UNIVERSITÉ DE STRASBOURG

DIPLÔME UNIVERSITAIRE
d'Actuaire de Strasbourg « DUAS »

Mémoire de fin d'études
- année universitaire 2008/2009 -

Julien MATHIS

Elaboration d'un zonier en assurance de véhicules par des méthodes de lissage spatial basées sur des simulations MCMC

➢ ETABLISSEMENT D'ACCUEIL

11 avenue Myron Herrick
75008 Paris

➢ MAITRE DE STAGE

Yannick APPERT-RAULLIN, Actuaire

➢ PERIODE DE STAGE

Du 02 mars au 31 août 2009

➢ MEMOIRE CONFIDENTIEL

Pôle européen de gestion
et d'économie (PEGE)
61 avenue de la Forêt Noire
F-67085 Strasbourg Cedex
http://sceco.u-strasbg.fr/actuariat/

Résumé

Michael Boskov et Richard Verrall ont proposé en 1994 une méthode de tarification par zones géographiques basée sur des modèles spatiaux à structure bayésienne. Celle-ci utilise des algorithmes de type Monte Carlo par Chaînes de Markov (MCMC) ainsi que l'échantillonnage de Gibbs pour obtenir une estimation du risque spatial sous-jacent aux régions étudiées.

Ce mémoire, dont la finalité est l'élaboration d'un zonier, développe principalement l'application de ce modèle dit de Boskov et Verrall. Il y est présenté dans un premier temps les éléments théoriques nécessaires à son utilisation avant de s'intéresser à son implémentation dans le logiciel statistique R. Son efficacité est testée en l'appliquant à une base de données de véhicules et en comparant les résultats avec ceux obtenus à l'aide d'une méthode plus pragmatique.

Abstract

Michael Boskov and Richard Verrall proposed in 1994 a method for premium rating by geographic area based on spatial models in a Bayesian framework. This method uses Markov Chain Monte Carlo methods and the Gibbs sampler to estimate the local risk in each geographical area.

This paper, whose purpose is the development of a zone system, mainly develops the application of the Boskov and Verrall model. The theoretical elements necessary for its use are first presented before considering its implementation in the statistical software R. Its efficiency is tested by applying it to a vehicle database and by comparing results with those obtained by means of a more pragmatic method.

Remerciements

Je tiens tout d'abord à remercier Michel PIERMAY, président de FIXAGE, de m'avoir accueilli au sein de son entreprise et d'avoir mis à ma disposition les moyens techniques nécessaires au bon déroulement de mon stage, ainsi que Emannuel TASSIN, directeur associé de FIXAGE, pour son encadrement régulier.

Je remercie tout particulièrement Yannick APPERT-RAULLIN, mon maître de stage, de m'avoir proposé ce sujet très intéressant, et de m'avoir apporté ses connaissances ainsi que son expérience, aussi bien pour la réalisation de ce mémoire que pour les missions que nous avons effectuées. La qualité de ses conseils a été une aide précieuse tout au long de mon travail.

Mes remerciements s'adressent également à l'ensemble des collaborateurs de FIXAGE, qui par leur chaleureux accueil, leur attention et leurs explications diverses, ont contribué à favoriser mon intégration, rendant ce stage très agréable.

Enfin, ce mémoire marquant la fin de mes études, je tiens à remercier ici l'ensemble du corps professoral de l'Université de Strasbourg pour leurs riches enseignements dispensés durant mes cinq années d'études supérieures.

Table des matières

Introduction

La tarification de produits d'assurance représente une des missions principales auxquelles un actuaire peut-être confronté durant sa carrière professionnelle. Dans le cas de l'assurance de véhicules [1] plus particulièrement, il convient de déterminer en premier lieu les variables dites explicatives, c'est-à-dire celles qui permettent d'expliquer la sinistralité observée. Elles répondent à un réel besoin de la part de l'assureur qui est de segmenter le tarif par facteurs de risque, permettant de répartir les individus de son portefeuille en sous-groupes homogènes. Il peut ainsi proposer un tarif adapté au risque auquel chaque assuré s'expose plutôt que de proposer un tarif moyen sur l'ensemble du portefeuille, et donc éviter que les assurés possédant un faible risque ne partent chez un concurrent car leur prime serait trop élevée. Cela permet également de répondre au problème lié à l'anti-sélection puisque seuls les assurés connaissent leur vrai risque et s'assurent uniquement s'ils trouvent le prix intéressant. Cependant, bien que le principe de la mutualisation des risques reste exact au niveau d'un portefeuille, il n'est plus applicable au niveau des segments, ce qui implique pour l'actuaire une utilisation de modèles mathématiques complexes. Les variables explicatives seront utilisées par la suite pour estimer la sinistralité à venir et ainsi proposer la prime la plus juste à chaque assuré en fonction de ses caractéristiques.

Il ressort de cette analyse préliminaire qu'en assurance de véhicules la variable correspondant au critère spatial est l'une des variables qui explique le mieux la sinistralité, autrement dit la fréquence des sinistres varie fortement en fonction du lieu de résidence de l'assuré. Cela peut être expliqué entre autres par les conditions météorologiques locales, le comportement des usagers ou encore la densité du trafic.

1. Le terme générique "véhicules" est employé ici et dans la suite de ce mémoire pour faire référence aussi bien aux produits de type deux roues que quatre roues.

Dès lors, un des enjeux majeurs est d'estimer au mieux le risque sous-jacent à chaque zone géographique de manière à l'intégrer dans la tarification de véhicules aux côtés des variables habituelles telles que l'âge du conducteur ou encore la puissance du véhicule.

Pour réaliser cette tarification, il convient de segmenter la zone géographique étudiée (la France dans notre cas) en un nombre de régions [2] adéquat afin d'estimer leur risque sous-jacent. Cette segmentation peut aller de la simple distinction entre zone rurale et urbaine (2 zones) jusqu'à un découpage par commune (36 686 zones).

L'estimation du risque géographique de chaque région est une tâche complexe. Nous disposons pour chaque assuré de son lieu de résidence mais aucune mesure nous permet de comparer ces lieux entre eux et de les classer par risque. Une première approche, tout à fait basique, consiste à utiliser la fréquence de sinistres observée. Supposons donc que celle-ci représente une estimation fiable du risque. Nous pouvons créer une variable polytomique représentant le critère spatial par l'intermédiaire d'un zonier en regroupant l'ensemble des régions étudiées en un nombre de classes prédéfini (souvent quatre ou cinq, allant de la moins risquée à la plus risquée), permettant d'intégrer le critère spatial au sein de la tarification.

Nous avons testé cette méthode sur un portefeuille d'assurance de véhicules d'une entreprise d'assurance française. Le zonier obtenu est représenté sur la FIGURE 1, en utilisant un découpage par départements et une répartition des fréquences observées en cinq zones de risques.

Le zonier obtenu permet de localiser des concentrations des risques, mais il est difficilement applicable tel quel lors d'une tarification car il présente de trop fortes disparités entre des régions proches géographiquement. Nous sommes effectivement confrontés ici à plusieurs problèmes. Premièrement, la méthode utilisée ignore la dépendance spatiale entre les régions voisines, du fait que chaque estimateur se base uniquement sur les données propres à sa région.
De plus, certaines régions sont sous-représentées (très peu d'assurés), leur fréquence

2. Nous utiliserons le terme "régions" dans la suite de ce mémoire pour désigner les zones géographiques issues de la segmentation.

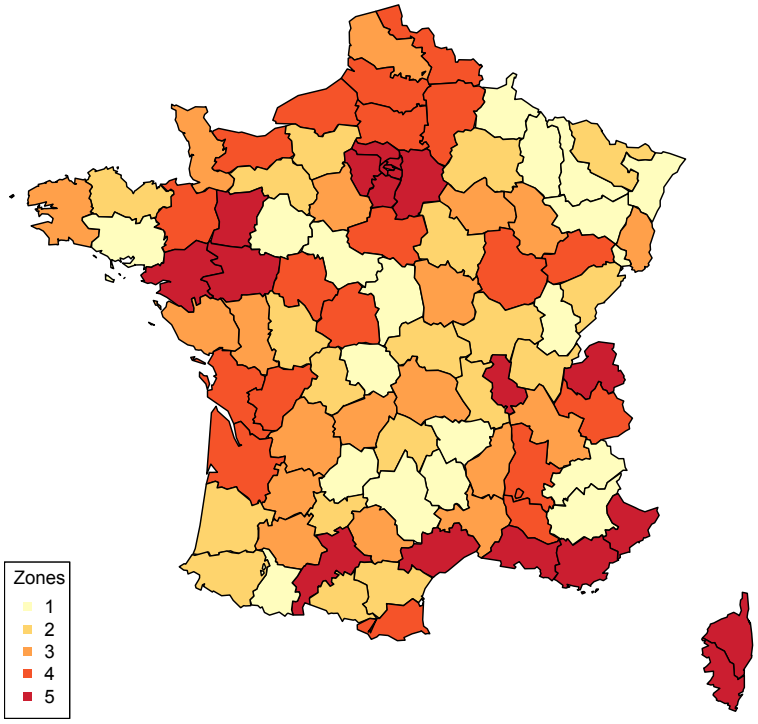


FIGURE 1 – Zonier réalisé en utilisant la fréquence observée

observée n'est pas significative. Enfin, nous obtenons ici un estimateur du risque global et non uniquement du risque géographique. Une forte fréquence peut être due à un autre effet que le critère spatial et, par exemple, une région qui serait placée en zone risquée pourrait l'être du simple fait que sa population ait des caractéristiques risquées (par exemple une population jeune).

Pour répondre à ces problèmes, nous nous intéresserons à des méthodes basées sur un lissage spatial des données permettant de dégager des tendances, mais surtout de réduire les variations locales brusques. En effet, il semble raisonnable de supposer que les assurés se déplacent dans un rayon plus ou moins grand et donc que leur risque dépend non seulement de leur lieu de résidence mais également des lieux avoisinants. Nous chercherons à utiliser un modèle mathématique permettant un transfert d'informations entre les régions voisines en supposant que les régions qui sont proches les unes des autres ont plus de chance d'avoir un risque spatial similaire que les régions qui sont éloignées. Cela ne nous empêche pas de regrouper des régions qui sont géographiquement éloignées mais qui possèdent le même risque. En outre, le lissage spatial nous permettra d'atténuer l'erreur d'échantillonnage.

Cette analyse permet de constater que nous sommes confrontés à trois problématiques majeures lors de l'intégration du critère spatial dans une tarification de produits d'assurance de véhicules, à savoir :

1. le choix de la segmentation ;
2. l'estimation du risque sous-jacent à chaque région ;
3. l'élaboration d'un zonier.

La première est fortement liée aux données disponibles. Bien que la segmentation la plus fine possible soit préférable, l'absence de données (surtout en assurance deux roues) oblige parfois à réaliser une segmentation intermédiaire. Nous illustrerons ce mémoire en choisissant une segmentation par départements, ce qui limite le nombre de régions étudiées, bien qu'une segmentation par code postaux soit privilégiée si nous sommes amenés à utiliser cet outil dans le cadre d'une tarification. En effet, le risque n'est pas identique à l'intérieur d'un même département, il y a par exemple des différences entre zones urbaines et rurales, ce qui peut conduire à une tarification moins juste qu'avec une segmentation par code postaux.

Quelques outils mathématiques et approches possibles d'estimation du risque spatial ont fait l'objet d'une application au domaine assurantiel, ils seront présentés en première partie. Par la suite, nous développerons l'approche initialement étudiée par Boskov et Verrall.

Après avoir présenté les fondements théoriques de celle-ci en deuxième partie, nous aborderons sa résolution qui passe par des techniques de Monte Carlo par Chaînes de Markov et plus particulièrement l'échantillonnage de Gibbs.

La quatrième partie de ce mémoire sera consacrée à l'application pratique à savoir l'implémentation dans le logiciel de statistiques R.

Enfin, nous étudierons différentes méthodes permettant de regrouper les estimations obtenues en classes, et donc de construire un zonier. Les résultats obtenus pourront être comparés au modèle pragmatique utilisé jusqu'à présent par Fixage afin de réaliser la critique du modèle utilisé et d'en constater les limites.

Chapitre 1

Les outils mathématiques comme réponse au découpage d'une région en zones de risque

Il existe de nombreuses méthodes statistiques de lissage des données mais très peu d'articles s'intéressent à leur application dans le domaine de l'assurance. Seuls quelques auteurs ont adapté des techniques utilisées dans d'autres domaines, tel que la restauration d'images spatiales, à la tarification par zones géographiques. Bien que le but de l'analyse diffère en fonction du domaine d'application, les outils mathématiques restent similaires. Dans le cadre de notre étude, nous avons décidé de présenter les méthodologies évoquées dans la littérature, de les expliciter et surtout de les critiquer afin de s'orienter vers notre propre méthodologie d'élaboration d'un zonier.

1.1 Lissage par splines bidimensionnels

Greg Taylor fut un des premiers à introduire une technique de lissage spatial pour la tarification de produits d'assurance dans un article [10] publié en 1989. Celle-ci est basée sur un ajustement d'une fonction de prime de manière à lisser les résultats. L'ajustement est réalisé par l'utilisation de splines bidimensionnels et testé sur un jeu de données d'un Etat d'Australie.

L'auteur considère la prime comme une fonction continue des coordonnées géographiques, pour tous les facteurs de risque constants à l'exception de ces coordonnées. Ainsi, une variation entre deux régions est expliquée uniquement par le risque spatial. La principale tâche consiste à ajuster la fonction de prime, dont la forme mathématique est inconnue, par des splines bidimensionnels. Les zones de tarifications seront obtenues en analysant les fortes variations de cette fonction.

L'utilisation de fonctions splines permet d'ajuster une fonction plus simple sur chaque sous-intervalle de la plage de la fonction à ajuster. Ces fonctions sont des polynômes, généralement de degré 2 (splines quadratiques) ou 3 (splines cubiques). Le découpage préalable de la plage de la fonction à ajuster est assez délicat et le raccordement des splines impose la continuité et l'égalité des pentes et des courbes aux nœuds. La méthode employée consiste à estimer dans un premier temps l'ensemble des facteurs discriminants [1] (y compris pour le facteur géographique) sur la base du modèle de tarification actuel.
Un index de risque représentant l'effet de la zone géographique et basé sur les valeurs standardisées de tous les autres facteurs est ensuite calculé. Il correspond en quelque sorte au coefficient par lequel la prime de la région étudiée a été multipliée par rapport à la prime qui aurait été demandée dans une région standard (avec un risque géographique neutre) ayant tous les autres facteurs au même niveau.
Enfin, cet index de risque est traité comme un estimateur de la fonction de coordonnées géographiques dont l'ajustement permet un lissage des résultats. Le plan est lié à la carte des codes postaux par une transformation de coordonnées. Une estimation de la prime est obtenue comme étant proportionnelle à cet index de risque.

L'avantage que présente cette méthode est que l'estimateur du risque géographique est décorrélé des autres facteurs de risque. En revanche, il est directement basé sur le niveau de la prime, donc à la fois sur la fréquence et le coût moyen, ce qui n'est pas optimal du fait que le risque spatial n'a pas le même effet sur ces deux composantes. De plus, le choix des sous-intervalles, et donc des nœuds, nécessite une étude a priori de la forme de la fonction, ce qui n'est pas évident, surtout pour un pays tout entier.

1. Utilisation des modèles linéaires généralisés présentés en partie 2.1

1.2 Lissage intégré dans un modèle bayésien hiérarchique

Michael Boskov et Richard Verrall ont introduit [1] une autre approche basée sur des modèles spatiaux à structure bayésienne. Ils prétendent que des résultats plus pertinents peuvent être obtenus en prenant un modèle basé sur le nombre de sinistres et leur montant séparément, contrairement à celui présenté juste avant qui ajuste directement la prime.

La méthode utilisée suppose qu'il existe une réelle structure du risque géographique et que les données sont une représentation de cette structure altérée par un bruit aléatoire. Le but est alors d'identifier cette structure.

L'utilisation d'un modèle bayésien permet de considérer le risque comme étant une variable aléatoire. Le problème d'optimisation consiste à trouver les facteurs de risque qui maximisent sa densité conditionnelle. Cela revient à estimer le risque géographique ainsi que le bruit aléatoire, qui peuvent être isolés du risque global, car les facteurs de risque (autres que spatial) sont supposés connus et estimés antérieurement par une régression de Poisson. La résolution de ce problème d'optimisation est réalisée par l'échantillonnage de Gibbs dont le fonctionnement sera détaillé dans le chapitre 3.3. La définition des densités permet d'introduire un transfert d'informations entre les régions voisines. Des résultats sont présentés dans l'article en réutilisant les mêmes données que Taylor [10].

La méthode proposée n'aboutit pas à l'élaboration d'un zonier mais uniquement à une estimation du risque géographique pour chaque région. Le zonier pourra être créé en fonction de ces valeurs. Elle présente l'avantage d'isoler le risque spatial et d'introduire nos hypothèses dans le modèle tout en se basant uniquement sur la fréquence et non directement sur la prime. Il faut tout de même rester vigilant au fait que l'introduction de ces hypothèses augmente le risque de spécificité.

Notons enfin qu'en 2002 un groupe de chercheurs, dont Richard VERRALL, a publié un article [2] proposant une tarification par zones géographiques basée sur la méthode de BOSKOV et VERRALL. Il s'agit d'une étude de cas concluant sur l'efficacité du modèle appliqué à un portefeuille d'assurance belge. Elle apporte également quelques précisons théoriques.

