag{5}$$

with

$$\frac{\partial^2 \ln A(y;\alpha)}{\partial \alpha^2} = \begin{cases} 2\sum_{k=0}^{z-1}(4k^2-\alpha^2)(\alpha^2+4k^2)^{-2} & \text{if } y = 2z, \\ -\alpha^{-2} + 2\sum_{k=0}^{z-1}[(2k+1)^2-\alpha^2][\alpha^2+(2k+1)^2]^{-2} & \text{if } y = 2z+1. \end{cases}$$

Estimates $(\widehat{\boldsymbol{\beta}}, \widehat{\alpha})$ of parameters are readily found via Newton-Raphson iteration (or scoring) algorithm, which is based on the Fisher information matrix $\mathbf{I} = \mathsf{E}(-\partial^2 l/\partial(\boldsymbol{\beta},\alpha)^2)$ evaluated at each iteration. The maximization of $l(\mathbf{y}|\mathbf{x};\boldsymbol{\beta},\alpha)$ will be done simultaneously for $\boldsymbol{\beta}$ and α (or via profile), because the dispersion parameter α is not orthogonal to $\boldsymbol{\beta}$ or $\boldsymbol{\mu} = (\mu_{rs})_{r=1,\cdots,R;s=1,\cdots,S}$ in the sense of Cox and Reid (1987), that is $\mathsf{E}(-\partial^2 l/(\partial\beta_i\partial\alpha)) \neq 0$ for $i = 1,\cdots,d$ from (5).

Note that it is possible to have the estimate $\widehat{\alpha} = 0$, implying that a Poisson regression model is best supported by the data. Thus, tests of the hypothesis $\alpha = 0$ are often of interest, since from (3) this corresponds to a Poisson model. Some approaches can be found in literature, such a test may be based on likelihood-ratio statistics (e.g., Chernoff, 1954) as

$$T = 2l\left(\mathbf{y}|\mathbf{x};\widehat{\boldsymbol{\beta}},\widehat{\alpha},\right) - 2l\left(\mathbf{y}|\mathbf{x};\widehat{\boldsymbol{\beta}}_0,0\right),$$

where $\widehat{\boldsymbol{\beta}}_0$ is a suitable estimate of $\boldsymbol{\beta}$ when $\alpha = 0$ like in the Poisson model.

2.3 Implementation

We perform the strict arcsine regression, essentially, with SAS software for its use and, also, because the negative binomial is already performed by SAS macro, based on Genmod procedure. However, we also develop a C++ program because the computation of log-likelihood value is not pleasant for us in SAS and we must do a second step to obtain it. So, the C++ program could perform an estimation of parameter and log-likelihood computation.

Indeed, in the previous section, we explained the two first derivatives of the log-likelihood for strict arcsine model. We could find the corresponding equations for negative binomial model in the paper of Lawless (1987).

Estimation of parameters can be obtained by maximization of $l(\mathbf{y}|\mathbf{x};\boldsymbol{\beta},\alpha)$ with respect to $\boldsymbol{\beta}$ and α, respectively. The maximization of $l(\mathbf{y}|\mathbf{x};\boldsymbol{\beta},\alpha)$ was solved by a Newton-Raphson iterative algorithm. The program reads data

for each combination of strata by regions (details on format are available in program's source which can be find on the web site of authors). Convergence was evaluated at each step by the difference between two successive estimations of parameters lower than 10^{-6} (which is the convergence criterion) or if the first derivation value at this step was lower than convergence criterion.

The program provides estimations of parameters, standard deviation and value of log-likelihood at the convergence point.

