Mémoire présenté le :

pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA et l'admission à l'Institut des Actuaires

Par : Charles Birenholz

Titre : Modélisation géographique des déviations de mortalité en vue d'optimiser la tarification de produits d'assurance

Confidentialité : ☒ NON ☐ OUI (Durée : ☐ 1 an ☐ 2 ans)

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de l'Institut des Actuaires *signature*

Entreprise :

Nom : RGA International Reinsurance Company dac

Signature : RGA International Reinsurance Company dac Succursale pour la France 31-33, rue de la Baume 75008 Paris Tél. : +33 (0)1 55 07 97 80 Siret 499 883 148 00078

Membres présents du jury de l'ISFA

Directeur de mémoire en entreprise :

Nom : Clara Cichowlas

Signature :

Invité :

Nom :

Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels *(après expiration de l'éventuel délai de confidentialité)*

Signature du responsable entreprise

Signature du candidat

Mémoire d'actuariat
Institut de Science Financière et d'Assurances

Modélisation géographique des déviations de mortalité en vue d'optimiser la tarification de produits d'assurance

Charles Birenholz

2021

Tuteur en entreprise : Clara Cichowlas

Tuteur universitaire : Stéphane Loisel

Sommaire

Résumé

Ce mémoire cherche à évaluer les risques de sur ou de sous-mortalité imputables à la géographie, en vue d'optimiser la tarification de produits d'assurance. Il propose pour ce faire une méthodologie originale, basée sur l'exploitation de données publiques. A l'issue de ce mémoire, nous serons en mesure de proposer des coefficients reflétant l'impact du lieu de domicile sur le risque de mortalité. Ces résultats seront dès lors exploitables à des fins de pilotage produit et, à l'avenir, pourront être utilisés lors de tarifications d'assurances décès. Ce mémoire a pour vocation de s'appliquer à l'ensemble des habitants de France métropolitaine entre 30 et 62 ans.

Dans un premier temps, nous estimerons le nombre de décès survenus sur le territoire métropolitain par des méthodes actuarielles. Nous viendrons ensuite comparer nos estimations avec les décès effectivement observés. Nous serons alors en mesure de cartographier les écarts de prédictions.

Nous chercherons alors, dans un second temps, à analyser ces divergences. L'idée sera ici de modéliser les écarts observés par le salaire moyen dans la ville ou la densité de population par exemple. Cela se fera au travers d'un modèle linéaire adapté. A l'issue de cette partie, nous disposerons d'une carte de France des déviations de mortalité, contextualisée et chiffrée. Nous proposerons également un lissage spatial de nos résultats.

Enfin, nous discuterons des visées et perspectives de ces travaux. Nous illustrerons nos résultats par leur l'application à des portefeuilles emprunteur et de prévoyance individuelle.

Abstract

This thesis seeks to assess the deviation of mortality attributable to geography, for the purpose of optimizing insurance pricing. To do this, it proposes an original methodology, based on the exploitation of public data. At the end of this thesis, we will have determined coefficients reflecting the impact of the place of residence on the risk of mortality. These results will therefore be usable for product management purposes and, in the future, can be used in pricing of death insurance. This thesis is intended to apply to all residents of metropolitan France between 30 and 62 years old.

First, we will estimate the number of deaths that have occurred in the metropolitan area by actuarial methods. We will then compare our estimates with the deaths actually observed. Thus, we will be able to map the prediction deviations.

We will then seek, in a second step, to analyze these divergences. The idea here will be to model the differences observed by the average salary in the city or the population density for example. This will be done through an adapted linear model. At the end of this part, we will have a map of France of mortality deviations, contextualized and quantified. We will also propose a spatial smoothing of our results.

Finally, we will discuss the aims and perspectives of this work. We will illustrate our results by applying them to both a credit insurance portfolio and an individual protection portfolio.

Remerciements

Je tiens à remercier l'ensemble des collaborateurs de RGA France.
Je remercie particulièrement Clara Cichowlas pour l'implication, le soutien et les conseils qu'elle m'a apportés. Je remercie également Gurvan Le Rhun pour sa confiance et sa participation.

Je remercie ma famille pour m'avoir toujours permis de poursuivre mes études dans les meilleures conditions.

Introduction

Le monde de l'assurance s'est beaucoup transformé depuis l'arrivée du numérique. La démocratisation d'internet a facilité la distribution des produits et les relations clients, si bien qu'aujourd'hui, on ne compte plus les acteurs proposant des solutions simplifiées sur le marché. Du point de vue de l'actuaire en revanche, c'est dans la construction des produits eux mêmes qu'ont eu lieu les changements les plus importants.

L'usage des bases de données et l'augmentation des puissances de calculs ont permis au secteur d'implanter des méthodes nouvelles comme des techniques de simulation ou de machine learning par exemple. Il n'a jamais été aussi aisé d'étudier un portefeuille d'assurés de manière précise et efficace, et pourtant les perspectives du secteur restent plus importantes que jamais.

Les applications de santé, les objets connectés et la promesse de capter toujours plus de données concentrent beaucoup d'attention. A terme, on peut légitimement s'attendre à ce que des données individuelles extrêmement précises arrivent un jour dans le monde de l'assurance (pourquoi pas par le biais des GAFAM). Il incombera aux actuaires de moderniser encore et toujours leurs méthodes pour s'adapter à ces changements.

Traiter de l'information à des fins d'optimisation et de calibrage produit n'est pas réservé aux assurances, mais on comprend facilement en quoi cela revêt une importance stratégique pour le secteur. Mieux qualifier le risque, pour une entreprise d'assurance, c'est :

— Pouvoir proposer des prix plus justes et donc plus compétitifs

— Mieux évaluer les provisions pour assurer l'avenir de l'entreprise, piloter les investissements et répondre aux engagements pris vis à vis des clients.

— Piloter la rentabilité et la vie des portefeuilles de clients.

Au delà des aspects commerciaux, l'assurance est un marché très réglementé, et il faut rappeler que cibler pertinemment les marchés est une exigence réglementaire, notamment au travers de la *Directive de Distribution des Assurances* (2016/97). Transposée en droit Français depuis 2018, cette directive européenne introduit le concept de *Product oversight and governance* qui inclut dans la validation produit l'identification du « marché cible » (pour lequel seront évalués « tous les risques pertinents ») [64].

Paradoxalement, on peut s'interroger sur l'impact de l'individualisation des données à terme. D'un point de vue légal, il n'est en effet pas certain que l'impact soit, lui, individuel. Cela a été montré par la directive 2004/113 de la Cour de justice de l'Union européenne [8], avec l'interdiction de prendre en compte le sexe de l'assuré dans les primes et prestations d'assurance.

Ces constats sur l'avenir du secteur se heurtent à une réalité bien différente. Aujourd'hui, les données utilisés par les assureurs restent relativement succinctes. En assurance décès, qui est le sujet motivant ce mémoire, l'essentiel des caractéristiques utilisées pour l'évaluation du risque de mortalité sont les suivante :

— L'âge

— Le sexe

— Le métier (ou la catégorie socioprofessionnelle (csp))

— Le statut fumeur / non fumeur

Dans ce mémoire, nous chercherons donc à ajouter le domicile de l'assuré à cette liste.

Aujourd'hui, seuls les instituts statistiques nationaux sont en mesure de nous fournir des données suffisamment vastes et fiables pour réaliser ce type d'étude. A ce titre, l'ensemble de nos travaux sera conditionné par les données publiques accessibles. Nous évaluerons la pertinence d'une telle démarche, ainsi que sa faisabilité. En effet, sans disposer d'une base d'exposition et d'une base

de décès associée, il peut être très complexe de prendre en compte de nouveaux paramètres sans introduire de biais. Nous prendrons ici soin de limiter l'influence des effets déjà pris en compte par les techniques de tarification classiques (l'âge, le sexe et la csp). L'étude des villes que nous proposons ici se veut complémentaire des études de portefeuilles d'assurances classiques.

Relativement aux perspectives évoquées plus haut, les données publiques sont donc à voir comme un moyen d'étendre les méthodes de tarifications classiques, tout en tenant compte de la réalité des données utilisées aujourd'hui.

L'open data permet de saisir de nombreux sujets à moindre coût mais son utilisation n'est pas exempte de problème. L'Insee sera notre source de données privilégiée tout au long de ce mémoire. S'il s'agit d'une source précieuse, car fiable et très détaillée, certaines subtilités (comme les différentes échelles géographiques) et contraintes (comme la profusion de bases et méthodes) ont nécessité un investissement important en termes de temps.

Le lecteur intéressé pourra donc trouver ici un condensé de concepts utilisés par l'Insee, ainsi qu'une liste de bases de données. Si elles ne se prétendent pas exhaustives, ces informations sont le fruit d'un long travail de recherche qui pourra donc être économisé.

Différents autres organismes mettant à disposition des données seront aussi utilisés et présentés succinctement.

La méthodologie originale que nous exposerons ici est analogue à celle qui est aujourd'hui appliquée pour évaluer les lycées [57]. Elle consiste à comparer la mortalité réelle et attendue ; plus précisément, nous opposerons, pour l'ensemble des 2094 cantons français, le nombre de décès réel et le nombre de décès théorique donné par les méthodes actuarielles classiques : c'est à dire basées sur l'âge, le sexe et la catégorie socioprofessionnelle. Notre étude sera centrée sur les adultes en âge de travailler.

Nous verrons en quoi la méthode utilisée ici présente de nombreux avantages dont, notamment, sa capacité à rendre compte de la diversité du territoire métropolitain et sa complémentarité par rapport aux méthodes actuarielles classiques. Nous noterons également les limites de notre approche et discuterons des perspectives de nos travaux.

Première partie

Enjeux et Contexte

Le présent mémoire a été rédigé au sein de l'entreprise américaine *Reinsurance Group of America* (RGA), l'un des plus importants réassureurs mondiaux sur le segment de la Vie. En effet, RGA est l'un des très rares réassureurs proposant exclusivement des traités Vie et Santé.

Le bureau français n'a été ouvert que récemment -en 2007- mais il a connu une importante croissance depuis. Il compte aujourd'hui une quarantaine de collaborateurs et s'est fait une place sur le marché français [1]. L'activité et l'expertise de RGA France portent aujourd'hui sur la réassurance de produits Emprunteur et de Prévoyance Individuelle : cela explique l'attention que nous porterons à ces deux secteurs d'activité. Néanmoins, nos travaux ont tout à fait le potentiel pour s'appliquer sur des sujets d'assurance collective tels que la prévoyance d'entreprise.

Dans cette partie, nous présenterons donc les marchés de l'emprunteur et de la prévoyance en France. Nous expliquerons également la pertinence de nos travaux dans le contexte de la réassurance, ainsi que l'intérêt de la géographie dans l'analyse du risque de décès.

Dans un second temps, nous réaliserons une première approche de ces aspects géographiques et motiverons la suite de nos travaux.

1 Risque et assurance décès

1.1 Marché et réglementation

1.1.1 L'assurance emprunteur

Comme nous l'apprend la Fédération Française de l'Assurance : "*L'assurance en cas de décès est un contrat d'assurance vie qui permet le versement d'un capital ou d'une rente à un bénéficiaire désigné, en cas de décès de l'assuré avant le terme du contrat. Ces contrats d'assurance vie peuvent être souscrits individuellement ou collectivement, par l'intermédiaire d'une entreprise ou d'une association. Ils peuvent être souscrits à l'occasion d'un emprunt.*" [15]

L'assurance emprunteur est une assurance temporaire, limitée à la durée du crédit et dont la prestation en cas de sinistre est le remboursement de la totalité ou d'une partie des mensualités du prêt, voire du capital restant dû. Cette garantie est régulièrement complétée par des garanties d'assurance de personnes couvrant les risques d'invalidité (totale ou partielle) et d'incapacité de travail. Le marché de l'assurance emprunteur est un marché important. En 2017, l'encours des crédits accordés aux ménages en France par les établissements de crédits était de 1 292 milliards d'euros, dont 74 % pour les prêts immobiliers. La même année, le montant des cotisations au titre des contrats d'assurance emprunteur était de 9,1 milliards d'euros. On compte en France environ un million de nouveaux emprunteurs par an [16].

Si elle était régulièrement souscrite par l'intermédiaire d'un établissement financier, l'assurance emprunteur est aujourd'hui un marché très concurrentiel. Ce sont en effet 3 grandes réformes qui ont permis cette libéralisation graduelle :

— La loi Lagarde (2010) qui permet de souscrire chez l'assureur de son choix au moment de contracter un prêt immobilier

— La loi Hamon (2014) qui permet à l'assuré de changer de contrat d'assurance n'importe quand durant la première année de son prêt immobilier

1. Bien que cela soit une part minoritaire de son portefeuille, RGA France opère également sur les marchés belges et luxembourgeois.

— L'amendement Bourquin (2018), parfois appelé loi Sapin 2, qui permet la résiliation et donc le changement d'une assurance emprunteur tous les ans à la date anniversaire du contrat.

Les attentes tarifaires des clients, poussées par les réglementations facilitant la résiliation, ont tiré les prix vers le bas depuis plusieurs années. Les données géographiques sont une manière d'affiner la vision du risque dans un marché compétitif.

Il faut néanmoins nuancer ces éléments. Le marché de l'emprunteur reste un marché assez captif. *L'Autorité de Contrôle Prudentiel et de Résolution* (ACPR) indique que, depuis 2008, le taux d'assurances emprunteur prises auprès d'un assureur alternatif a très peu évolué : 13% en 2008 et 15% en 2018. Les chiffres que nous avons cités, importants et en augmentation constante, demeurent trustés par les bancassureurs. Les 5 premiers vendeurs d'assurances emprunteur sont ainsi, à l'exception de la CNP, des filiales totalement intégrées aux groupes bancaires, à savoir dans l'ordre : Crédit Agricole Assurances, Groupe des assurances du Crédit Mutuel, BNP Paribas et BPCE [56].

1.1.2 Prévoyance individuelle

La prévoyance individuelle fait également partie des types d'assurance notables à laquelle s'appliquent les travaux réalisés ici. Assurance non obligatoire, elle s'adresse principalement aux Travailleurs Non Salariés (TNS), c'est à dire essentiellement aux professions libérales, aux artisans et aux commerçants. Globalement, ces trois catégories représentent deux tiers des TNS. Le dernier tiers est composé d'une manière plus diverse avec, par exemple, des emplois liés à l'agriculture, au monde du spectacle ou à la construction [22].

Les travailleurs non salariés peuvent avoir des besoins spécifiques et ne sont pas toujours suffisamment pris en charge par les régimes généraux. Depuis le 1er janvier 2020 les TNS relevant du Régime Social des Indépendants ont été rattachés à la sécurité sociale. La loi 94-126 du 11 février 1994 dite *loi Madelin* est une des lois importantes de ce secteur. Elle a été créée dans le but d'inciter les TNS à se constituer eux-mêmes leur propre protection sociale en rendant leur souscription fiscalement avantageuse. Elle permet en effet à l'assuré de déduire de son revenu imposable les cotisations versées au titre d' un contrat Madelin.

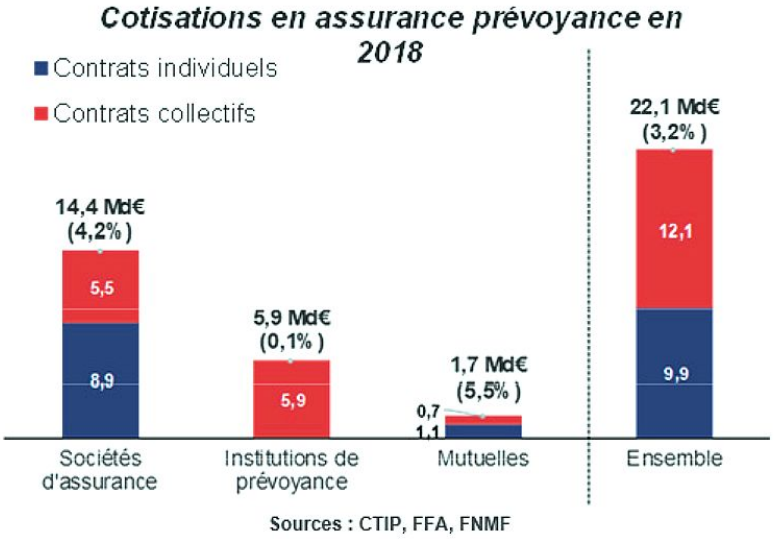


FIGURE 1 – Cotisations d'assurance prévoyance en 2018 - FFA [17]

Avec 3,2% d'augmentation et 22,1 milliards d'euros de cotisations en 2018, la prévoyance est un marché très volumineux. La prévoyance individuelle a porté une part importante de cette somme

avec 9,9 milliards d'euros en 2018, mais porte aussi cette augmentation, avec +4,5% de croissance la même année. Les acteurs sont également plus divers. Sur les 22 milliards évoqués : près de 2/3 sont captés par les compagnies d'assurances, prés d' 1/4 le sont par les institutions de prévoyance, et le reste, environ 7%, l'est par les mutuelles.

La prévoyance individuelle est un marché attractif, mais avec un taux de pénétration encore assez faible. Sur les 3 millions d'indépendants français seulement 1,7 million sont couverts par un contrat Prévoyance Madelin fin 2017 [17]. La diversité des contrats de prévoyance individuelle fait qu'il est aujourd'hui difficile d'énumérer les garanties d'un contrat type ; les grandes lignes restent néanmoins les garanties décès, incapacité temporaires et invalidité.

1.2 L'analyse géographique dans le contexte de la réassurance

Comme nous l'avons mentionné, le présent mémoire a été réalisé au sein de l'entreprise de réassurance vie *Reinsurance Group of America* (RGA). Quelles motivations peuvent pousser une compagnie de réassurance à étudier le risque géographique de mortalité ?

Il est connu que les entreprises de réassurance prennent en charge une partie des sommes sous risque des compagnies d'assurances. Par ce biais, elles aident leurs partenaires à capter des parts de marché ou à monitorer leur risque. Ce qui est moins connu en revanche, c'est que l'accompagnement qu'elle propose n'est pas que financier.

Les entreprises de réassurance endossent régulièrement un rôle d'expertise et d'accompagnement de l'activité, lors du lancement de produits ou de leur pilotage. Si les entreprises de conseil proposent une facturation à l'heure pour leurs services, les compagnies de réassurance, elles, se rémunèrent lorsqu'elles prennent des parts sur les contrats qu'elles aident à construire ou à piloter ; essentiellement sur des *traités proportionnels* où le réassureur est donc partie prenante.

Les assureurs et réassureurs n'ont pas nécessairement d'immenses bases de données prêtes à l'emploi lorsqu'ils démarrent ou étendent leur activité. Le manque d'informations lorsque l'on construit de nouveaux produits (par exemple, un assureur spécialisé dans l'emprunteur qui chercherait à proposer des contrats de prévoyance individuelle) impose très souvent de prendre des hypothèses à priori.

Dans un contexte de co-construction de produit, on comprend donc l'intérêt de maîtriser la donnée géographique. Ce type d'information peut aider à spécifier les hypothèses et donc à mieux calibrer un nouveau produit.

Malgré le rôle de partenaire que nous avons mis en avant ici, il faut savoir que plusieurs réassureurs se partagent régulièrement un même traité (essentiellement sur des *traités non-proportionnels*). Cela a des effets sur la confiance que les assureurs peuvent accorder. Il est même parfois possible que seul l'apériteur (i.e. le réassureur majoritaire sur le traité) ait accès aux données clients. D'un autre coté, les entreprises d'assurance n'ayant pas toutes les mêmes méthodologies et exigences en termes de remontées d'information, les données reçues ne se valent pas toujours.

Le réassureur reste donc tributaire des données qu'il reçoit. Les travaux présentés ici n'ont pas vocation à s'appliquer à l'ensemble de l'activité de l'entreprise. Néanmoins, il se trouve que la résidence de l'assuré est une donnée dont les assureurs disposent souvent et qu'il n'est pas absurde d'avoir.

Les travaux réalisés dans ce mémoire pourront permettre de mieux appréhender le risque pris dans ce type de situation.

Voici un exemple de questions concrètes que pourrait se poser un réassureur et auquel ce mémoire propose d'apporter des réponses :

— Une cédante, dont l'essentiel des clients est concentré à Paris, prétend que la faible mortalité dans la capitale française justifierait une baisse de prix. Les catégories socioprofessionnelles

étant déjà prises en compte dans la tarification [2], est-ce justifié ou cela relèverait-il du geste commercial ?

— Un courtier lyonnais cherche à mettre en place un nouveau produit. Ce dernier sera essentiellement distribué dans ses agences en région Rhônes-Alpes ; est-ce une information pertinente à prendre en compte pour définir les hypothèses de portefeuille ?

— Le réassureur souscrit un traité en quote-part sur un risque de prévoyance collective dans le nord de la France. Doit-il s'attendre à ce que cette concentration géographique ait un impact sur sa rentabilité attendue ?

— Le réassureur doit annoncer la part qu'il souhaite prendre sur un traité emprunteur, mais le taux de réassurance sur lequel il doit se positionner est inférieur à son taux technique. Sachant que ce portefeuille a été essentiellement souscrit par des résidents d'Alsace-Moselle, sa perception du risque de décès est-elle aggravée ou au contraire atténuée ?

En définitive, si ce mémoire ne s'intéresse pas aux structures de réassurance et à leurs méthodes de tarification, il est néanmoins au cœur de nombreuses problématiques réassurantielles, en particulier celles du conseil, de la vision marché et du calibrage produit.

1.3 L'intérêt de la géographie dans l'analyse du risque décès

Comme nous le mentionnons en introduction, le numérique permet aujourd'hui une vente décentralisée des contrats d'assurance. A mesure que la souscription par internet se banalise, le rôle des agences devient moins prépondérant. Néanmoins, elles ont toujours un impact dans la constitution des portefeuilles. D'autant plus qu'une partie des clients est encore intéressée par avoir un interlocuteur réel. De même, les compagnies n'ayant ni les mêmes méthodes de distribution ni la même image, chaque compagnie capte un type de client différent et il est ainsi possible de voir apparaître des phénomène de concentration, notamment géographique.

On pourrait croire que la modernisation du pays a uniformisé les territoires mais il n'en est rien. Le lieu de vie de tout un chacun rythme encore et toujours les interactions sociales, l'exposition aux pesticides ou à la pollution, les temps de trajets, les métiers exercés et les perspectives d'avenir. Ces éléments ont des implications directes sur le risque de mortalité des individus et continueront d'en avoir à l'avenir.

L'idée d'utiliser le domicile d'un individu pour affiner notre vision du risque de mortalité semble donc prometteuse. Par ce biais, on sera en mesure d'exploiter les caractéristiques des territoires (et les nombreuses bases de données qui les décrivent). Cela reviendra à capter de l'information supplémentaire d'une manière peu intrusive et, ainsi, d'enrichir à moindre coût les données de portefeuilles. Néanmoins, on peut s'interroger. Le lieu de résidence permet-il réellement de mieux comprendre les risques individuels ? D'autant plus que les individus se déplacent au cour de leur vie.

Les villes présentent des caractéristiques globales qui évoluent lentement. Même si les effets que nous décrirons ne s'appliquent pas à tous les individus (par exemple aux nouveaux résidents), sur un grand nombre de personnes cela sera globalement vrai. C'est par ce même argument que l'on applique un taux de mortalité q_x aux personnes d'âge x, alors que tous les individus n'ont pas la même probabilité de décéder. Sur le temps long, les villes peuvent -de plus- être pensées comme structurantes pour les populations qui y vivent.

En ce qui concerne les trajectoires de vie, les individus sont relativement sédentaires. Selon une étude de l'Insee parue fin 2019, près d'une personne sur deux meurt dans son département de naissance [54].

2. Nous détaillerons par la suite en quoi la proportion de cadre à Paris (notamment) empêche de se faire un avis immédiat sur la question

Ce constat est néanmoins à relativiser pour les diplômés du supérieur et étudiants du supérieur nés en France. Une autre étude de l'Insee nous apprend que 4,9 millions d'entre eux résident dans une région différente de celle où ils sont nés ; cela correspond à 37 %, contre 25 % pour les personnes disposant au plus d'un niveau de formation équivalent au baccalauréat [39]. On comprend ainsi que la mobilité inter-régionales des diplômés du supérieur contribue aux contrastes observés entre les territoires ; ce qui est à rejoindre avec le phénomène de concentration des cadres.

A priori, on peut s'attendre à capter dans nos analyses des différences de mortalité résultantes d'effets de groupe comme la pratique du sport, l'alimentation, la fréquence des consultations médicales et la prévention des maladies. Ces différences résultent de l'effet de facteurs exogènes comme la pollution urbaine ou agricole, ou de facteurs structurels comme l'usage de la voiture et la prépondérance d'activités industrielles à risque dans une région.

2 Première approche du risque de décès

Les Français bénéficient d'une espérance de vie élevée. En France métropolitaine, elle s'élève à 85,7 ans pour les femmes et à 79,8 ans pour les hommes en 2019. Pourtant, cette réalité flatteuse ne reflète pas la diversité du territoire : la même année, on constate par exemple que dans les Hauts-de-France l'espérance de vie est de 77 ans, tandis qu'en Ile-de-France elle est de 81 ans ; soit une différence de d'environ 4 ans.

Comme nous allons le voir, les disparités augmentent à mesure que l'on segmente la population française. Mais a quel point sont-elles importantes ? Et, finalement, ne sont-elles pas déjà captées par les techniques de tarification classique ? Dans un cadre assurantiel, ces effets pourraient de plus sembler simples à intégrer dans nos estimations de mortalité.