1.3 Lissage de Whittaker

Greg TAYLOR publia un deuxième article [11] concernant la tarification par zones géographiques en utilisant cette fois-ci un lissage par la méthode de Whittaker en dimension deux. Le principe de cette méthode est d'obtenir une valeur lissée qui minimise une combinaison linéaire entre une déviance (mesurant la perte due au lissage) et une variable représentant une pénalité pour manque de lissage. Nous obtenons donc un compromis entre la fidélité aux données et l'ajustement.

Ici aussi, l'auteur isole le facteur spatial de la tarification tout en se basant sur un schéma bayésien. En effet, il suppose que la fréquence de sinistres est une variable aléatoire dont l'espérance est fonction, entre autre, du risque géographique. Les autres facteurs de risque ont également été mesurés lors d'une étape antérieure.

La variable à ajuster est le risque dans chaque région, standardisé des autres facteurs. C'est elle qui isole le facteur spatial du fait qu'il est égal à l'espérance de celle-ci.

Bien que cette approche présente également l'avantage d'estimer le risque géographique de chaque région à partir des fréquence décorrélées des autres facteurs, elle parait plus difficile à mettre en place pour un pays entier et est considérée comme mal adaptée aux fortes variations locales du risque par l'auteur de l'article.

Suite à la recherche puis l'analyse des différentes méthodes existantes, nous avons décidé d'établir notre étude sur un zonier élaboré au travers de l'approche décrite par Michael Boskov et Richard Verrall.

Il apparaît en effet, suite à des études que nous avons réalisées en parallèle à ce mémoire, que c'est la fréquence qui est la plus influencée par l'effet spatial. Le coût moyen subit également cet effet mais il est beaucoup moins fort que ceux liés au véhicule (son prix ou sa puissance par exemple). De plus, les effets peuvent être expliqués par des phénomènes complètement indépendants (climat pour la fréquence et tribunal administratif pour le coût).

Nous souhaitons ainsi développer un modèle basé sur la fréquence, et devons au préalable la décorréler des autres effets que l'effet spatial (utilisation de modèles linéaires généralisés). Enfin, bien qu'une approche par simulation de distribution ajoute des niveaux d'hypothèses concernant les paramètres, elle permet de générer des états de la nature jamais observés.

Les éléments théoriques nécessaires à l'utilisation d'un tel modèle seront décrits dans les prochains chapitres.

Nous avons sélectionné une approche de lissage basée sur un modèle mathématique avancé, permettant d'obtenir un estimateur du risque spatial pour chaque zone géographique de notre segmentation. Il faut à présent étudier la théorie sous-jacente à ce modèle de manière à pouvoir l'implémenter et l'utiliser.

Chapitre 2

Présentation de l'approche retenue

Le modèle utilisé est un modèle mathématique de lissage spatial basé sur l'inférence bayésienne hiérarchique. Il convient donc, avant de présenter ce modèle, d'en préciser le principe. Etant donné que cette méthode de lissage nécessite l'utilisation des modèles linéaires généralisés afin d'isoler l'effet spatial des données observées, nous commencerons par rappeler leur but ainsi que les choix que nous avons effectués.

2.1 Les modèles linéaires généralisés

Comme nous l'avons vu au cours du chapitre précédent, il peut être intéressant d'analyser uniquement les estimations du risque spatial lors de l'élaboration d'un zonier. Pour ce faire, nous allons dans un premier temps estimer tous les autres facteurs de risque afin de travailler sur les fréquences estimées en tenant compte de tous les facteurs autres que le facteur spatial. Ces estimations sont obtenues au travers des modèles linéaires généralisés qui permettent d'étudier le lien entre une variable réponse et un ensemble de variables explicatives. Ils sont formés de trois composantes :

1. une composante aléatoire Y, à laquelle nous associons une loi de probabilité ;
2. une composante déterministe η, ou prédicteur linéaire ;
3. une fonction de lien $g(\cdot)$, décrivant la relation entre l'espérance de la variable réponse et la combinaison linéaire des variables explicatives.

Le modèle linéaire généralisé est de la forme :

$$g\big[\mathbb{E}(Y)\big] = \beta X$$

où

g	est la fonction de lien ;
$Y = (Y_1, \ldots, Y_n)'$	est le vecteur des variables aléatoires ;
$\beta = (\beta_1, \ldots, \beta_n)'$	est le vecteur des paramètres à estimer ;
X	est la matrice des variables explicatives de taille n x p.

Dans notre cas, la variable à expliquer est la sinistralité dans chaque région (dont nous disposons des observations $y_1, \ldots, y_n$), nous supposerons qu'elle est distribuée selon une loi de Poisson [1]. La partie déterministe est composée de l'ensemble des variables explicatives utilisées pour la tarification à l'exception, dans un premier temps, du facteur spatial. Les modèles linéaires généralisés seront une nouvelle fois utilisés après l'élaboration du zonier pour obtenir la tarification complète. Enfin, étant donné que nous souhaitons réaliser une régression de Poisson, adaptée à une estimation de la sinistralité, nous choisissons une fonction de lien logarithmique [1] $\big(\log\big[\mathbb{E}(Y)\big] = \beta X\big)$. Cette fonction permet d'obtenir une structure tarifaire multiplicative. Le risque est modélisé par un produit d'effets multiplicatifs permettant d'obtenir des primes pures toujours positives.

L'utilisation d'un modèle linéaire généralisé, par rapport à un modèle linéaire classique, présente plusieurs avantages pratiques lors d'une tarification. Premièrement, la loi de Y doit appartenir à la famille exponentielle (qui englobe la plupart des lois standards), alors que le modèle linéaire classique se cantonne à la loi normale. Ensuite, la variance des données n'est plus supposée constante mais dépendante de l'espérance.

Notons enfin que les estimations des paramètres sont obtenues par la méthode du maximum de vraisemblance.

1. Choix réalisés lors d'études antérieures

2.2 L'inférence bayésienne hiérarchique

En inférence classique, dite aussi fréquentiste, le paramètre θ de la densité $p(y|\theta)$, utilisé pour représenter la distribution de la variable aléatoire observée y, est considéré comme fixé mais inconnu. Une estimation de celui-ci est obtenu à partir des données tirées des échantillons. Par exemple, l'estimation par le maximum de vraisemblance est une méthode fréquentiste.

L'approche bayésienne permet d'ajouter à l'information contenue dans les données observées des informations provenant d'autres sources. Elle consiste à probabiliser le paramètre inconnu en lui associant une loi $p(\theta)$, dite loi *a priori* car antérieure aux observations, précédant toute information sur y. Elle permet de placer dans le modèle l'ensemble des hypothèses, et avis d'experts, que nous connaissons ou supposons a priori. Les paramètres de cette loi sont appelés *hyperparamètres* et sont fixés initialement.
Comme son nom l'indique, l'inférence bayésienne est basée sur le théorème de Bayes, à savoir :

$$p(\theta|y) = \frac{p(y|\theta)\ p(\theta)}{p(y)}$$

Cette équation traduit le fait que la densité obtenue après l'observation des données (*a posteriori* car dépendant directement de y) est basée sur les données observées (*fonction de vraisemblance*) modifiées par nos croyances (*loi a priori*). Le problème reposant sur une estimation de θ, le dénominateur représente qu'une normalisation et ne sera pas utilisé ce qui nous amènera à définir des densités à une constante près :

$$p(\theta|y) \propto p(y|\theta)\ p(\theta)$$

Où le signe " $\propto$ " signifie " proportionnelle à ". On cherchera alors à représenter cette densité afin d'obtenir une estimation de θ.

L'inférence bayésienne hiérarchique consiste à introduire un niveau d'incertitude supplémentaire. En effet, il peut arriver que l'on ne dispose pas de suffisamment d'informations concernant les hyperparamètres pour définir convenablement le prior. Les hyperparamètres ne sont alors plus fixés mais à leur tour considérés comme des variables aléatoires de loi donnée et de paramètre ϵ inconnu. Les modèles bayésiens hiérarchiques

peuvent donc se résumer comme suit, pour une variable aléatoire observée y de paramètre θ et d'hyperparamètre ϵ :

$$\begin{aligned} y|\theta &\sim p\big(y|\theta\big) \\ \theta|\epsilon &\sim p\big(\theta|\epsilon\big) \\ \epsilon &\sim p\big(\epsilon\big) \end{aligned}$$

L'inférence bayésienne estime les hyperparamètres par rapport aux distributions conditionnelles des observations alors que l'inférence bayésienne hiérarchique le fait en ajoutant un deuxième niveau de priors. L'objectif est alors de modéliser ces paramètres inconnus. L'inférence bayésienne hiérarchique permet d'introduire beaucoup de niveaux d'hypothèses a priori mais présente en contrepartie un risque de spécification élevé.

2.3 Le modèle mis en œuvre

Le modèle que nous développons fournit un estimateur du risque de chaque région en utilisant leur sinistralité observée ainsi que celle des régions qui leur sont proches. En effet, l'hypothèse de base que nous faisons est que les régions qui sont proches les unes des autres ont plus de chance de présenter un risque similaire que celles qui sont éloignées. Concernant le zonier, cela se traduit par le fait que l'on s'attende à ce que deux régions adjacentes soient de la même couleur, ou proches, de manière à avoir une carte lissée. L'intégration de cette hypothèse dans le modèle mathématique est réalisée par l'utilisation de l'inférence bayésienne hiérarchique décrite plus haut qui permet cette intégration dans la définition de la densité a priori du facteur de risque.

Nous supposons que la région étudiée (la France) soit découpée en n sous-régions de telle sorte que nous disposions du nombre de sinistres observés, y_i, ainsi que de l'exposition au risque, r_i, pour chacune de ces sous-régions. Soit x_i le paramètre inconnu que l'on cherche à estimer représentant le véritable risque dans la région i.

La densité postérieure de x sachant y peut alors s'écrire, selon le théorème de Bayes :

$$p\big(x|y\big) \propto p\big(y|x\big)\ p\big(x\big)$$

où x est le vecteur des x_i et y le vecteur des y_i.

Le niveau de risque inconnu est donc considéré comme une variable aléatoire dont l'espérance sera déduite à partir de la définition de sa densité a posteriori conditionnellement aux données observées. L'utilisation d'un modèle linéaire généralisé, tel que celui présenté dans la partie 2.1, nous permet d'obtenir une estimation de la sinistralité sur la base des facteurs usuels. En notant $\eta = \beta X$ le prédicteur linéaire du modèle, l'utilisation d'une fonction de lien logarithmique implique que la fréquence de sinistres estimée est égale à $\exp(\eta_i)$, et donc que le nombre de sinistres prédit est $e_i = r_i \exp(\eta_i)$, où r_i correspond à l'exposition au risque dans la région i. En intégrant la variation de risque due au critère spatial ainsi que la variation inexpliquée, nous pouvons décomposer le niveau de risque comme suit :

$$\begin{aligned} x_i &= e_i \exp(u_i) \exp(v_i) \\ &= r_i \exp(\eta_i + u_i + v_i) \end{aligned}$$

où

- r_i est une constante connue mesurant l'exposition au risque ;
- η_i est un prédicteur linéaire basé sur les facteurs connus ;
- u_i représente un composant avec une structure spatiale significative ;
- v_i représente la variation inexpliquée.

Les paramètres r_i et η_i étant connus, ou du moins estimés lors d'une première étape, seuls les paramètres u_i et v_i présentent un caractère aléatoire, ce qui nous amène à redéfinir la densité a posteriori de x sachant y :

$$p\big(u, v|y\big) \propto p\big(y|x\big)\ p\big(u, v\big)$$

où u est le vecteur des u_i et v le vecteur des v_i.

L'objectif est alors de déterminer cette densité à partir des observations afin d'évaluer, entre autre, la moyenne des u_i par région qui sera utilisée pour construire le zonier. Cette étape peut être réalisée par la méthode de Monte Carlo qui consiste à simuler un nombre important de réalisations d'une densité afin d'obtenir une estimation de l'espérance à partir de la densité empirique simulée. Pour la déterminer, il convient donc de définir dans un premier temps les expressions de la vraisemblance, $p\big(y|x\big)$, et de la densité a priori, $p\big(u, v\big)$.

2.3.1 La fonction de vraisemblance

Conditionnellement à x_i, nous supposons que les y_i sont mutuellement indépendantes et distribuées selon une loi de Poisson de paramètre x_i :

$$p\big(y_i|x_i\big) = \exp\left(-x_i\right) \frac{x_i^{y_i}}{y_i!} \ , \quad y_i \in \mathbb{N}$$

D'où

$$\begin{aligned} p\big(y|x\big) &= \prod_{i=1}^{n} p\big(y_i|x_i\big) \\ &= \prod_{i=1}^{n} \exp\left(-x_i\right) \frac{x_i^{y_i}}{y_i!} \end{aligned}$$

En effet, y est un échantillon du nombre de sinistres observés dont une distribution de Poisson est appropriée.

2.3.2 Les distributions a priori

Nous avons vu que le niveau de risque dépendait de deux paramètres inconnus, u et v correspondant respectivement à la variation de risque due au critère spatial et la variation inexpliquée par l'ensemble des facteurs du modèle. Ces paramètres étant indépendants, $p\big(u,v\big) = p\big(u\big)\, p\big(v\big)$. Nous cherchons alors à définir séparément leur densité a priori.

v_i étant un bruit blanc dans notre modèle, nous supposons une densité normale a priori pour v_i de moyenne nulle et de variance inconnue λ (hyperparamètre du modèle bayésien hiérarchique) :

$$p\big(v_i|\lambda\big) \propto \lambda^{-\frac{1}{2}} \exp\left(-\frac{1}{2\lambda}\, v_i^2\right)$$

De plus, nous pouvons supposer que les v_i sont mutuellement indépendantes ce qui nous amène à écrire la densité jointe de v comme :

$$\begin{aligned} p\big(v|\lambda\big) &= \prod_{i=1}^{n} p\big(v_i|\lambda\big) \\ &\propto \lambda^{-\frac{n}{2}} \exp\left(-\frac{1}{2\lambda} \sum_{i=1}^{n} v_i^2\right) \end{aligned}$$

La définition de la densité a priori de u_i va nous permettre d'intégrer la notion de voisinage que nous avons supposée pour notre modèle. Nous introduisons donc δ_i comme étant l'ensemble des régions comprises dans le voisinage de i et l'utilisons pour définir la densité conditionnelle a priori de u_i :

$$p\big(u_i|u_{-i},\tau\big) \propto \tau^{-\frac{1}{2}} \exp\left(-\frac{1}{2\tau}\sum_{j\in\delta_i}(u_i-u_j)^2\right)$$

où $u_{-i} = u_1,\ldots,u_{i-1},u_{i+1},\ldots,u_n$.

Cette relation intègre bien le fait que le risque dans la région i dépende que du risque dans les régions avoisinantes ($j \in \delta_i$). De plus, elle modélise un lissage spatial par le fait que plus la valeur de u_i sera éloignée de celle des u_j, plus la valeur de cette expression sera réduite. En d'autres termes, nous associons une probabilité plus faible à des valeurs de risque géographique qui sont éloignées de celles des régions avoisinantes.

En utilisant la définition de la densité conditionnelle :

$$\begin{aligned} p\big(u|\tau\big) &= p\big(u_n|u_1,\ldots,u_{n-1},\tau\big)\, p\big(u_1,\ldots,u_{n-1}|\tau\big) \\ &= p\big(u_n|u_1,\ldots,u_{n-1},\tau\big)\, p\big(u_{n-1}|u_1,\ldots,u_{n-2},\tau\big)\, p\big(u_1,\ldots,u_{n-2}|\tau\big) \\ &= p\big(u_n|u_1,\ldots,u_{n-1},\tau\big)\ \ldots\ p\big(u_2|u_1,\tau\big)\, p\big(u_1|\tau\big), \end{aligned}$$

nous pouvons écrire la densité jointe de u comme :

$$p\big(u|\tau\big) \propto \tau^{-\frac{n}{2}} \exp\left(-\frac{1}{2\tau}\sum_{i\sim j}(u_i-u_j)^2\right)$$

où $i \sim j$ signifie que nous sommons sur l'ensemble des régions voisines, en ne différenciant pas le couple (i,j) du couple (j,i). Ainsi, chaque paire de régions voisines ne sera utilisée qu'une seule fois, comme indiqué dans la définition de la densité conditionnelle.

Ne disposant pas de suffisamment d'informations concernant les hyperparamètres τ et λ, qui déterminent les variances de u et v, nous les considérons à leur tour comme étant des variables aléatoires dont nous devons fixer leur distribution. Un choix approprié pour celle-ci, qui est proche de la distribution non informative habituelle mais qui évite des difficultés techniques, est :

$$p\big(\tau,\lambda\big) \propto \exp\Big(-\frac{\epsilon}{2\tau}-\frac{\epsilon}{2\lambda}\Big)$$

où ϵ est une petite constante positive (nous prendrons $0,01$ dans un premier temps avant d'effectuer des tests de sensibilité).