In the SAS macro, we used options about the user's specification of Variance (3) and Deviance. We used the log-likelihood of strict arcsine, explained in (4), and we obtain the deviance as

$$\begin{aligned} D_{\mathsf{SA}} &= 2l\left(\mathbf{y}|\mathbf{x};\widehat{\boldsymbol{\mu}},\widehat{\alpha}\right) - 2l\left(\mathbf{y}|\mathbf{x};\mathbf{y},\widehat{\alpha}\right) \\ &= \sum_{\mathsf{r,s}} y_{\mathsf{rs}} \ln\left(\frac{y_{\mathsf{rs}}^2(1+\widehat{\mu}_{\mathsf{rs}}^2\widehat{\alpha}^2)}{\widehat{\mu}_{\mathsf{rs}}^2(1+y_{\mathsf{rs}}^2\widehat{\alpha}^2)}\right) \\ &\quad -\frac{2}{\widehat{\alpha}}\sum_{\mathsf{r,s}}\left[\arcsin\sqrt{\frac{y_{\mathsf{rs}}^2\widehat{\alpha}^2}{1+y_{\mathsf{rs}}^2\widehat{\alpha}^2}} - \arcsin\sqrt{\frac{\widehat{\mu}_{\mathsf{rs}}^2\widehat{\alpha}^2}{1+\widehat{\mu}_{\mathsf{rs}}^2\widehat{\alpha}^2}}\right], \end{aligned}$$

where $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_{\mathsf{rs}})_{\mathsf{r}=1,\cdots,\mathsf{R};\mathsf{s}=1,\cdots,\mathsf{S}}$ and $\mathbf{y} = (y_{\mathsf{rs}})_{\mathsf{r}=1,\cdots,\mathsf{R};\mathsf{s}=1,\cdots,\mathsf{S}}$.

The convergence criteria were defined by a modification of Deviance lower than 10^{-6}. In order to compare the different models, we used the value of log-likelihood, with the final parameters, as criterion. For negative binomial, we directly performed the computation with SAS software. We used the SAS Output Display System (ODS) option to obtain the predictor (μ_{ks}) and the dispersion parameter (α).

The specific C++ program uses this table to compute the log-likelihood of the model. This operation is necessary because the analysis expression of this indicator is too complex to be computed in SAS Software.

3 Simulations

We performed simulations in order to validate our model, strict arcsine, by comparison with a negative binomial regression which is the classical model for overdispersion. These simulations also permit us to validate estimations by the two programs that we propose previously. We also simulated one dataset based on the strict arcsine and another one on the negative binomial. The parameter values were fixed to different values which could be

found in epidemiological studies. The model performed had a baseline risk (probability of event for non exposed person) and two variables X_1 and X_2.

The strict arsine regression presented in section 2, Poisson and negative binomial models were performed on each simulated dataset and the results are presented in the table 1 and table 2.

Table 1: Simulation of strict arcsine distribution with $R = 100$, $S = 12$, $\beta_0 = -3$, $\beta_1 = -0.2$, $\beta_2 = -0.4$ and $\alpha = 0.25$. Estimations of parameters and variance of parameters with Poisson, negative binomial and strict arcsine models.

	Poisson		Negative	binomial	Strict	arcsine
	$\widehat{\beta}$	$\widehat{\sigma}(\widehat{\beta})$	$\widehat{\beta}$	$\widehat{\sigma}(\widehat{\beta})$	$\widehat{\beta}$	$\widehat{\sigma}(\widehat{\beta})$
Intercept	−3.056	0.109	−3.071	0.456	−3.086	0.662
X_1	−0.121	0.077	−0.126	0.323	−0.129	0.468
X_2	−0.420	0.078	−0.400	0.320	−0.381	0.456
a and α	−		1.24	$= \widehat{a}$	0.20	$= \widehat{\alpha}$
Log-likelihood	−14621.86		−4371.35		−4195.06	

Table 2: Simulation of negative binomial distribution with $R = 100$, $S = 12$, $\beta_0 = -4$, $\beta_1 = -0.4$, $\beta_2 = 0.2$ and $a = 0.6$. Estimations of parameters and variance of parameters with Poisson, negative binomial and strict arcsine models.