Dans cette partie, nous allons chercher à nous faire une idée de la diversité du territoire français face au risque de décès. Nous verrons alors que ces effets sont plus subtils qu'il n'y parait et évoquerons les biais qui pourraient résulter d'une analyse mal construite. Nous comprendrons alors pourquoi il est justifié d'étudier cette diversité géographique et en quoi la méthode que nous proposons ici est adaptée.

2.1 Approche de la géographie par les taux standardisés

On définit le taux de mortalité comme la probabilité de décès. Soit :

a l'âge de l'individu.

T la variable aléatoire représentant l'instant de décès.

On a alors le taux de mortalité des individus d'âge a :

$$q_a = \mathbb{P}(T < a + 1 \backslash T \geq a)$$

Ce taux peut être raffiné en définissant, par exemple, le sexe s de l'individu. On notera $q_{a,s}$ le taux de mortalité des individus d'âge a et de sexe s.

Proposons, dans un premier temps, une méthode simple pour évaluer l'impact du lieu de domicile sur la mortalité. Nous verrons alors à quelle condition très restrictive cette méthode peut être utilisée dans les techniques de tarification.

Les différences d'âge et de sexe dans la population française sont des paramètres dont l'effet peut être assez facilement isolé. Cela peut notamment se faire par le biais des taux standardisés de mortalité. Il s'agit du taux de mortalité que l'on observerait dans la population étudiée (ici les départements) si elle avait la même structure d'âge que la population de référence (ici la France métropolitaine).

En reprenant les notations précédentes, on pose maintenant :

A_{min} l'âge minimum considéré.

A_{max} l'âge maximum considéré.

j la sous-population considérée *(ici les habitants du département)*.

$N_{a,s}$ le nombre d'individus d'âge a et de sexe s dans la population de référence.

N le nombre d'individus dans la population de référence.

$q_{j,a,s}$ le taux de mortalité observé dans la sous-population j chez les individus d'âge a et de sexe s.

Ce qui donne le taux de mortalité standardisé de la sous-population j :

$$\tilde{Q_j} = \frac{1}{N} \sum_{s \in (H,F)} \sum_{a=A_{min}}^{A_{max}} q_{j,a,s} N_{a,s}$$

Il est ainsi possible de calculer cet indicateur par département comme on peut le voir sur cette carte réalisée par l'Insee [27] :

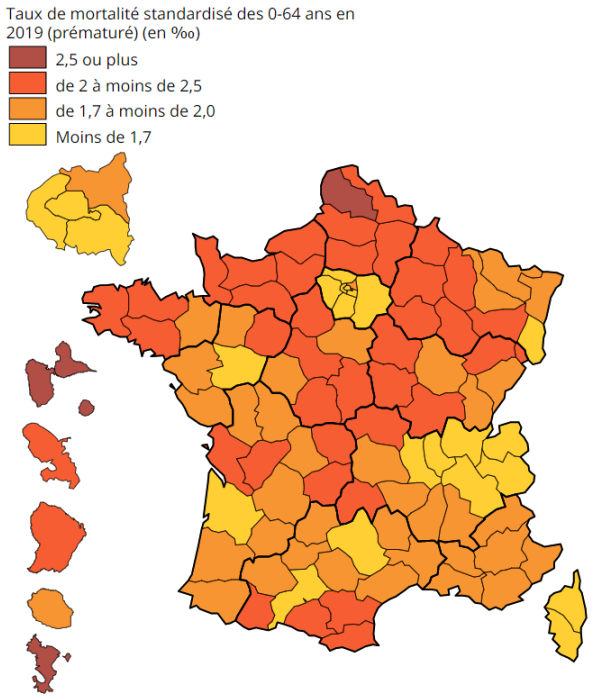


FIGURE 2 – Taux de mortalité standardisé des 0-64 ans

Les taux présentés ici ont été calculés en population générale, mais ils existent également pour les hommes et pour les femmes pris séparément.

On peut donc d'ores et déjà se faire une idée des différentiels de mortalité que l'on peut rencontrer sur le territoire, une fois l'âge et le sexe traités.

Il faut remarquer ici que, très souvent, les départements comportant des grandes villes ressortent positivement dans cette représentation. C'est notamment le cas pour la Haute Garonne (Toulouse) et la gironde (Bordeaux), ainsi que pour les régions Rhones Alpes (Lyon) et Ile de France (Paris).

On pourrait être tenté d'exploiter ces taux directement. En divisant chacun de ces taux par le taux standardisé national, on disposerait d'une pondération du risque de mortalité dans chaque département :

$$Y_{Reg}^{Insee} = \frac{\tilde{Q_{Reg}}}{\tilde{Q_{France}}}$$

En appliquant cette pondération aux taux de mortalité, on capterait donc un semblant de risque géographique très simplement.

Néanmoins, cette méthode interdirait de prendre en compte les csp (ou encore le niveau de vie ou le niveau de diplôme) dans notre tarification.

Il est en effet plus que probable que la concentration d'activités de cadres à Paris ou d'activités industrielles dans le nord de la France (par exemple) soient à l'origine des différences que nous observons. Introduire ces ratios reviendrait, dès lors, à comptabiliser deux fois l'impact des csp. Pour reprendre l'exemple de Paris, il n'est pas évident (à priori) qu'un cadre parisien ait des chances de mourir différentes d'un cadre d'une autre ville. C'est pourtant ce que traduirait l'implémentation de ces ratios.

2.2 Catégories socioprofessionnelles et biais statistiques

La volonté de proposer ici une analyse fine de l'exposition des villes est parfaitement illustrée par le paradoxe de Simpson.

Nous proposons de l'illustrer rapidement.

Prenons l'exemple de deux villes A et B peuplées de cadres et d'ouvriers. On dispose cette année du nombre de décès et du nombre d'individus par csp. On se propose donc de calculer des taux de mortalité :

Ouvriers		Cadres	
Ville A	Ville B	Ville A	Ville B
8 décès / 2 700 hab	23 décès / 8 700 hab	20 décès / 9 000 hab	4 décès / 1 900 hab
$\hat{q_{ouvriers}} = 0.00296$	$\hat{q_{ouvriers}} = 0.00264$	$\hat{q_{cadres}} = 0.00222$	$\hat{q_{cadres}} = 0.00210$

Table 1 – Illustration du paradoxe de Simpson - 1

Dans les deux cas, la ville B affiche un meilleur taux de mortalité.

Toutefois, si l'on construit un résultat global en additionnant naïvement nos effectifs, on trouve que la ville A a un meilleur taux de mortalité :

Ville A	Ville B
28 décès / 11 700 hab	27 décès / 10 600 hab
$\hat{q_A} = 0.00239$	$\hat{q_B} = 0.00254$

Table 2 – Illustration du paradoxe de Simpson - 2

Tenir compte uniquement de cette dernière analyse pourrait faire conclure, à tort, que les cadres et les ouvriers bénéficient de meilleurs conditions de vie dans la ville A. On comprend bien le risque d'introduire ce type de biais dans une tarification.

Ce paradoxe statistique nous enseigne que des proportions mal manipulées peuvent, dans certains cas, donner une fausse image de la réalité.

— Dans un contexte assurantiel classique, il faut comprendre qu'une mauvaise segmentation des populations peut faire apparaître un effet trompeur.

— Dans notre analyse, il faut comprendre qu'une ville comportant une forte proportion d'individus peu à risque pourrait apparaître comme ayant un effet bénéfique sur la mortalité ; quand bien même cette ville aurait *toutes choses égales par ailleurs* une influence négative.

On pourrait, par exemple, supposer que ce soit la prépondérance de cadres à Paris qui tire le risque de décès parisien vers le bas. Un cadre Parisien pourrait ainsi très bien avoir un risque de mortalité supérieur à celui d'un cadre habitant une zone rurale.

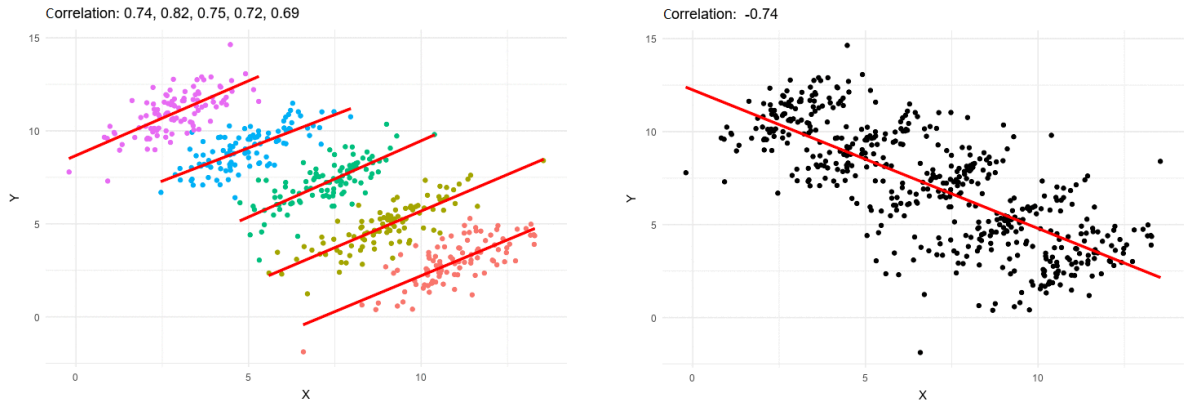


FIGURE 3 – Illustration du paradoxe de Simpson dans le cadre de modèles linéaires

C'est en ce sens qu'il convient de s'assurer que l'on compare des groupes comparables. Dans le cas de nos estimations de décès, nous prendrons donc soin de différencier nos taux de mortalité par catégories socioprofessionnelles, en accord avec les techniques de tarifications usuelles.

Lorsque nous constaterons une sur (ou sous) mortalité dans une ville par rapport à nos estimations, ce ne sera ni l'effet de l'âge, du sexe ou des proportions de catégorie socioprofessionnelle. Nous capterons, autant que possible, des effets pertinents et ne pénaliserons (ou n'avantagerons) pas les individus plus que de raison.

2.3 La diversité du territoire illustrée par un modèle de Cox

L'Insee propose, dans une de ses études, un modèle de Cox dont l'analyse se révélera très intéressante [40]. Cette partie sera d'ailleurs l'occasion d'évoquer la portée de nos travaux, conditionnés par les données librement accessibles.

Ce mémoire n'a pas vocation à traiter du modèle de Cox et nous renvoyons donc vers les sources suivantes pour de plus amples explications : [9] [6]. Néanmoins, pour contextualiser l'étude dont nous parlons, il convient de le décrire succinctement :

Le modèle de Cox[3] *est un modèle où l'on ajuste un "risque de base" μ_x (dépendant du temps) par l'exponentielle d'un modèle linéaire (indépendant du temps) qui vient pondérer μ_x par les caractéristiques de l'individu ou du groupe étudié. Ce modèle présente plusieurs intérêts et permet, en l'occurrence, d'évaluer l'impact d'un paramètre toutes choses égales par ailleurs sur le risque de base. L'Insee propose donc de mesurer l'effet des différentes régions sur le niveau de mortalité.*

L'étude nous confirme les mises en garde que nous avons formulées : traiter les individus par âge et sexe ne suffit pas à gommer les différences de mortalité entre les régions. Par exemple, entre 2012 et 2016, les Hauts-de-France auraient avec cette méthode un niveau de mortalité 22% plus important que l'Auvergne-Rhône-Alpes (utilisée comme référence) et l'Ile-de-France un niveau relatif 6% plus faible.

Plus encore, cette étude montre que, toutes choses égales par ailleurs (en l'occurrence en tenant compte du sexe, de l'âge, du niveau de vie, du diplôme et de la catégorie sociale) des écarts entre régions subsistent. La tendance est alors renversée pour l'Ile-de-France qui présente dans ce cas un niveau relatif de mortalité 6% plus important.

Figure 4 - Rapport de risque de décès par région de résidence

A âge et sexe donnés			Toutes choses égales par ailleurs	
Île-de-France	0,94 ***		Occitanie	0,94 ***
Occitanie	0,97		Pays de la Loire	0,95 *
Provence-Alpes-Côte d'Azur	0,99		Nouvelle-Aquitaine	0,98
Pays de la Loire	0,99		Centre-Val de Loire	0,99
Auvergne-Rhône-Alpes	*Réf.*		Bourgogne-Franche-Comté	0,99
Nouvelle Aquitaine	1,01		Provence-Alpes-Côte d'Azur	1,00
Centre-Val de Loire	1,01		*Auvergne-Rhône-Alpes*	*Réf.*
Bourgogne-Franche-Comté	1,02		Île-de-France	1,06 **
Bretagne	1,11 ***		Bretagne	1,08 ***
Grand Est	1,13 ***		Grand Est	1,10 ***
Normandie	1,16 ***		Normandie	1,11 ***
Hauts-de-France	1,22 ***		Hauts-de-France	1,14 ***

Note : il s'agit d'un modèle de Cox - modèle de durée à risque instantané proportionnel. Sans indication le rapport de risque n'est pas significatif, * s'il est significatif au seuil de 10 %, ** au seuil de 5 %, *** au seuil de 1 %. Les données pour les DOM et la Corse ne sont pas intégrées ici : les résultats ne sont pas suffisamment robustes en raison de la faiblesse des effectifs.

Lecture : il s'agit du rapport entre le risque instantané de décès et le risque instantané de décès de référence. Entre 2012 et 2016, « toutes choses égales par ailleurs », c'est-à-dire à sexe, âge, niveau de vie, catégorie sociale et diplôme donnés, les personnes résidant en Occitanie ont en moyenne un risque de décès inférieur de 6 % (0,94-1) à celui des personnes résidant en Auvergne-Rhône-Alpes.

Champ : France métropolitaine (hors Corse).

Source : Insee-DGFIP-Cnaf-Cnav-CCMSA, Échantillon démographique permanent.

FIGURE 4 – Modèle de Cox - Insee [40]

On constate donc à nouveau que des effets importants existent et qu'ils ne sont pas simples à capter. Le type d'étude que propose l'Insee est malheureusement impossible à reproduire ici, car basé sur l'Échantillon Démographique Permanent [30]. Cette base retranscrit, pour un échantillon de personnes, les informations des recensements depuis 1968. Elle a l'avantage de permettre le suivi des individus jusqu'au décès. Elle autorise ainsi ce type d'analyse, car ici la base de décès est directement rattachée à son exposition.

Nous évoquerons de nouveau cette étude pour mettre en perspective nos résultats, dans le sens où elle permettra de distinguer ce que nos travaux seront en mesure de capter et les effets subtils dont ils ne pourront pas rendre compte.

Les données dont nous disposerons par la suite sont plus hétérogènes. Elles nécessiteront beaucoup d'attention pour ne pas introduire de biais dans notre analyse. Il faudra également évaluer les imprécisions qui pourraient résulter des concessions que nous seront amenés à faire. Néanmoins, de par les travaux réalisés et la méthode que nous nous sommes proposés d'appliquer, nous serons finalement en mesure de proposer une évaluation solide des effets que nous venons d'évoquer et à une maille fine.

La méthode que nous proposons d'appliquer est la suivante :

3. (ou modèle de durée à risque instantané proportionnel)

1. Comparer la réalisation des décès à l'estimation que l'on pourrait en faire par les méthodes de tarifications classiques. En prenant en considération les effets de l'âge, du sexe et des csp des populations, nous capterons uniquement des effets qui échappent à ces trois paramètres.
2. Pour ne pas tenir compte des différences de mortalité qui seraient dues au seul hasard, nous construirons alors un modèle linéaire prédisant ces écarts. Nous disposerons alors finalement de valeurs fiables et résultant d'un modèle compréhensible.

Par cette méthodologie, nous serons finalement en mesure de proposer une analyse pertinente, malgré les difficultés inhérentes à ce type d'étude.

Deuxième partie

Calcul de l'impact de la géographie sur la mortalité

Nous allons maintenant commencer la mise en place de notre analyse. Cette partie sera divisée en deux sections :

— Dans un premier temps nous construirons notre variable d'intérêt. Nous détaillerons divers concepts, les sources et les méthodes employées.

— Dans un second temps, nous mettrons en place nos données prédictives. Nous réaliserons alors une première analyse de leur structure et de leur interaction avec notre variable d'intérêt.

Cette partie permettra de donner un cadre à la suite de nos travaux.

1 Construction de ratios de mortalité par canton français

Cette section détaille la construction des estimations de décès et leur comparaison avec la mortalité observée. Nous y détaillons également les sources et les données utilisées pour mener à bien nos travaux. Diverses notions nécessaires à la bonne compréhension de ce mémoire seront présentées. On rendra également compte d'un certain nombre de problèmes pratiques rencontrés.

L'Institut National de la Statistique et des Études Économiques (Insee) sera largement cité dans ce mémoire. Que ce soit pour ses méthodes, sa fiabilité et par la profondeur de son historique, il s'agit d'une source d'information précieuse. Une présentation détaillée est disponible sur son site [43]. L'ensemble des définitions est centralisé [33].

1.1 Maille et découpage du territoire

L'Insee étudiant le territoire national, il décline ses découpages statistiques sur un certain nombre d'échelles. Certains découpages sont naturels, comme lorsque les données concernent le pays, les régions, les départements ou les communes. D'autres le sont moins, comme l'échelle infra-communale (IRIS) et celle du Canton-Ville (CV).

L'échelle communale [CODGEO]

Les bases de données de l'Insee ne sont pas toujours disponibles aux mêmes échelles mais l'échelle communale n'est pas rare.

Comme le définit lui-même l'Insee [32], la commune est la plus petite subdivision administrative française. C'est par elle que sont réalisées un certain nombre de formalités administratives, ce qui en fait une antenne de choix pour mailler le territoire. Au 1er janvier 2010 on comptait 36 682 communes, dont 36 570 en métropole. L'Insee note cette codification par CODGEO (pour "code géographique").

On peut se donner une idée de la précision de cette maille par la cartographie suivante :

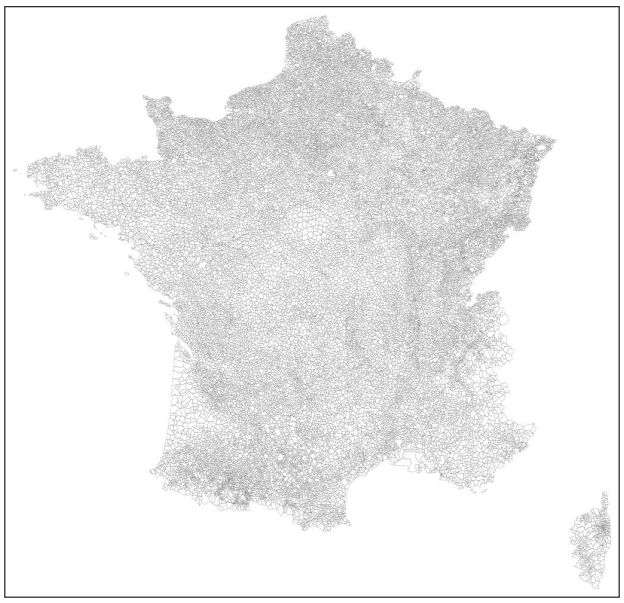


FIGURE 5 – Maille communale
Obtenu à partir de www.CommunesFrance.fr

L'échelle cantonale [CV]

Une maille plus large est la maille "Canton-ou-ville" ou "pseudo-canton", noté CV. Ce découpage sera celui utilisé tout au long de ce mémoire, pour des raisons que nous détaillerons par la suite. On peut se faire une idée de sa précision grâce à la représentation suivante :

FIGURE 6 – Maille du pseudo-canton
Obtenu à partir de www.france-decouverte.geoclip.fr [20]

Les pseudo-cantons (que nous désignerons plus simplement par "cantons") sont des communes ou regroupement de communes [31]. En général, il s'agit simplement de considérer que les communes périphériques d'une ville (Par exemple Villeurbanne pour Lyon) en font partie. Néanmoins, dès lors que les villes sont d'une taille raisonnable, les mailles ville et canton coïncident l'immense majorité du temps [4].

4. Lyon et sa banlieue font figure d'exception en l'occurrence.

Les codifications CODGEO et CV sont plutôt intuitives dès lors que l'on sait à quoi elles correspondent :

— Chaque département est codifié par son numéro (ex : 01 pour l'Ain ; 2A pour la corse du sud ; 972 pour la Martinique).

— A l'intérieur de chacun d'eux, on numérote ensuite le maillage en question, c'est à dire :

- par 3 chiffres pour l'échelle communale [CODGEO] (ex : Bourg-en-Bresse dans l'Ain (01) porte le numéro 053 ; elle est donc codifiée par 01053).
- 2 chiffres pour l'échelle cantonale [CV] (ex : Bourg-en-Bresse dans l'Ain (01) est désigné par le numéro 99 ; elle est donc codifiée par 0199).

L'usage d'un format standardisé est très appréciable, mais un certain nombre de subtilités rend parfois leur utilisation complexe.

Du fait du caractère administratif des découpages, ces derniers changent en effet au fur et à mesure des années. Il arrive que des villes soient regroupées au sein d'un même canton ou qu'au contraire elles s'en dissocient. Si ces changements restent minimes, leur prise en compte impose néanmoins la manipulation de tables de passage [50] via un gestionnaire de base de données.

On trouve deux types de bases rendant compte de ces évolutions :

- La table de passage : qui décline pour chaque commune l'historique de ses codifications
- Les tables d'appartenance : qui résument chaque année les différentes codifications (CODGEO, CV, Zone d'Emploi, etc...) de chaque commune.

En toute rigueur, manipuler des CV sur plusieurs années doit donc passer par la manipulation des CODGEO.

Finalement, il ne semble pas exister de base faisant correspondre les villes et leurs coordonnées sur le site de l'Insee. Cette information a donc été récupérée et intégrée par ailleurs [10].

L'ensemble des difficultés que nous venons d'évoquer ne se rencontrent pas lorsque l'on étudie une seule année et un seul type de base. Dans ce cas, chaque base de données se suffit à elle-même et aucun retraitement n'est nécessaire. De par les perspectives que nous nous sommes proposées d'explorer, certains retraitements ont ici été nécessaires. La maille que nous avons jugée pertinente étant le CV, il a notamment fallu convertir l'ensemble des bases de données à la maille CODGEO en bases de données à la maille CV (cette dernière n'étant jamais proposée). Le retraitement s'est fait au cas par cas et a consisté à grouper chaque ville selon son canton grâce à un logiciel de base de données [5].

La gestion de l'ensemble des bases de données aura été un point chronophage mais nécessaire.

1.2 Décès français localisés au domicile de l'individu

L'Insee met à disposition l'historique des décès survenus en France (disponible ici [37]).

L'intérêt de ces bases est évident :

— Elles sont exhaustives.

— Elles sont disponibles pour chaque année depuis 1970.

— L'âge, le sexe, le lieu de naissance et le lieu de décès des individus sont renseignés (à la maille CODGEO).

Idéalement, on souhaiterait disposer des informations comme la csp (qui aurait autorisé des analyses plus fines encore) mais ce point n'est pas rédhibitoire en soi.

En revanche, le domicile du défunt n'est pas renseigné. Or, notre étude suppose de pouvoir rattacher l'exposition des villes à la réalisation effective du risque. Ces bases ne peuvent donc pas être utilisées ici.

5. en l'occurrence Microsoft Access

Le domicile du défunt est une information dont dispose l'Insee mais qui n'est pas partagée ici. Ainsi, l'utilisation de ces bases (qui profitent d'un bon historique) nous est rendue impossible. Comme les villes possédant des hôpitaux cumulent les décès au dépend des autres villes, le lieu de décès est inexploitable. Étudier les décès au niveau du département aurait éventuellement été envisageable en posant l'hypothèse que la majorité des décès ont lieu à distance raisonnable du domicile.

Si elles n'ont pas été retenues, l'intérêt évident de ces bases fait qu'elles méritent d'être mentionnées ici.

Les travaux que nous proposons ici n'auraient en fait pas été possibles sans la mise à disposition de fichiers de décès quotidiens pendant la pandémie de la Covid-19 [29]. Ces données -presque au même format que celles que nous venons de mentionner- ont l'avantage d'indiquer le domicile du défunt. Ce sont donc ces bases qui ont été finalement retenues.

La principale réserve que l'on peut émettre à propos de cette base opportune est qu'elle est extrêmement récente. Elle porte sur les décès survenus après le 1er janvier 2018 uniquement. Nous ne serons donc pas en mesure de proposer une étude sur le temps long.

Une inquiétude légitime serait la non-exhaustivité de cette base. L'Insee indique dans la description de ces données que : *"les données relatives aux années 2018 et 2019 ont été extraites à la date du 15 octobre 2020 et sont désormais figées. Elles sont considérées comme définitives."* Entre le moment où ont été démarrés nos travaux (été 2020) et la finalisation, nous avons actualisé ces bases (21 janvier 2021) : pour l'année 2018 un seul décès avait été ajouté.

On ne s'attend donc pas à subir d'importantes déformations dues à une mauvaise actualisation des décès.

Nous avons finalement fait le choix de retenir les années 2018 et 2019 pour :

— Disposer d'un nombre robuste de décès

— Ne pas intégrer de données sujettes à caution

— Ne pas impacter notre étude des effets de la Covid-19 [6]

Ainsi, nous disposons maintenant de deux années de sinistres (2018 et 2019). Nos données sont localisées à la ville et comportent l'âge et le sexe des individus.