Le modèle bayésien hiérarchique peut donc se résumer ainsi :

$$\begin{aligned} y|u,v &\sim p\left(y|u,v\right) \\ u|\tau &\sim p\left(u|\tau\right) \\ v|\lambda &\sim p\left(v|\lambda\right) \\ \tau,\lambda &\sim p\left(\tau,\lambda\right) \text{ d'hyperparamètre } \epsilon \text{ fixé} \end{aligned}$$

2.3.3 La distribution a posteriori

Nous avons à présent tous les éléments pour définir la densité a posteriori de notre modèle :

$$\begin{aligned} p\big(u,v,\tau,\lambda|y\big) &\propto p\big(y|u,v,\tau,\lambda\big)\, p\big(u,v,\tau,\lambda\big) \\ &\propto \left[\prod_{i=1}^{n} p\big(y_i|x_i\big)\right] p\big(u|\tau\big)\, p\big(v|\lambda\big)\, p\big(\tau,\lambda\big) \\ &\propto \left[\prod_{i=1}^{n} \exp\left(-x_i\right) \frac{x_i^{y_i}}{y_i!}\right] \tau^{-\frac{n}{2}} \exp\left(-\frac{1}{2\tau} \sum_{i\sim j} (u_i - u_j)^2\right) \\ &\qquad \lambda^{-\frac{n}{2}} \exp\left(-\frac{1}{2\lambda} \sum_{i=1}^{n} v_i^2\right) \exp\Big(-\frac{\epsilon}{2\tau} - \frac{\epsilon}{2\lambda}\Big) \end{aligned}$$

L'estimateur bayésien usuel est l'estimateur dit du maximum a posteriori. Notre objectif est alors de chercher les paramètres qui maximisent cette densité a posteriori et de tirer des conclusions concernant la structure spatiale des données à partir de ceux-ci. Cependant, la maximisation n'est pas réalisable directement en pratique du fait qu'il s'agit d'une fonction non linéaire définie sur plusieurs centaines de variables ($2n + 2$ exactement, avec n le nombre de régions).
De plus, l'utilisation d'un algorithme de type Monte Carlo est également délicat du fait que la définition de la densité ne corresponde à aucune fonction de distribution standard. Nous ne pouvons donc pas générer un échantillon de variables aléatoires indépendantes et identiquement distribuées de la densité a posteriori afin d'obtenir une densité empirique comme le voudrait cette méthode.

La résolution de ce problème nécessite l'utilisation de méthodes dites de Monte Carlo par Chaînes de Markov que nous détaillerons dans le chapitre suivant.

L'approche développée peut se résumer ainsi : en se basant sur la sinistralité de chaque région et en décomposant le risque global en fonction de plusieurs paramètres, nous cherchons la valeur la plus probable de ce dernier dans chaque région. Ayant défini le modèle bayésien, le principal problème est de trouver les paramètres qui maximisent sa densité a posteriori, basée sur un transfert d'informations entre les régions voisines. Nous retiendrons alors la valeur du paramètre représentant le risque géographique (u_i) utilisée.

Problème : il s'agit de la maximisation d'une fonction non linéaire et multivariée empêchant l'utilisation de toutes méthodes d'optimisation classiques. De plus, la simulation d'une densité empirique est difficilement réalisable.

Nous devons donc passer par des méthodes de simulations de types Monte Carlo par Chaînes de Markov.

Chapitre 3

Résolution du modèle bayésien

Nous nous intéressons à présent aux méthodes dites de Monte Carlo par Chaînes de Markov, particulièrement adaptées aux problèmes bayésiens. Celles-ci sont basées sur certaines propriétés spécifiques des chaînes de Markov que nous commencerons par détailler. Nous présenterons alors une méthode en particulier, l'échantillonnage de Gibbs, qui sera employée par la suite.

3.1 Les chaînes de Markov

Nous mentionnons ici quelques notions essentielles de la théorie des chaînes de Markov sur lesquelles repose l'algorithme de Gibbs.

Soit I un ensemble fini ou dénombrable. Tout élément i de I est appelé un état. On dit que la suite $\lambda = (\lambda_i)_{i \in I}$ est une **distribution de probabilité** si, $\forall i \in I$:

- $0 \leq \lambda_i \leq 1$
- $\sum_{i \in I} \lambda_i = 1$

On appelle **matrice stochastique** $P = (p_{i,j})_{(i,j) \in I^2}$ toute matrice carrée doublement indexée sur I, où chaque ligne est une distribution.

Un ensemble de variables aléatoires $\{X_0, X_1, \ldots\}$, à valeur dans I, est une **chaîne de Markov** si, $\forall n \in \mathbb{N}$ et $\forall (x_0, \ldots, x_{n+1}) \in I^{n+2}$:

$$\Pr\big(X_{n+1} = x_{n+1} | X_n = x_n, \ldots, X_0 = x_0\big) = \Pr\big(X_{n+1} = x_{n+1} | X_n = x_n\big)$$

Autrement dit, le futur ne dépend du passé qu'à travers le présent.

Nous nous intéresserons uniquement aux chaînes de Markov dites **homogènes**, où la probabilité de passer d'un état i à un état j ne dépend pas de l'instant n :

$$\Pr\big(X_{n+1} = j | X_n = i\big) = \Pr\big(X_1 = j | X_0 = i\big)$$

Cette probabilité de transition en une étape est notée $p_{i,j}$.

Ces valeurs sont placées dans une matrice stochastique $P = (p_{i,j})_{(i,j)\in I^2}$, appelée **matrice de passage** de $(X_n)_{n\in\mathbb{N}}$.

En outre, la probabilité de passer d'un état i à un état j en n étapes est :

$$\Pr\big(X_n = j | X_0 = i\big) = p_{i,j}^{(n)}$$

Nous notons $\mu^{(n)} = (\mu_i^{(n)})_{i\in I}$ la distribution des états de la chaîne de Markov X après n étapes. Cette distribution correspond à la loi de X_n, avec $\mu_i^{(n)} = \Pr\big(X_n = i\big)$. En particulier, la distribution initiale de X est $\mu^{(0)}$.

Nous avons donc :

$$\begin{aligned} \mu_i^{(n)} &= \Pr\big(X_n = i\big) \\ &= \sum_{j\in I} \mu_j^{(n-1)} p_{j,i} \end{aligned}$$

Soit, sous forme matricielle :

$$\begin{aligned} \mu^{(n)} &= \mu^{(n-1)} P \\ &= (\mu^{(n-2)} P) P \\ &\vdots \\ &= \mu^{(0)} P^n \end{aligned}$$

Autrement dit, la loi de X_n dépend uniquement de la loi initiale de X et de la matrice de passage d'un état à un autre.

Nous nous intéressons alors au comportement de $\mu^{(n)}$, et en particulier nous voulons savoir sous quelles conditions X converge vers un état stable qui ne dépend plus de X_0. Pour cela, nous présentons ici quelques définitions qui nous permettront d'aboutir au théorème central utilisé par la suite. Nous notons $X = (X_n)_{n\in\mathbb{N}}$ une chaîne de Markov de matrice de transition P et prenons $i \in I$ et $j \in I$ deux états quelconques.

- X est **irréductible** si tous les états de I communiquent entre eux, $\forall n, m > 0$:

$$p_{i,j}^{(n)} > 0 \text{ et } p_{j,i}^{(m)} > 0$$

- X est **périodique** s'il existe une périodicité des états (par exemple, on revient au même état toutes les 3 étapes). Sinon elle est **apériodique** (le plus grand commun diviseur des nombres d'étapes possibles pour revenir à l'état initial est 1).

- X est **récurrente positive** si le temps moyen mis pour revenir à l'état i sachant que l'on est parti de i est fini.

- Une chaîne est dite **ergodique** si elle est irréductible, apériodique et récurrente positive.

P étant une matrice stochastique, sa plus grande valeur propre vaut 1 et il existe une distribution π telle que

$$\pi = \pi P$$

π est appelée la **distribution stationnaire** de la chaîne de Markov, elle ne change pas suite au passage d'un état à un autre.

Si P est une chaîne de Markov ergodique, alors sa distribution stationnaire π est unique et P^n converge vers une matrice dont chaque ligne est égale à cette unique distribution. Ainsi, la loi de X_n converge vers la distribution stationnaire.

Théorème ergodique des chaînes de Markov : Soit $(X_n)_{n\in\mathbb{N}}$ une chaîne de Markov ergodique de distribution stationnaire π et f une fonction réelle définie sur l'espace des états I de la chaîne. Alors :

$$\lim_{T\to\infty} \frac{1}{T+1}\sum_{t=0}^{T} f\big(X_t\big) = \sum_{i\in I} \pi_i f(i) = \mathbb{E}_\pi\big(f(x)\big) \quad \text{p.s.}$$

Ainsi, l'espérance de X peut être approchée par la moyenne empirique des états d'une chaîne de Markov ergodique lorsque le nombre d'états est suffisamment grand.

3.2 Les méthodes de Monte Carlo par Chaînes de Markov

Les techniques de Monte Carlo par Chaînes de Markov sont utilisées lorsque la simulation d'une densité est impossible à réaliser en pratique. Bien que leur utilisation soit idéale dans de nombreux cas, leur application est devenue courante que depuis une vingtaine d'années grâce aux progrès informatiques.

Elles offrent une solution basée sur le fait qu'il est plus facile de construire une chaîne de Markov ergodique ayant π comme distribution stationnaire que de générer directement un échantillon depuis π. Cela provient du fait que l'on part d'une chaîne arbitraire dont les états sont ajustés au cours du processus de manière à assurer la stationnarité.

L'approche de Monte Carlo par Chaînes de Markov fournit plus d'informations à l'actuaire que l'estimateur du maximum a posteriori, puisque nous obtenons désormais un échantillon représentatif de la densité a posteriori pour chaque zone géographique. On peut alors estimer u, v, τ et λ par la moyenne de leur densité empirique simulée.

Une méthode basée sur cette technique d'estimation et particulièrement efficace en présence de modèles bayésiens hiérarchiques est l'algorithme de Gibbs.

3.3 L'échantillonnage de Gibbs

L'échantillonnage de Gibbs est une des techniques de Monte Carlo par Chaînes de Markov les plus simples à mettre en œuvre. Elle fût introduite par GEMAN et GEMAN en 1984 dans le domaine de la restauration d'images et a depuis été réutilisée de nombreuses fois dans des domaines très variés.

L'algorithme de Gibbs est basé sur des propriétés particulières des chaînes de Markov énoncées plus haut, et permet d'obtenir des réalisations de la densité a posteriori sans avoir à la calculer. La méthode exploite uniquement l'ensemble des densités conditionnelles et le problème de simulation d'une fonction complexe et multivariée se transforme donc en un problème de simulations de fonctions plus simples et univariées.

Dans notre étude, l'algorithme de Gibbs peut se décrire comme suit. Notons $X = \big(u_1, \ldots, u_n, v_1, \ldots, v_n, \tau, \lambda, y\big)$ un vecteur aléatoire dont certains composants sont des paramètres du modèle, alors que d'autres représentent des données observées. Supposons que sa densité conditionnelle $p\big(u, v, \tau, \lambda | y\big)$ soit tellement compliquée et analytiquement insoluble qu'elle ne permette pas d'obtenir des échantillons indépendants.

Nous utilisons alors l'échantillonnage de Gibbs. Il s'agit d'une méthode itérative où à chaque étape de l'algorithme, la valeur actuelle de chaque paramètre est remplacée par une nouvelle qui est choisie de manière aléatoire depuis sa distribution conditionnelle totale. Tous les autres paramètres sont supposés fixés à leur valeur actuelle durant cette étape.

L'idée principale est de simuler des réalisations d'une chaîne de Markov ergodique qui a $p\big(u, v, \tau, \lambda | y\big)$ comme distribution stationnaire. Les valeurs initiales des paramètres doivent être fournies, elles sont placées dans le vecteur $X^{(0)}$.

Chaque itération procède par tirage au sort depuis les distributions conditionnelles comme suit :

$$\begin{aligned}
\tau^{(t+1)} &\sim p\big(\tau|u^{(t)}, v^{(t)}, \lambda^{(t)}, y\big) \\
\lambda^{(t+1)} &\sim p\big(\lambda|u^{(t)}, v^{(t)}, \tau^{(t+1)}, y\big) \\
u_1^{(t+1)} &\sim p\big(u_1|u_{-1}^{(t)}, v^{(t)}, \tau^{(t+1)}, \lambda^{(t+1)}, y\big) \\
&\vdots \quad \vdots \quad \vdots \\
u_n^{(t+1)} &\sim p\big(u_n|u_{-n}^{(t+1)}, v^{(t)}, \tau^{(t+1)}, \lambda^{(t+1)}, y\big) \\
v_1^{(t+1)} &\sim p\big(v_1|v_{-1}^{(t)}, u^{(t+1)}, \tau^{(t+1)}, \lambda^{(t+1)}, y\big) \\
&\vdots \quad \vdots \quad \vdots \\
v_n^{(t+1)} &\sim p\big(v_n|v_{-n}^{(t+1)}, u^{(t+1)}, \tau^{(t+1)}, \lambda^{(t+1)}, y\big)
\end{aligned}$$

Une étape est terminée lorsque les $2n+2$ valeurs de $X^{(t+1)}$ ont été obtenues. Elles permettent alors de définir un nouvel état de la chaîne de Markov. Nous voyons ici que lorsqu'une nouvelle valeur est créée, elle remplace directement l'ancienne.

Après chaque étape, nous disposons d'un nouveau vecteur de valeurs $X^{(t+1)} = \big(u_1^{(t+1)}, \ldots, u_n^{(t+1)}, v_1^{(t+1)}, \ldots, v_n^{(t+1)}, \tau, \lambda, y\big)$ dépendant uniquement de celui créé à l'étape t.

Habituellement, la chaîne va converger vers la distribution stationnaire qui sera la densité cherchée $p\big(u, v, \tau, \lambda|y\big)$ après quelques milliers d'itérations (cela dépend de l'application et est défini par la pratique). Les valeurs des états suivants permettent alors de construire la densité empirique conditionnelle de X.

La chaîne de Markov ergodique simulée à T états est donc :

$$\big\{(u^{(0)}, v^{(0)}, \tau^{(0)}, \lambda^{(0)}, y), \ldots, (u^{(T)}, v^{(T)}, \tau^{(T)}, \lambda^{(T)}, y)\big\}$$

En supposant que cette chaîne atteigne son état stationnaire après la k^{e} simulation, et en utilisant le théorème ergodique des chaînes de Markov, nous pouvons prendre comme estimateur du risque :

$$\hat{X} = \frac{1}{T-k} \sum_{t=k+1}^{T} X^{(t)}$$

L'avantage d'une telle méthode est qu'elle est bien adaptée aux modèles bayésiens hiérarchiques dont les définitions des priors sont connues. Nous passons de la simulation d'une loi multivariée à des simulations de loi univariées. En revanche, elle nécessite la définition des distributions conditionnelles totales pour tous les paramètres.

3.4 Les densités conditionnelles du modèle

On désire construire une chaîne de Markov dont la distribution stationnaire sera la loi conditionnelle a posteriori des paramètres u, v, τ et λ. Pour cela, il est nécessaire de définir les distributions conditionnelles totales de chaque paramètre.

La densité conditionnelle de u_i est donnée par :

$$\begin{aligned}
p\big(u_i|u_{-i}, v, \tau, \lambda, y\big) &\propto p\big(y_i|x_i\big)\, p\big(u_i|u_{-i}, \tau\big) \\
&\propto \exp(-x_i)\, x_i^{y_i}\, \exp\left(-\frac{1}{2\tau}\sum_{j\in\delta_i}(u_i - u_j)^2\right) \\
&\propto \exp\Big(-e_i\, \mathrm{e}^{(u_i+v_i)}\Big)\, \exp\Big(y_i \log\big(-e_i\, \mathrm{e}^{(u_i+v_i)}\big)\Big)\, \exp\left(-\frac{\#\delta_i}{2\tau}\,(u_i - \overline{u}_j)^2\right) \\
&\propto \exp\Big(-e_i\, \mathrm{e}^{(u_i+v_i)}\Big)\, \exp\Big(y_i \log(-e_i) + y_i(u_i + v_i)\Big)\, \exp\left(-\frac{\#\delta_i}{2\tau}\,(u_i - \overline{u}_j)^2\right) \\
&\propto \exp\left(-e_i \exp(u_i + v_i) + u_i\, y_i - \frac{\#\delta_i}{2\tau}\,(u_i - \overline{u}_j)^2\right)
\end{aligned}$$

où $\#\delta_i$ est le nombre de voisins de i et $\overline{u}_j$ est la moyenne des u_j pour $j \in \delta_i$.

De manière similaire, celle de v_i est de la forme :

$$\begin{aligned}
p\big(v_i|v_{-i}, u, \tau, \lambda, y\big) &\propto p\big(y_i|x_i\big)\, p\big(v_i|\lambda\big) \\
&\propto p\big(y_i|x_i\big) \exp\left(-\frac{1}{2\lambda}v_i^2\right) \\
&\propto \exp\left(-e_i \exp(u_i + v_i) + v_i\, y_i - \frac{1}{2\lambda}v_i^2\right)
\end{aligned}$$

La densité conditionnelle totale de τ est :

$$\begin{aligned} p\big(\tau|u,v,\lambda,y\big) &\propto p\big(u|v,\tau,\lambda,y\big)\, p\big(\tau|v,\lambda,y\big) \\ &\propto \tau^{-\frac{n}{2}} \exp\left(-\frac{1}{2\tau}\sum_{i\sim j}(u_i-u_j)^2\right) \exp\Big(-\frac{\epsilon}{2\tau}-\frac{\epsilon}{2\lambda}\Big) \\ &\propto \tau^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\tau}\left(\epsilon+\sum_{i\sim j}(u_i-u_j)^2\right)\right\} \end{aligned}$$

Et celle de λ est :

$$\begin{aligned} p\big(\lambda|u,v,\tau,y\big) &\propto p\big(v|u,\tau,\lambda,y\big)\, p\big(\lambda|u,\tau,y\big) \\ &\propto \lambda^{-\frac{n}{2}} \exp\left(-\frac{1}{2\lambda}\sum_{i=1}^{n} v_i^2\right) \exp\Big(-\frac{\epsilon}{2\tau}-\frac{\epsilon}{2\lambda}\Big) \\ &\propto \lambda^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\lambda}\left(\epsilon+\sum_{i=1}^{n} v_i^2\right)\right\} \end{aligned}$$

Le problème rencontré ici est que les distributions conditionnelles de u_i et v_i ne sont pas standards et il est donc difficile de simuler des variables aléatoires à partir de celles-ci. Nous utiliserons pour cela la méthode d'*Adaptive Rejection Sampling*.