	Poisson		Negative	binomial	Strict	arcsine
	$\widehat{\beta}$	$\widehat{\sigma}(\widehat{\beta})$	$\widehat{\beta}$	$\widehat{\sigma}(\widehat{\beta})$	$\widehat{\beta}$	$\widehat{\sigma}(\widehat{\beta})$
Intercept	−3.99	0.019	−3.99	0.045	−3.99	0.048
X_1	−0.39	0.013	−0.39	0.029	−0.40	0.031
X_2	0.18	0.013	0.18	0.029	0.17	0.030
a and α	−		0.538	$= \widehat{a}$	0.078	$= \widehat{\alpha}$
Log-likelihood	−7158.25		−5076.75		−5088.71	

With respect to the negative binomial model, we are in the presence of higher-overdispersion and lower-overdispersion in the table 1 and table 2, respectively. The first conclusion concerns the good quality of the estimation of parameters by the strict arcsine regression model. With the table 1, we could conclude to the better estimation of parameters by our program. The two different programs which are presented here provide the same results, that validate them.

The log-likelihood appears to be a good indicator of goodness-of-fit. We saw with the simulations that the approach by the two models concludes in

both cases to choice of the correct distribution. We could show that estimations of the parameters were identical in the two models, but the estimation of variance was better in the adequate model (strict arcsine in table 1 and negative binomial in table 2). Nevertheless, estimation of the overdispersion parameter (scale parameter) was different in the two models, which could be explained by the difference in the variance form and no comparison is possible between them. We could also choose the better fit by the log-likelihood, which provides the best result in both cases.

Finally, we showed the complementarity of the two models and we evaluated the log-likelihood as an indicator of goodness-of-fit in the case of overdispersion.

4 Example

Our sample was composed by the population living in the 75 ($= R$) parishes included in PAQUID, which is a prospective cohort in the south-west of France. These parishes were drawn at random among all parishes of Gironde and Dordogne according to a design based on the size of the parish. We considered the population between 01/01/1990 and 31/12/1996, aged over 65 years and we collected data on all deaths occurring in this population. The number of person-years was calculated using the average of the population between the two dates, for sex and age (5-years classes). So we had 6 classes for each sex that provides 12 ($= S$) age-sex strata. We obtained the distribution of the population of the parishes by age and sex from the "Institut National de la Statistique et des Etudes Economiques" (INSEE). The causes of deaths were provided by the SC8 (a specialized department of the "Institut Nationale de la Santé Et de la Recherche Médicale") from the deaths certificates. For each death, we recorded sex, age at death (five-year precision), parish of residence and main cause of death (coding on 4 numbers accorded to ICD-9). We observed 14,311 deaths and the total number of person-years was 777,493.

We here present the results of Poisson, negative binomial and strict arcsine regressions on the previously presented data. On the table 3, we present estimators and standard errors of these estimators for each variable. Furthermore, we present the log-likelihood in order to compare the three models and, for negative binomial and strict arcsine, the overdispersion (a) and dis-

persion (α) parameters respectively.

Table 3: Poisson, negative binomial and strict arcsine models on population of ($75 = R$) parishes of Gironde and Dordogne (with $12 = S$), between 1990 and 1997.

	Poisson		Negative	binomial	Strict	arcsine
	$\widehat{\beta}$	$\widehat{\sigma}(\widehat{\beta})$	$\widehat{\beta}$	$\widehat{\sigma}(\widehat{\beta})$	$\widehat{\beta}$	$\widehat{\sigma}(\widehat{\beta})$
Women vs men	$-0.321\dagger$	0.012	$-0.392\dagger$	0.038	$-0.443\dagger$	0.035
Calcium	$-0.001*$	0.0003	-0.001	0.0006	-0.001	0.0005
Magnesium	0.002	0.002	$0.006*$	0.003	$0.005*$	0.002
Rurality	$0.141\dagger$	0.027	$0.109*$	0.040	$0.102*$	0.037
a and α	$-$		0.111	$= \widehat{a}$	0.062	$= \widehat{\alpha}$
Log-likelihood	-2254.58		-2061.38		-2015.69	