1.3 Le recensement comme base d'exposition

Le type d'approche que nous proposons ici suppose de mettre en parallèle la réalisation du risque et une exposition. Cette étude est rendue possible -d'une part- par l'utilisation de la base de décès réels que nous venons de présenter et -d'autre part- par les fichiers de recensement de la population française.

Comme le définit lui-même l'Insee [44], "*le recensement de la population a pour objectifs le dénombrement (...) de la population résidant en France et la connaissance de leurs principales caractéristiques : sexe, âge, activité, professions exercées, ...*". Il s'agit d'une source importante pour l'Insee. Les informations qu'elle contient sont déclinées en de nombreuses bases de données.

Nous allons détailler ici l'exploitation du recensement de la population française. Au terme de cette partie, nous disposerons d'une base d'exposition détaillant, pour l'ensemble des cantons français, le nombre d'individus par âge, sexe et csp.

6. Il pourra être envisagé, à l'avenir, d'étudier son impact via les méthodes développées dans ce mémoire, mais il s'agit d'un objectif différent de celui que nous nous sommes fixé ici.

Le recensement présente le double avantage de se vouloir exhaustif et d'être librement accessible sous la forme de "fichiers de détail" annuels [46]. Du fait de leur taille, la manipulation de ces fichiers n'est pas aisée sans passer par un logiciel de bases de données, mais pour une exploitation année par année, l'utilisation des fichiers .txt localisé par zone est néanmoins efficace [7].

En termes de données, on y trouve notamment les informations suivantes :

- Le canton/la ville de domicile
- L'année de naissance
- L'âge au moment du recensement
- Le sexe
- La catégorie socioprofessionnelle
- Le département de naissance
- Le diplôme le plus élevé
- Le statut conjugal
- Le nombre d'enfants
- etc...

La liste complète des variables est disponible sur la page de téléchargement des fichiers de détails.

Les fichiers de détail se déclinent à différentes échelles, dont :

1. Les fichiers "Logement", qui contiennent des données localisées par logement (au niveau CODGEO et IRIS) et qui décrivent les caractéristiques des logements et celles des ménages qui les occupent. Cette base a été étudiée mais pas retenue. [8]
2. Les fichiers "Canton-Ville", qui contiennent les caractéristiques de chaque personne recensée (âge, sexe et csp notamment) localisée au canton, ainsi que les caractéristiques de son ménage et de sa résidence principale. C'est cette base qui a finalement été retenue.

Les fichiers sont anonymisés mais permettent d'évaluer avec précision les populations des cantons. Le recensement le plus récent étant celui de 2017 (calculé sur 2015-2019), c'est donc celui que nous avons utilisé.

De plus amples détails sur les fichiers de recensement sont disponibles ici : [36]

Considérations statistiques

Le recensement est réalisé ainsi :

— Les communes de moins de 10 000 habitants sont recensées exhaustivement tous les cinq ans.

— Les communes de 10 000 habitants ou plus font, elles, l'objet d'une enquête annuelle auprès d'un échantillon de 8% de leur population. Au bout de 5 ans, les résultats du recensement sont calculés à partir de l'échantillon de 40% de leur population ainsi constitué.

Les résultats de recensement sont ainsi produits à partir des cinq enquêtes annuelles les plus récentes. Les informations issues de l'enquête la plus ancienne sont abandonnées, et celles de l'enquête la plus récente sont prises en compte. De ce fait, il n'existe pas de recensement décrivant la population française en 2018 et 2019 à l'heure actuelle. Nous utiliserons donc le recensement 2017, constitué sur la période 2015-2019.

Il faut noter que la variable âge (*"aged"*) représente l'âge des personnes (en différence de millésime) au moment où elles ont été enquêtées, et non celui à l'année médiane des 5 années d'enquête.

7. Le logiciel R a été utilisé tout au long de cette partie, que ce soit pour manipuler ces bases ou pour leur appliquer les taux de mortalité que nous définirons plus loin.

8. Elle utilise en effet la "personne de référence du ménage" comme donnée. Plus encore, certaines informations sont régulièrement incomplètes comme le sexe de la personne, ce qui est rédhibitoire.

Nous avons choisi de conserver cette variable plutôt que de calculer l'âge à partir de l'année de naissance [9].

L'Insee précise que "les effectifs supérieurs à 500 peuvent normalement être utilisés en toute confiance". La totalité des cantons répondant à cette description, ils sont donc issus d'un calcul statistique satisfaisant.

Un ensemble de conseils d'utilisation des fichiers de recensement est disponible ici [28].

On comprend en définitive que l'exposition utilisée dans ce mémoire ne sera pas le parfait reflet de la population en 2018 et 2019. Si ces réserves sont à mentionner, elles restent acceptables.

L'utilisation de données publiques contraint nécessairement les travaux, car il faut se contenter de ce qui est accessible. Si chacune des bases que nous avons exploitées comporte des limites relatives, nous avons tâché de comprendre les biais induits et les avons évalués autant que possible. L'ensemble de ces limitations a été jugé acceptable ou a été contrôlé au mieux. D'autre part, elles ne sauraient remettre en cause notre analyse qui, justement, a été construite pour proposer des résultats robustes malgré l'absence de base de données parfaitement adaptée.

La plupart des problèmes rencontrés sont de fait la conséquence des caractéristiques des bases de données dont nous disposons. Compte tenu de tout ce qui a été mentionné précédemment, l'idéal serait encore et toujours de disposer d'une base d'exposition et d'une base de mortalité, liées et détaillées.

Ces considérations faites nous disposons, au terme de cette partie, d'une base d'exposition de bonne qualité.

1.4 Construction des taux de mortalité

1.4.1 Table du moment

On utilise comme référence les *Tableaux de séries longues* proposés par l'Insee [51]. On trouve sur cette page un ensemble de tables de mortalité dont voici la décomposition :

- Tableau 67 : Tables de mortalité annuelles par groupe d'âge de 5 ans
- Table 68 : Tables de mortalité, fonctions de survie et Espérances de vie résiduelles annuelles détaillées par âges
- Tables 69 : Reprend les éléments de la table 68, mais décomposés en trois tables (QMORT pour les taux, SUR pour les fonctions de survie, ESP pour l'espérance de vie résiduelle)
- Table 70 : Tables de mortalité infantile annuelles.

La table de mortalité que nous considérerons est celle contenue dans les tables T68 (ou de façon équivalente dans la table T69QMORT).

Ces taux, aussi appelés "Taux du moment", sont calculés chaque année sur des périodes de 3 ans. La table que nous avons choisi est la plus récente, c'est à dire 2018. Elle a donc été calculé sur les années 2016, 2017 et 2018, ce qui coïncide avec le recensement retenu, c'est à dire 2017.

D'un point de vue technique, ces taux sont calculés par âge atteint dans l'année (i.e. en différence de millésime) via un estimateur de Hoem défini ainsi :

$$\hat{q_a} = \frac{D_{(a,n)}}{N_{(a,n])}}$$

9. Prenons l'exemple d'une ville de moins de 10 000 habitants recensée en 2015. Pour estimer sa population de trentenaire en 2018, est-il plus raisonnable de vieillir les habitants de 27 ans ou de conserver la dernière image dont on dispose ? Les décès, les départs et les arrivées dans la villes sont moins bien pris en compte dans le premier cas ; les subtilités de la pyramide des âges moins bien pris en compte dans le second. Il n'y a donc pas de solution idéale. La précision des deux méthodes pour l'estimation des décès a été évaluée et les résultats sont extrêmement proches au niveau national : de l'ordre de 0.5% de différence sur notre population d'étude (les 30-62 ans).

$D_{(a,n)}$ représente le nombre des décès intervenus l'année n chez les personnes de la génération née en $n-a$ et $N_{(a,n)}$ l'effectif de cette génération au 1er janvier. A noter que l'Insee prend soin de corriger les effectifs des soldes migratoires.

Pour plus de détails sur la façon donc sont calculés les indicateurs démographiques proposés par l'Insee, un fichier de documentation est disponible ici : [35].

On dispose maintenant d'un taux de mortalité calculé sur trois ans, différencié par âge et sexe. Ce taux étant "centré" sur l'année 2017, on lui appliquera également les bases d'amélioration de mortalité RGA afin de le projeter et le ramener aux année 2018 (taux amélioré de 1 an) et 2019 (taux amélioré de 2 ans).

1.4.2 Introduction des catégories socioprofessionnelles

En termes de ciblage de risque, considérer les csp (ou une catégorisation analogue) est une étape importante dans la gestion du risque. Au même titre que la souscription médicale, son but est d'homogénéiser le risque. Ce type de procédés justifie, entre autre, l'hypothèse de risque identiquement distribué.

Le métier est, comme l'âge et le sexe, représentatif du risque de mortalité d'un individu. La vie professionnelle conditionne notre mode de vie et notre environnement, autant qu'il reflète notre réalité sociale. Plus encore, si un fort impact des conditions de travail sur l'état de santé a été depuis longtemps démontré, la réciproque est également vrai ; l'état de santé impacte également sur les trajectoires professionnelles.

Les csp sont un moyen simple de différencier les types d'activités et leurs spécificités. Elles répondent à un besoin de stratification du risque de mortalité et sont, logiquement, très utilisées dans les techniques de tarification actuarielles.

La nomenclature des Professions et Catégories Socioprofessionnelles que nous utiliserons dans ce mémoire sera celle proposée par l'Insee. Elles sont au nombre de huit et sont également stratifiées en sous-catégories dont il ne sera pas fait mention ici. Elles sont définies ainsi :

1. Agriculteurs exploitants
2. Artisans, commerçants et chefs d'entreprise
3. Cadres et professions intellectuelles supérieures
4. Professions Intermédiaires
5. Employés
6. Ouvriers
7. Retraités
8. Autres personnes sans activité professionnelle

Les définitions que donne l'Insee de ces catégories sont accessibles dans les sources [45]

Malgré ses qualités le découpage des csp reste une simplification ; certaines catégories recouvrent parfois des réalités bien différentes tandis que d'autres (comme les ouvriers et les employés) se recoupent presque par endroit. Ce découpage reste néanmoins satisfaisant et le meilleur dont nous disposons.

Afin d'introduire cette information dans notre analyse, nous nous basons sur une étude de l'Insee parue en 2016 [52]. Elle propose une estimation des taux de mortalités $q_{a,s,c}$ par âge, sexe et catégories socioprofessionnelles. L'étude se restreint en revanche aux adultes de plus de trente ans. D'un autre côté, on ne dispose pas de données sur la catégorie socioprofessionnelle anciennement exercée par les personnes retraitées dans le recensement. Pour ces raisons, nous avons choisi de restreindre notre étude à une plage d'âge de 30 à 62 ans.

Cette plage centrée sur les actifs coïncide globalement avec celle que l'on retrouve généralement dans les portefeuilles emprunteurs, l'âge moyen des emprunteurs étant en France d'environ 37 ans. Il ne semble néanmoins pas y avoir de raison à ce que les moins de 30 ans ne soient pas concernés par les résultats de notre étude. Les personnes retraitées en revanche, du fait de la difficulté à les catégoriser, doivent être exclues des conclusions que nous pourrons tirer.

Les méthodes utilisées pour estimer la mortalité par csp sont détaillées dans un rapport de l'Insee [23]. On y apprend que c'est la *méthode de Brass* qui a été utilisée. Cette méthode permet de construire plusieurs tables de mortalité à partir d'une seule (servant de référence). On admet l'existence d'une relation affine entre les logits des quotients cumulés de deux tables et on ajuste ensuite cette relation linéaire. Ici, la référence est une table de mortalité pour les hommes ou les femmes résidant en France métropolitaine entre 2009 et 2013. Elle est basée sur l'Échantillon Démographique Permanent.

Si l'utilisation de ces tables de mortalité présente un intérêt certain, elle sont en revanche plutôt anciennes [10]. Il n'est en effet pas évident que les différences relatives entre les taux soient restées les mêmes. Le rapport de l'Insee que nous venons de mentionner propose à ce titre une représentation intéressante. Après avoir détaillé les deux méthodes principales de standardisation de la mortalité (directe et indirecte), le rapport retient celui que vous avons évoqué dans la première partie de ce mémoire (standardisation directe). Il l'utilise alors pour calculer un ratio de taux standardisés (RTS), en anglais « comparative mortality » ou « standardized rate ratio », obtenu en divisant les taux standardisés de deux populations :

$$RTS_{csp} = \frac{\tilde{Q_{csp}}}{\tilde{Q}}$$

où $\tilde{Q_{csp}}$ est le taux standardisé présenté plus haut (en considérant ici des csp) et $\tilde{Q}$ le taux standardisé de la population de référence (ici les habitants de France métropolitaine, toutes csp confondues). L'intérêt de ces ratios, une fois calculés sur plusieurs périodes, est qu'ils démontrent une certaine stabilité dans le comportement des taux de mortalité entre les csp. Les écarts constatés dans les années 90 sont presque les mêmes 20 ans plus tard :

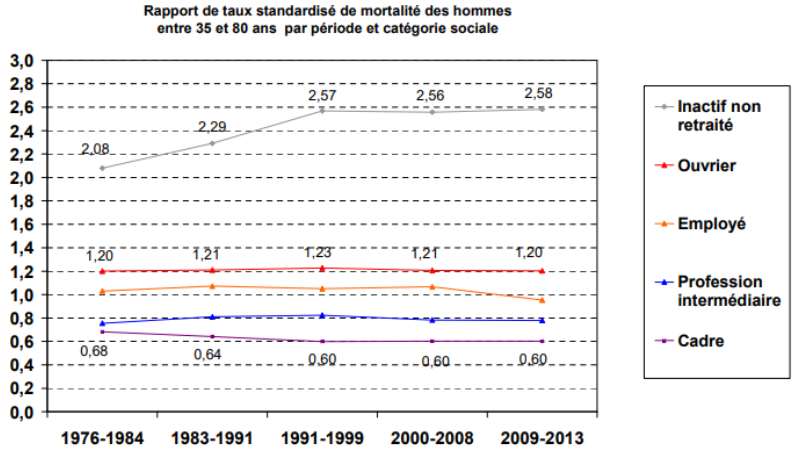


FIGURE 7 – Stabilité des écarts de mortalité entre les csp (Hommes)

10. Nous avons à ce titre contacté l'Insee qui nous a confirmé que l'étude ne serait pas renouvelée d'ici la fin de nos travaux.

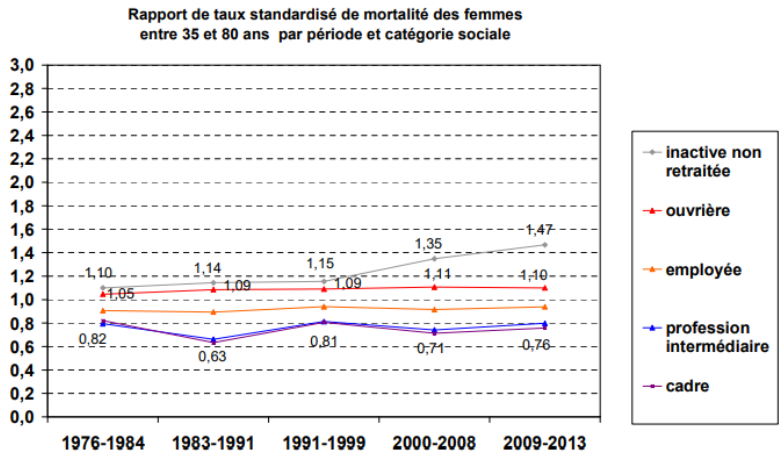


FIGURE 8 – Stabilité des écarts de mortalité entre les csp (Femmes)

Malgré la standardisation, qui gomme par principe les subtilités, ces représentations mettent en évidence que les différentiels de mortalité entre les csp sont relativement constants. Nous poserons donc l'hypothèse que ces écarts se sont maintenus ou ont dévié uniformément selon l'âge depuis 2010.

La détérioration du ratio chez les femmes inactives de 1990 à 2010 rappelle néanmoins que ces effets peuvent toujours évoluer. L'hypothèse de travail que nous posons ici sera mise en perspective dans la suite de nos travaux.

On pose donc le coefficient suivant :

$$C_{csp,a,s} = \frac{q_{csp,a,s}}{q_{a,s}}$$

Où $q_{a,s}$ est le taux de mortalité de référence ; ici le taux du moment 2016-2018. Ce coefficient viendra pondérer les taux de mortalité du moment. Les rares individus retraités présents dans notre plage d'âge se verront appliquer un taux non différencié.

Graphiquement, on peut observer l'impact relatif qu'aura l'introduction des csp :

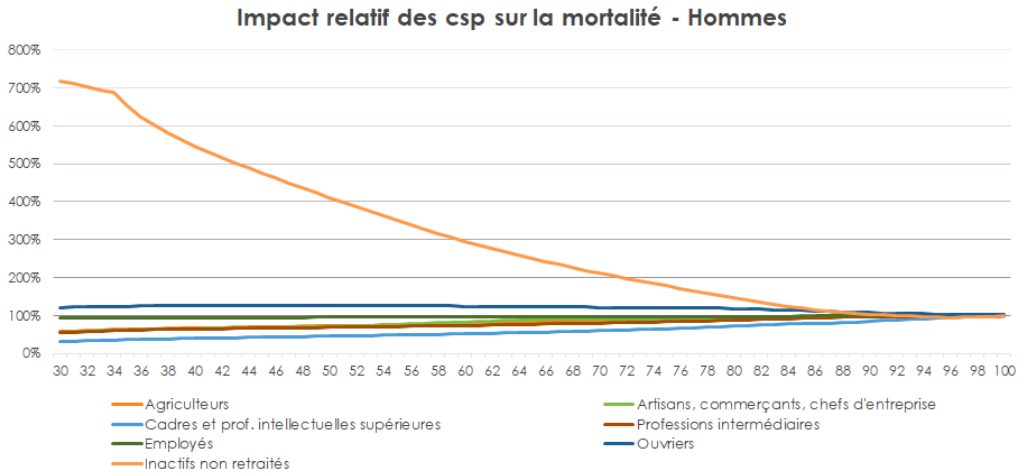


FIGURE 9 – Coefficients appliqués par csp - Hommes

Impact Femmes en annexe (figure 58)

Il faut noter ici que les coefficients attribués aux inactifs sont extrêmement forts. Une fois retirés de notre représentation, les différenciations appliquées aux actifs sont plus visibles :

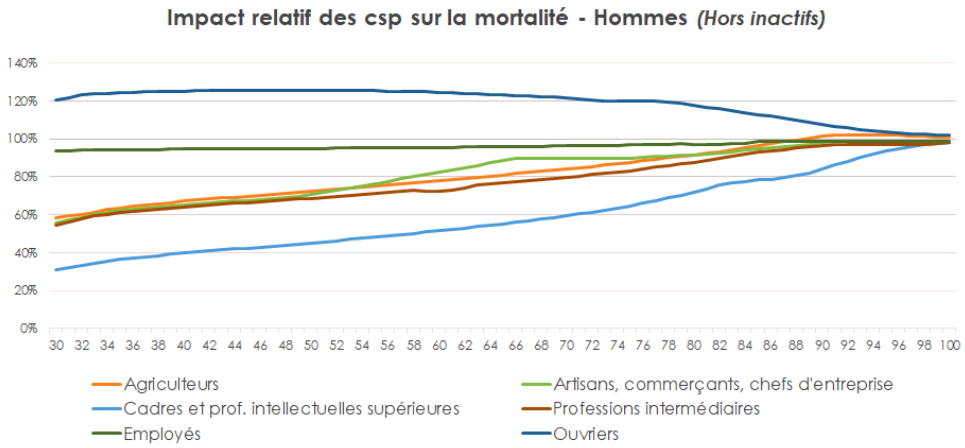


FIGURE 10 – Coefficients appliqués par csp - Hommes *(Hors inactifs)*

Impact Femmes en annexe (figure 59)

Au terme de cette partie nous disposons donc d'un taux de mortalité différencié par âge, sexe et catégorie socioprofessionnelle.

1.5 Nombre de décès théoriques par canton

1.5.1 Estimation

Notre estimation des décès dans les cantons se base sur les techniques de tarification classique, chaque canton étant assimilé à ce que pourrait être un portefeuille d'assurance.

On pose :

A_{min}, A_{max} l'âge minimum et maximum considéré *(ici 30 et 62 ans respectivement)*.

$N^{cv}_{a,s,csp}$ le nombre d'individus d'âge a, de sexe s et de csp csp dans le canton cv considéré.

$q_{a,s}$ le taux de mortalité observé dans la population de référence *(ici les habitants de France métropolitaine)* chez les individus d'âge a et de sexe s

Cl l'ensemble des csp *(ici au nombre de 8)*.

$C_{csp,a,s}$ le coefficient pondérant le taux de mortalité $q_{a,s}$ selon a, s et csp.

$$\hat{DC}_{cv} = \sum_{s\in(H,F)} \sum_{csp\in Cl} \sum_{a=A_{min}}^{A_{max}} C_{csp,a,s} q_{a,s,} N^{cv}_{a,s,csp}$$

Dans un premier temps, nous étudierons également l'estimation n'intégrant pas la différenciation par csp. On notera donc dans ce chapitre :

— $\hat{DC}^{csp}_{cv}$ l'estimation de décès tenant compte de la csp.

— $\hat{DC}_{cv}$ l'estimation de décès ne tenant pas compte de la csp.

Pour rappel, afin de proposer, pour chaque canton français une estimation de décès en 2018 et 2019, nous avons donc utilisé :

— Le recensement 2017

— La table de décès calculée sur 2016-2018 (différenciée par sexe et âge) améliorée (via les bases d'améliorations RGA) de :
  — 1 an pour la prédiction 2018
  — 2 ans pour la prédiction 2019
— Une étude étudiant le niveau relatif de mortalité entre les csp.

Nous avons étudié l'intérêt d'utiliser des bases plus anciennes (Recensement 2016 et taux 2015-2017), mais il se trouve que ces données donnent de moins bonnes estimations, que ce soit pour 2018 ou 2019.

Une fois nos estimations par canton réalisées, nous obtenons les résultats suivants (pour la France métropolitaine, chez les 30-62 ans) :

Année	**Décès th.**	**Décès th. (csp)**	**Décès réels**	*Biais*	*Biais - csp*
2018	75 588,80	80 472,33	71 506	*+5.70%*	*+12,54%*
2019	74 552,25	79 349,28	69 745	*+6,89%*	*+13,77%*
2018+2019	150 141,05	159 821,61	141 251	+6,29%	+13,15%

TABLE 3 – Décès réels et décès théoriques

On remarque un biais dans notre estimation non différenciée par csp, avec une surestimation des décès de +6.29% sur 2018-2019. On remarque également que ce biais augmente lorsque l'on introduit les csp. Nous allons maintenant contextualiser ces résultats et expliquer en quoi il restent pertinents pour notre analyse.

Estimation hors csp

Lorsque l'exposition est suffisamment importante [11] (comme c'est le cas pour le recensement) on peut poser l'approximation :

$$\hat{q_a} \sim \mathbf{N}(q_a, \sigma_a) = \sqrt{\frac{q_a(1-q_a)}{N_a}}$$

En conséquence, on définit l'intervalle de confiance d'ordre α % d'un estimateur de Hoem par :

$$\hat{q_a} \pm u_{\frac{\alpha}{2}} \sigma_a$$

où $u_{\frac{\alpha}{2}}$ est le quantile de la loi normale d'ordre $\frac{\alpha}{2}$ [61].

Étant donné que les taux de mortalité 2016-2018 utilisés dans cette étude on été calculés par ce biais, nous avons déduit un intervalle de confiance par âge pour nos qx (en utilisant les pyramides des âges 2016-2018). Dans ces conditions, nous estimons que le demi intervalle de confiance d'ordre $\alpha = 2.5\%$ est de 3% en moyenne.

Notre estimation présentant un biais supérieur sur un périmètre à peu près équivalent (i.e. la population métropolitaine), nous pouvons déduire qu'il ne résulte pas uniquement du hasard, mais également des différentes approximations (nécessaires) qui ont été faites jusqu'ici. A savoir : l'ancienneté relative du recensement et sa construction, l'ancienneté relative des taux utilisés couplée à une amélioration de mortalité probablement trop prudente pour la population générale.

Estimation avec csp

Notre seconde estimation (différenciée par csp) dévie plus significativement. Il faut rappeler ici que l'étude utilisée pour estimer nos coefficients de mortalité par csp a été réalisée sur une exposition différente (l'Échantillon Démographique Permanent). De plus, malgré la relative stabilité des effets par csp que nous avons constaté, ces données restent relativement anciennes et ont pu évoluer depuis.

11. C'est à dire $N_a q_a > 5$ ou $N_a(1-q_a) > 5$

Au vu des coefficients appliqués, nous supposons que cette surestimation est probablement due à la pénalisation des inactifs. Ces derniers ont en effet des taux de mortalité considérés comme très supérieurs à la moyenne et sont donc les plus susceptibles de faire monter nos estimations. Pour observer cet effet, il suffirait que la proportion d'inactifs ait augmenté et que la composition de ce groupe se soit, de fait, améliorée en termes de mortalité depuis 2013.

Malgré tout, les csp améliorent nos estimations et restent donc une information pertinente à introduire. Comme nous allons le voir, elles permettent en effet une réduction significative de la variance des résidus.

1.5.2 Apports successifs des différents paramètres

Nous allons maintenant évaluer les apports successifs des différents paramètres entrant dans notre estimation, à savoir l'âge, le sexe et la csp des individus.