En revanche, des échantillons provenant des distributions conditionnelles totales des paramètres τ et λ peuvent être obtenus plus facilement. Il suffira de réaliser une simulation à partir d'une loi de khi-deux. En effet, une variable aléatoire X suit une loi de khi-deux à d degrés de liberté si sa fonction de densité est définie par :

$$p\big(x\big) \propto x^{\frac{d}{2}-1} e^{-\frac{x}{2}}$$

En posant $Z = \frac{a}{X}$, la fonction de répartition de Z est :

$$\begin{aligned} F\big(z\big) &= \mathbb{P}\big(Z \leq z\big) \\ &= \mathbb{P}\big(\tfrac{a}{X} \leq z\big) \\ &= \mathbb{P}\big(X \geq \tfrac{a}{z}\big) \\ &\propto \int_{\frac{a}{z}}^{+\infty} x^{\frac{d}{2}-1} e^{-\frac{x}{2}}\, dx \end{aligned}$$

Nous utilisons alors la règle de Leibniz permettant de dériver une intégrale dont les bornes dépendent de la variable de dérivation. En notant :

$$I(z) = \int_{g(z)}^{+\infty} f(x)\,dx$$

où $f(\cdot)$ et $g(\cdot)$ sont dérivables, nous avons :

$$I'(z) = -f\big(g(z)\big)\,g'(z)$$

Nous posons $f(x) = x^{\frac{d}{2}-1}e^{-\frac{x}{2}}$ et $g(z) = \frac{a}{z}$, et en appliquant la règle de Leibniz, nous obtenons :

$$\begin{aligned} f\big(z\big) &= F'\big(z\big) \\ &\propto -\left(\frac{a}{z}\right)^{\frac{d}{2}-1} e^{-\frac{a}{2z}} \left(-\frac{a}{z^2}\right) \\ &\propto \left(\frac{1}{z}\right)^{\frac{d}{2}+1} e^{-\frac{a}{2z}} \end{aligned}$$

Cette densité correspond aux distributions conditionnelles totales de τ et λ, pour autant que les paramètres a et d soient définis de façon convenable.

Pour simuler τ, nous prendrons $d = n-2$ et $a = \epsilon + \sum_{i\sim j}(u_i - u_j)^2$, alors que pour simuler λ nous prendrons $d = n-2$ et $a = \epsilon + \sum_{i=1}^{n} v_i^2$.

Lors de l'estimation de ces deux paramètres, il faudra donc dans un premier temps simuler une réalisation x d'une loi de khi-deux à $n-2$ degrés de liberté puis poser τ et λ comme étant égal à $\dfrac{a}{x}$ où a aura été calculée en conséquence.

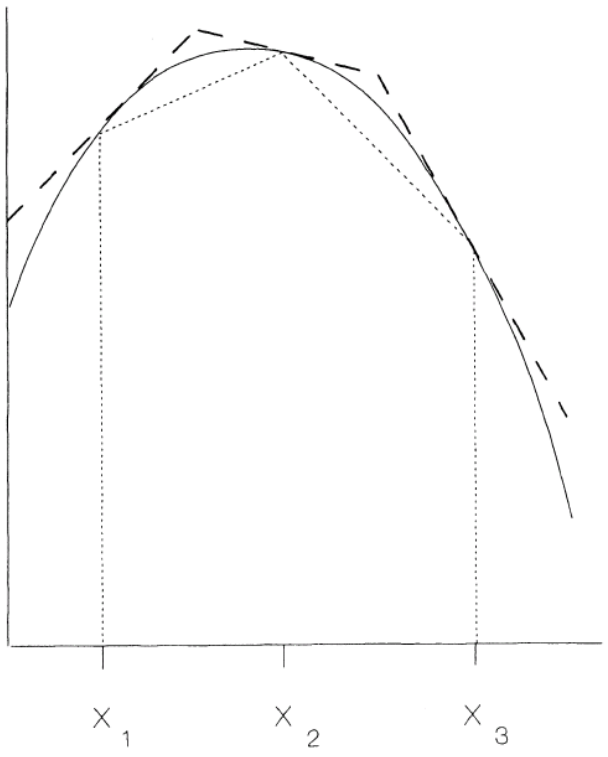
3.5 La méthode d'Adaptive Rejection Sampling

Cette méthode a été proposée par Gilks et Wild [4] en 1992. Elle permet de générer un échantillon de variables aléatoires indépendantes et identiquement distribuées à partir d'une fonction de densité $f(x)$ qui ne présente pas une forme standard, pour autant qu'elle soit log-concave et univariée.

La méthode de rejet adaptatif utilise la log-concativité pour envelopper la fonction de log-densité cible par des enveloppes “supérieure” et “inférieure”. Ces dernières sont ajustées tout au long du processus pour converger vers cette densité.

Posons $h(x) = \log g(x)$, où $g(h)$ est une fonction proportionnelle à la densité que l'on souhaite simuler. Supposons que $h(x)$ ait déjà été évaluée en k points de $T_k = \{x_i : i = 1, \ldots, k\}$.

- L'enveloppe “supérieure” est définie par $h_u(x) = \exp u(x)$, où $u(x)$ est une fonction linéaire par morceaux, formée des tangentes à $h(x)$ aux points de T_k.
- L'enveloppe “inférieure” est définie par $h_l(x) = \exp l(x)$ où $l(x)$ est une fonction linéaire par morceaux, formée des segments de droites reliant les points $h(x_i)$.



Ces fonctions sont représentées sur ce graphique, tiré de l'article cité au début de la section, avec trois valeurs initiales. La courbe pleine représente la fonction $g(x)$ alors que les fonctions “supérieure” et “inférieure” sont respectivement $u(x)$ et $l(x)$. La concativité de $g(x)$ assure le fait que $l(x) \leq g(x) \leq u(x)$ pour tout x.

Concrètement, l'algorithme consiste à simuler tout d'abord une valeur x^* provenant de $h_u(x)$ normalisée, et une valeur w de la loi uniforme entre 0 et 1. La probabilité de garder le x^* qui a été tiré va dépendre de la distance entre les deux fonctions enveloppes.

Test en trois étapes :

1) Si

$$w \leq \frac{h_l(x^*)}{h_u(x^*)}$$

alors x^* est acceptée comme valeur provenant de la densité cible.

2) Sinon, si

$$w \leq \frac{h(x^*)}{h_u(x^*)}$$

alors x^* est acceptée.

3) Sinon x^* est rejetée et on l'ajoute aux points définissant les enveloppes.

Ainsi, au fur et à mesure de l'algorithme, les enveloppes vont se rapprocher et converger vers la log-densité cible, baissant par la même occasion la probabilité de rejeter les prochains points. Le test est effectué jusqu'a l'obtention du nombre de valeurs désiré.

Cette méthode réduit le nombre d'évaluations de la fonction cible lors de la simulation d'une valeur, par rapport aux méthode d'acceptation-rejet classiques. En effet, $f(x)$ étant concave, il n'est pas nécessaire de localiser son maximum. De plus, après chaque rejet les enveloppes sont mises à jour en intégrant les nouvelles informations.

Nous utiliserons cet algorithme afin de simuler une valeur des densité conditionnelles de u_i et v_i. En effet, nous vérifions facilement que qu'il s'agit de fonctions log-concaves :

$$\begin{aligned}
\frac{\partial^2 \log\Big(p\big(u_i|u_{-i},v,\tau,\lambda,y\big)\Big)}{\partial u_i^2} &\propto -e_i \exp(u_i+v_i) - \frac{\#\delta_i}{\tau} < 0 \\
\frac{\partial^2 \log\Big(p\big(v_i|v_{-i},u,\tau,\lambda,y\big)\Big)}{\partial v_i^2} &\propto -e_i \exp(u_i+v_i) - \frac{1}{\lambda} < 0
\end{aligned}$$

car, pour tout i, $e_i \exp(u_i+v_i)$ est positif (nombre de sinistres), $\#\delta_i$ est positif (nombre de voisins) et τ et λ sont positifs (variances).

L'approche développée peut se résumer ainsi :

- observer la sinistralité dans chacune des régions étudiées ;
- estimer leur nombre de sinistres sans l'effet spatial par les modèles linéaires généralisés ;
- décomposer le risque en plusieurs facteurs permettant de l'expliquer (dont le facteur spatial) ;
- déterminer les fonctions de densités conditionnelles de ces facteurs inconnus ainsi que de leur paramètre ;
- simuler successivement les paramètres du modèle (soit à partir d'une loi du khi-deux, soit en utilisant l'algorithme de rejet adaptatif), un grand nombre de fois, en mettant à jour les informations lors de chaque étape selon l'échantillonnage de Gibbs ;
- se baser sur les u_i obtenus pour estimer le risque géographique.

Nous pouvons à présent essayer d'implémenter ce modèle afin d'analyser les résultats puis voir comment les utiliser pour l'élaboration d'un zonier. Il sera intéressant de traiter la problématique du découpage des zones géographiques en zones de risque homogènes sur la base de ces valeurs.

Chapitre 4

Transposition de la théorie à la pratique

Nous disposons à présent de tous les éléments théoriques nécessaires à l'implémentation du modèle sous R. Pour cela, nous commencerons par présenter les données que nous avons à notre disposition. Le but de l'implémentation est d'avoir un résultat interprétable et robuste pour la tarification. Dans cette mesure, nous devons rester vigilant et cherchons également à vérifier la véracité du modèle établi (par exemple la convergence de la chaîne de Markov vers la distribution a posteriori ou encore la concordance des résultats avec nos hypothèses), à travers la validation des hypothèses et le test de sensibilité par rapport aux paramètres. Cela nous permettra de disposer d'un outil informatique efficace et utilisable lors de futures tarifications de produits d'assurance de véhicules.

4.1 Présentation des données

Dans le cadre d'une précédente mission pour l'un de ses clients, FIXAGE a effectué une tarification sur un portefeuille d'assurance de véhicules. Le modèle utilisé lors de cette mission pour créer le zonier sera présenté dans la partie 6.1.1. L'objet de mon mémoire est également de voir si l'utilisation d'un modèle mathématique plus avancé, par rapport à celui utilisé, apporte une réelle amélioration et s'il peut résoudre les problèmes rencontrés.

Nous appliquons donc ici le modèle de tarification par zones géographiques, présenté en première partie, à ce même portefeuille afin de pouvoir comparer les résultats obtenus. Pour des raisons de présentation et d'illustration nous choisissons ici une segmentation par départements, indexée sur i avec $i \in [1, 95]$.

Pour chacun de ces départements, nous disposons du nombre de sinistres observés sur une période de cinq ans, y_i, ainsi que de l'exposition au risque, r_i, à savoir le nombre d'années-police. Par exemple, si un individu est présent sur toute la durée d'observation, son nombre d'années-police est égale à 5 alors que s'il n'était présent que 6 mois, celui-ci sera de 0,5. Les valeurs observées dans chaque département correspondent à la somme des valeurs observées pour chaque individu présent dans le département en question.

La fréquence de sinistres estimée, $\exp(\eta_i)$, est obtenue par l'utilisation des modèles linéaires généralisés sous le logiciel SAS (*PROC GENMOD*). Cette estimation est réalisée sur la base des variables tarifaires sélectionnées par FIXAGE au cours de la précédente tarification, à l'exception du facteur correspondant au risque spatial qui est ignoré. Nous avons donc à notre disposition le nombre de sinistres estimés par département, à savoir $r_i \exp(\eta_i)$.

Le modèle bayésien utilisé nécessitant la définition de voisinage, δ_i, nous cherchons pour chaque département l'ensemble des ses voisins. Nous choisissons de définir un voisin comme étant un département adjacent (frontière commune). Le nombre de voisins varie alors de 2 à 11.

Toutes ces données seront nécessaires à l'implémentation du modèle mathématique sous R, elle sont placées dans un fichier Excel sous la forme suivante :

Dép.	Nb Observé	Nb Estimé	Nb Voisins	Voisin 1	Voisin 2	Voisin 3	...
1	65	89.34	6	74	73	38	
2	28	30.75	7	59	62	80	
3	28	40.37	7	18	58	71	
4	29	38.84	6	5	6	83	
5	10	15.63	4	4	73	38	
⋮							

Notons que l'ordre des voisins n'a pas d'importance et ne respecte pas de logique (le premier voisin n'est pas nécessairement plus proche de i que le second). Cette approche de la méthode peu d'ailleurs être critiquée car on s'attendrait à donner moins d'importance aux voisins plus lointains. La seule logique à respecter est la symétrie, si j est voisin de i alors i doit être voisin de j.

Le fait d'observer un nombre de sinistres supérieur (resp. inférieur) au nombre de sinistres estimés signifie que le risque géographique à un effet positif (resp. négatif) sur le risque global. Sans l'intégration du critère spatial dans le modèle, nous risquons de sous-estimer (resp. surestimer) le risque sous-jacent à ce département. On s'attend donc à ce que le modèle attribue une valeur de u_i positive (resp. négatif) à ce département (facteur multiplicatif du modèle de tarification, $\exp(u_i)$, supérieur (resp. inférieur) à 1). Cependant, l'estimation obtenue ici ne prenant pas encore en compte les données des départements voisins, il se peut que les effets s'annulent si ces derniers présentent un effet en sens inverse.

Nous voulons que l'outil créé fournisse en sortie les valeurs des u_i et v_i pour chaque département afin d'élaborer un zonier et de l'incorporer dans le système de tarification. Nous choisissons d'implémenter le modèle sous le logiciel R, environnement mathématique utilisé pour le traitement de données et l'analyse statistique, qui présente de nombreux avantages :

- simulations performantes de lois ;
- multiples représentations graphiques (dont cartographie) ;
- plate-forme multi OS (FIXAGE travaillant sous Mac) ;
- gratuité ;
- forte communauté d'universitaires, ...

4.2 Etapes de l'implémentation sous R

Nous commençons par charger le fichier Excel contenant les données dans une matrice D de taille n x 15, où n est égale au nombre de départements soit 95 et 15 correspond au nombre de colonnes du fichier source.

Nous devons ensuite initialiser les paramètres (u, v, τ et λ) de manière à pouvoir simuler le premier état de la chaîne. Les valeurs choisies lors de l'initialisation n'ont pas une grande importance étant donnée que la chaîne de Markov générée convergera vers sa distribution stationnaire, pour un nombre significatif de simulations. Nous choisissons tout de même des valeurs du même ordre de grandeur que les valeurs attendues, à savoir que nous initialisons tous les paramètres à 0 sauf *epsilon* à 0,01, dont la valeur n'évoluera pas au cours de l'algorithme. Des tests de sensibilité seront effectués par la suite permettant de justifier ces choix.

Nous souhaitons garder en mémoire les valeurs des états simulés de la chaîne de Markov afin de constater et de dater sa convergence éventuelle vers sa distribution stationnaire. Nous prenons $N = 10\,000$ simulations et créons :

- U, une matrice nulle de taille n x N
- V, une matrice nulle de taille n x N
- tau, un vecteur ligne nul de taille N
- $lambda$, un vecteur ligne nul de taille N

Ensuite, nous récupérons les données du fichier initial :

- $y = D[\quad, (c-3)]$ nombre de sinistres observés, vecteur colonne de taille n
- $e = D[\quad, (c-2)]$ nombre de sinistres estimés sans la variable zone, vecteur colonne de taille n
- $nb = D[\quad, (c-1)]$ nombre de voisins, vecteur colonne de taille n

où c est le numéro de la colonne contenant le premier voisin.

Ces valeurs permettront d'alléger le code lors du passage de l'algorithme à la codification.

L'algorithme implémenté peut être schématisé comme suit :

Pour chaque état de la chaîne (k allant de 2 à N) faire :

1. Récupérer les valeurs générées à l'étape précédente

$$\begin{aligned} U[\ ,k] &= U[\ ,k-1] \\ V[\ ,k] &= V[\ ,k-1] \end{aligned}$$

2. Simuler $\tau^{(k)}$ selon

$$p\left(\tau^{(k)}|u^{(k)},v^{(k)},\lambda^{(k)},y\right) \propto \left(\tau^{(k)}\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\tau^{(k)}}\left(\epsilon+\sum_{i\sim j}\left(u_i^{(k)}-u_j^{(k)}\right)^2\right)\right\}$$

3. Simuler $\lambda^{(k)}$ selon

$$p\left(\lambda^{(k)}|u^{(k)},v^{(k)},\tau^{(k)},y\right) \propto \left(\lambda^{(k)}\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\lambda^{(k)}}\left(\epsilon+\sum_{i=1}^{n}\left(v_i^{(k)}\right)^2\right)\right\}$$

4. Pour i allant de 1 à n, simuler $u_i^{(k)}$ selon

$$p\left(u_i^{(k)}|u_{-i}^{(k)},v^{(k)},\tau^{(k)},\lambda^{(k)},y\right) \propto \exp\left(-e_i\,\mathrm{e}^{\left(u_i^{(k)}+v_i^{(k)}\right)}+u_i^{(k)}\,y_i-\frac{\#\delta_i}{2\tau^{(k)}}\left(u_i^{(k)}-\overline{u}_j^{(k)}\right)^2\right)$$

5. Pour i allant de 1 à n, simuler $v_i^{(k)}$ selon

$$p\left(v_i^{(k)}|v_{-i}^{(k)},u^{(k)},\tau^{(k)},\lambda^{(k)},y\right) \propto \exp\left(-e_i\,\mathrm{e}^{\left(u_i^{(k)}+v_i^{(k)}\right)}+v_i^{(k)}\,y_i-\frac{1}{2\lambda^{(k)}}\left(v_i^{(k)}\right)^2\right)$$

Un autre avantage de R est que la simulation selon l'algorithme de rejet adaptatif est déjà implémentée dans le package *Runuran*. Nous utilisons la fonction *ars.new*$(\cdot)$ qui prend comme paramètres le logarithme de la fonction de densité conditionnelle de u_i (ou v_i) ainsi que la dérivée de celle-ci. Cette fonction créée un objet dont l'utilisation conjointe avec la fonction $ur(ars.new(\cdot))$ permet d'obtenir une valeur simulée depuis la densité cherchée. Il ne reste alors plus qu'à calculer en amont la moyenne des u_j $(j \in \delta_i)$ de manière à pouvoir l'intégrer dans la définition du logarithme de la densité de u_i.