$*p$-value ≤ 0.05
$\dagger p$-value ≤ 0.001

We are in the case of lower-overdispersion with respect to the negative binomial model: $\widehat{a} = 0.111$. We can compare goodness-of-fit in these models by the log-likelihood (l) criterion, the greater been the better. The table 3 shows that strict arcsine is the better model, with $l_{\mathsf{SA}} = -2016$ when $l_{\mathsf{BN}} = -2061$. We can conclude with a better fit for strict arcsine even if estimations are slightly close. According to the previous results obtained by Marque et al. (2003), the latter confirm the effects of these two minerals on cardiovascular mortality. Calcium appears to have a protective effect on cardiovascular and Magnesium seems as the opposite effect. We can denote also that the statistical significance disappears in Calcium case when we use a valid model although Magnesium becomes slightly significance (p-value $\simeq 0.05$). Even if the estimations of parameters are similar (Cox, 1983), strict arcsine approach provides a better fit than negative binomial and Poisson (the value of the log-likelihood of this model was -2255).

5 Conclusion

The first conclusion concerns the nearness between the results of negative binomial and strict arcsine. For instance, this conclusion is avaible for lower-

overdispersion, middle-overdispersion and weak higher-overdispersion with respect to negative binomial model. It would be interesting to construct a test of the degree of overdispersion with respect to the negative binomial model to choose any adequate overdispersed model. The estimation of parameters in these two models are very close even if the estimations of variance are smaller for strict arcsine than negative binomial model. We also showed slight differences between Poisson and strict arcsine or negative binomial estimations. Even if the estimations of parameters are close, the variance of the Poisson model are clearly smaller than the negative binomial or the strict arcsine. So, we could conclude that Poisson is the worst model in this application.

According to simulations and applications presented below, the major avantage of this approach is to provide an alternative to the negative binomial regression to take into account overdispersion in count data. But, these two approachs are complementary and the authors advice using the two regressions in a practical use. In presence of overdispersion, a complementary approach can be used by performing strict arcsine and negative binomial models and, in this case, the log-likelihood estimator criterion seems to be a good indicator of choice between them. Indeed, actually the negative binomial approach is the only regression available in classical software, but we provide two programs in order to estimate the different parameters of regression. the different simulations show that the algorithm converges rapidly.

Finally, we presented a method based on the simultaneous maximum likelihood method for estimating both interest and nuisance parameters. Nevertheless, estimation can be performed by profile-likelihood on the dispersion parameter traited as nuisance parameter.

Acknowledgements. The authors thank Dr Jacqmin-Gadda and Prs Dartigues and Commenges for providing the data of the application.

References

Chernoff, H. (1954). On the distribution of the likelihood ratio. Ann. Math. Statist. **25**, 38-44.

Cox, D.R. (1983). Some remarks on overdispersion. Biometrika **70**, 269-274.

Cox, D.R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. J. Roy. Statist. Soc. Ser. **B49**, 1-39.

Dean, C., Lawless, J.F. and Willmot, G.E. (1989). A mixed Poisson-inverse Gaussian regression model. Canadian J. Statist. **17** (2), 171-181.

Jørgensen, B. (1997). The Theory of Dispersion Models. Chapman & Hall, London.

Jourdan, A. and Kokonendji, C.C. (2002). Surdispersion et modèle binomial négatif généralisé. Rev. Statistique Appliquée **L** (3), 73-86.

Kokonendji, C.C. and Khoudar, M. (2002). On strict arcsine distribution. Preprint LMA-Pau No.**2002/14** (CSTM: submitted for publication).

Lawless, J.F. (1987). Negative binomial and mixed Poisson regression. Canadian J. Statist. **15**, 203-225.

Letac, G. and Mora, M. (1990). Natural real exponential families with cubic variance functions. Ann. Statist. **18**, 1-37.

Marque, S., Jacqmin-Gadda, H., Dartigues, J.-F. and Commenges, D. (2003). Cardiovascular mortality and calcium and magnesium in drinking water: an ecological study in elderly people. European J. Epidemiology (to appear).

McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models. Chapman & Hall, London, 2nd edition.

Wei, B.-G. (1998). Exponential Family Nonlinear Models. Lecture Notes in Statistics, No.130, Springer-Verlag, Singapore.

Sébastien MARQUE
Université de Bordeaux2 - INSERM U330
& Stat'Aids - Tour Montparnasse
33, avenue du Maine - 75755 Paris cedex 15, France
E-mail: smarque@stataids.com

Célestin C. KOKONENDJI
Université de Pau et des Pays de l'Adour
Laboratoire de Mathématiques Appliquées - CNRS FRE2570
Avenue de l'Université - 64000 Pau, France
E-mail: celestin.kokonendji@univ-pau.fr

Annexe D

Codes sources des programmes permettant l'estimation de la régression SA sous SAS

```
%macro asinus(dsin=, yvar=, xvars=, clsvars=, offvar=, ithist = 1,
                expected = 0 );
                                  * Uncomment the following for a debug
trace;
* options mprint;
title Arc Sinus Regression ;
* Turn off printing;
%global _print_;
%let _print_ = OFF;
%let maxiter = 50;
%let iter = 1;
%let conv = 0;
/* Recupere et formate les parametres du modele */
%if %upcase(&offvar) ne
%then %let offstmt = OFFSET=&offvar;
%else %let offstmt= ;
%if %upcase(&clsvars) ne %then
%let clsstmt = %str(CLASS &clsvars;);
%else %let clsstmt= ;
%if(&expected=1) %then %let expstmt = EXPECTED ;
%else %let expstmt= ;
/* Premier passage avec une regression de Poisson */
proc genmod data=&dsin;
&clsstmt   make 'ModelFit' out=A;
   model &yvar = &xvars / dist   = poisson  &offstmt;
   run;

/* Affiche les results du 1er passage */
data _NULL_;
%if( &ithist = 1 ) %then
  %str(file print;);
  set A;
   if _N_ = 3 then      do;
        call symput( 'disp', put( valuedf, best10.6 ) );
        %if( &ithist = 1 ) %then
            %do;
               temp2 = 1/valuedf;
               put 'Iteration number: ' "&iter";
               put 'Pearson Chi2/DF:  ' valuedf;
               put 'Alpha:            ' temp2;
            %end;
      end;
run;

%let alpha = 1 / &disp;
```

```
/* Modele ArcSinus */
/* Convergence par le parametre de dispersion */
%do %while( &conv = 0 ) ;
/* Modele avec modif de la varaince et deviance */
proc genmod data=&dsin ;
      &clsstmt
      make 'ModelFit' out = A;
      _K = &alpha;
      _A = _MEAN_;
      _Y = _RESP_;

/* Variance AS */
      variance _VAR = _A+_K*_A*_A*_A;

/* Deviance AS */
      _E=(_Y/2)*(log((_Y*_Y*_K*_K)/(1+_Y*_Y*_K*_K))
            -log((_A*_A*_K*_K)/(1+_A*_A*_K*_K)));

      _F=(1/_K)*(arsin(sqrt((_Y*_Y*_K*_K)/(1+_Y*_Y*_K*_K)))
           -(arsin(sqrt((_A*_A*_K*_K)/(1+_A*_A*_K*_K)))));

      _D=2*(_E-_F);

      deviance _DEV = _D;
      model   &yvar   =   &xvars     /   &offstmt   link=log   scoring=1
covb/*itprint*/;
run;

%let iter = %eval( &iter + 1 );
```

```
data _NULL_;
      %if( &ithist = 1 ) %then %str(file print;);
      set A;
      if _N_ = 1 then
    do;
        call symput( 'deviance', put( value, 10.4 ) );
        call symput( 'devdf', put( valuedf, 10.4 ) );
      end;

      if _N_ = 3 then
      do;
      call symput( 'PX2', put( value, 10.4 ) );
        call symput( 'PX2df', put( valuedf, 10.4 ) );
        temp3 = &alpha;
        if ( ABS( valuedf - &disp ) <= 1.e-6 OR &iter > &maxiter )
        then  call symput( 'conv', '1' ) ;
      else do;
      temp3 = valuedf * temp3;
      call symput( 'disp', put( valuedf, best10.6 ) );
      call symput( 'alpha', put( temp3, best20.10 ) );
      end;
    %if( &ithist = 1 ) %then
      %do;
        put 'Iteration number: ' "&iter";
        put 'Pearson Chi2/DF:  ' valuedf;
        put 'Alpha:            ' temp3;
    %end;
  end;
run;

/* Affiche modele final */
%if ( &conv = 1 )  %then
%do;
      proc genmod data=&dsin;
    &clsstmt         make 'ModelFit'  out = A;
    make 'parmest' out = B;
      _K = &alpha;
      _A = _MEAN_;
      _Y = _RESP_;

/* Variance AS */
      variance _VAR = _A+_K*_A*_A*_A;
```

```
/* Deviance AS */
      _E=(_Y/2)*(log((_Y*_Y*_K*_K)/(1+_Y*_Y*_K*_K))
            -log((_A*_A*_K*_K)/(1+_A*_A*_K*_K)));

      _F=(1/_K)*(arsin(sqrt((_Y*_Y*_K*_K)/(1+_Y*_Y*_K*_K)))
           -(arsin(sqrt((_A*_A*_K*_K)/(1+_A*_A*_K*_K)))));

      _D=2*(_E-_F);

      deviance _DEV = _D;
    model &yvar = &xvars  / covb &offstmt &expstmt link = log ;
run;

data _NULL_;
file print;
set A;
if _N_ = 5 then do;
      iter10 = &iter;
    alpha10 = &alpha;

    put 'Number of iterations: ' iter10  10.0 ;
    put 'Alpha:                ' alpha10 10.4;
    put 'Deviance:             ' "&deviance    "
        'Deviance/DF:          ' "&devdf";
    put 'Pearson Chi2:         ' "&PX2    "
        'Pearson Chi2/DF:      ' "&PX2df";
    put 'LogLikelihood:        ' value 10.4;

end;
run;
%end; /* Fin boucle DO */
%end;
proc print data = B;   run;
%let _print_ = ON;
title;
options nomprint;
%mend asinus;
```