On note ε_{cv} l'écart de prédiction constaté canton par canton :

$$\varepsilon_{cv} = DC_{cv} - \hat{DC}_{cv}$$

Pour observer l'évolution de la variance de cet indicateur selon la méthode de calcul, on estime $\hat{DC}_{cv}$ des manières suivantes :

— En utilisant un taux de mortalité indifférencié par âge, sexe, csp (i.e. $Q_I = \frac{Deces}{Population}$). On multiplie simplement ce taux général par le nombre d'habitants du canton pour réaliser notre estimation.

— En utilisant un taux de mortalité indifférencié par sexe q_a.

— En utilisant le taux $q_{a,s}$.

— En utilisant le taux $q_{a,s,csp}$.

On peut alors constater que les ajouts successifs d'informations réduisent la variance de nos résidus :

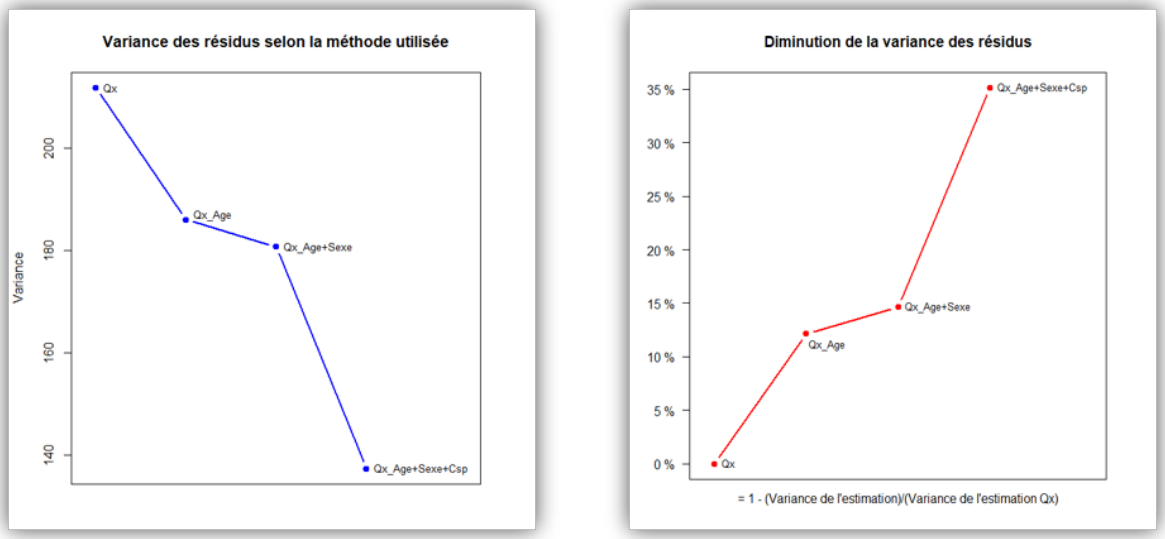


FIGURE 11 – Amélioration de la variance par l'intégration d'informations

L'intérêt de mieux caractériser les populations est ici clair, avec in fine une diminution de 35% de la variance des résidus de nos estimations.

On peut tirer plusieurs autres constats de ces représentations. Pour la population métropolitaine des 30-62 ans :

— Les différences de proportions hommes-femmes entre les cantons sont relativement peu impactantes. La proportion 50-50 est à peu près respectée la majorité du temps.

— Les catégories socioprofessionnelles ont un pouvoir de discrimination plus fort que l'âge sur la mortalité. C'est à dire que les écarts socio-économiques sont plus impactant sur la mortalité que les déformations de la pyramide des âges d'un canton à l'autre.

En conclusion, la surestimation de décès évoquées précédemment n'est pas rédhibitoire.

Si nous partons du principe que le biais est uniformément réparti entre les villes françaises, l'impact est même théoriquement nul une fois nos estimations redressés. Cette hypothèse a été jugée raisonnable. Nous nous permettrons donc d'ajuster nos estimation de décès par canton. On définit le coefficient :

$$\alpha = \frac{DC}{\hat{DC}}$$

Où :

— DC est le nombre de décès constatés en France métropolitaine sur 2018 et 2019.

— $\hat{DC}$ est le nombre de décès estimé en France métropolitaine sur 2018 et 2019.

Ce qui nous permet de corriger notre estimation $\hat{DC}_{cv}$ du nombre de décès survenus dans le canton cv :

$$\hat{DC}^*_{cv} = \alpha \hat{DC}_{cv}$$

Cette manipulation nous permet de redresser nos estimations et de centrer nos ratios de mortalité autour de 1. Le biais introduit pas l'ajout des csp est donc vraisemblablement peu impactant et entièrement compensé par l'amélioration globale des estimations.

Nous disposons donc, au terme de cette partie, d'une estimation efficiente du nombre de décès par canton. Cette estimation a été corrigée de son biais et bénéficie d'une variance aussi faible que possible.

1.6 Construction et étude des ratios de mortalité par Canton

En appliquant les différents éléments évoqués jusqu'ici, nous sommes maintenant capable de calculer un ratio de mortalité pour l'ensemble des cantons français :

$$Y_{cv} = \frac{DC_{cv}}{\hat{DC}^*_{cv}}$$

où DC_{cv} est le nombre de décès réel survenus dans le canton cv et $\hat{D^*_{CV}}$ le nombre de décès théorique différencié par csp et redressé.

On choisit de traiter des ratios de mortalité plutôt que des résidus pour plusieurs raisons :

— Cela permet d'homogénéiser notre variable d'intérêt autant que possible.

— Cela permettra de disposer de ratios et non de valeurs brutes. L'interprétation et l'utilisation des résultats s'en trouvera simplifiée.

A noter que le modèle que nous construirons sera pondéré par le nombre d'habitants de 30-62 ans dans le canton considéré. Les ratios n'auront donc pas tous le même impact malgré cette uniformisation : on ne perd donc pas d'information ici.

On peut observer les résultats obtenus sur les représentations suivantes :

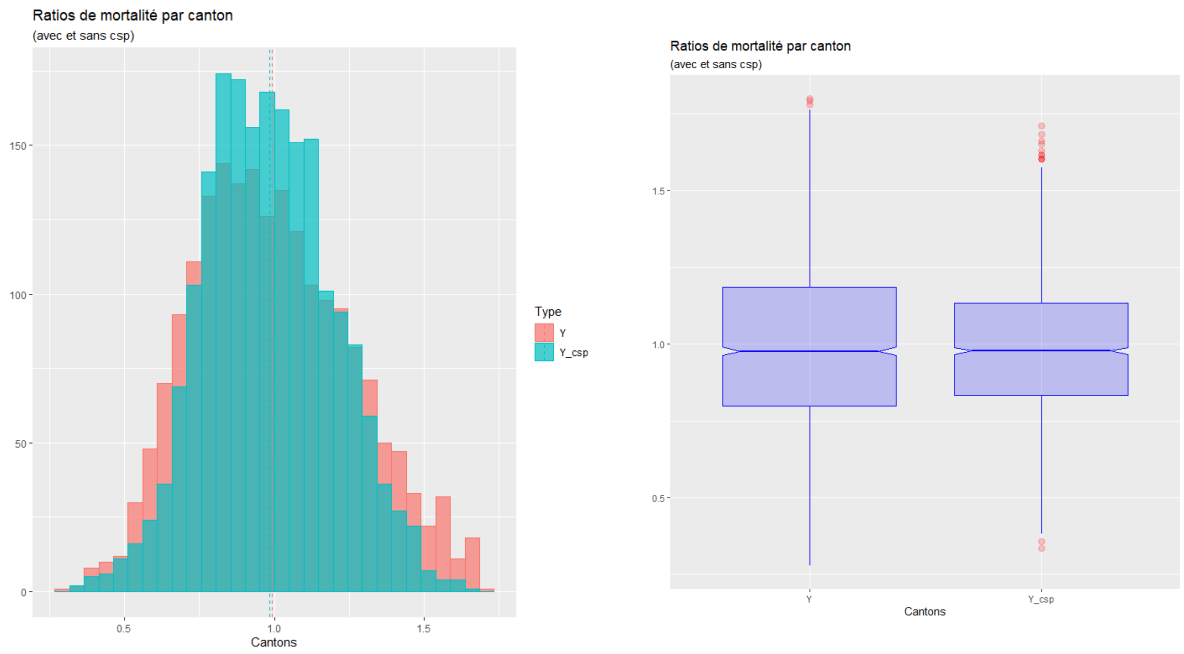


FIGURE 12 – Distributions des ratios de mortalité par canton selon la méthode employée

On constate, comme nous l'avions déjà fait remarquer, une amélioration de la qualité des prédictions lors de l'ajout de la csp.

A partir d'ici nous ne considérerons plus que les ratios résultant de la différenciation par csp.

Une fois ces ratios mis en carte, nous obtenons la représentation suivante :

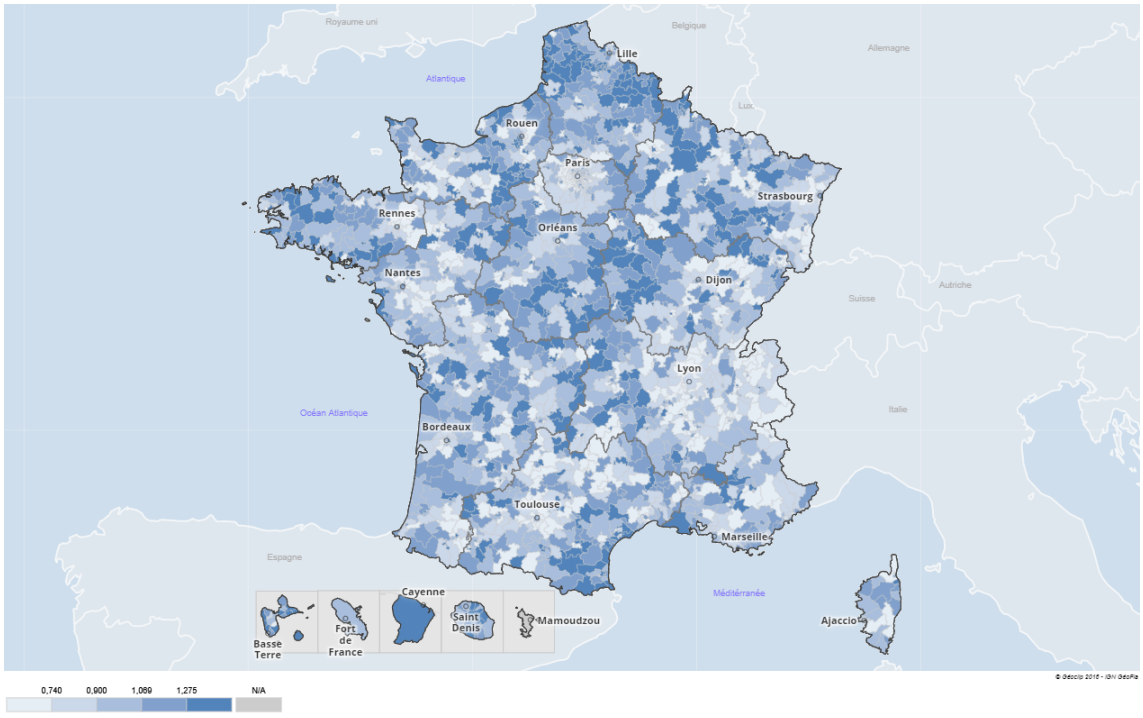


FIGURE 13 – Ratios de mortalité par canton français (après introduction de l'âge, du sexe et des csp)
Obtenu à partir de www.france-decouverte.geoclip.fr - Données importées [20]

Peu d'effets clairs apparaissent dans cette première représentation. On remarque que, malgré l'introduction des csp, l'Ile de France continue de ressortir positivement. La région Rhône-Alpes semble également présenter un effet positif. Le nord et le centre de la France ressortent plutôt négativement. A noter que les effets que l'on remarque restent, malgré deux années de sinistres, très hétérogènes.

Il semble complexe au vu de cette représentation de considérer seul le critère géographique. Notre analyse permettra de mettre en évidence des effets plus clairs et moins sujets à la variabilité que nous observons ici.

Au terme de cette partie, nous disposons donc d'une évaluation du risque de sur ou de sous mortalité pour chaque canton de France métropolitaine. Nous avons ainsi pu constater des différences de mortalité sur le territoire métropolitain dont les techniques de tarifications classiques ne rendent pas compte.

La prochaine partie introduira à ce titre une série de bases de données localisées au canton et mettra en application des techniques d'analyse de données. Cela nous permettra d'étudier et de contextualiser les différences que nous venons de mettre en évidence, et ce afin de comprendre ce qui se joue ici.

Du fait de leur spécificité, les territoires d'outre-mer n'ont pas été introduits dans la suite de notre étude.

2 Analyse des différences géographiques mesurées

2.1 Introduction des données prédictives

On se propose maintenant d'introduire un certain nombre de données Insee à la maille du pseudo-canton. Cela nous permettra dans un premier temps d'identifier les variables qui représentent le mieux les effets que nous mettons en évidence. On donnera ici le nom des bases de données utilisées, suivi d'une référence bibliographique où l'on pourra trouver un lien vers ces dernières. La source utilisée par l'Insee est également indiquée.

Base de données		**Exemples de variables**	**Variables**	*Source Insee*
Base de comparateur	[24]	Superficie, Nombre d'habitants, Nombre de naissances	7	*Diverses*
Diplômes et formation	[34]	Nombre de personnes scolarisées (par sexe, tranche d'âge), Niveau de diplôme des non scolarisés	26	*Recensement*
Emploi *Population active*	[41]	Nombre d'actifs, Nombre de chômeurs	30	*Recensement*
Emploi *Caractéristiques*	[25]	Nombre de salariés/non salariés, Nombre d'actifs en temps partiel/complet	43	*Recensement*
Entreprises	[26]	Nombre d'entreprises actives, Nombre d'emplois (par secteur)	42	*Connaissance locale de l'appareil productif (Clap)*
Logement	[38]	Nombre de résidences principales et secondaires	8	*Recensement*
Revenus	[47]	Taux de pauvreté, Part des impôts, Part des revenus d'activité	19	*Dispositif sur les revenus localisés sociaux et fiscaux (Filosofi)*
Salaires	[48]	Salaire net horaire moyen (par sexe et csp)	17	*Déclaration Annuelle de Données Sociales (DADS)*
Structure population	[49]	Hommes/femmes de plus de 65 ans	15	*Recensement*

TABLE 4 – Liste des bases de données - Insee

Nous n'avons pas toujours utilisé toutes les variables proposées dans chacune des bases. Nous avons cherché à éviter les redondances et nous sommes restreint aux données susceptibles d'apporter une information pertinente sur la mortalité.

Une liste exhaustive des variables retenues est disponible en annexe *(Table 18)*.

Ces bases sont toutes à la maille communale et ont nécessité un retraitement conséquent pour être ramenées à l'échelle du pseudo-canton. Néanmoins, l'introduction de ces données nous permet de disposer d'un espace vectoriel de dimension conséquente (près de 150 variables une fois les variables redondantes retirées).

Ces données sont déterminantes ; la prochaine étape de nos travaux consistera ainsi à comprendre leur dynamique et les liens qui les unissent à notre variable d'intérêt.

Autre sources

Afin de compléter notre approche, divers indicateurs complémentaires ont été ajoutés. Ils sont essentiellement environnementaux (pollution de l'eau, densité du réseau routier...) et sociaux. Ces données proviennent de sources annexes, rarement utilisables sans retraitements :

— D'une part, elles sont en général proposées à des échelles moins fines (typiquement le département [12]). .

— D'autre part, ces bases comportent parfois trop de variables pour décrire un même effet. Par exemple, la base de données décrivant le niveau d'équipement des départements détaille le nombre de commerces (pharmacies, épiceries, écoles, postes de police, etc...).

Nous avons donc cherché à calibrer ces données pour qu'elles soient plus simples et plus adaptées ; c'est à dire essentiellement en les résumant par des variables catégorielles. En reprenant l'exemple précédent, les variables « Nombre de boulangeries, de boucheries, de pharmacies, etc... » sont devenues une variable catégorielle « Taux d'équipement : [élevé ; moyen ; faible]

La méthode utilisée a toujours été la même dans ce cas :

— Étude par ACP de l'espace des variables (lorsqu'elles sont nombreuses).

— Création d'une variable catégorielle par K-means sur la base de critères statistiques.

La construction des variables ne sera pas décrite ici dans un souci de concision. En revanche, cette méthodologie sera utilisée pour explorer notre base de données et sera donc détaillée plus amplement dans la suite de ce mémoire.

Au terme de cette prospection et des traitements complémentaires réalisés, nous disposons donc de 14 nouvelles variables qui peuvent être résumées par le tableau suivant :

Source	**Résumé des variables**
DREES [12]	Accès à un médecin généraliste Coûts et taux d'occupation des logements Indicateur de l'intensité du chômage Part des élèves de 6ème ayant déjà redoublé Indicateur de pauvreté Âge moyen de la mère à l'accouchement Pollution de l'eau au nitrate et aux pesticides
EIDER [14]	Type de climat (nord ou sud), équipement (commerces, écoles, etc...), importance du réseau de transport
ODICER [58]	Indicateur d'impact de l'alcool et du tabac sur la mortalité

Table 5 – Liste des bases de données - Complément

Le lecteur intéressé pourra se référer à l'annexe de ce mémoire pour plus d'informations sur les données construites et les bases utilisées *(Tables 19 à 22)*.

En résumé, pour l'ensemble des 2094 Cantons Français [13], nous disposons de maintenant de 12 bases de données, pour un total de 167 variables.

Des techniques spécifiques sont nécessaires pour manipuler un tel nombre de variables, mais nous verrons que le temps passé à rassembler et à traiter ces données présente un intérêt certain. Nous sommes d'ores et déjà capable de décrire avec précision l'ensemble des cantons français, que ce soit du point de vue de l'emploi, du logement, ou plus généralement du cadre de vie.

Dans la prochaine partie, nous serons ainsi en mesure de proposer une première approche du lien entre les dynamiques du territoire et la mortalité. Nous catégoriserons également les cantons français pour préparer notre modélisation.

La base de données construite ici sera utilisée tout au long de ce mémoire. Elle pourra également servir de support dans le cadre de travaux ultérieurs, que ce soit pour étendre notre étude ou en développer de nouvelles.

12. Certaines bases à la maille région ont été étudiées mais jugées inutilisables
13. En comptant les arrondissements de Paris, Marseille et Lyon

2.2 Exploration de la base de données

2.2.1 Analyse en composantes principales

L'analyse en composantes principales (ACP) est une méthode de statistique multivariée. Elle permet, pour un espace vectoriel X de dimension n, de se ramener à un espace orthonormé $\tilde{X}$ de dimension très inférieure [55]. L'idée est la suivante : il s'agit d'identifier les directions qui résument le mieux notre nuage de point (i.e. qui portent le plus de variance) pour résumer un grand nombre de variables en quelques axes.

La construction de cette méthode fait appel à des connaissances d'algèbre linéaire et ne sera pas détaillée ici. De façon générale, la recherche des composantes principales revient à calculer les vecteurs propres de la matrice de corrélation de notre nuage.

L'ACP a pour la première fois introduite par Pearson (en 1901) [59], un statisticien à qui l'on doit notamment le test du $\tilde{\chi}^2$ et le coefficient de corrélation.

Cette méthode présente plusieurs intérêts. En résumant notre espace vectoriel de grande dimension à quelques axes, nous serons en mesure d'analyser facilement la structure de nos données. C'est à dire :

— Étudier les corrélations entre les variables.

— Identifier les axes portant l'amplitude du nuage de points.

Nous pourrons alors :

— Comprendre plus facilement la répartition des cantons dans cet espace.

— Identifier des groupes atypiques le cas échéant.

On applique ainsi l'ACP à notre espace vectoriel. Le graphique suivant nous donne une idée de la façon dont notre base de données est résumée :

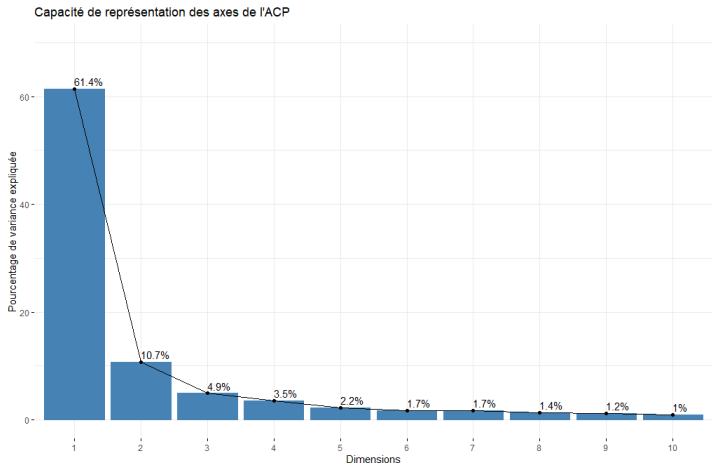


FIGURE 14 – Représentativité des axes

Ici, on attribue à chaque axe de l'ACP la proportion de variance du nuage de points qu'il capte.

Dans le cas de notre premier axe, cela signifie que ce dernier porte à lui seul 60% des variations de notre nuage de points. Il résume donc plus de la moitié des variations que l'on peut observer.

Les cantons se répartissent ainsi dans cet espace simplifié :

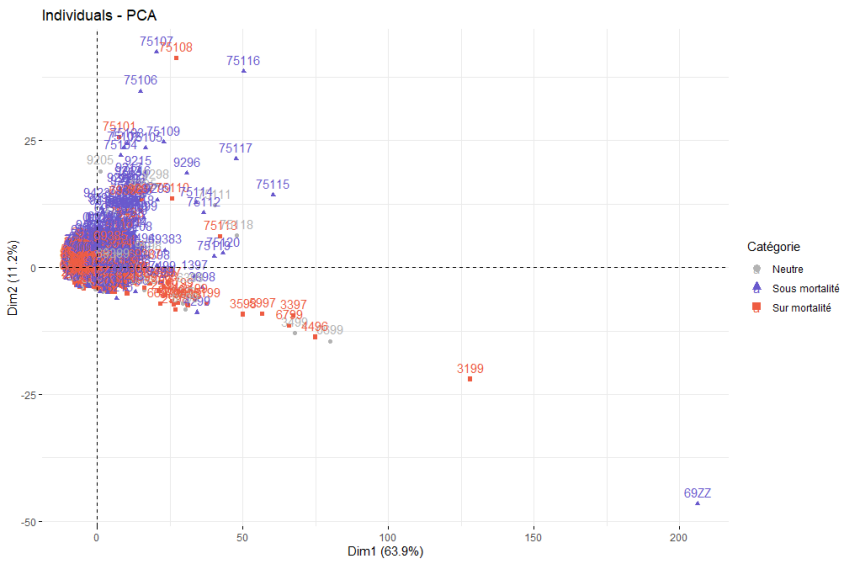


FIGURE 15 – Cantons projetés sur les axes 1 et 2 - Affichage général

Afin d'améliorer la lecture et l'interprétation de notre analyse, retirons dans un premier temps la banlieue de Lyon (comptabilisée comme un seul ensemble, elle compte ainsi 800 000 habitants) et Toulouse (qui compte 480 000 habitants) de nos représentations. Les arrondissements de Paris sont difficiles à interpréter à ce stade, nous les retirons donc également.

On obtient ainsi la représentation suivante :

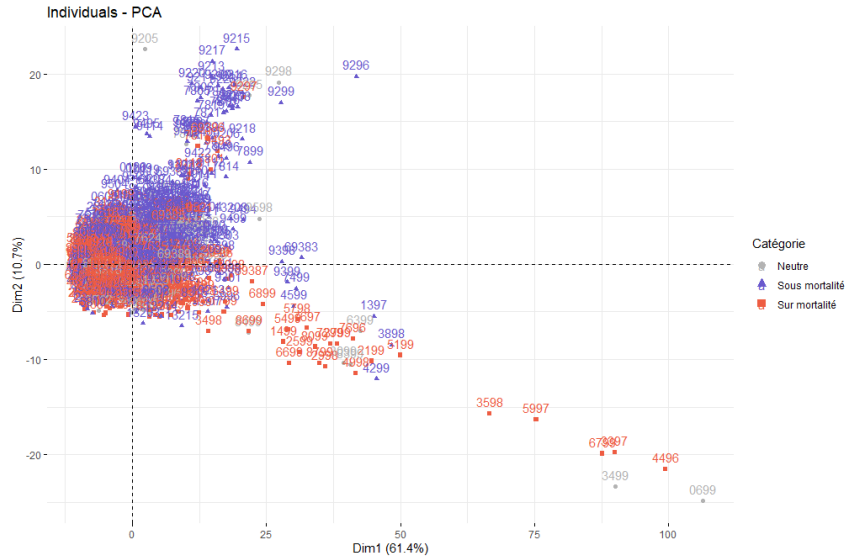


FIGURE 16 – Cantons projetés sur les axes 1 et 2

Nous allons donc chercher à comprendre la signification de chacun de ces axes. Cela nous permettra de proposer une première analyse. Par la suite, nous commencerons à catégoriser les cantons afin de proposer des développements plus fins.

Axes 1 et 2

Sur la représentation précédente, on remarque que nos cantons se répartissent vers la droite de l'image. Ce type d'effet est souvent désigné comme un "effet taille" et est révélateur des effets que nous allons maintenant décrire.