De même, les valeurs de τ et λ sont obtenues en simulant une réalisation d'une loi de chi-deux à $n-2$ degrés de liberté en utilisant la fonction $urchisq(\cdot)$ présente dans ce même package. Nous constatons ici le réel avantage dû à l'utilisation du logiciel R. En revanche, les packages étant en open source, ils présentent un risque d'erreur. Nous nous sommes donc contentés de les utiliser au minimum en faisant appel uniquement aux fonctions permettant la simulation et en programmant tout le reste.

Nous disposons à présent d'un outil informatique répondant à l'approche théorique initialement énoncée. Nous devons en analyser les résultats, voir s'ils sont conformes à nos attentes et si les hypothèses peuvent être validées. Enfin, nous devons étudier différentes méthodes de regroupement des valeurs en classes permettant d'aboutir à la définition de notre zonier.

Chapitre 5

Présentation des résultats

L'outil informatique élaboré permet de récupérer en outputs les matrices U et V remplies, avec dans chaque ligne les 10 000 valeurs simulées pour chaque département. La première chose que nous souhaitons vérifier est alors la convergence de la chaîne de Markov générée vers sa distribution stationnaire. Il faut ensuite retenir un critère permettant de définir qu'elle valeur retenir comme estimateur de risque par département avant de regarder comment obtenir un zonier à partir ce celles-ci.

5.1 Convergence vers la distribution stationnaire

Nous souhaitons évidemment constater si la chaîne de Markov simulée converge bien. Nous avons donc gardé en mémoire les 10 000 états simulés de u, v, τ et λ pour les 95 départements. Il est toutefois difficile de savoir quand la chaîne de Markov a atteint son état stationnaire. Nous basons alors notre analyse sur les moments d'ordre de u et en particulier sa moyenne. En effet, nous souhaitons uniquement obtenir une estimation des paramètres, or une estimation de l'espérance de la densité simulée est plus stable que son élément le plus probable.

L'évolution de la moyenne empirique au fil des états construits pour les départements 75 et 67 est représentée sur les FIGURES 5.1 et 5.2 qui font partie respectivement des départements les plus et les moins risqués. Nous pouvons constater la convergence de la moyenne des états générés au vu de ces graphiques, mais elle reste délicate à dater avec précision.

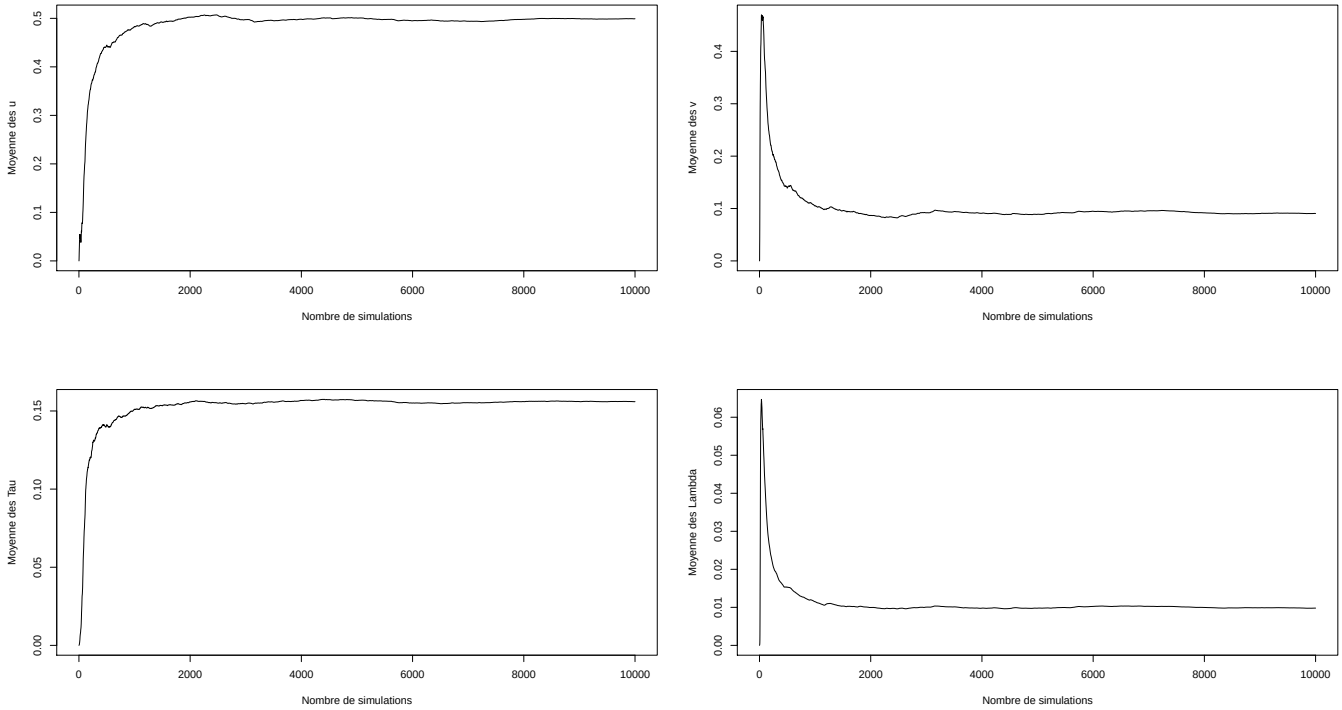


FIGURE 5.1 – Moyennes des valeurs générées au fil des états pour le département 75

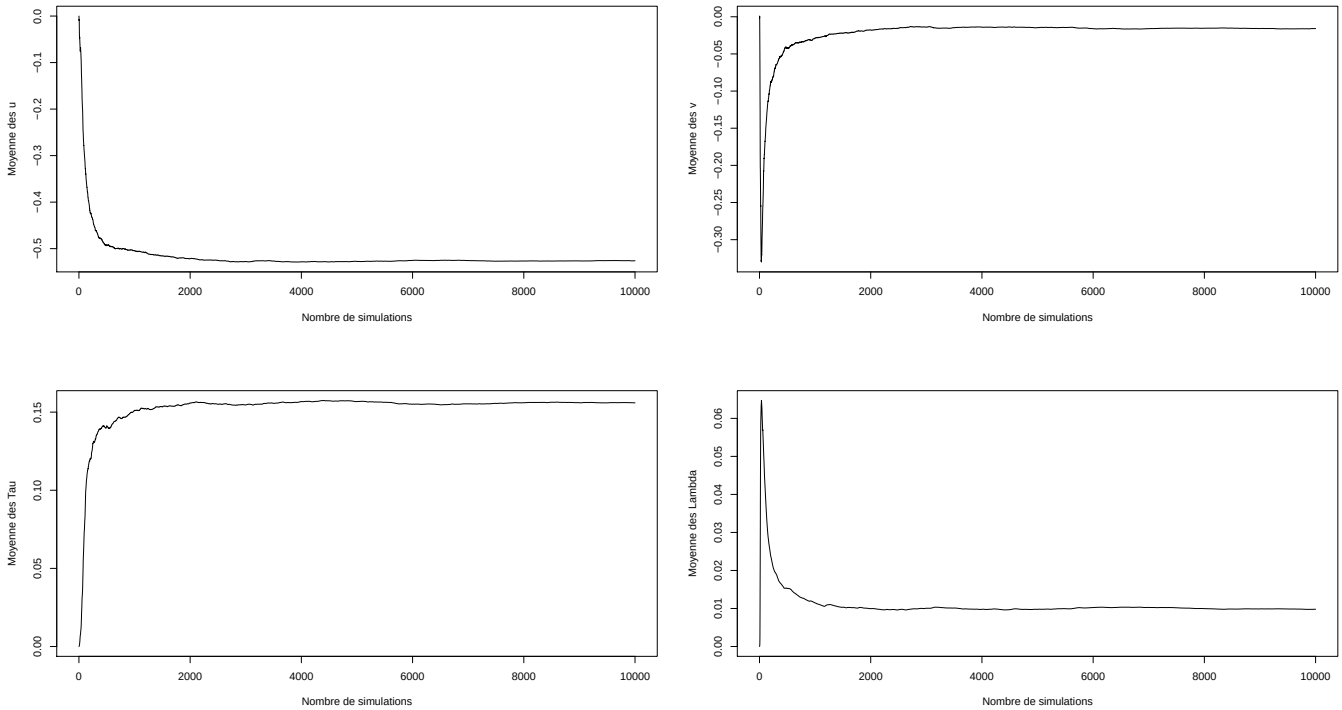


FIGURE 5.2 – Moyennes des valeurs générées au fil des états pour le département 67

Pour ce faire, nous lançons plusieurs fois la simulation, mais en changeant les valeurs initiales des paramètres. Nous avions initialement pris 0. Nous comparons les résultats avec 0.5 et -0.5 correspondant environ aux valeurs extrêmes de u obtenues lors de la première simulation ainsi que -0.2 qui correspond à la moyenne. Les résultats sont une nouvelle fois illustrés pour les départements 75 (FIGURE 5.3) et 67 (FIGURE 5.4).

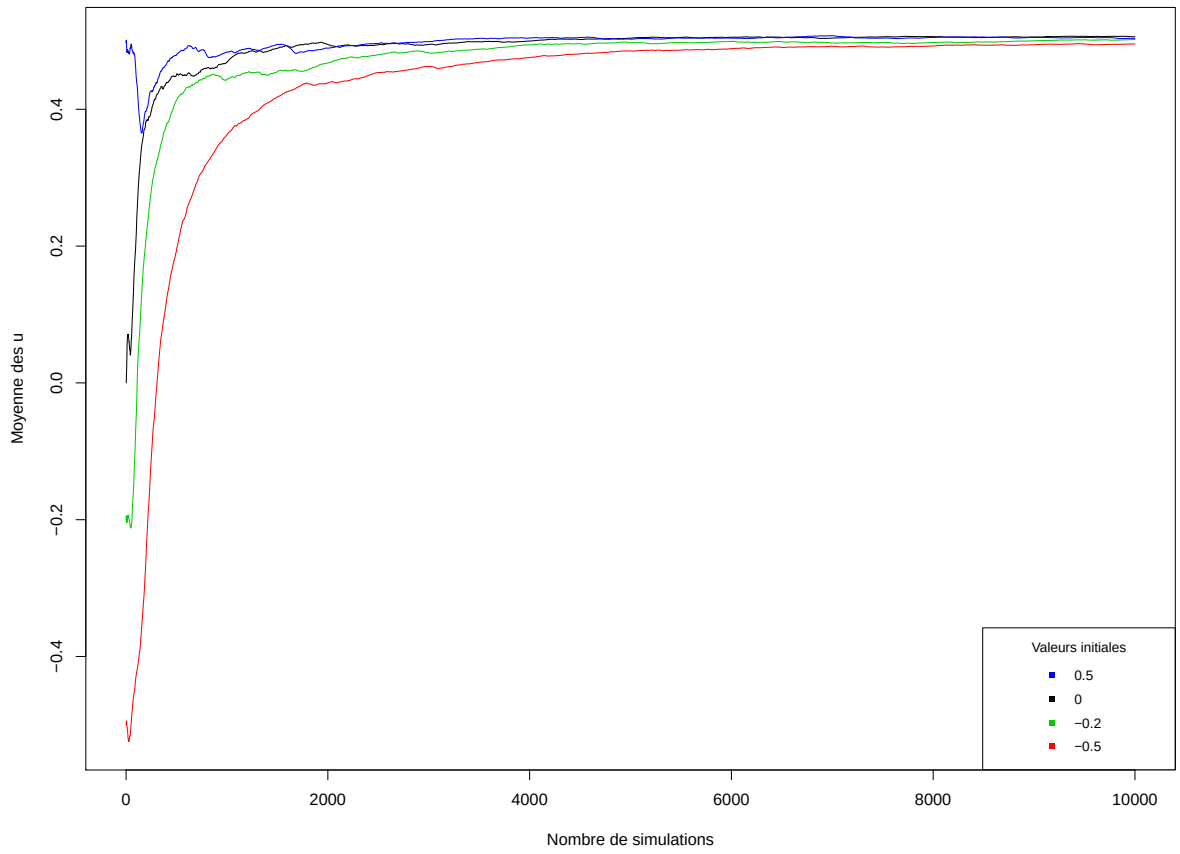


FIGURE 5.3 – Moyenne des u_i générées au fil des états pour le département 75

La lecture de ces graphiques permet de confirmer la convergence puisque quelles que soient les valeurs initiales des paramètres (en restant cohérent avec l'ordre de grandeur), les moyennes des valeurs simulées aboutissent aux mêmes résultats. Nous pouvons cependant noter que le fait d'initialiser les paramètres en prenant une valeur éloignée de l'estimateur attendu ralenti la convergence (courbe rouge pour le département 75 et bleue pour le département 67). Nous choisissons de conserver une initialisation des paramètres à 0 qui correspond au milieu de la plage des u_i obtenus et qui permet une convergence plus rapide sur l'ensemble des départements. Pour rester prudent, nous disons que la chaîne de Markov atteint sa distribution stationnaire après environ 5 000

états (période dite de “ *burn in* ”) et supposons que les u_i générés après la 5 000^{e} simulation peuvent être vus comme un échantillon provenant de la densité a posteriori de u_i.

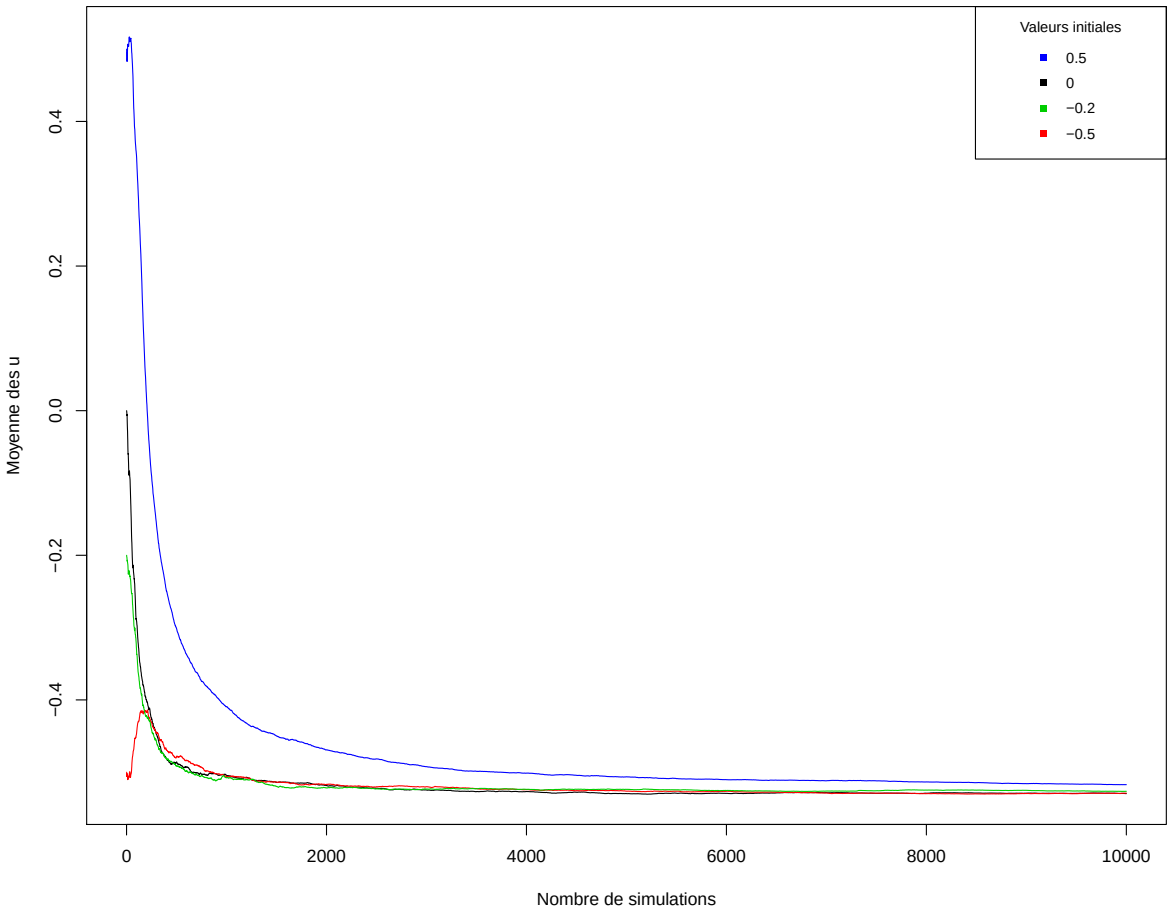
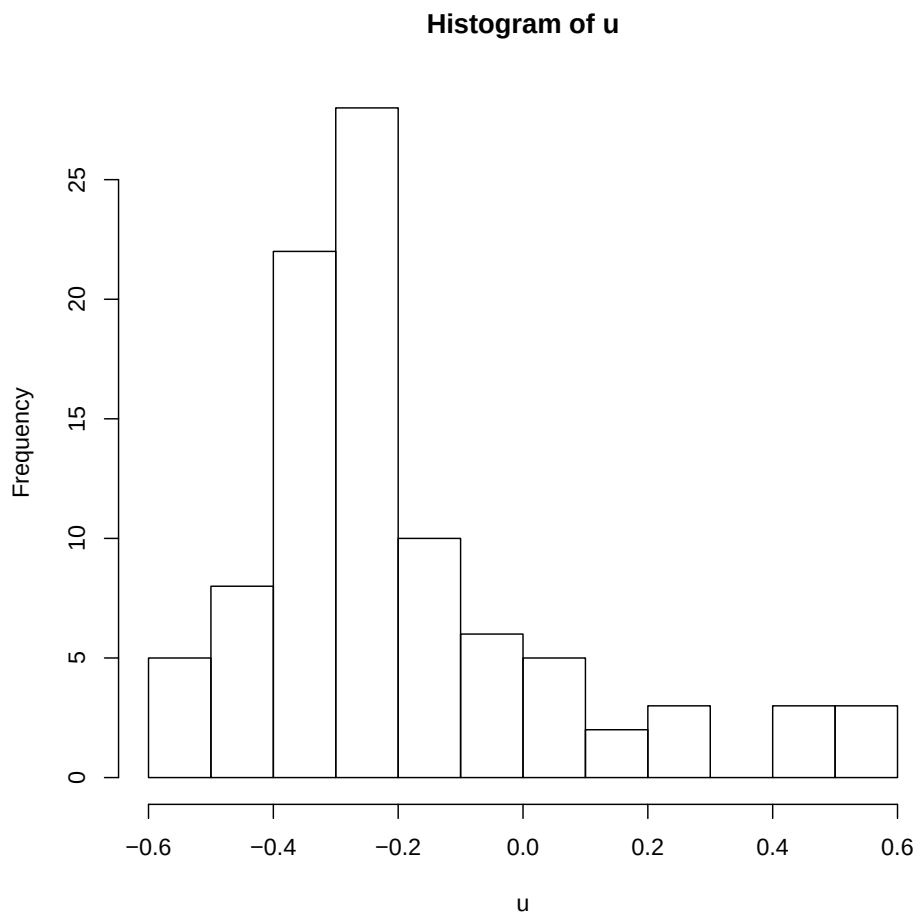