Bibliographie

[1] S Marque, H Jacqmin-Gadda, J Dartigues, and D Commenges. Cardiovascular mortality and calcium and magnesium in drinking water : an ecological study in elderly people. *Eur J Epidemiol*, 18(4) :305–9, 2003.

[2] S Marque and CC Kokonendji. A strict arcsine regression. *Communication in Statistics : Theory and Methods*, 2003. (Submitted).

[3] S Marque and D Commenges. Prise en compte de l'homogénéité de la population par un modèle de poisson surdispersé. *XXXIIè Congrès de la Société française de Statistiques*, 15-19 Mai 2000.

[4] S Marque, D Gillet, and D Commenges. Relation entre le calcium de l'eau de boisson et la mortalité cardio-vasculaire chez les personnes âgées. *Congrès de l'ADELF*, 12-14 Octobre 2000.

[5] S Marque and D Commenges. Utilisation de données individuelles dans une étude écologique : application au rôle du calcium dans la mortalité cardiovasculaire. *XXXIIIè Congrès de la Société Française de Statistiques*, 14-18 Mai 2001.

[6] S Marque and D Commenges. Modélisation de la surdispersion dans un modèle de poisson par une structure spatiale : application à la mortalité cardiovasculaire en aquitaine. *XXXIVè Congrès de la Société Française de Statistiques*, 13-17 Mai 2002.

[7] S Marque and CC Kokonendji. La régression arcsinus : Une alternative à la loi binomial-négative : la régression arcsinus. *XXXVè Congrès de la Société Française de Statistiques*, 2-6 Juin 2003.