Comme on peut le voir sur ce graphique, le premier axe porte à lui seul un nombre très important de variables :

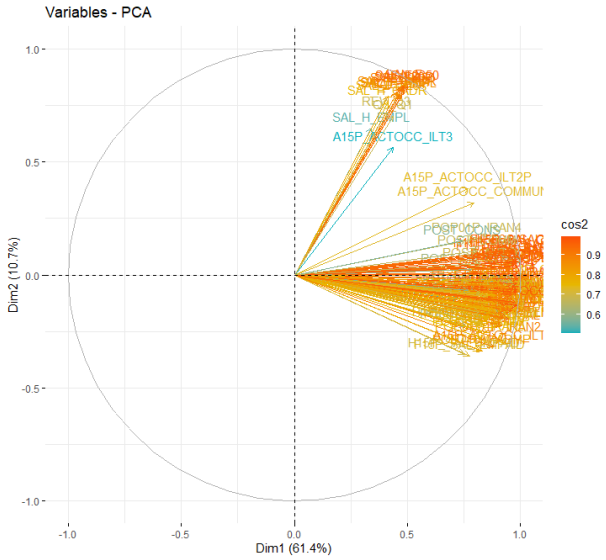


FIGURE 17 – Représentations des vecteurs dans l'ACP - Axes 1 et 2

Pour bien comprendre cette représentation, il faut savoir que :

— Plus les variables sont proches les unes des autres, plus elles sont corrélées.

— La couleur/la longueur d'une variable correspond à la qualité de sa représentation (cos^2) sur le plan factoriel.

On remarque ainsi qu'un nombre très important de variables est porté par le premier axe. On peut lister les variables qui ont le plus contribué à sa construction :

— F2554 : Nombre de femmes entre 25 et 54 ans

— H2554 : Nombre d'hommes entre 25 et 54 ans

— A15P_ACTOCC : Nombre d'actifs occupés (> 15 ans)

— LOG_RP : Nombre de résidences principales

— POP01P_IRAN1 : Nombre d'individus habitant un an avant dans le même logement

— H15P_NSCOL : Nombre d'hommes non scolarisés (> 15 ans)

Il est maintenant clair que le premier axe représente essentiellement la taille de la ville. Plus une ville est loin sur cet axe, plus elle comptera un nombre important d'habitants.

L'effet taille mentionné plus haut et la forte corrélation entre nos variables s'explique donc parfaitement. Que ce soit son nombre d'habitants, de salariés ou de diplômés, ou son nombre de logements, la majorité des données que l'on pourrait introduire pour décrire un canton ou une ville sont de nature à décrire sa taille.

Notre second axe est construit à partir du second groupe de variables (diagonale vers le haut et la droite). Si nous listons les variables utilisées pour le construire, on obtient le résultat suivant :

— SAL_F : Salaire horaire net moyen des femmes

— SAL : Salaire horaire net moyen

— SAL_F2650 : Salaire horaire net moyen des femmes entre 26 et 50 ans

— SAL_F50P : Salaire horaire net moyen des femmes de plus de 50 ans

— SAL_H2650 : Salaire horaire net moyen des hommes entre 26 et 50 ans

— SAL_H : Salaire horaire net moyen des hommes

On voit ici que c'est le revenu qui est mesuré ici. Le fait que le revenu soit en partie corrélé à la taille de la ville fait que l'on obtient cet effet de diagonale, ce qui compliquera légèrement l'interprétation.

Un point est important à noter ici. Si nous revenons à notre répartition des villes dans le plan factoriel, on observe un lien important entre l'axe 2 et la mortalité. Les villes en sous-mortalité (c'est à dire présentant un ratio de mortalité inférieur à 0.97) sont clairement réparties vers le haut de notre espace alors qu'à l'inverse, les villes en surmortalité sont réparties vers le bas. Cet axe semble être une piste prometteuse pour discriminer les cantons français.

On comprend en fait que, après la taille de la ville, le revenu est la variable qui discrimine le mieux les différentes villes françaises et que, plus encore, cette discrimination est pertinente pour décrire le niveau de mortalité.

Le lien entre la mortalité est le revenu est connu. Il est par exemple décrit par une étude de l'Insee de 2018 [40]. Cet effet apparaît clairement dans notre analyse et semble ainsi exploitable à partir de données publiques.

Le sens de la variable revenu net moyen est à relativiser ; il ne faudrait pas en conclure que le revenu limite en lui même les risques de décès : nous observons un lien de corrélation. Le revenu est ici la donnée rendant le mieux compte de la diversité des territoires. Et nous observons que cette diversité n'est pas sans conséquence sur la mortalité.

Axes 3 et 4

Si nous avons facilement pu constater des liens entre nos deux premiers axes et la mortalité, les axes 3 et 4 sont en revanche moins prometteurs, comme on peut le voir sur cette représentation :

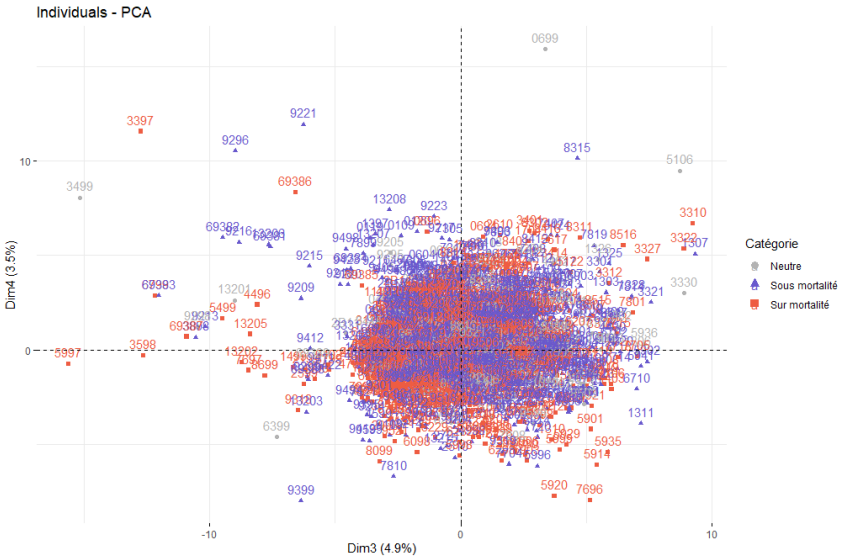


FIGURE 18 – Cantons projetés sur les axes 3 et 4

Nous passerons plus rapidement sur les représentations suivantes, leur interprétation étant plus aisée. Comme précédemment, on n'affiche ici que les variables suffisamment bien représentées :

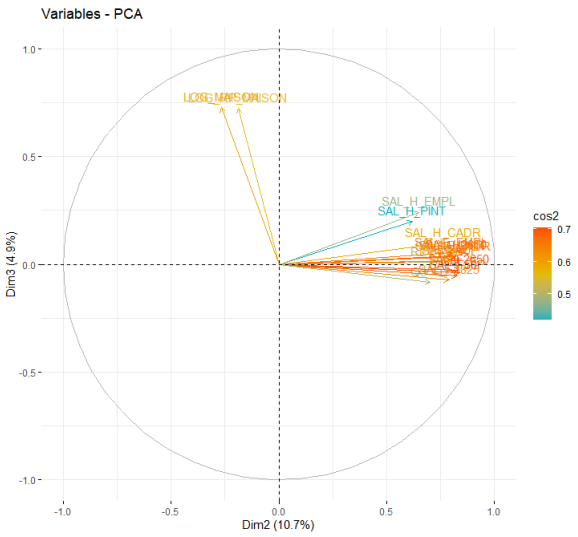


FIGURE 19 – Axes 2 et 3

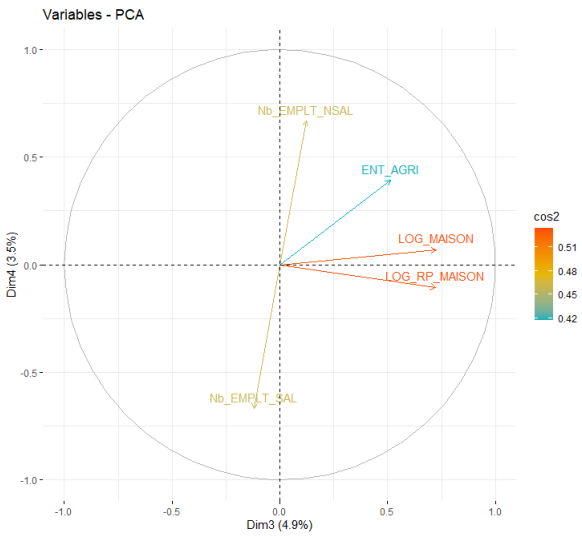


FIGURE 20 – Axes 3 et 4

FIGURE 21 – Représentations des vecteurs dans l'ACP - Axes 3 et 4

On comprend donc ici que l'axe 3 porte une partie de l'information sur le niveau d'urbanisation (le nombre de maisons est légèrement corrélé au nombre d'entreprises agricoles). Le quatrième axe porte sur le type d'emploi et montre une opposition entre les cantons où le travail salarié domine et ceux où le travail non salarié occupe une part importante des emplois.

Nous comprenons mieux, à l'issue de cette partie, les effets qui sous-tendent la répartition des cantons dans notre espace vectoriel. Nous allons maintenant chercher à exploiter les effets que nous venons d'évoquer.

2.2.2 Clustering

Nous nous proposons ici d'appliquer la méthode K-means aux cantons répartis dans l'espace ACP.

K-means est une méthode de partitionnement de données relativement classique. Son principe est le suivant :

On cherche à partitionner n points $(x_1, ..., x_n)$ en k ensembles $S = (S_1, ..., S_k)$ avec $(k \leq n)$, souvent appelés clusters. Pour ce faire, on va chercher à minimiser la distance entre les points à l'intérieur de chaque partition :

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

où μ_i est le barycentre des points dans S_i.

L'algorithme utilisé est en lui-même est itératif. Après avoir initialisé au hasard k points (la position moyenne des partitions $m1^{(1)}, ..., mk^{(1)}$) on réalise les étapes suivantes jusqu'à obtenir une convergence :

1. On affecte chaque observation à la partition la plus pertinente (par exemple la plus proche).
2. On met à jour le centre de chaque cluster.

Dans les faits, il existe plusieurs algorithmes K-means utilisant ce principe général. Nous avons utilisé ici la méthode Hartigan-Wong [21].

A noter que nous avons également tenté d'implémenter d'autres techniques de clustering (Classification hiérarchique et clustering basés sur les modèles de mélange gaussien notamment), mais cette technique simple a donné les meilleurs résultats.

Afin de choisir un nombre k adéquat de clusters, on fait varier le nombre de groupes et on surveille l'évolution d'un indicateur de qualité. Nous avons étudié la part :

— La part d'inertie expliquée : Le rapport entre la variance inter-groupe et la variance totale du nuage.

— L'indice de Calinski-Harabasz : Le rapport entre la variance inter-groupes et la variance intra-groupe [5].

Graphiquement, nous obtenons les représentations suivantes :

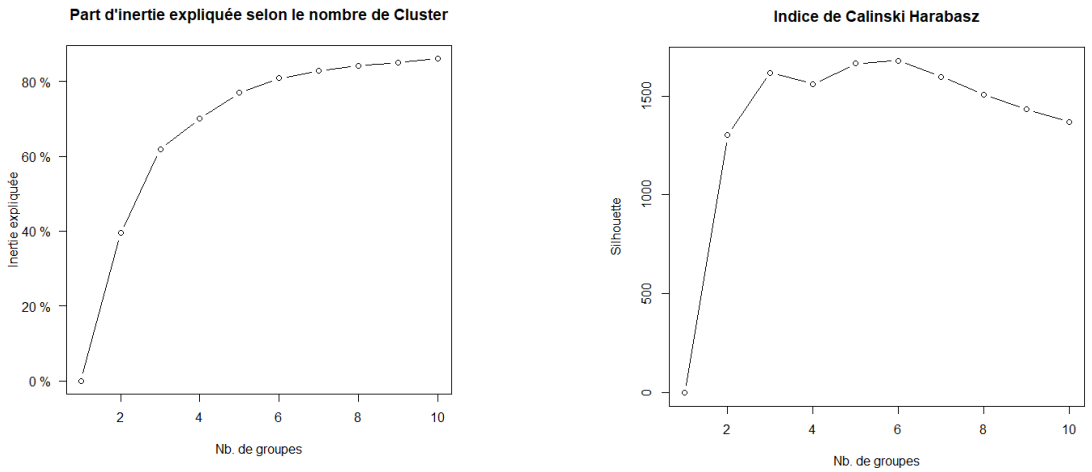


FIGURE 22 – Part d'inertie expliquée et indice de Calinski-Harabasz

On remarque qu'à partir de 6 groupes, la proportion de variance expliquée a tendance à plafonner. On comprend donc qu'il y a peu d'intérêt à continuer de partitionner notre nuage à partir de ce moment. On voit également que l'indice de Calinski-Harabasz est maximisé avec 3 et 6 groupes.

Nous faisons donc le choix de retenir 6 groupes. Une fois K-means appliquée au nuage de point construit précédemment pour $k = 6$, on obtient la représentation suivante :

FIGURE 23 – Catégorisation des cantons dans l'espace ACP

Cette représentation est déterminante pour la suite de notre analyse. Nous identifions donc 6 types de villes (i.e. canton) :

- `Très grandes villes`
- `Grandes villes`
- `Moyennes villes`
- `Moyennes villes aisées`
- `Petites villes`
- `Très petites villes`

Cette terminologie sera utilisée tout au long de nos travaux.

Il faut remarquer plusieurs choses sur cette représentation :

— Les *Moyennes*, *Petites* et *Très petites villes* forment un ensemble relativement homogène.

— Les *Moyennes villes aisées* se distinguent par un salaire net moyen élevé et un nombre d'habitants proche de celui des *Moyennes villes*.

— Les *Grandes* et *Très grandes villes* forment un groupe à part et sont relativement peu nombreuses.

Nous comprenons donc que ces trois "ensembles" devront être traités séparément, car ils présentent des dynamiques différentes. Les deux derniers sont moins denses et peuvent d'ores être déjà être analysés.

Avant de venir "zoomer" sur ces deux ensembles, nous pouvons afficher nos résultats sous forme de carte. La représentation obtenue est relativement intuitive :

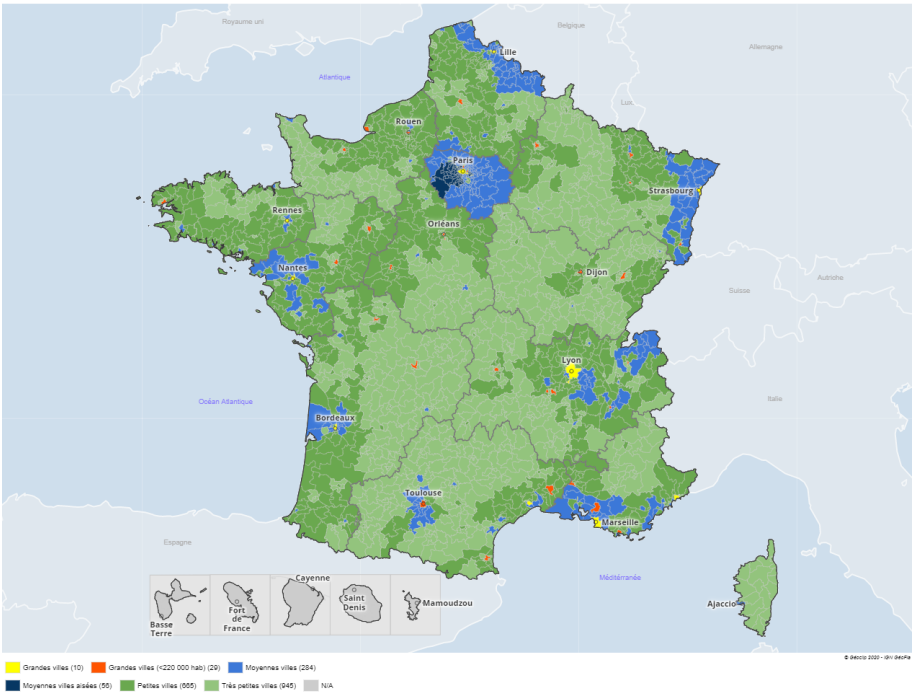


FIGURE 24 – Cartographie des catégories de canton construites
Obtenu à partir de www.france-decouverte.geoclip.fr - Données importées [20]

On identifie globalement neuf aires très urbanisées :

— L'île de France, avec notamment Paris (très dense) et la banlieue sud ouest de Paris (i.e. la quasi totalité du cluster *Moyennes villes aisées*[14]) qui se distinguent.

— Lyon et Marseillle.

— Les alentours de Nantes, Toulouse et Bordeaux.

— Les frontières belge, allemande et suisse.

Nous allons maintenant analyser spécifiquement les deux ensembles atypiques que nous avons mis en évidence, à savoir les *Moyennes villes aisées* et les *Grandes* et *Très grandes villes*.

14. Ces villes exceptées, seuls certains arrondissements de Lyon et Marseille rentrent dans cette catégorie)

Étude des Moyennes villes aisées

Lorsque l'on affiche spécifiquement le groupe des *Moyennes villes aisées* sur un nouvel espace ACP, on observe clairement une sous-mortalité, avec des ratios de mortalité très faibles :

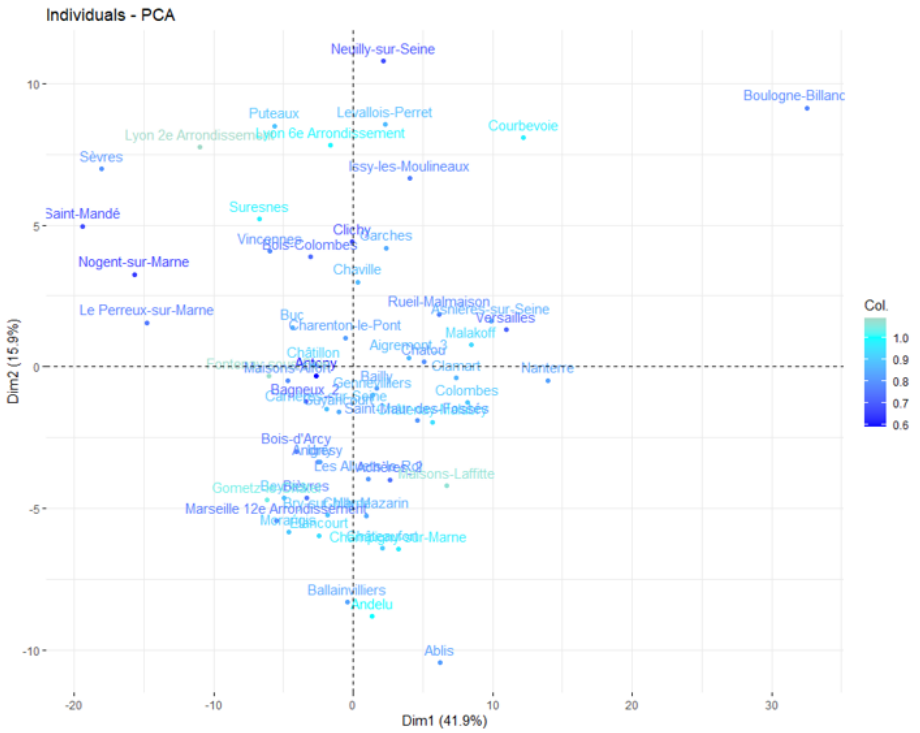


FIGURE 25 – Représentation des Moyennes villes aisées dans un espace ACP - Axes 1 et 2

Comme pour notre première représentation, les deux premiers axes décrivent toujours, respectivement, la taille de la ville et le niveau de vie (salaire). Si l'on réalise une première régression linéaire simple sur cet espace, on voit que notre espace de variables semble toujours porter de l'information :

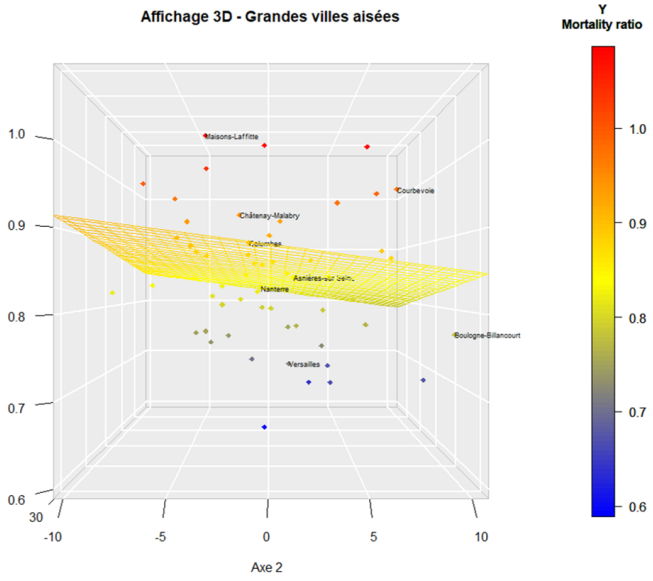


FIGURE 26 – Régression sur les deux premiers axes de l'ACP

Néanmoins, il est clair que cet ensemble est très homogène.

Il n'est en général pas pertinent de proposer de modélisation lorsque cela n'est pas nécessaire. La distribution des ratios est en première approximation i.i.d. (i.e. indépendants et identiquement distribués). Nous pouvons voir sur ce graphique que retenir le ratio de mortalité médian est suffisamment efficace pour catégoriser ces villes :

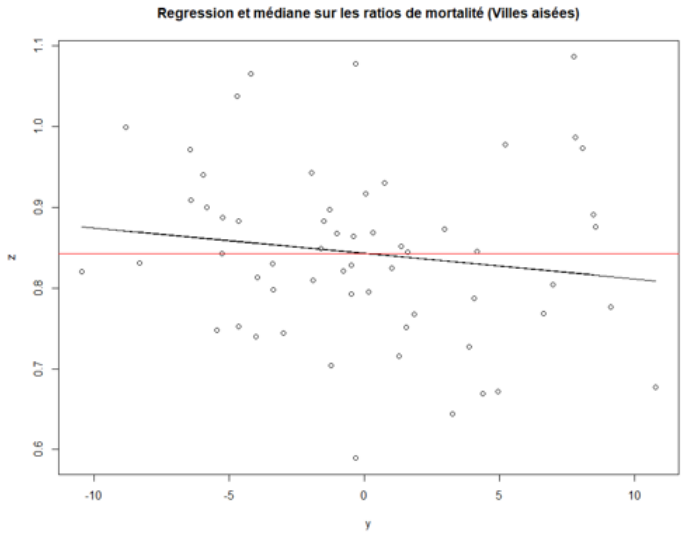


FIGURE 27 – Médiane et régression des ratios de mortalité sur le second axe de l'ACP

Le second axe utilisé ici semblait être le plus prometteur pour modéliser ces ratios et, pourtant, il n'a pas un pouvoir de représentation beaucoup plus fort que la médiane, ce qui nous conforte dans notre choix.

Nous choisissons de retenir une médiane plutôt qu'une moyenne car elle est plus robuste pour traiter des ratios et est moins sensible aux données extrêmes.

Le ratio ainsi estimé est de l'ordre de 0.837. Autrement dit, nous estimons donc que l'application des taux de mortalité nationaux différenciés par âge, sexe et csp conduirait, sur cette population, à une surestimation des décès de l'ordre de 16.3%.

Si ces cantons sont simples à analyser et à modéliser, ce n'est en revanche pas le cas du reste des cantons français. Étudions maintenant les grandes villes françaises.