FIGURE 5.4 – Moyenne des u_i générées au fil des états pour le département 67

5.2 Valeurs retenues

Sur la base de cette convergence, la première partie de la simulation est rejetée et nous décidons de prendre comme estimateur de chacun des paramètres la moyenne des 5 000 dernières valeurs simulées dans le but de réduire l'effet des valeurs de départ. Par ailleurs, nous pouvons signaler que nous n'avons pas besoin d'une très grande précision dans les valeurs retenues, étant donnée qu'elles ne servent qu'à ordonner les départements, nous les regrouperons finalement en classes tarifaires.

Résumé des u_i retenus :

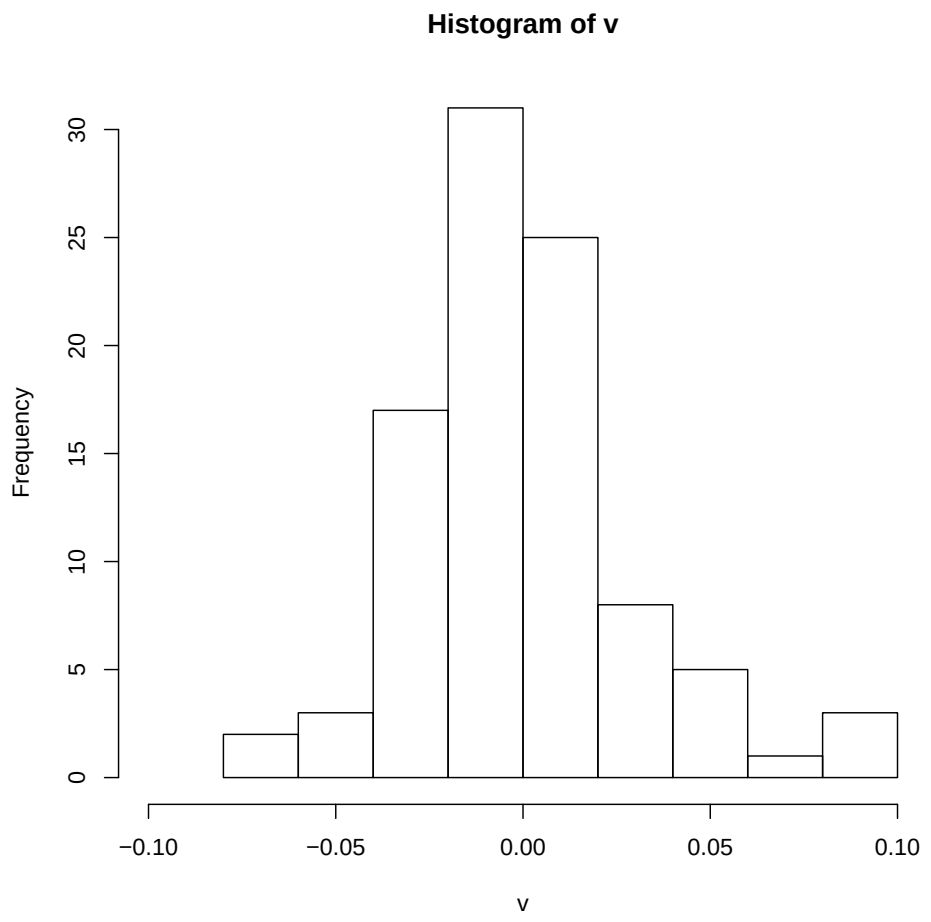


Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.5604	-0.3422	-0.2528	-0.1936	-0.1392	0.5412

et $\tau = 0.1594$

Nous pouvons constater ici la dominance de u par rapport à v, ce qui conforte notre hypothèse concernant l'existence d'une structure spatiale du risque. Celle-ci sera d'autant plus mise en relief lors de la cartographie des u_i.

Résumé des v_i retenus :



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.076570	-0.017630	-0.002186	-0.000634	0.013390	0.086840

et $\lambda = 0.00917$

Parmi les 95 estimations des u_i retenues, seul un département obtient un facteur de risque de signe opposé, mais très proche de zéro tout de même, par rapport au signe que nous attendions. Tous les autres ont bien un facteur de risque positif (ou négatif) là où nous avions estimé un nombre de sinistres inférieur (ou supérieur) au nombre de sinistres observés.

Disposant d'estimateurs comparables pour chacun des départements, nous pouvons à présent élaborer notre zonier.

5.3 Utilisation des résultats pour créer un zonier

L'outil créé permet d'associer à chaque département une estimation a posteriori de son risque sous-jacent. Les valeurs obtenues peuvent être comparées entre elles afin de classer les départements par ordre de risque. Cependant, ne pouvant effectuer une tarification propre à chaque département, nous souhaitons les regrouper en fonction de leur risque en un nombre de classes réduit. Ce regroupement est alors effectué au travers d'un zonier.

Là encore nous rencontrons plusieurs problèmes, à commencer par le choix du nombre de classes tarifaires à utiliser. Il faut suffisamment de zones pour que la tarification soit efficace sans pour autant considérer un nombre trop important en raison de la perte de l'effet de mutualisation par segment. Nous voulons que deux régions de la même zone de risque soient le plus similaires possible et que deux régions de deux zones de risque différentes soient le plus différenciées possible. Nous présentons ici trois méthodes pouvant être utilisées pour découper la plage des u_i en un nombre de classes déterminé.

5.3.1 Découpage par bandes régulières

Une première méthode consiste simplement à découper la plage des valeurs des u_i obtenues en z zones de longueur identique.

En notant cette longueur :

$$d = \frac{(u_{max} - u_{min})}{z}$$

nous pouvons définir les z classes par :

$$\begin{aligned} c_1 &= \big[u_{min}\ ;\ u_{min} + d\big[\\ &\vdots \\ c_i &= \big[u_{min} + (i-1)\,d\ ;\ u_{min} + i\,d\big[\\ &\vdots \\ c_z &= \big[u_{min} + (z-1)\,d\ ;\ u_{max}\big] \end{aligned}$$

Il ne reste alors plus qu'à associer à chaque département sa zone de risque par rapport à la valeur de son u_i.

Comme nous pouvons le voir, cette méthode de découpage est très facile à mettre en place et est aussi la plus intuitive. Le problème engendré par celle-ci est que nous risquons très probablement de créer des classes disproportionnées. En effet, étant de même longueur, les zones extrêmes seront composées de très peu de départements alors que les zones moyennes seront surreprésentées.

5.3.2 Découpage par quantiles

Pour répondre à ce problème d'effectif des classes, le plus simple semble être de former des zones de risque comportant exactement le même nombre de départements. Ainsi, chaque zone de risque sera composée de

$$a = \frac{n}{z}$$

éléments (à un élément près, dû aux arrondis).

Les classes sont donc définies par :

$$\begin{aligned} c_1 &= \left[u_{(1)} \ ; \ u_{(\lfloor a \rfloor)}\right] \\ &\vdots \\ c_i &= \left[u_{(\lfloor (i-1)\, a \rfloor + 1)} \ ; \ u_{(\lfloor i\, a \rfloor)}\right] \\ &\vdots \\ c_z &= \left[u_{(\lfloor (z-1)\, a \rfloor + 1)} \ ; \ u_{(n)}\right] \end{aligned}$$

où $\lfloor \cdot \rfloor$ est la fonction partie entière et $u_{(.)}$ est la statistique d'ordre $\left(u_{(1)} \leq \cdots \leq u_{(n)}\right)$.

Cette méthode est également très facile à mettre en place (en utilisant la fonction *rank*$(\cdot)$ sous R) mais, à l'inverse de la précédente, elle ne tient pas compte de la dispersion des u_i. Une même classe peut contenir deux u_i très différents (surtout les classes extrêmes), autrement dit, certaines classes seront définies sur un intervalle très petit alors que d'autres le seront sur un intervalle pouvant être beaucoup plus grand. Ce qui peut nous amener à regrouper des départements présentant des risques peu similaires.

5.3.3 Découpage par classification hiérarchique ascendante

La classification (clustering) est une technique empruntée à l'analyse de données visant à partitionner un ensemble de données en plusieurs classes homogènes. Parmi ces méthodes, la classification hiérarchique ascendante consiste à partir de la classification la plus fine (n classes) puis d'effectuer des regroupements jusqu'à l'obtention d'une seule classe contenant tous les départements. Les étapes successives regroupent à chaque fois les deux départements les plus proches par rapport à leur niveau de risque géographique (c'est-à-dire les deux qui ont les u_i les plus proches).

Une classification en k classes $G_1, \ldots, G_k$ des u_i, de moyenne respective $\bar{G}_1, \ldots, \bar{G}_k$, est caractérisée par :

- son inertie totale : $\sum_{i=1}^{n} d^2(u_i; \bar{u})$
- son inertie inter-classes : $\sum_{i=1}^{k} n_i \, d^2\big(\bar{G}_i; \bar{u}\big)$
- son inertie intra-classes : $\sum_{i=1}^{k} \sum_{j \in G_k} d^2\big(u_j; \bar{G}_i\big)$

où $d^2(\cdot;\cdot)$ est la distance euclidienne entre deux points.

L'inertie totale est égale à la somme entre l'inertie inter-classes et l'inertie intra-classes. Si l'inertie totale est constante tout au long du processus de classification, il est facile de comprendre que l'inertie inter-classes diminue (il y a de moins en moins de classes) alors que l'inertie intra-classes augmente. La qualité de la classification en k classes est mesurée par la proportion de l'inertie totale expliquée par l'inertie inter-classes. On cherche donc à minimiser la perte d'inertie inter-classes lors de chaque regroupement. Or, la diminution de l'inertie inter-classes (et donc l'augmentation de l'inertie intra-classes) suite au regroupement de deux classes peut être mesurée par la distance de Ward.

Soient deux classes $A = (u_i, i = a_1, \ldots, a_{n_A})$ et $B = (u_{i'}, i' = b_1, \ldots, b_{n_B})$. La distance de Ward entre ces deux classes est définie par :

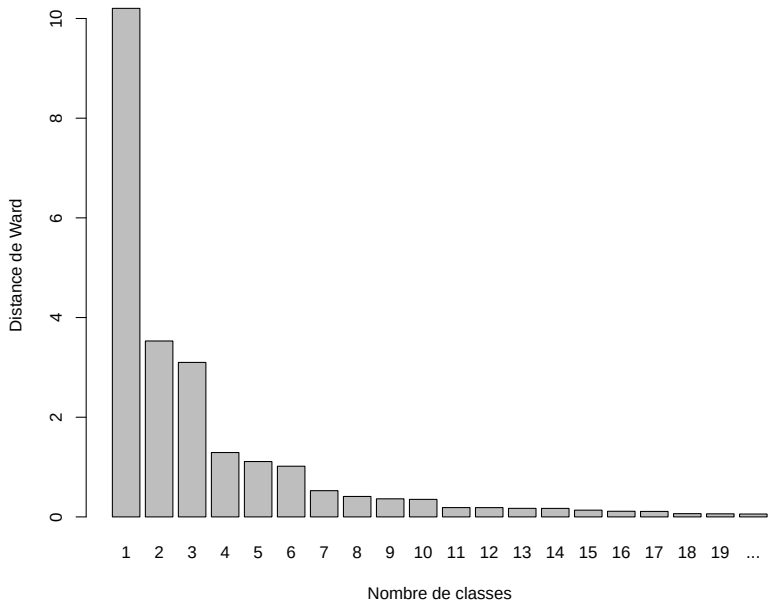
$$D_{Ward}\,(A; B) = \frac{n_A \, n_B}{n_A + n_B} \, d^2\big(\bar{A}; \bar{B}\big)$$

A chaque étape, le regroupement est donc effectué entre les deux classes qui minimisent le critère de Ward, c'est-à-dire tel que l'inertie inter-classes diminue le moins possible.

Les étapes sont représentées à travers un arbre appelé dendrogramme dont les sections permettent de définir les partitions. Il sera optimal de couper le dendrogramme à un niveau où le regroupement entre classes entraîne une perte d'inertie intra-classes importante.

5.3.4 Choix d'un zonier

Avant toutes choses, nous devons choisir le nombre de zones de risques que nous allons utiliser afin d'appliquer, et de comparer, ces trois méthodes de construction de zonier. Pour ce faire, nous étudions la qualité de la classification suite aux regroupements successifs par la méthode de classification hiérarchique ascendante.



Nous représentons ici les distances de Ward entre les deux classes dont le regroupement aboutit à une classification en k classes. Ainsi, la distance entre les deux dernières classes de l'algorithme (correspondant au 94^{e} regroupement) est de $10,20$. Le nombre de zones finalement choisi est un nombre pour lequel nous avons une chute importante de la distance de Ward (par exemple : 2, 4, 7 ou encore 11). En effet, cela signifie que

le passage de 1 zone à 2, par exemple, permet d'augmenter significativement l'inertie inter-groupe, or nous voulons que celle-ci soit maximale. Nous ne pouvons cependant pas non plus prendre un nombre de classes trop important car sinon la classification n'aurait plus de sens (nous voulons regrouper les départements en quelques zones de risque seulement).

Finalement, au vu de la FIGURE 5.5, nous décidons de regrouper l'ensemble des départements en sept zones de risque. Cette typologie explique $84,58\%$ de l'inertie totale. Quatre zones semble être trop petit et n'explique que $70,30\%$ de l'inertie totale.

Le dendrogramme obtenu à l'aide de la fonction *hclust* du package *cluster* sous R est présenté en FIGURE 5.6.