Étude des Grandes et Très grandes villes

Affichons maintenant un nouvel espace ACP pour les *Grandes* et *Très grandes villes*. Les axes 1 et 3 sont cette fois ceux qui donnent la représentation la plus visuelle :

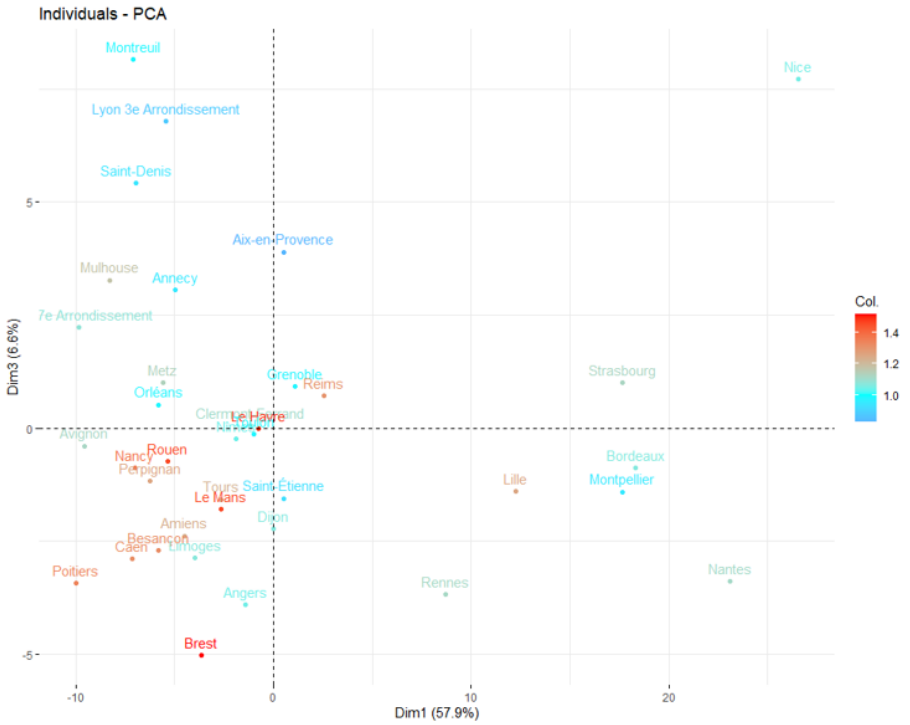


FIGURE 28 – Représentation des Grandes et Très grandes villes dans un espace ACP - Axes 1 et 3

Ici, l'interprétation est plus complexe et l'on est clairement confronté à un ensemble non homogène. Les axes construits par l'ACP le sont ainsi :

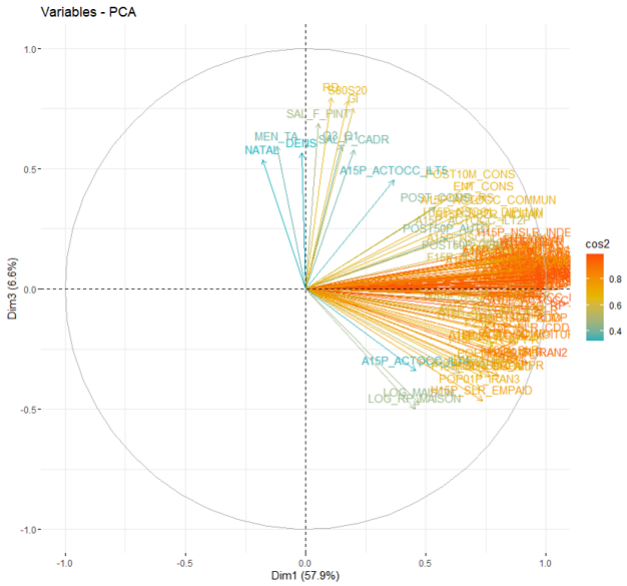


FIGURE 29 – Représentations des vecteurs dans l'ACP - Axes 1 et 3

On comprend que cet espace est construit de la façon suivante :

— Axe 1 : Représente le nombre d'actifs et d'emplois

— Axe 3 : Représente la "dynamique" des populations (densité, natalité, taille des ménages, etc...)

A noter que le fait de vivre en maison semble ici corrélé au niveau d'activité (axe 1) et inversement corrélé aux variables ayant construit l'axe 3.

On voit donc que :

— Les *Très grandes villes* se dispersent à droite de cet espace, et semblent disposer de ratios de mortalité positifs (à l'exception de Lille).

— Les *Grandes villes*, sur la gauche, se distinguent essentiellement par la "dynamique" de leur population, avec vers le haut des villes plutôt denses et dynamiques, apparemment avec de bons ratios de mortalité.

Outre ces considérations générales, il faut essentiellement retenir que l'étude de ces villes est bien plus complexe que celle que nous avons réalisée sur la banlieue parisienne et ne pourra pas se résumer à un ratio médian. Nous nous orienterons donc plutôt vers une modélisation.

Le reste des villes, en revanche, peut raisonnablement être traité en une seule fois. En excluant ces deux ensembles particuliers, le risque que des villes grandes et/ou aisées tirent l'analyse vers elles est maintenant moindre. Dans la prochaine partie, nous allons continuer à modéliser nos ratios en tirant parti des données que nous avons introduites. Cela se fera au travers de modèles linéaires.

Troisième partie

Modélisation des déviations de mortalité sur le territoire métropolitain

Dans la partie précédente, nous avons catégorisé les cantons français et identifié trois ensembles devant être traités séparément :

— Les *Moyennes*, *Petites* et *Très petites villes*.

— Les *Moyennes villes aisées* (essentiellement banlieue parisienne).

— Les *Grandes* et *Très grandes villes*.

Nous avons considéré pertinent de traiter les *Moyennes villes aisées* comme un groupe homogène et disposons d'ores et déjà d'un ratio de mortalité médian jugé représentatif.

Nous proposons donc, dans cette partie, de modéliser les ratios de mortalité des deux ensembles restant par des modèles adaptés. Les *Très petites*, *Petites* et *Moyennes villes* étant nombreuses, nous construirons un modèle relativement complexe, dont nous évaluerons la viabilité et la robustesse.

Pour les *Grandes* et *Très grandes villes*, les ratios dont nous disposons sont logiquement plus robustes, car calculés sur des populations plus grandes. Le modèle linéaire simple que nous construirons permettra de les fiabiliser. Les ratios que nous retiendrons seront ainsi une pondération de la réalité et de cette modélisation.

1 Construction du modèle linéaire principal

Nous allons dans un premier temps considérer les cantons appartenant aux groupes *Moyennes*, *Petites* et *Très petites villes*. Ce modèle linéaire sera notre principale source de résultats et, basé sur un grand nombre de cantons, sera le plus complexe. Sa construction sera donc détaillée et challengée en profondeur.

Nous commencerons par introduire la notion de modèle linéaire avant de détailler sa construction. Dans un second temps nous évaluerons ses performances et sa robustesse.

L'application des modèles linéaires se faisant sous un certain nombre d'hypothèses, nous les challengerons au fur et à mesure. Les choix réalisés dans notre modélisation seront également justifiés dès que nécessaire.

1.1 Définition du modèle

Pour tout canton $i (i = 1, ..., n)$, le modèle linéaire que nous allons construire se note ainsi [60] :

$$Y_i = \sum_{j=1}^{p} \beta_j X_i^j + U_i$$

où Y_i est une variable aléatoire réelle (ici le ratio de mortalité du canton i). On suppose que Y_i se décompose en une partie déterministe linéaire et une variable aléatoire d'erreur U_i supposée suivre une loi normale $\mathcal{N}(0, \sigma^2)$ (les U_i étant indépendants entre eux). Les β_j sont des coefficients (inconnus) que nous souhaitons estimer. Les X_i^j sont les valeurs des variables explicatives (exemple : le nombre d'habitants dans le canton j). Cette écriture peut être donnée de façon matricielle :

$$Y = X\beta + U$$

où Y et U sont des vecteurs aléatoires de $\mathbb{R}^n$, X une matrice n x p et β un vecteur de $\mathbb{R}^p$. U étant considérée normale, l'espérance de Y que nous souhaitons modéliser se fait au travers de $X\beta$.

L'estimation du vecteur des paramètres β peut être calculée par la méthode des moindres carrés ou par celle du maximum de vraisemblance et, dans le cas gaussien, donne dans les deux cas le résultat suivant :

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Cette estimation permet de calculer $\hat{Y}_i$, une estimation du ratio de mortalité moyen du canton i

$$\hat{Y} = X\hat{\beta}$$

Une fois notre modèle construit, nous disposerons donc, pour chaque canton i, d'une estimation du ratio de mortalité moyen $\hat{Y}$.

Ici, nous faisons le choix d'introduire une pondération dans notre modèle pour deux raisons :

— La variance de notre variable d'intérêt n'est pas constante. Or, il s'agit d'une hypothèse essentielle dans la construction d'un modèle linéaire.

— Malgré l'introduction de ratios, nous cherchons à donner plus d'importance aux villes peuplées.

Il semble raisonnable de considérer que la variance de nos ratios est inversement proportionnelle au nombre d'habitants. Nous posons donc l'hypothèse :

$$\sigma_i^2 = \frac{1}{N_i}$$

Où σ_i^2 est la variance du ratio Y_i et N_i le nombre d'habitants de 30-62 ans dans le canton i.

En posant cette hypothèse, nous sommes en mesure de satisfaire ces deux exigences, à savoir homogénéiser la variance et valoriser les villes peuplées dans notre modélisation. Nous introduisons comme pondération le nombre d'habitants de 30-62 ans dans le canton.

L'estimateur de $\hat{\beta}$ que nous avons présenté plus haut minimise le carré des résidus. Il doit donc être adapté pour minimiser le carré des résidus pondéré. Lorsque les erreurs ne sont pas corrélées et que l'on pose W la matrice de pondération, l'estimateur optimal se réécrit :

$$\hat{\beta} = (X'WX)^{-1}X'WY$$

Ici, W est diagonale où $W_{ii} = \frac{1}{\sigma_i^2}$.

Ces modèles mathématiques sont, dans les faits, optimisés par des méthodes itératives. Notre modèle linéaire est ainsi optimisé par l'algorithme IWLS (Iterative weighted least square) [13].

Afin de fiabiliser la construction de notre modèle nous retirons également de notre analyse les cantons présentant :

— Un nombre de décès insuffisant (< 4)

— Un ratio de mortalité extrême (< 0.3 ou > 1.7)

Ce qui correspond à 77 cantons problématiques sur les 1976 étudiés, soit environ 3.8% de la base. Ces cantons pourraient présenter des ratios non représentatifs et nuire à l'ajustement de notre modèle. Ils seront en revanche réintroduits lorsque nous évaluerons notre modèle définitif.

Ces considérations faites, il reste maintenant à construire le modèle en lui-même.

1.2 Construction du modèle

Une des étapes les plus importantes dans la construction d'un modèle linéaire est la sélection de variables prédictives.

Nous disposons d'une base de données très conséquente et allons donc ici chercher à mettre en évidence un nombre restreint de variables d'intérêt, parmi lesquelles nous ferons finalement notre sélection.

La mise en évidence de ces variables candidates se fera essentiellement par deux procédés statistiques :

— Classement des variables corrélées à notre variable d'intérêt.

— Étude des résultats proposés par des algorithmes de construction automatique.

A chacune de ces étapes, nous détaillerons les données qui nous ont semblé pertinentes. La liste des variables utilisées est disponible en annexe de ce mémoire *(Table 18)*.

Variables corrélées aux ratios de mortalité

Dans un premier temps, nous allons observer quelles sont les variables les plus corrélées à nos ratios de mortalité. On se propose de calculer le coefficient de corrélation linéaire de Pearson entre chaque X^j et Y. Sa formule est la suivante :

$$r = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

$\mathrm{Cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}[X])\,(Y - \mathbf{E}[Y])]$ désigne la covariance des variables X et Y.
σ_X et σ_Y désignent leurs écarts types.

En triant les résultats obtenus, on peut observer quelles variables sont les plus susceptibles de décrire nos ratios de mortalité. On regroupe ici les variables de façon à faciliter la compréhension, mais l'ordre d'importance des corrélations avec Y est pour l'essentiel respecté :

1. MEN_TA :
   Taille moyenne des ménages
2. REV_Q3, Q3_Q1, S80S20, RD, GI :
   Indicateurs d'inégalité
3. SAL_H2650, SAL_F2650, SAL_H, SAL_H50P, SAL, SAL_F, SAL_F_EMPL SAL_F50P SAL_F_PINT :
   Salaire net horaire moyen dans le canton (par âge, sexe et/ou csp).
4. MIGR1116 :
   Le solde migratoire de la population entre 2011 et 2016.
5. A15P_ACTOCC_ILT2P :
   Le nombre d'actifs travaillant dans une commune autre que leur commune de résidence.
6. Prop_H_Cadr :
   La proportion d'hommes cadres chez les 30-62 ans.

La taille moyenne des ménages est la variable explicative la plus corrélée à nos ratios. Comme nous le verrons, plus cette variable est élevée (ménages nombreux), plus le ratio du canton a tendance à être faible. On peut avancer plusieurs raisons intuitives à ces effets, comme la présence d'individus isolés ou de familles monoparentales concentrant des difficultés et présentant un risque de décès plus important.

Les variables portant sur *les inégalités, le solde migratoire et la proportion de cadres* se révèlent en fait peu exploitables, tout comme les proportions de csp (Prop_H_Cadr en l'occurrence). Les ratios, de manière générale, ont donné de mauvais résultats une fois implantés dans notre modèle et on leur préférera presque toujours un indicateur brut.

Le salaire net horaire moyen est une variable très efficace pour évaluer le risque de décès comme nous l'avions déjà fait remarquer lors de notre étude par ACP. Cela n'a encore une fois rien de surprenant, car elle est très représentative du niveau et du cadre de vie.

L'indicateur de mobilité domicile-travail des actifs, enfin, est une variable très pertinente également. Elle est même révélatrice des effets que nous mettrons en évidence et sera développé dans une partie ultérieure.

Variables sélectionnées par les modèles automatiques

Il existe plusieurs techniques de construction automatique de modèles. Ces algorithmes cherchent à maximiser un critère statistique et permettent de disposer d'un modèle performant rapidement. Nous allons ici les utiliser pour disposer d'une pré-sélection de variables. Les modèles construits nous serviront également de référence une fois notre modélisation faite.

Les techniques que nous avons utilisées sont les suivantes :

— Selection Backward : part d'un modèle maximal et retire la variable la moins pertinente à chaque étape.

— Selection Forward : part d'un modèle minimal et ajoute la variable la plus pertinente à chaque étape.

— Selection Stepwise : part d'un modèle minimal puis ajoute et/ou retire des variables à chaque étape.

En appliquant cette technique, nous pouvons disposer d'une liste de variables pertinente et de trois pré-modèles très efficaces. On peut ainsi lister les variables [15] apparaissant dans :

— Les trois modèles :
DENS MEN_TA NATAL DECE1116 H65P F15P_NSCOL_CAPBEP LOG_MAISON
H15P_SLR A15P_ACTOCC_MARCHE A15P_ACTOCC_ILT3 AlcTbc

— Deux modèles :
H15P_NSLR_EMPLOY AgeACC POST10M POP01P_IRAN5
LOG_RP F1517_SCOL POP01P_IRAN3

— Un modèle :
NAIS1116 H1824 H5564 F0217 SAL_F_EMPL SAL_F_OUVR
H1824_SCOL H15P_NSCOL LOG LOG_APPART etc...

Certaines variables déjà mises en évidence à l'étape précédente sont de nouveau sélectionnées, notamment la mobilité domicile-travail.

La densité et le taux de natalité [16] sont, comme la taille des ménages, des variables structurantes. Le *type de logement* (appartement ou maison) peut également correspondre à cette description.

Le niveau d'éducation et les types d'emploi sembleraient être des variables pertinentes. Or, elles sont très corrélées à la taille de la ville (le nombre de salariés est proportionnel au nombre d'habitants). Choisir un niveau d'étude spécifique (exemple : F15P_NSCOL_CAPBEP) ou une catégorie de travailleurs (exemple : H15P_NSLR_EMPLOY) pourrait limiter ce problème mais relèverait de l'arbitraire.

L'âge moyen de la mère à l'accouchement est une variable catégorielle également. Cette variable est un indicateur social qui, nous le verrons, a de bonne propriété pour prédire nos ratios de mortalité.

La consommation de tabac et d'alcool est une variable catégorielle que nous avons construite. On comprend facilement sa pertinence dans l'étude du risque mortalité.

Ces modèles sont très performants et réalisent leur objectif : maximiser un critère statistique. Néanmoins :

15. La liste détaillée des variables est disponible en annexe *(Table 18)*
16. Nombre de naissances divisé par la population.

— Ils sont relativement complexes à interpréter et nous préférerions comprendre les données que nous utilisons.

— Ils comportent une vingtaine de variables en moyenne.

En effet, bien que le critère maximisé ici tienne compte du nombre de variables introduites, il nous faudra privilégier un modèle simplifié.

Le risque de sur-interprétation

Lorsque l'on construit un modèle, un risque important est la sur-interprétation ; c'est à dire le risque de modéliser des variations dues au hasard et non pas à des effets réels.

Une pratique classique pour éviter ce phénomène est la validation croisée, qui consiste à séparer l'échantillon en une base d'apprentissage et une base de contrôle. Nous faisons ici le choix de ne pas appliquer ce partitionnement, car nous souhaitons ajuster notre modèle sur la totalité du territoire. Nous calculerons en revanche des méthodes de validation croisée dans la partie *performance*, pour contrôler tout de même le risque de sur-interprétation.

Ici, il est en fait moins grave de sur-interpréter dans le sens où les données à prédire sont (et resteront) les données qui servent à la construction du modèle : le territoire français.

Néanmoins, ce modèle n'aura pas pour autant vocation à s'appliquer aux années 2018 et 2019 uniquement et devra donc tout de même conserver une capacité de généralisation. Plus encore, malgré ces deux années d'exposition, la variabilité des ratios que nous avons calculés reste importante. Notre modèle aura d'autant plus tendance à capter cette amplitude qu'il sera complexe.

Le nombre important de données autoriserait, techniquement, une certaine complexité. Pour toutes les raisons évoquées, nous privilégierons malgré tout la parcimonie et n'introduirons pas trop de variables prédictives. Un modèle simple sera par nature moins sujet à la variabilité.

1.3 Modèle linéaire principal

Nous avons pu identifier les variables explicatives susceptibles d'apporter une information efficace et pertinente.

Le choix définitif des variables que nous faisons maintenant est motivé par des éléments pratiques et statistiques :

— Nous réalisons un arbitrage entre la performance du modèle et le nombre de variables (pour éviter la sur-interprétation).

— Nous n'ajoutons que des variables dont nous pouvons (relativement) interpréter les effets.

— Nous n'utilisons que des variables présentant des niveaux de corrélations suffisamment faibles.

Après sélection des variables, nous obtenons ainsi le modèle suivant :

```
Call:
glm(formula = Y_FR ~ NATAL + MEN_TA + DENS + SAL + A15P_ACTOCC_ILT2 +
    LOG_MAISON + AgeACC + AlcTbc, family = gaussian(link = "identity"),
    data = BASE_FR, weights = Nbr)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-65.879  -11.569   -0.485   11.112   75.658

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.109e+00  6.439e-02  32.757  < 2e-16 ***
NATAL               5.907e+00  4.187e-01  14.109  < 2e-16 ***
MEN_TA             -5.259e-01  2.361e-02 -22.274  < 2e-16 ***
DENS               -1.294e-05  2.130e-06  -6.077 1.48e-09 ***
SAL                -1.102e-02  4.188e-03  -2.631  0.00859 **
A15P_ACTOCC_ILT2   -9.445e-06  1.292e-06  -7.312 3.88e-13 ***
LOG_MAISON          1.080e-05  9.393e-07  11.493  < 2e-16 ***
AgeACCPlutôt agé   -8.660e-02  1.299e-02  -6.664 3.48e-11 ***
AgeACCPlutôt jeune -9.213e-02  1.097e-02  -8.398  < 2e-16 ***
AlcTbcMoyenne      -8.736e-02  9.270e-03  -9.424  < 2e-16 ***
AlcTbcTrès élevée   4.583e-02  1.879e-02   2.439  0.01483 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 296.6688)

    Null deviance: 967956  on 1898  degrees of freedom
Residual deviance: 560111  on 1888  degrees of freedom
AIC: -1146.8

Number of Fisher Scoring iterations: 2
```

FIGURE 30 – Modèle linéaire principal

Nous détaillons dans un premier temps les variables que nous avons choisi de conserver et en proposons une interprétation *toutes choses égales par ailleurs*[17] :

— Variables structurantes :
Ces variables servent à catégoriser les cantons. Elles ne sont pas interprétables quantitativement et l'on doit se contenter d'analyser leur effet (positif ou négatif).

1. NATAL : Le taux de natalité d'un canton a plutôt une influence négative sur la perception du risque de mortalité.
2. MEN_TA : A l'inverse, la taille moyenne des ménages a plutôt une influence positive, pour les raisons déjà évoquées.
3. DENS : La densité a plutôt une influence positive et représente, très probablement, un indicateur du niveau d'urbanisation.

17. Tous les constats que nous tirons doivent être interprétés dans le cadre de notre modèle et ne sauraient être extrapolés pour tirer des conclusions générales sur le risque de mortalité.

4. LOG_MAISON : Le nombre de logements de type maison a, de la même façon, une influence plutôt négative car représente, très probablement, un indicateur du niveau de ruralité.

— Variables socio-économiques :
Ces variables décrivent les populations. Elles sont en général interprétables quantitativement dans le cadre du modèle. [18]

1. SAL : Le salaire horaire net moyen a une influence positive sur le risque de mortalité. Ici, un euro de salaire horaire moyen en plus (d'un canton à un autre) se traduit par une diminution de 1.1 points du ratio de mortalité.
2. A15_ACTOCC_ILT2 : Le nombre d'actifs travaillant dans une commune différente et dans le même département permet, quant à lui, de décrire la mobilité domicile-travail. L'impact positif que nous constatons sur les ratios de mortalité s'expliquent et sera détaillé lors de l'interprétation de nos résultats [19].
3. AgeACC : Par rapport à un département catégorisé comme présentant des âges à l'accouchement "Jeunes", le ratio de mortalité diminue de 8.6 à 9 points pour un département "Plutôt âgé" ou "Plutôt jeune". [20]
4. AlcTbc : Par rapport à un département catégorisé comme ayant un taux de consommation d'alcool et de tabac "Élevée", un département avec une consommation dite "moyenne" aura un ratio diminué de 8 points et un avec une consommation "Très élevée" une mortalité augmentée de 4.5 points.

Comme nous pouvons le voir sur ce graphique, les relations entre chacune des variables (prises individuellement) et les ratios de mortalité sont en fait plutôt respectées :

18. On utilisera le terme "point" comme unité. Un ratio passant de 1 à 0.89 sera décrit comme "diminué de 11 points."

19. Nous proposons également une analyse rapide des relations entre la mobilité des actifs et les ratios de mortalités en annexe de ce mémoire.

20. Seuls Paris et le département des Hauts-de-Seine sont catégorisés comme "Âgés". Leurs cantons étant toujours "Grands" ou "Aisés" cette catégorie n'apparaît pas ici.

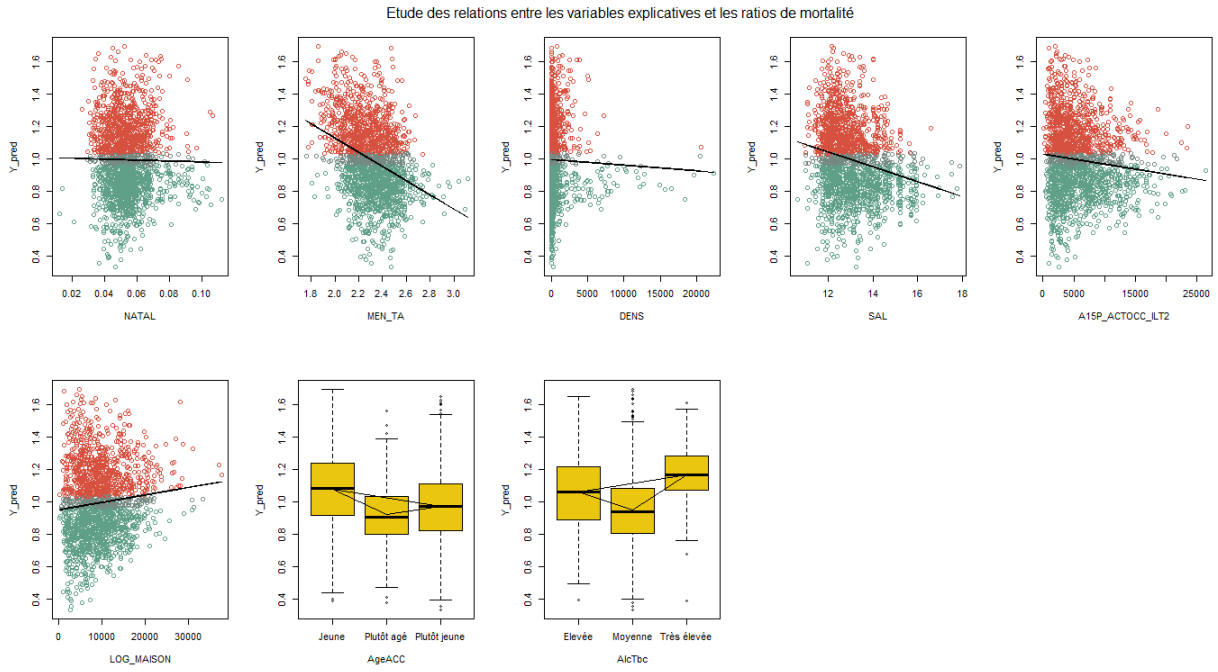


FIGURE 31 – Relations entre les variables prédictives et les ratios de mortalité

Comme cela a déjà été mentionné, la taille des ménages et le salaire horaire moyen sont les variables prédictives les plus corrélées à nos ratios de mortalité :

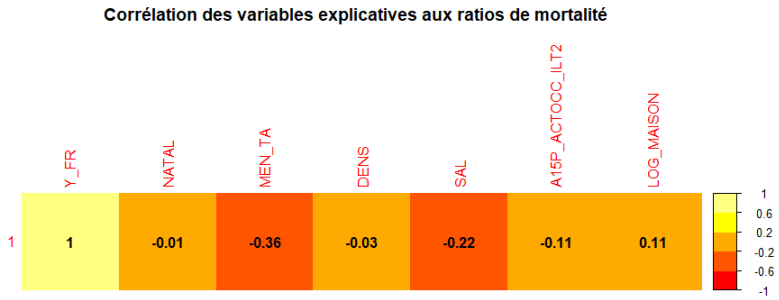


FIGURE 32 – Corrélations entre les variables prédictives et les ratios de mortalité

D'autre part, nous constatons que le test de Student était significatif pour chacune des variables. Nous n'avons donc pas introduit d'informations superflues. D'une façon générale, ce test est réalisé en construisant (à partir des données du modèle) une statistique suivant une loi de student. Il permet d'évaluer si une variable est statistiquement différente de 0 ; c'est à dire si elle est représentative d'un véritable effet ou si elle permet simplement de stabiliser le modèle. [60] Nous confirmerons plus loin que l'hypothèse de normalité de résidus (nécessaire pour appliquer ce test) est bien vérifiée.