Nombre de classes	Inertie intra-classes	Inertie inter-classes	% de l'inertie totale expliquée	Distance de Ward
95	0.000	23.945	100.000	
94	0.000	23.944	100.000	0.000
⋮	⋮	⋮	⋮	⋮
12	1.854	22.090	92.256	0.186
11	2.041	21.903	91.475	0.187
10	2.393	21.552	90.006	0.352
9	2.756	21.189	88.490	0.363
8	3.166	20.778	86.777	0.410
7	3.692	20.252	84.580	0.526
6	4.709	19.235	80.332	1.017
5	5.820	18.124	75.693	1.111
4	7.112	16.833	70.300	1.291
3	10.211	13.733	57.355	3.100
2	13.741	10.204	42.614	3.530
1	23.945	0.000	0.000	10.204

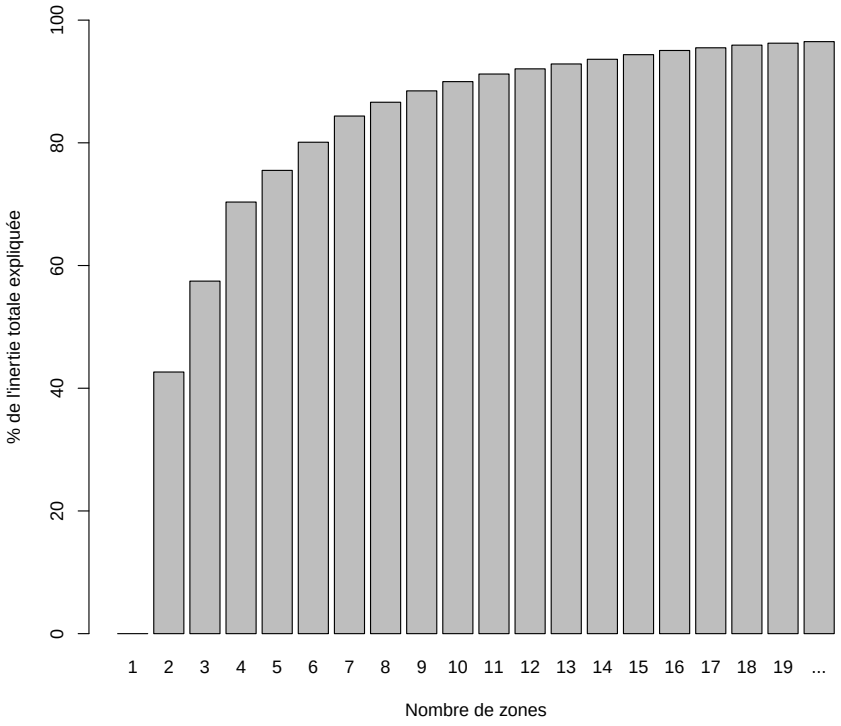


FIGURE 5.5 – Qualité de la classification

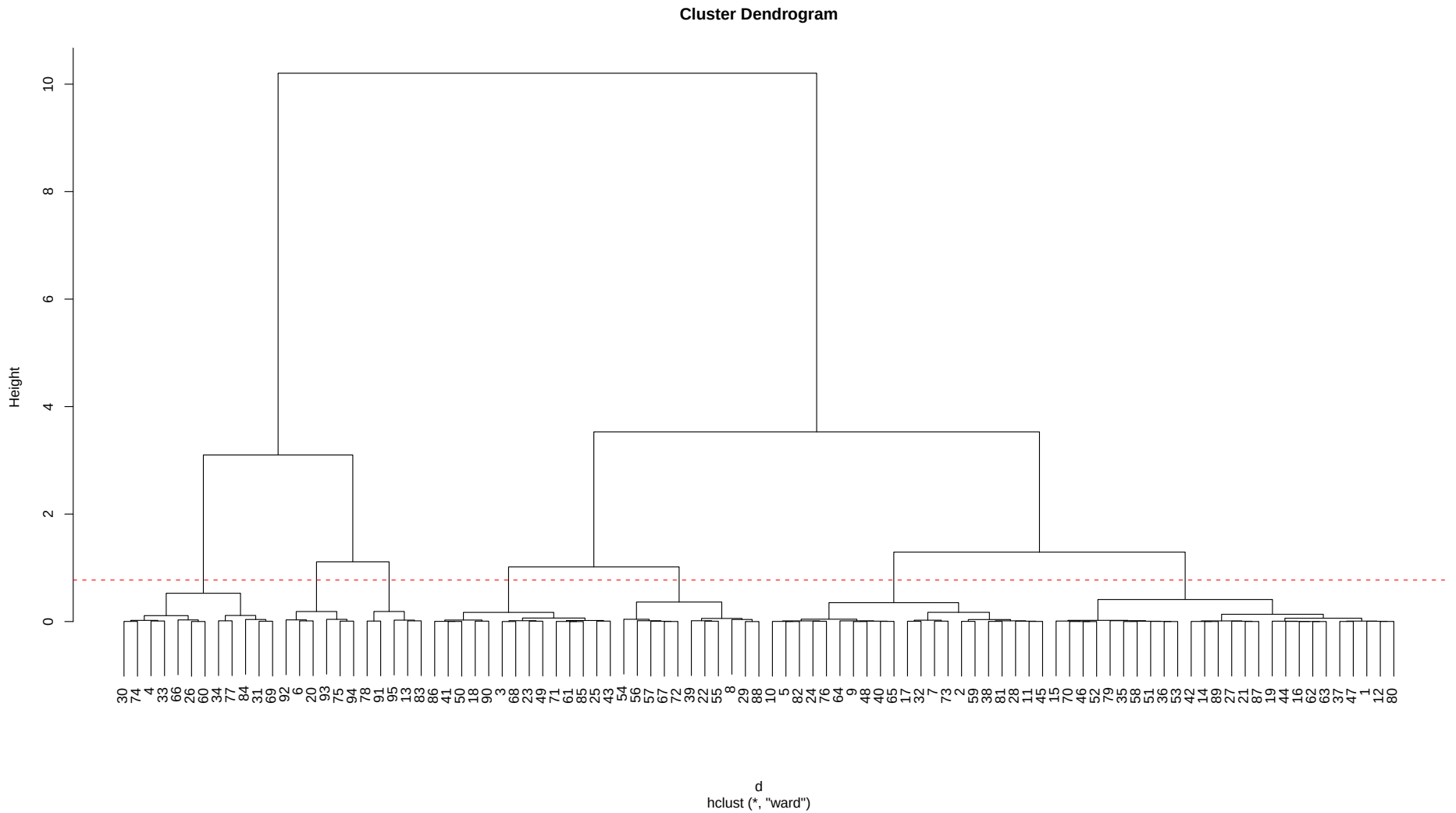


FIGURE 5.6 – Dendrogramme

Une fois le nombre de zones de risque fixé, nous pouvons appliquer les trois méthodes de découpage proposées, aux résultats obtenus avec l'outil d'estimation du risque géographique. Il n'est pas évident de comparer à première vue les résultats. Nous représentons donc chacun des zoniers sous forme de carte géographique. Cette analyse nous permet dès lors de déceler la structure spatiale du risque que nous supposions exister. Les régions du Sud-Est ainsi que l'Ile-de-France semblent globalement présenter un risque plus important sur le portefeuille étudié. Les différentes cartes sont réalisées en utilisant la fonction *map* de R. Si la première carte (découpage par bandes régulières) semble présenter des résultats un peu trop lissés (pas beaucoup de forts risques), la seconde (découpage par quantiles) ne lisse pas suffisamment nos résultats et on aperçoit de nombreux sauts de zones entre des départements voisins.

Afin de compléter notre analyse, nous représentons la courbe des $u_{(i)}$ de manière à constater d'éventuels groupes homogènes de risque et de voir comment chacune de ces méthodes les a regroupés. Nous constatons que quasiment tous les départements présentant un u_i positif se retrouvent dans la même classe, avec un découpage par quantiles, alors que les deux autres découpages semblent bien différencier les petits paquets de points que l'on peut apercevoir.

Enfin, un histogramme présentant le nombre de départements placés dans chaque zone de risque est présenté de manière à voir si les méthodes utilisées concordent avec la distribution des u_i générés.

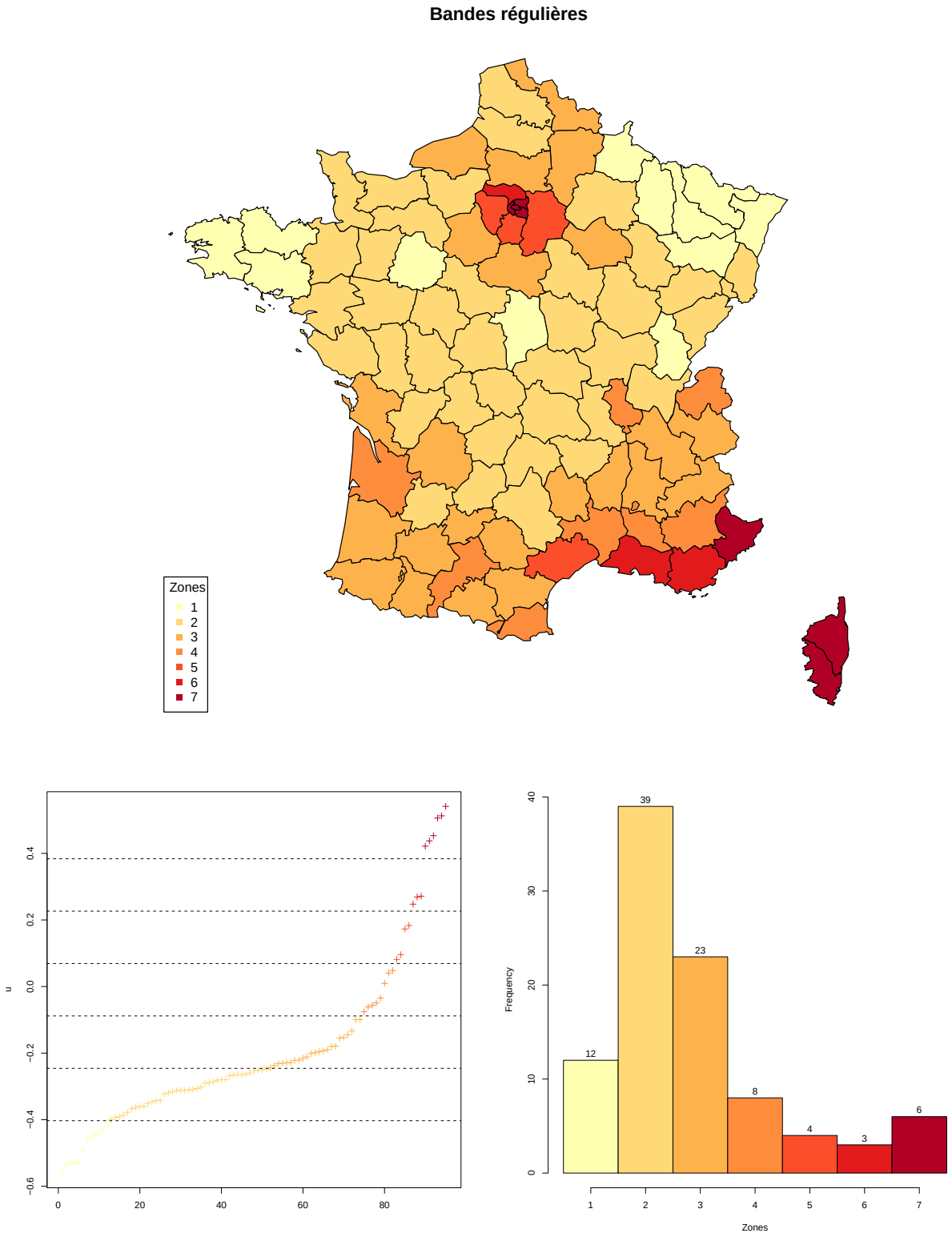


FIGURE 5.7 – Découpage par bandes régulières

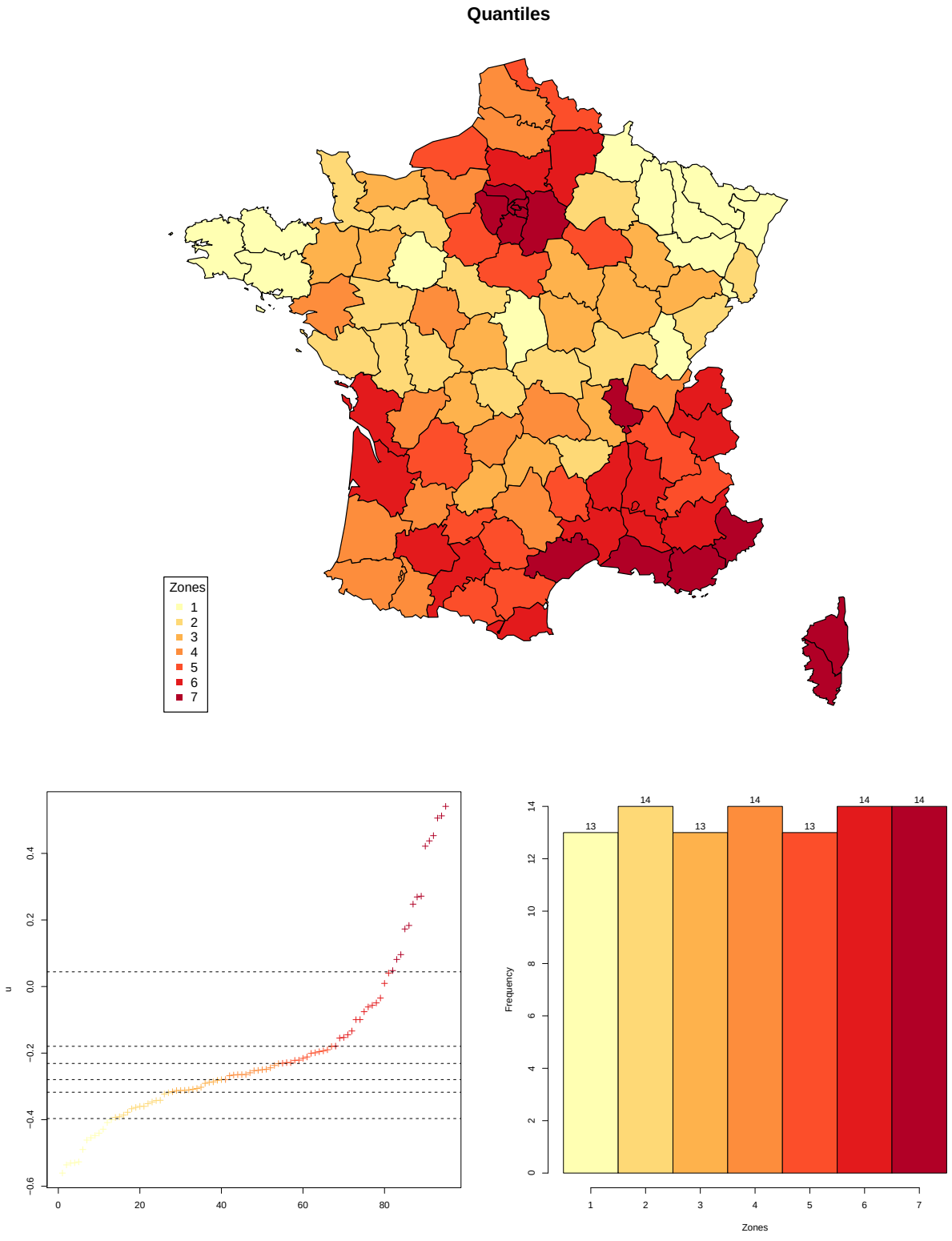


FIGURE 5.8 – Découpage par quantiles

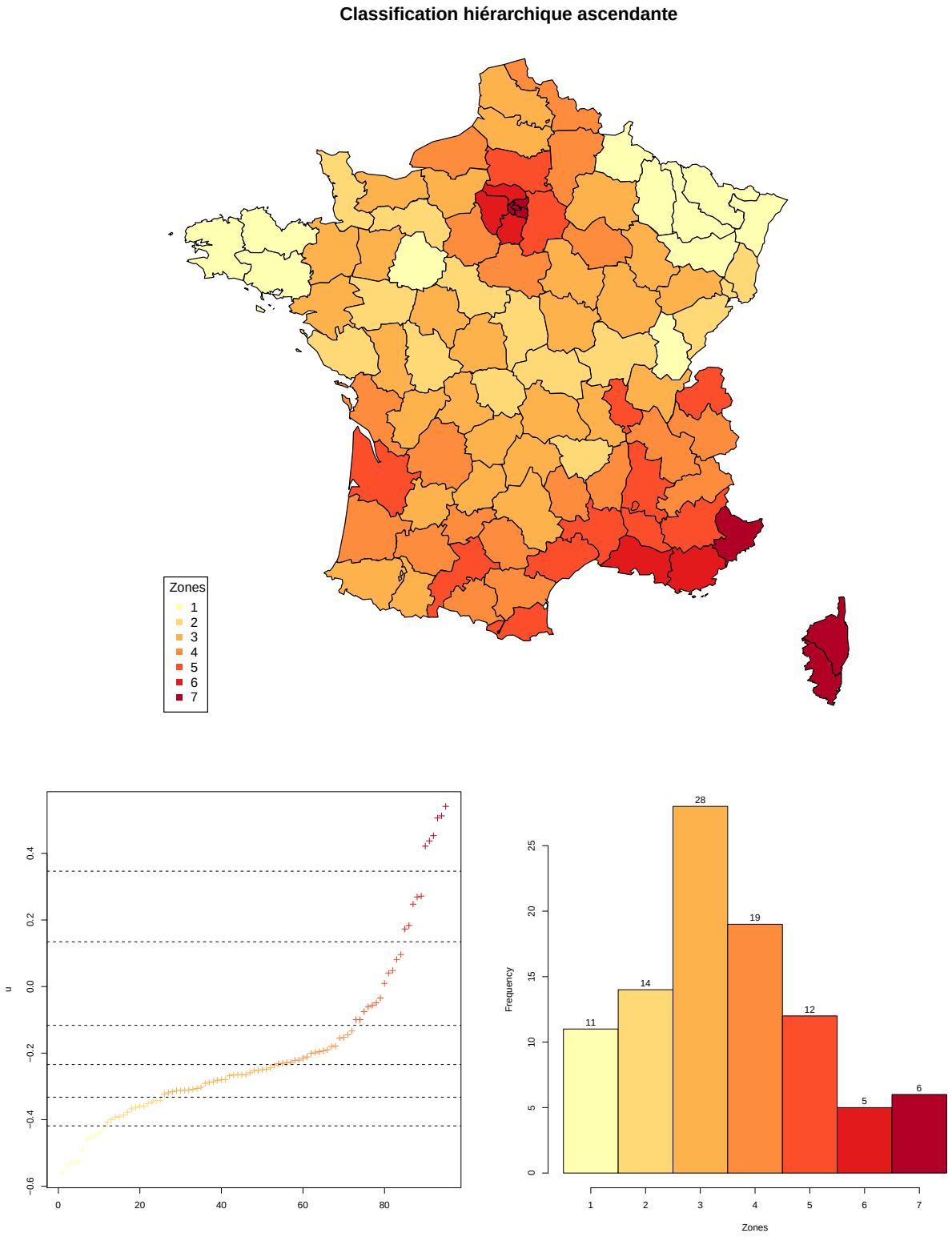


FIGURE 5.9 – Découpage par classification hiérarchique ascendante

Comme nous pouvions le prévoir, la méthode de découpage de la plage des u_i par bandes régulières donne beaucoup d'importance à une classe en particulier (ici la zone 2), au détriment des zones à forts risques. De même, avec un découpage par quantiles, les groupes constitués ne sont pas homogènes et il y a de fortes différences entre les départements présents dans la zone la plus risquée.

Nous choisissons de retenir le zonier créé suite à un découpage par une classification hiérarchique ascendante en sept zones de risque. Celle-ci semble présenter un bon compromis entre effectif, longueur des classes et regroupement en zones de risque homogènes.

L'outil informatique que nous avons créé fournit un estimateur du risque géographique pour chacune des régions sur la base du modèle de lissage spatial développé dans ce mémoire.

Pour autant, les valeurs ne sont pas utilisables telles quelles lors d'une tarification et nous devons trouver une méthode complémentaire à notre modèle, permettant de regrouper ces valeurs en un petit nombre de classes homogènes.

La liste des méthodes présentées ici est bien sûr non exhaustive, mais celles-ci présentent l'avantage d'être soit facilement misent en place, soit de répondre efficacement au problème posé.

Bien que notre critère de sélection soit essentiellement visuel, nous décidons de garder la méthode de regroupement par classification ascendante hiérarchique qui répond efficacement à nos attentes.

L'ensemble des régions est finalement placé dans sa zone de risque correspondante, et une tarification complète peut être menée en intégrant le zonier obtenu à l'intérieur d'une variable tarifaire.

Chapitre 6

Critique du modèle développé

Nous avons à présent terminé la réalisation du zonier par la méthode de lissage spatial basé sur un modèle à structure bayésienne. Les résultats obtenus semblent satisfaisant (choix du nombre de zones indépendant des estimations, obtention d'une carte lissée, découpage des régions en zones de risque homogènes, ...). Pour terminer l'analyse de ce modèle, nous souhaitons le comparer au modèle pragmatique utilisé par FIXAGE. On se penchera entre autre sur les sauts tarifaires que ces modèles génèrent afin de constater le problème des effets de bords. Nous aborderons enfin les limites que présente le modèle développé.

6.1 Comparaison avec le modèle pragmatique

6.1.1 Présentation du modèle pragmatique

La méthode actuellement utilisée par FIXAGE lors de la construction d'un zonier permet de réaliser, au cours d'un même processus, la segmentation des codes postaux en zones géographiques, l'estimation du risque dans ces zones géographiques ainsi que leur affectation à une zone de risque.

Cette méthode pragmatique consiste à calculer tout d'abord les fréquences observées à partir du nombre de sinistres et de l'exposition au risque pour chaque code postal $\left(\text{fréquence} = \frac{\text{nombre de sinistres}}{\text{exposition au risque}}\right)$. Les zones de risque sont ensuite créées au vu de ces fréquences.

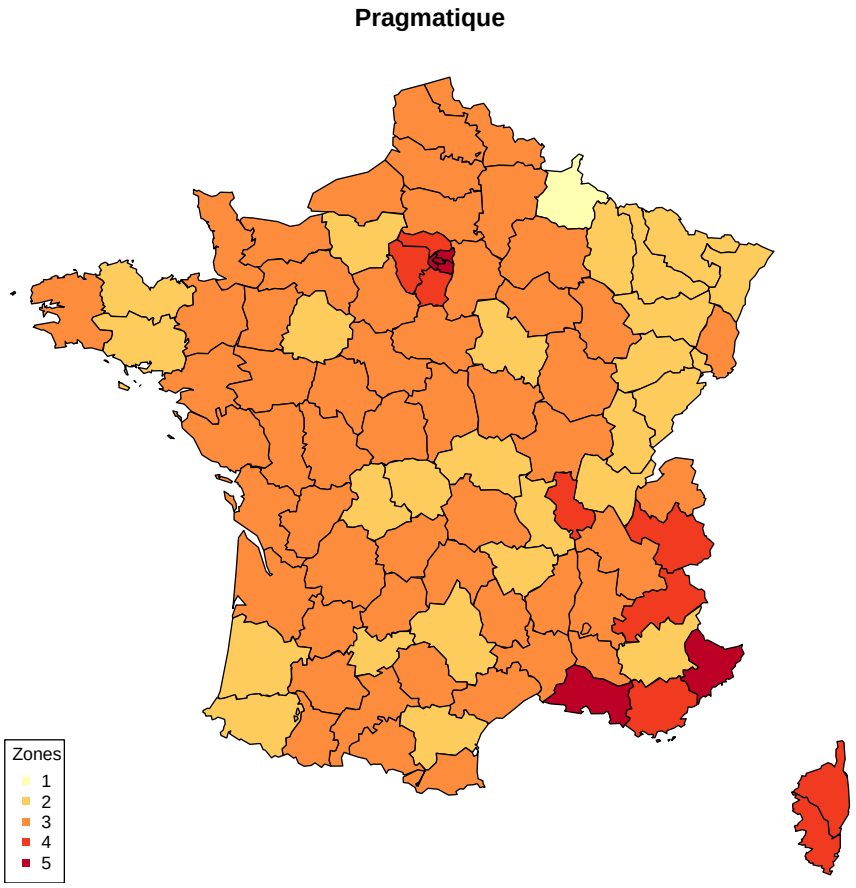
Exemple : si la fréquence moyenne pour l'ensemble des codes postaux est de 5%, nous allons créer les zones suivantes :

Zone	1	2	3	4	5
Fréquence	$< 2\%$	2% à 4%	4% à 6%	6% à 8%	$> 8\%$

L'affectation des codes postaux dans ces zones est réalisée en utilisant un algorithme permettant de regrouper les zones géographiques par proximité et par " indice de crédibilité " de la fréquence observée. En effet, nous associons à chaque fréquence observée un intervalle de confiance à 95%. Si l'intervalle de confiance touche plus de deux zones (ex : fréquence de 5% avec un intervalle de confiance $[3; 7]$), alors la fréquence pour ce code postal n'est pas considérée comme suffisamment significative. Nous le regroupons alors avec le code postal du même département, puis de la même région, ayant la plus grande population.
En revanche, si l'intervalle de confiance touche que deux zones (ou une seule), la commune est affectée à la zone correspondante à sa fréquence observée.

Au final, très peu de fréquences se révèlent significatives et cela revient souvent à regrouper toutes les communes d'un même département, voire d'une même région, ensemble. De plus, trop de communes ne présentent aucun sinistre voire même aucune exposition au risque (pas d'assuré dans le portefeuille) ce qui pose problème lors de l'évaluation du risque sous-jacent. Nous devons alors les regrouper automatiquement avec la plus grande ville du département, puis de la région, de manière à pouvoir calculer un intervalle de confiance associé à la fréquence observée. Le modèle présenté dans ce mémoire contourne ce problème en intégrant les données des régions voisines à l'intérieur d'un modèle bayésien permettant un transfert d'informations.

Les zones obtenues par la méthode pragmatique sont cartographiées comme suit, en gardant la zone majoritaire de chaque département :



Nous retrouvons ici la structure spatiale du risque avec les régions à fort risque, ainsi qu'une dominance de la troisième zone de risque.

Bien que toutes les étapes soient regroupées dans un même algorithme, le principal problème est que le choix des bornes lors de la définition des zones de risque est primordial et influence significativement les résultats obtenus. Nous pouvons difficilement créer beaucoup de zones car plus les intervalles sont petits, plus le test de significativité va être rejeté (intervalle de confiance réparti sur deux zones au plus). Les zones sont ainsi ajustées plusieurs fois avant d'obtenir un résultat satisfaisant.

A l'inverse, les estimations du risque géographique obtenues par le modèle bayésien et utilisées pour élaborer un zonier sont totalement indépendantes de celui-ci. Nous pouvons définir le nombre et les intervalles des zones de risque comme nous le voulons.

6.1.2 Comparaison des résultats

Nous souhaitons à présent constater l'apport d'un modèle mathématique avancé par rapport à un modèle plus pragmatique. Pour cela, il est délicat de se limiter aux seules cartes représentant les estimations de risque, bien que la première semble mieux répartir les régions par rapport aux zones de risque.

Nous regardons alors l'influence de ces deux méthodes sur la tarification finale des produits d'assurance. En effet, un des problèmes majeurs auquel nous sommes confrontés est celui des effets de bords. Ce problème fait référence à un manque de lissage et consiste à observer une différence tarifaire importante entre deux régions voisines qui semblent présenter le même risque.

Nous intégrons donc les zoniers obtenus dans une nouvelle régression de Poisson afin d'estimer la fréquence de sinistres sur la base de toutes les variables précédemment sélectionnées. Nous nous intéressons aux facteurs obtenus pour chaque zone de risque.

Grille tarifaire obtenue suite à l'estimation du risque par le modèle bayésien :

Zone	1	2	3	4	5	6	7
Coef. tarifaires	0.4428	0.5473	0.7331	1.0775	1.3459	1.8515	2.2931
Hausse		23.6%	34.0%	47.0%	24.9%	37.6%	23.8%

Grille tarifaire obtenue suite à l'estimation du risque par le modèle pragmatique :

Zone	1	2	3	4	5
Coef. tarifaires	0.3151	0.5635	0.7295	1.0327	1.5880
Hausse		78.8%	29.5%	41.6%	53.8%

Le coefficient tarifaire correspond au nombre par lequel la prime pure doit être multipliée en fonction de la zone de risque dans laquelle vit l'assuré. On constate logiquement que les zones à faible risque entraînent une diminution de la prime pure (coefficients inférieurs à un) alors que les zones à fort risque l'augmentent.

Il est intéressant de comparer les augmentations de prime lors du passage d'une zone de risque à une autre. Nous voulons bien sûr que celle-ci soit la plus petite possible afin de ne pas pénaliser les assurés habitant aux frontières de nos zones géographiques. Ils ne comprendraient pas une différence de prix par rapport au village voisin, ayant un risque a priori similaire au leur, qui lui se trouverait dans une autre zone géographique rattachée à une zone de risque inférieure. Ne pouvant techniquement pas empêcher cette distinction, nous gardons la méthode qui limite le plus possible les variations entre zones voisines.

Le premier zonier obtenu présente une hausse tarifaire plus homogène, due aussi au fait que nous avons utilisé sept zones de risque. Nous ne pouvons cependant pas prendre plus de zones, car cela fausserait la régression.

Les résultats obtenus semblent plus justes, au sens de l'équité tarifaire, suite à l'utilisation du zonier obtenu par la méthode développée dans ce mémoire.

6.2 Limites du lissage spatial

Le modèle mis en œuvre et implémenté parait satisfaisant et applicable au vu des résultats obtenus qui sont cohérents avec les observations. Il reste bien sûr perfectible et affiche certaines limites.

Tout d'abord, nous somme en présence d'un risque de modèle, sa définition étant basée sur un nombre important d'hypothèses. Celles-ci paraissent toutefois justifiées au vu de l'expérience apportée par des experts dans ce domaine et des résultats obtenus.

Le modèle dépend de plusieurs niveaux d'adéquation à des distributions théoriques, intégrant donc un risque de spécification. La définition de la fonction de densité conditionnelle du paramètres u_i permet d'intégrer le lissage spatial des valeurs. La définition faite attribue le même poids à tous les voisins d'une région donnée. Il serait intéressant d'intégrer une pondération tenant compte de l'éloignement lors du passage à une segmentation par codes postaux.

Nous sommes également soumis à un risque de paramètre, puisque le modèle utilise un hyperparamètre fixé. Nous avons évaluer ce risque en changeant la valeur de *epsilon* initialement fixée à 0,01 et en comparant les résultats obtenus. Nous avons lancés l'outil en fixant successivement *epsilon* à 0,05 puis 0,1. Nous constatons de légères différences mais décidons de garder 0,01 qui semble permettre d'atteindre l'état stationnaire plus rapidement.

Les autres paramètres doivent également être initialisés. L'utilisation de l'échantillonnage de Gibbs permet de réduire l'erreur due à une mauvaise initialisation en se basant sur les chaînes de Markov. L'idéal serait de simuler plus d'états, par exemple 15 000, de manière à ne prendre en compte que les 5 000 derniers et ainsi limiter d'avantage l'influence de l'initialisation des variables sur les résultats obtenus, mais cela allongerait le temps de calcul.

Nous avons réalisé une étude préalable afin de déterminer le nombre de zones de risque que nous allions utiliser. Il semble cependant difficile d'en évaluer un nombre optimal car nous ne pouvons pas quantifier les résultats obtenus suite à différents découpages. Ce problème est général et non borné à cette approche d'estimation par zones géographiques.

Enfin, bien que l'outil créé mette environ une heure pour fournir les résultats d'une simulation à 10 000 états et 95 zones géographiques, le temps calcul est beaucoup plus important (environ une semaine) lors du passage à une segmentation par code postaux (36 686 zones géographiques). Nous pourrons cependant le modifier légèrement pour ne plus garder en mémoire les 5 000 premiers états qui servent à atteindre un état stationnaire.

Conclusion

Nous sommes finalement arrivés, non sans mal, à l'objectif initial qui était de fournir un outil de base pour l'élaboration de futurs zoniers. Nous nous sommes tournés vers utilisation de techniques statistiques appliquées à l'actuariat afin de contourner les problèmes pratiques. Le modèle mis en œuvre a nécessité une compréhension d'outils mathématiques complexes afin de bien schématiser le processus d'estimation du risque et de pouvoir l'implémenter sous R.

Nous disposons à présent d'un outil efficace et réutilisable permettant d'associer à chaque région étudiée une valeur de risque, et ainsi les regrouper en sous-groupes homogènes de risque afin d'élaborer une tarification. Bien que cette méthode paraisse assez efficace, elle reste sans doute peu connue et donc peu utilisée par les actuaires français lors de leurs tarifications.

Le processus de tarification à travers l'approche mise en œvre peut se résumer en quatre étapes principales :

1. Utiliser une première fois les modèles linéaires généralisés afin d'estimer les facteurs de risque non spatiaux.
2. Déterminer un estimateur du risque géographique pour chaque région à l'aide de l'outil développé.
3. Regrouper les régions en un nombre de classes défini par rapport aux valeurs obtenues afin de créer un zonier.
4. Intégrer ce zonier dans une variable tarifaire et utiliser une deuxième fois les modèles linéaires généralisés afin d'établir la tarification.

Bien qu'il présente ses limites, l'utilisation d'un modèle mathématique puissant offre une alternative intéressante aux approches pragmatiques (fréquences brutes, fréquences significatives) en répondant aux principales contraintes initiales.

L'outil devrait cependant encore être optimisé de manière à intégrer une segmentation en zones géographiques plus fine (par code postaux) et fournir un estimateur de leur risque géographique dans des temps raisonnables.

Enfin, il est intéressant de mentionner le fait qu'il s'agit uniquement d'un modèle technique et qu'il n'a donc pas fait l'objet de retraitements commerciaux. Des ajustements devront être réalisés en interne par l'entreprise d'assurance en raison de nombreux facteurs non captés par le modèle : effets de bords très particuliers, modification du portefeuille assuré ou encore modification des techniques de commercialisation (augmentation des points de distribution en zone rurale, ...).

Bibliographie

[1] Boskov, M., and Verrall, R. Premium rating by geographic area using spatial models. *ASTIN Bulletin 24* (1994), No 1, 131–143.

[2] Brouhns, N., Denuit, M., Masuy, B., and Verrall, R. Ratemaking by geographical area in the Boskov and Verrall model : A case study using belgian car insurance data. *actu-L 2* (2002), 3–28.

[3] Gilks, W., Richardson, S., and Spiegelhalet, D. *Markov chain Monte Carlo in practice.* CCR Press, 1995.

[4] Gilks, W., and Wild, P. Adaptive rejection sampling for Gibbs sampling. *Appl. Statist. 41* (1992), No 2, 337–348.

[5] Guillou, A. Modèles linéaires. *Support de cours, Université de Strasbourg* (2007-2008).

[6] Klutchnikoff, N. Modélisation stochastique. *Support de cours, Université de Strasbourg* (2007-2008).

[7] Mahy, S., and Denuit, M. Découpage géographique par zones de Voronoï en assurance automobile. *Université Catholique de Louvain* (2002).

[8] Maumy-Bertrand, M., and Nobelis, P. Introduction à l'analyse de données multidimensionnelles. *Support de cours, Université de Strasbourg* (2008-2009).

[9] Planchet, F. Modèles de durée : Méthodes de lissage et d'ajustement. *Support de cours, ISFA* (2008-2009).

[10] Taylor, G. Use of spline functions for premium rating by geographic area. *ASTIN Bulletin 19* (1989), No 1, 91–122.

[11] Taylor, G. Geographic premium rating by Whittaker spatial smoothing. *Research paper 38, University of Melbourne* (1996).