Nous pouvons nous faire une première idée du pouvoir prédictif de notre modèle par le graphique suivant :

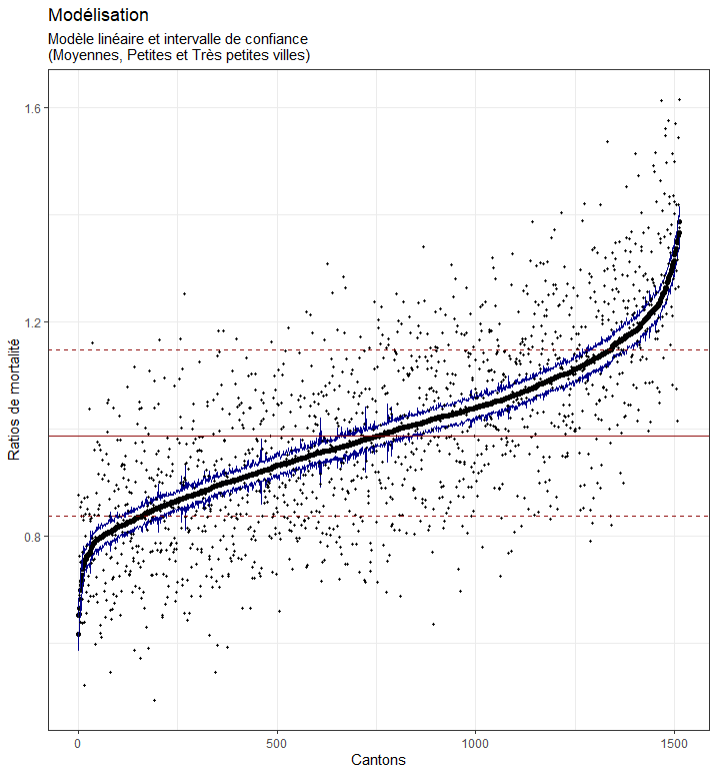


FIGURE 33 – Modèle linéaire principal - Représentation

Lors de la construction de notre modèle, nous pouvions déjà constater que la variance résiduelle de notre modèle est forte. Ce graphique montre à nouveau la sur-dispersion de nos résidus. Néanmoins, étant donné la grande variabilité de notre variable de départ et les considérations évoquées concernant la sur-interprétation, ce résultat est très satisfaisant.

Nous allons maintenant évaluer la viabilité et la fiabilité de notre modèle.

1.4 Contrôle des hypothèses et de la viabilité du modèle

Absence de multicolinéarité

Une des hypothèses des modèles linéaires est l'absence de relation de colinéarité entre les variables.

Une base orthonormée calculée par ACP permettrait notamment de coller facilement aux hypothèses des modèles linéaires (non corrélation des variables). Néanmoins, la perte en terme d'interprétabilité/de traçabilité des effets étant trop importante, nous n'avons pas fait ce choix.

Les *facteurs d'inflation de la variance* (en anglais Variance Inflation Factor), notés VIF, sont une mesure du niveau de multicolinéarité entre les variables d'un modèle de régression [19].

On le définit comme :

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

où R_i^2 est le coefficient de détermination (que nous définissons plus loin) du modèle linéaire prédisant X_i à partir du reste des variables explicatives.

Lorsque le VIF d'une variable î est élevé, cela indique qu'il existe une relation de dépendance avec les autres variables explicatives. On considère généralement qu'un VIF inférieur à 5 est satisfaisant.

Notre modèle incluant des variables catégorielles, cette notion doit être étendue et on parlera alors de facteurs généralisés, notés GVIF [18]. Pour les variables continues ($\mathrm{Df}^{21} = 1$), ce nouvel indicateur reste néanmoins égal au VIF.

Pour comparer les variables entre elles, il est préférable d'utiliser le critère $GVIF^(1/(2*Df))$.

On peut voir sur le tableau suivant que les niveaux de multicolinéarité sont satisfaisants :

Variable	$GVIF$	Df	$GVIF^(1/(2*Df))$
NATAL	2.375443	1	1.541247
MEN_TA	1.661203	1	1.288877
DENS	2.013096	1	1.418836
SAL	1.821082	1	1.349475
A15P_ACTOCC_ILT2	2.639624	1	1.624692
LOG_MAISON	2.153781	1	1.467576
AgeACC	2.418670	2	1.247080
AlcTbc	1.976355	2	1.185677

TABLE 6 – VIF - Modèle linéaire principal

Graphiquement, on peut se faire une idée des relations entre nos variables en affichant les coefficients de corrélation de Pearson :

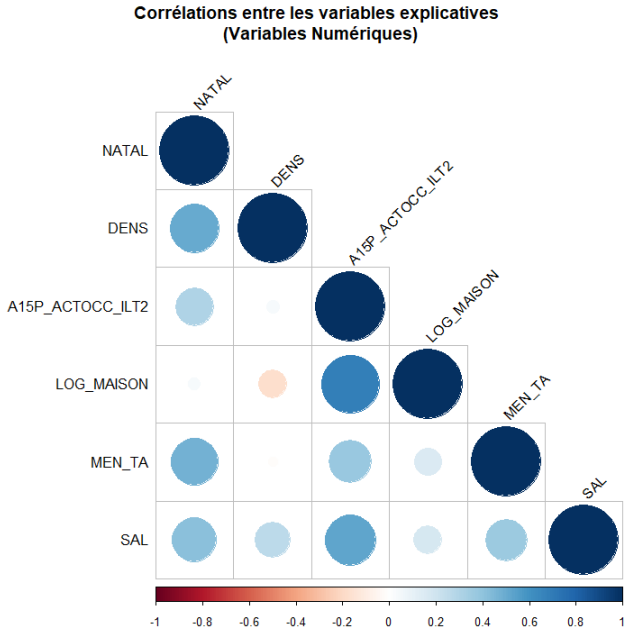


FIGURE 34 – Corrélations entre les variables explicatives

La relation entre les variablesA15P_ACTOCC_ILT2, LOG_MAISONetSALpourrait sembler préoccupante, mais ces dernières ne présentent en fait pas un VIF trop important. Lorsque l'on observe les relations dans le détail, nous observons qu'il n'existe pas de relation de colinéarité rédhibitoire :

21. Le degré de liberté de la variable

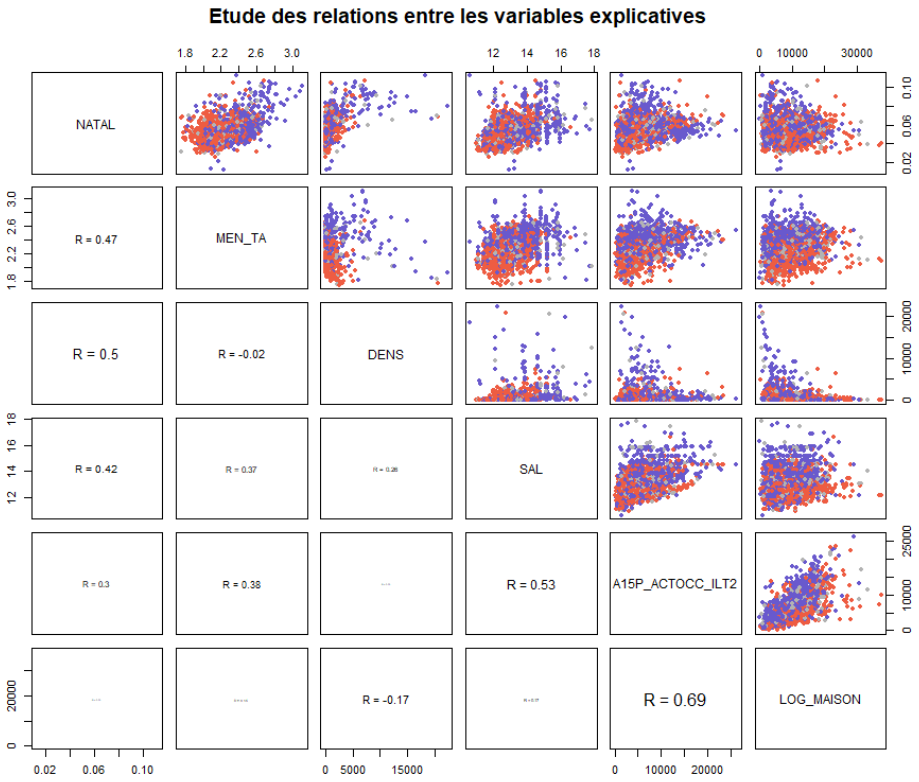


FIGURE 35 – Relations entre les variables explicatives

Ce graphique permet également de se faire une meilleure idée du pouvoir prédictif de nos variables une fois couplées. On remarque encore une fois l'effet de la taille des ménages sur le niveau de mortalité du canton.

Homoscédasticité et normalité des résidus

Nous allons maintenant évaluer le respect de l'homoscédasticité des résidus, c'est à dire la constance de leur variance.

Comme nous pouvons le voir, le test de Breusch-Pagan évaluant la variance des résidus est concluant : on ne rejette donc pas l'hypothèse de constance de la variance des résidus.

```
        studentized Breusch-Pagan test

data:  fit
BP = 13.554, df = 10, p-value = 0.1943
```

FIGURE 36 – Test de Breusch-Pagan - Modèle linéaire principal

D'une façon générale, ce test est utilisé exclusivement dans les modèles linéaire et est réalisé en construisant (à partir des données du modèle) une statistique suivant une loi du Khi 2 [3].

Nous pouvons également constater que, sur le graphique suivant, nos résidus sont distribués de manière uniforme :

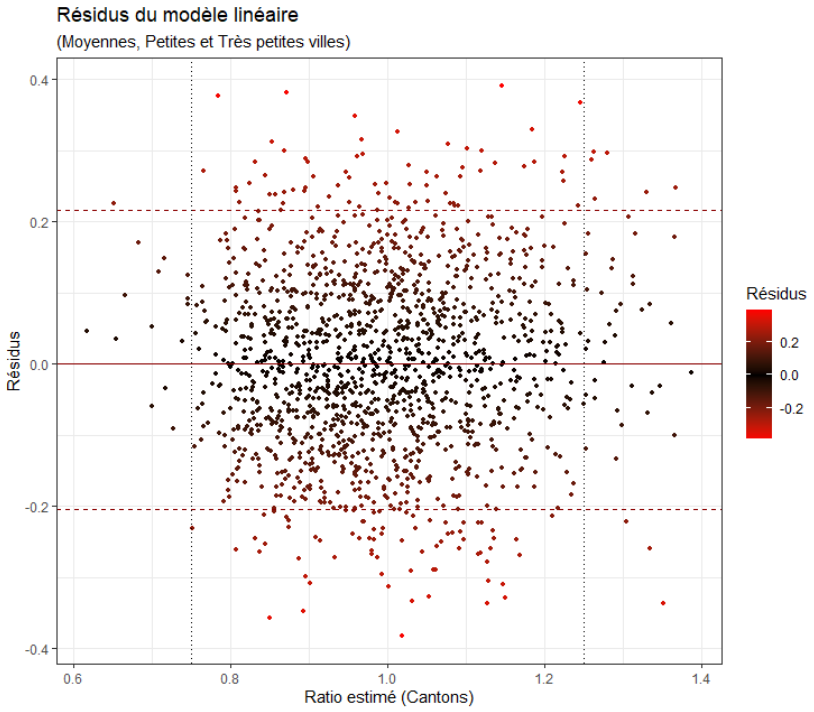


FIGURE 37 – Constance de la variance des résidus

La concentration des points autour de 1 a tendance à légèrement complexifier la lecture. On remarque tout de même que quelques points s'éloignent significativement, mais ils restent minoritaires sur les 1900 cantons étudiés ici.

Nous considérons l'hypothèse vérifiée.

Une hypothèse plus forte que l'homoscédasticité des résidus est leur normalité. Cette hypothèse n'est pas indispensable, mais souhaitable car elle justifie notamment l'usage des tests statistiques de significativé que nous avons détaillés plus avant.

Le test de Sharpiro-Wilk est concluant : on ne rejette pas l'hypothèse de normalité des résidus.

```
        Shapiro-Wilk normality test

data:  fit$residuals
W = 0.99882, p-value = 0.4154
```

FIGURE 38 – Test de Sharpiro-Wilk - Modèle linéaire principal

D'une façon générale, ce test est réalisé en construisant (ici sur les résidus du modèle) une statistique suivant une loi de probabilité spécifique. Il permet d'évaluer si un échantillon a été généré par une loi normale [63].

Nous pouvons confirmer sur les graphiques suivants que cette hypothèse est globalement vérifiée :

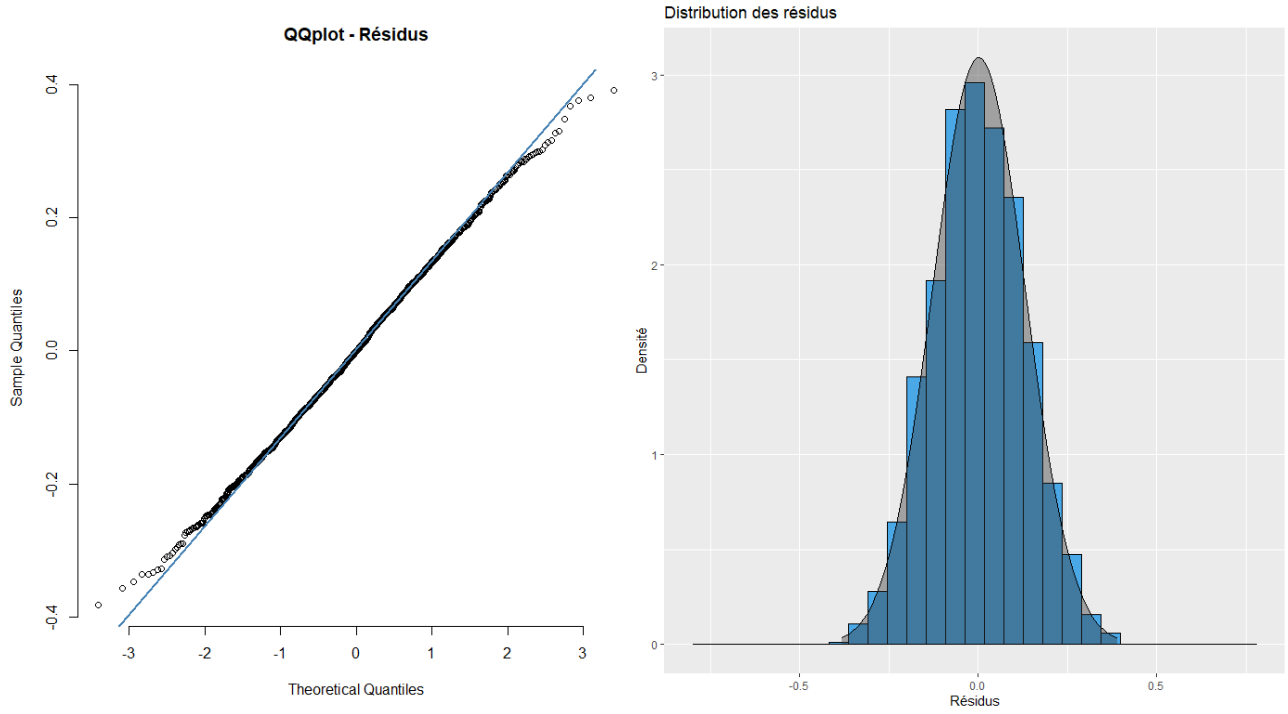


FIGURE 39 – Normalité des résidus

On note néanmoins une légère déviation des valeurs extrêmes.

Également, nous constatons une très légère asymétrie des résidus qui s'explique par la prépondérance des ratios inférieurs à 1 dans nos données. Notre modèle étant pondéré par le nombre d'habitants des cantons, cette sur-représentation est cohérente car elle concerne des ensembles de petite ou moyenne taille, naturellement plus nombreux.

Influence des cantons

Afin de terminer l'évaluation de notre modèle, nous pouvons évaluer l'influence de chacun des cantons dans sa construction. On calcule, pour ce faire, la distance de Cook [7] :

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{k \text{ MSE}}$$

où :

$\hat{Y}_j$ est la prédiction pour le canton j.

$\hat{Y}_{j(i)}$ est la prédiction pour le canton j à partir d'un modèle où le canton i a été omis.

MSE est l'erreur quadratique moyenne du modèle de régression (que nous définirons plus loin).

Une fois cette valeur calculée pour chacun des cantons, on peut tracer l'histogramme des valeurs pour vérifier qu'aucune valeur aberrante ne vient nuire à notre modélisation :

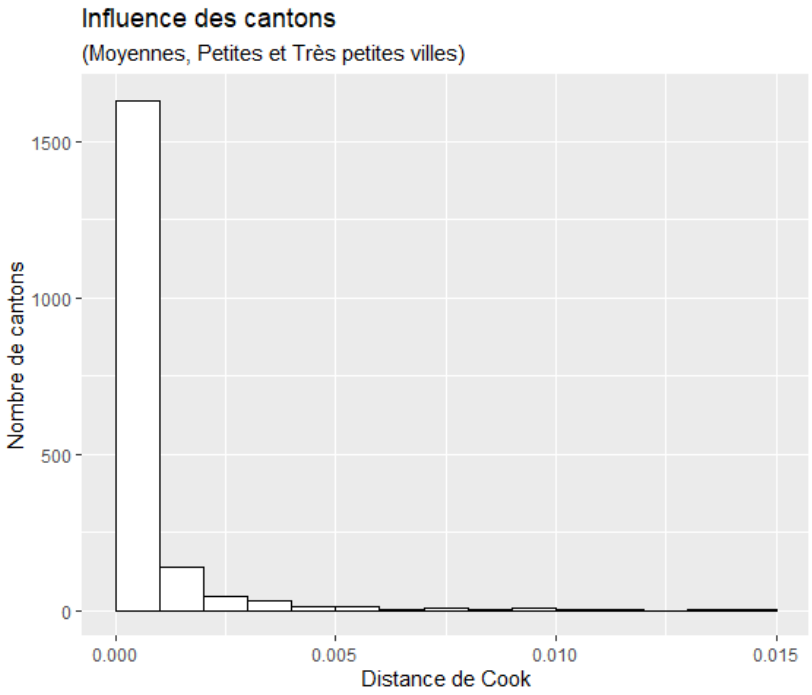


FIGURE 40 – Histogramme des distances de Cook par canton

Le faible nombre de données présentant une distance de Cook élevée nous conforte dans notre choix d'avoir traité les *Moyennes*, *Petites* et *Très petites villes* ensemble.

1.5 Performance du modèle

Nous allons maintenant évaluer la capacité de notre modèle à rendre compte des effets observés. Cela se fera au travers de plusieurs indicateurs dont nous décrirons le fonctionnement. Afin de disposer de points de comparaison, nous indiquerons également les résultats obtenus avec :

— Un modèle simpliste basé sur le salaire net horaire moyen uniquement.

— Les modèles complexes proposés par les algorithmes de construction automatique que nous avons présentés plus haut.

Les résultats que nous obtiendrons avec ces modèles serviront de borne inférieure et supérieure (respectivement) et permettront de situer les performances de notre modèle.

Critères basés sur la vraisemblance

Le critère d'information d'Akaike (en anglais *Akaike Information criterion*) est noté AIC [1]. Plus sa valeur est basse, plus le modèle est efficace. En règle générale, l'ajout de nouveaux paramètres a tendance à augmenter la vraisemblance du modèle et donc à améliorer ce type d'indicateurs. Le critère AIC pénalise les modèles ayant un grand nombre de paramètres pour limiter les effets de sur-ajustement que nous avons évoqués plus avant.

En notant k le nombre de paramètres et $\hat{L}$ la vraisemblance du modèle maximisée, le critère d'information d'Akaike s'écrit :

$$\text{AIC} = 2k - 2\ln(\hat{L})$$

Pour rappel, dans le cadre du modèle linéaire gaussien utilisé ici, la vraisemblance s'écrit :

$$L(y, \beta, \sigma^2) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma^n} \exp -\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)$$

Le critère d'information bayésien (en anglais *Bayesian Information Criterion*) est noté BIC [62]. Il s'agit d'un critère dérivé de l'AIC, où la pénalité dépend de la taille de l'échantillon et pas seulement du nombre de paramètres. En reprenant les notations précédentes et n le nombre d'observations, le critère d'information bayésien s'écrit :

$$\mathrm{BIC} = -2\ln(\hat{L}) + k\ln(n)$$

En appliquant ces critères, on obtient les résultats suivants :

Critère	**Modèle** *(Salaire)*	**Modèle**	**Modèle Auto.** *(Backard)*	**Modèle Auto.** *(Forward)*	**Modèle Auto.** *(Stepwise)*
BIC	-275.1424	-1080.231	-1264.34	-1309.615	-1322.116
AIC	-291.79	-1146.82	-1464.106	-1459.44	-1460.843

TABLE 7 – Critères de performance - Modèle linéaire principal

Théoriquement, choisir entre plusieurs modèles consiste simplement à privilégier celui affichant le critère le plus bas. Ici, ils nous permettent d'évaluer les performances de notre modèle manuel. Nous pouvons constater que :

— Le gain par rapport au modèle constitué uniquement du salaire est très net.

— La perte par rapport aux modèles automatiques (construits pour minimiser le critère AIC) est modérée.

Nous estimons donc que notre modèle affiche une performance satisfaisante.

Biais et variance des erreurs

On définit le biais comme l'écart entre la moyenne du modèle et celle de la variable à prédire.

L'erreur absolue moyenne (en anglais Mean Absolute Error) est noté MAE. Comme son nom l'indique, elle s'écrit :

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

où e_i sont appelés les résidus du modèle.

La racine carrée de l'erreur quadratique moyenne (en anglais Root Mean Square Error) est notée RMSE. Elle renseigne également sur la précision du modèle et, pour un estimateur non biaisé, elle mesure l'écart type de nos résidus.

$$\mathrm{RMSE}(\hat{Y}) = \sqrt{\mathrm{E}((\hat{Y} - Y)^2)}$$

En appliquant ces indicateurs, on obtient les résultats suivants :

Critère	**Modèle** *(Salaire)*	**Modèle**	**Modèle Auto.** *(Backard)*	**Modèle Auto.** *(Forward)*	**Modèle Auto.** *(Stepwise)*
Biais	0.01469613	-0.0001361969	-0.0008109321	0.0006770943	0.0002369886
MAE	0.1741477	0.14309	0.1338101	0.1338814	0.1338597
RMSE	0.2192372	0.1838045	0.1734018	0.173494	0.1737208

TABLE 8 – Caractéristiques de la régression - Modèle linéaire principal

Lors de la construction de notre modèle, nous avons pu constater que la médiane de la variance résiduelle était proche de zéro. Nous pouvons constater ici aussi que notre modèle n'est pas biaisé. En ce qui concerne les critères MAE et RMSE, nous pouvons constater que l'augmentation de la

variance des erreurs (par rapport aux modèles automatiques) est globalement raisonnable. Dans le cas des résidus de notre modèle, l'erreur absolue de majoration du risque serait donc de l'ordre de 1%, ce qui est raisonnable ; d'autant plus que nous considérons notre modèle comme plus robuste.

Pour l'ensemble de ces critères, le gain par rapport au modèle constitué uniquement du salaire est très net.

Coefficient de détermination linéaire de Pearson

Le coefficient de détermination linéaire, noté R^2, est une mesure de la qualité de la prédiction d'une régression linéaire [2]. Il se calcule de la manière suivante :

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_{i=0}^{n}(Y_i - \hat{Y})^2}{\sum_{i=0}^{n}(Y_i - \bar{Y})^2}$$

Où SSR est la moyenne des carrés des résidus telle que : $SST = \sum_{i=0}^{n}(Y_i - \hat{Y})^2$ et SST est la variance totale telle que : $SST = \sum_{i=0}^{n}(Y_i - \bar{Y})^2$

Avec Y_i l'estimation du ratio de mortalité du canton i ; Y_i le ratio de mortalité réel et $\bar{Y_i}$ la moyenne des ratios de mortalité réels.

Dit autrement, le R^2 représente la part de variance expliquée par le modèle :

$$R^2 = \frac{\sum_{i=0}^{n}(\hat{Y_i} - \bar{Y})^2}{\sum_{i=0}^{n}(Y_i - \bar{Y})^2}$$

Le coefficient de détermination peut aussi être vu comme le coefficient de corrélation linéaire au carré, pour les valeurs prédites $\hat{y_i}$ et les mesures y_i :

$$R^2 = corr(\hat{y}, y)^2$$

De ce fait, il peut être vu comme une généralisation du coefficient de corrélation au cas d'une régression linéaire multivariée. Ici, on utilisera le R^2 ajusté qui tient compte du nombre de variables utilisées :

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

Le nombre de variables dont nous disposons étant très important, cet indicateur est ici quasiment équivalent au R^2 non ajusté.

Nous proposons ici de calculer également un R2 par validation croisée, ce qui permettra de mettre en perspective notre évaluation classique et nous renseignera sur la robustesse de notre modèle. Pour ce faire, 100 fois, on répète cette opération :

1. On sépare notre jeu de données en 10 groupes.
2. Pour chacun de ces groupes, on ajuste notre modèle linéaire et on évalue le R^2 sur les 9 autres groupes.
3. On moyenne les R^2 obtenus pour en obtenir un représentatif.

Une fois nos 100 simulations réalisées on moyenne les 100 R^2 moyens obtenus.

Une fois nos calculs réalisés, nous obtenons ainsi les résultats suivants :

Critère	**Modèle** *(Salaire)*	**Modèle**	**Modèle Auto.** *(Backard)*	**Modèle Auto.** *(Forward)*	**Modèle Auto.** *(Stepwise)*
R^2 Ajusté	0.04912343	0.3285218	0.4032417	0.4029272	0.401118
R^2 Ajusté (Validation Croisée)	0.05283879	0.3253033	0.3903529	0.393358	0.3927015
Évolution relative	*7.563%*	*-0.979%*	*-3.196%*	*-2.374%*	*-2.098%*

TABLE 9 – Qualité de la régression - Modèle linéaire principal

Le R^2 du modèle salaire semble s'améliorer mais reste néanmoins très bas. Nous pouvons constater que la perte de capacité à porter la variance est quasi nulle dans notre modèle, ce qui indique une certaine stabilité dans sa construction. Nous constatons également que la diminution du R^2 existe sur les modèles automatiques, ce qui a tendance à indiquer une certaine variabilité. Encore une fois, la perte de pouvoir explicatif par rapport aux modèles automatiques est voulue et maîtrisée.

2 Approfondissements et ajustements de la modélisation

2.1 Construction des modèles complémentaires

2.1.1 Métropoles

Nous allons maintenant modéliser spécifiquement les *Grandes villes*. Plus précisément, il s'agit des villes comptant plus de 30 000 habitants entre 30 et 62 ans (ce qui correspond -globalement- aux villes de plus de 100 000 habitants). Parmi elles, seule une dizaine comptent plus de 100 000 habitants entre 30 et 62 ans, le maximum étant atteint pour Toulouse et la banlieue de Lyon, avec respectivement 190 000 et 350 000 habitants dans ces tranches d'âge.

Il y a deux éléments importants à prendre en compte ici :

— Les ratios étant calculés sur des populations plus grandes, ils sont logiquement plus représentatifs que ceux utilisés précédemment.

— Le nombre de grandes villes en France est relativement restreint (37 en l'occurrence).

Partant de ces constats, nous souhaitons :

— Conserver en partie l'information apportée par les ratios.

— Proposer une modélisation relativement simple.

Il a été envisagé dans un premier temps de conserver directement les ratios de mortalité bruts, car nous disposons d'un nombre de décès *relativement* important (entre 200 et 400 décès pour prés de trois quarts des villes et plus de 400 pour le reste d'entre elles). Néanmoins, nous disposons de deux ans d'exposition et pouvons donc supposer d'une certaine variabilité.

Les ratios que nous retiendrons seront donc une pondération à 50-50 des ratios constatés et de notre modélisation.

Comme pour le modèle général, nous pouvons lister les variables sélectionnées automatiquement et apparaissant dans :

— Les trois modèles : SAL

— Deux modèles POST10M_CONS AgeACC H15P_NSLR_INDEP MIGR1116

— Un modèle RD H15P_NSCOL_DIPLMIN A15P_NSCOL_DIPLMIN Equip AlcTbc

Le salaire est la seule variable quantitative que nous retiendrons ici, car efficace et simple d'interprétation. Nous retiendrons également les trois variables catégorielles proposées, car elles permettent d'apporter de l'information sans complexifier notre modèle plus que de raison. Les autres variables ont été étudiées, mais ne permettent pas de proposer une modélisation satisfaisante.

Nous obtenons donc le modèle suivant :

```
Call:
glm(formula = Y_GV ~ SAL + AgeACC + AlcTbc + Equip, family = gaussian(link = "identity"),
    data = BASE_GV, weights = Nbr)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-50.259  -14.646    2.605   12.323   47.730

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.990033   0.237490   8.379 4.09e-09 ***
SAL                     -0.054043   0.018828  -2.870 0.007722 **
AgeACCPlutôt agé        -0.214081   0.057314  -3.735 0.000851 ***
AgeACCPlutôt jeune      -0.222769   0.057361  -3.884 0.000574 ***
AlcTbcMoyenne           -0.111467   0.042123  -2.646 0.013205 *
AlcTbcTrès élevée       -0.214169   0.084504  -2.534 0.017137 *
EquipRural bien équipé   0.245342   0.103799   2.364 0.025272 *
EquipUrbain              0.216885   0.088043   2.463 0.020171 *
EquipUrbain bien équipé  0.005694   0.087748   0.065 0.948720
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 583.2488)

    Null deviance: 92861  on 36  degrees of freedom
Residual deviance: 16331  on 28  degrees of freedom
AIC: -58.861

Number of Fisher Scoring iterations: 2
```

FIGURE 41 – Modèle linéaire secondaire (Métropoles)

Encore une fois, nous avons choisi nos variables de façon à comprendre les effets que nous introduisons. Ce modèle est, néanmoins, essentiellement utilisé dans une vision pratique. Nous retrouvons plusieurs effets déjà constatés sur notre première modélisation :

— Le salaire a tendance à faire diminuer les ratios modélisés.

— Variable AgeAcc [Catégorie de référence : Élevée]
Les villes catégorisées "Plutôt âgé" ou "Plutôt jeune" présentent un effet équivalent, positif par rapport la référence.

— Variable AlcTbc [Catégorie de référence : Élevée] :
Les villes catégorisées comme ayant une incidence "Moyenne" ont des taux de mortalité plus faibles. Il faut noter que Lille et Amiens sont les seules villes catégorisées comme "Très élevée" et, donc, les seules à être concernées par cet effet statistique contre-intuitif qui présente une consommation "Très élevée" comme moins a risque.

— Variable Equip [Catégorie de référence : Ile-de-France]
On voit que les villes catégorisées comme "Urbain bien équipé" ne sont pas statistiquement différente de l'Ile-de-France. Les autres villes en revanche présentent un effet équivalent, négatif par rapport la référence.

Cette considération faite, nous pouvons nous faire une idée de la modélisation proposée par le graphique suivant :

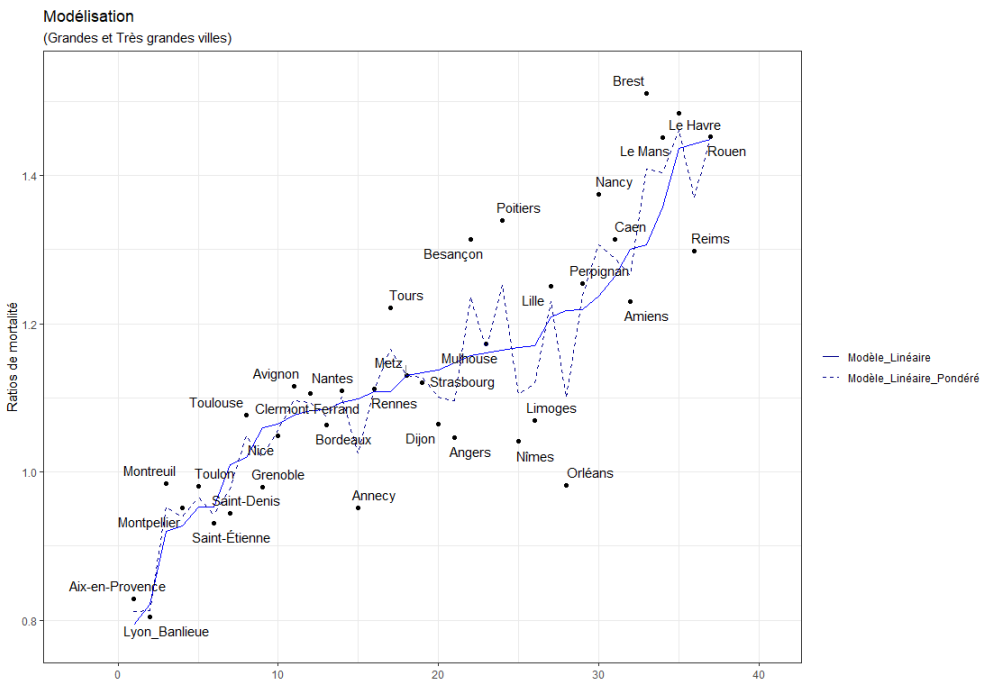


FIGURE 42 – Modèle linéaire secondaire (Métropoles) - Représentation

Sur ce graphique, nous pouvons donc observer notre modélisation, ainsi que les ratios retenus une fois la pondération suivante réalisée :

$$\hat{Y_{cv}^{pond}} = \frac{\hat{Y_{cv}}}{2} + \frac{Y_{cv}}{2}$$

Où :

— $\hat{Y_{cv}}$ est le ratio que nous venons de modéliser pour la ville cv

— Y_{cv} est le ratio brut de la ville cv

A noter que l'axe horizontale ne représente pas une valeur numérique, mais numérote simplement les villes modélisées. Une représentation triée par les valeurs du modèle pondéré est également proposée en annexe *(figure 60)*.

Nous pouvons désormais nous faire une idée des différents niveaux de mortalité dans les grandes villes françaises.

Lorsque nous interpréterons notre modélisation finale, il sera utile de faire référence à cette représentation. Il faut rappeler que ces effets tiennent compte de l'impact des catégories socioprofessionnelles.

On note que certaines villes présentent des ratios extrêmement élevés, notamment les villes de Bretagne et Normandie. Nous verrons que, en vue de notre application assurantielle, cet effet sera par la suite discuté et encadré.

Nous allons maintenant contrôler notre modélisation en reprenant, en partie, les étapes suivies pour évaluer le modèle principal. Nous serons ici plus succincts, ce modèle restant relativement simple.

Hypothèses du modèle linéaire

Comme pour notre premier modèle, l'hypothèse d'homoscédasticité peut être retenue :

```
        studentized Breusch-Pagan test

data:  fit_GV
BP = 12.804, df = 8, p-value = 0.1188
```

FIGURE 43 – Test de Breusch-Pagan - Modèle linéaire secondaire (Métropoles)

De la même façon, l'hypothèse de normalité des résidus n'est pas rejetée.

```
        Shapiro-Wilk normality test

data:  fit_GV$residuals
W = 0.97936, p-value = 0.7095
```

FIGURE 44 – Test de Sharpiro-Wilk - Modèle linéaire secondaire (Métropoles)

On conclu donc que notre modèle vérifie les principales hypothèses des modèles linéaires.

Performances

Étant donné le faible nombre de points dont nous disposons, nous allons privilégier ici le critère d'information d'Akaike corrigé, noté AICc.

$$\text{AICc} = \text{AIC} + \frac{2k^2 + 2k}{n - k - 1}$$

Critère	**Modèle** *(Salaire)*	**Modèle**	**Modèle Auto.** *(Backard)*	**Modèle Auto.** *(Forward)*	**Modèle Auto.** *(Stepwise)*
BIC	-11.56529	-42.75161	-50.27465	-49.51251	-49.51251
AICc	-16.0451	-52.19412	-58.21659	-58.95502	-58.95502

TABLE 10 – Critères de performance - Modèle linéaire secondaire (Métropoles)

Ici, les variables sélectionnées par les méthodes Forward et Stepwise sont les mêmes. Les constats sont globalement les mêmes que pour notre modélisation principale. Avec une amélioration nette par rapport au modèle basé sur le salaire et une perte modérée par rapport aux modèles automatiques. Il faut noter que notre modèle présente un R^2 de 0.734, ce qui est déjà élevée ; ainsi malgré notre volonté de nous rapprocher des données initiales, nous privilégions tout de même notre modèle manuel aux modèles automatiques.

2.1.2 Paris

Ayant retiré les arrondissements de Paris très tôt dans notre étude, il est intéressant de développer une analyse rapide de la capitale française. Comme nous l'avons vu, les arrondissements de Paris apparaissaient d'une manière assez particulière dans notre ACP et semblaient difficiles à interpréter. Typiquement, le lien entre le nombre d'habitants et le salaire net horaire moyen suit une dynamique différente d'une ville classique. Une fois isolés, les ratios de mortalité des arrondissements sont néanmoins relativement simples à modéliser.

Lorsque l'on affiche les variables corrélées aux ratios de mortalités parisiens, la majorité sont liées au salaire : On s'aperçoit très vite que le salaire net horaire moyen constaté dans un arrondissement donne une bonne idée de son ratio de mortalité.

Comme pour les Grandes et Très grandes villes, nous accordons une certaine crédibilité au ratios bruts. Néanmoins, nous le faisons avec plus de prudence, les arrondissements étant très inégalement peuplés. En effet, en population générale :

— Les dix premiers arrondissements peinent à dépasser les 100 000 habitants (le moins peuplé étant le 1er arrondissement qui, avec environ 15 000 habitants, compte a peine plus de 7500 individus en 30 et 62 ans.)

— Les dix arrondissements suivants comptent entre 150 000 et 200 000 habitants en moyenne.

La pondération que nous choisissons ici est d'un tiers pour les ratios bruts et de deux tiers pour notre modélisation. Comme pour les deux premiers modèles, nous appliquerons une pondération par nombre d'habitants.

Nous obtenons donc le modèle suivant :

```
Call:
glm(formula = Y_Pa ~ SAL, family = gaussian(link = "identity"),
    data = BASE_Pa, weights = Nbr)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-38.466   -8.705    2.401   17.008   44.337

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.146790   0.081569  14.059 3.79e-11 ***
SAL         -0.010674   0.003457  -3.087  0.00635 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 491.4945)

    Null deviance: 13532.0  on 19  degrees of freedom
Residual deviance:  8846.9  on 18  degrees of freedom
AIC: -25.554

Number of Fisher Scoring iterations: 2
```

FIGURE 45 – Modèle linéaire secondaire (Paris)

La régression que nous obtenons est ici très visuelle, d'autant plus que nous pouvons afficher la totalité du modèle :

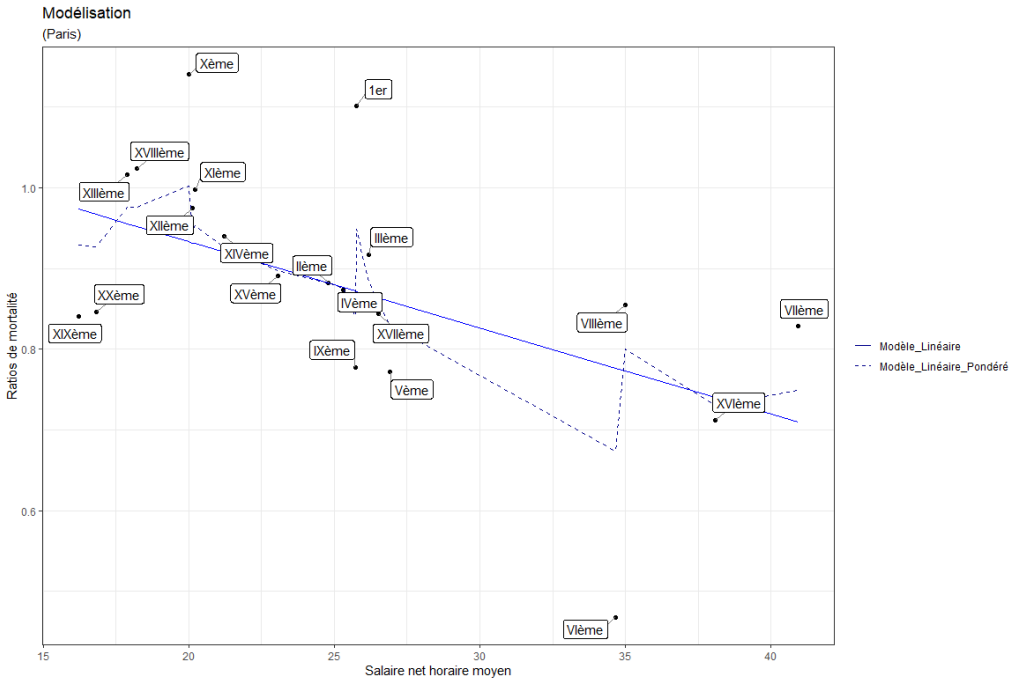


FIGURE 46 – Modèle linéaire secondaire (Paris) - Représentation

Il est clair que la capitale bénéficie d'un niveau de mortalité très avantageux par rapport à la moyenne nationale. Certains points semblent s'écarter significativement de notre régression pondérée, mais les effets sont globalement encadrés :

— Le premier et le sixième arrondissement étant peu peuplés, les ratios calculés sont sujets à une variabilité assez importante. Notre modélisation nous semble donc plus juste.

— Le dixième, le dix-neuvième et le vingtième arrondissement semblent s'écarter de notre modélisation et il est difficile de savoir si cela est justifié ou non. Les ratios modélisés semblent néanmoins raisonnables, notamment pour le vingtième dont le ratio ne modifie par une estimation classique (car égal à 1).

Il ne nous semble pas nécessaire de nous étendre plus amplement sur ce modèle dans le sens où il reste simple et pondéré ; donc peu sujet à un risque de modélisation. Nous donnons donc simplement le R^2 du modèle non pondéré qui est de 0.33 et qui, après pondération, permet d'atteindre une modélisation satisfaisante.

2.2 Encadrement des ratios modélisés

Nous disposons, au terme de ces travaux, de différentes modélisations que nous sommes maintenant en mesure d'agréger.

Nos ratios étant par nature très volatiles, notre modèle s'autorise des variables très élevées. Malgré nos efforts pour éviter la sur-interprétation, il semble prudent de limiter les excès générés par cette variabilité. Nous tirons donc les constats suivants :

— Il apparaît peu probable que l'on trouve de meilleurs niveaux de mortalité que ceux constatés dans notre cluster "Moyennes villes aisées".

— On peut supposer que plus les ratios sont élevés, plus ils sont représentatifs d'une prépondérance de population très à risque (où le taux de pauvreté serait élevée). Potentiellement, ils sont donc moins représentatifs de la population assurée.
On pense notamment ici à notre modélisation des grandes villes qui présente Rouen, le Havre,

etc... comme très à risque alors que, dans un cadre assurantiel, il semblerait peu réaliste d'appliquer à ces assurés une pénalisation de 40%.

Nous faisons donc le choix :

— De minorer les ratios modélisés par le ratio "Moyennes villes aisées". Notre ratio minimal est donc de 0.8370 ; 171 cantons sont concernés, soit 8% du nombre total de cantons.

— De majorer les ratios modélisés par le quantile à 97.5%. Notre ratio maximal est donc de 1.2534 ; pour 52 cantons sont concernés.

Dans le cadre d'une application en condition réelle, cette précaution semble pertinente et limite le risque de modélisation.

2.3 Lissage spatial

Avant de présenter nos résultats, nous allons compléter notre modélisation en proposant différents lissages spatiaux du modèle que nous avons construit.

En uniformisant géographiquement nos ratios, nous cherchons à atteindre plusieurs objectifs :

— Améliorer la lisibilité des cartes que nous construirons.

— Homogénéiser les ratios modélisés lorsque les cantons sont proches.

— Rendre notre modélisation moins dépendante du découpage cantonal.

A l'issue de cette partie, nous disposerons de deux nouvelles cartographies proposant :

— Un lissage fort : qui pourra immédiatement servir dans de l'aide à la décision.

— Un lissage modéré : potentiellement applicable dans le cadre d'une tarification.

L'usage de ratios pondérés (et leur applicabilité dans un cadre assurantiel) sera en fait l'occasion de réfléchir à un certain nombre de problématiques et sera plus amplement discuté lorsque nous présenterons nos résultats.

La technique de lissage spatial que nous appliquons est un lissage par noyau (Kernel smoothing). Le principe est relativement simple : on ajuste chacun des points en réalisant une moyenne sur lui-même et les points alentours. L'influence du voisinage est alors décidée par deux éléments :

— Le noyau : qui décrit la façon dont le voisinage est appréhendé.

— La bande passante : qui décrit la taille du voisinage.

De façon générale, les noyaux sont construits de façon à donner moins d'influence aux points éloignés.

Les lissage par noyau (typiquement noyaux gaussien ou quadratique) sont des lissages moyens dans le sens où ils sont fondés sur des calculs locaux de moyennes. Cette notion peut néanmoins être étendue pour définir des statistiques locales fondées sur des quantiles (médiane, déciles, etc...) comme cela est proposé dans cette étude : [4]. Nous utiliserons ici un lissage médian.

Il faut noter que les noyaux moyens sont en général à éviter pour les taux car il donne un poids identique à des territoires inégalement peuplés ; malheureusement pour des ratios de mortalité, il semble hasardeux de séparer numérateur et dénominateur, les risques de déformations étant trop importants. Le choix de la médiane nous semble ici pertinent dans le sens où il aura tendance à ne lisser que les zones très hétérogènes. Nous serons néanmoins attentifs à ne pas appliquer ces ratios sans contrôler leur pertinence.

Bande passante et tests de bande passante optimale

Ayant défini le type de lissage utilisé, il nous faut maintenant choisir une bande passante. La bande passante est un paramètre fondamental de l'analyse spatiale et conditionne l'aspect plus ou moins «lissé» de notre estimation. Plus la bande passante est importante, plus le nombre de points participant au calcul des estimations est important. D'une façon générale, une bande passante élevée a tendance à augmenter le biais et à réduire la variance.

Il est intéressant d'utiliser plusieurs niveaux de lissage pour appréhender les effets spatiaux de différentes manières.

Nous avons choisi d'utiliser des tests de bande passante optimale suivant :

— Critère par erreur quadratique moyenne :
Ce critère choisit une bande passante qui minimise un critère $M(\sigma)$ basé sur l'erreur quadratique moyenne. Nous obtenons une bande passante de 53.69.
Ce lissage sera retenu et désigné comme **lissage minimal**.

— Critère par maximum de vraisemblance :
Ce critère choisit une bande passante en calculant un estimateur de vraisemblance et applique une validation croisée.
Nous obtenons une bande passante de 121.22.
Ce lissage sera retenu et désigné comme **lissage modéré**.

— Critère de Scott :
Ce critère choisit une bande passante est choisie est proportionnelle à $n^{(-1/(d+4))}$ où n est le nombre de points et d la dimension (ici 2). Nous obtenons une bande passante en deux dimension : x = 389.91 , y = 430.68.
Ce lissage sera retenu et désigné comme **lissage maximal**.

— Critère géométrique :
Ce critère choisit une bande passante selon un critère géométrique. La bande passante est un quantile de la distance entre deux points indépendants, pris au hasard dans la fenêtre. Nous avons utilisé le quantile proposé par défaut. Nous obtenons une bande passante de 3561.961.
Ce lissage ne sera retenu car il propose un niveau de lissage considéré comme irréaliste.

On observe une grande variabilité des bandes passantes. Il faut encore une fois noter qu'aucune bande passante n'est optimale et que toutes sont susceptibles d'apporter une représentation du monde pertinente. A l'issue de cette partie, nous avons donc retenu trois niveaux de lissage.

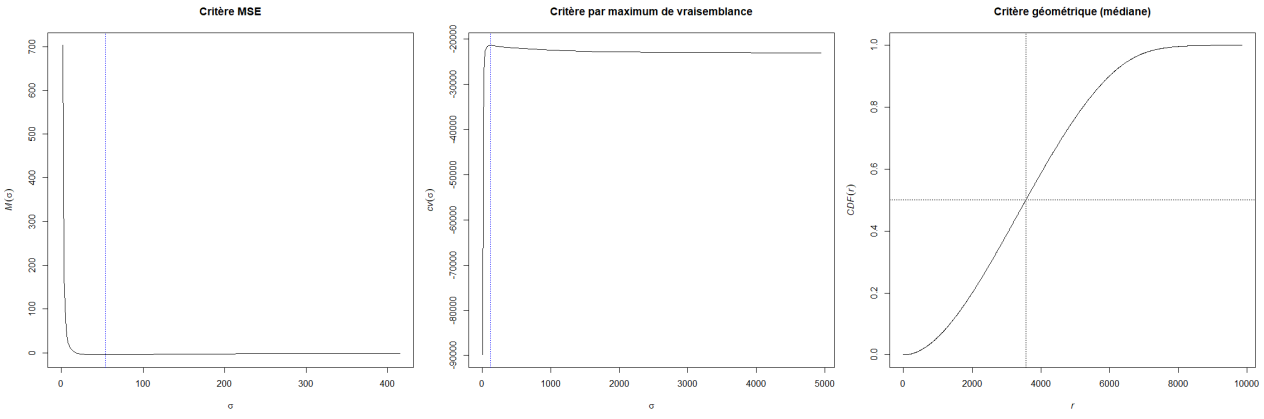


FIGURE 47 – Visualisation des critères de lissage

Quatrième partie

Présentation, analyse et application des résultats

Au terme de nos travaux, nous disposons finalement de ratios de mortalité pour l'ensemble des 2094 [22] cantons français. Nous proposons également un lissage spatial de nos résultats.

Cette partie vise essentiellement à présenter nos résultats sous forme de cartographies. Nous discuterons ici des effets mis en évidence, ainsi que des liens entre le cadre de vie et la mortalité. Nous verrons que nos travaux, au-delà de leur intérêt pratique, sont révélateurs d'un certain nombre de constats sur la mortalité en France métropolitaine.

Par la suite, nous tâcherons de prendre du recul sur le travail réalisé. Nous chercherons à identifier l'ensemble des limites et interrogations inhérentes à l'étude que nous avons réalisée. En réponse, nous apporterons des solutions et des pistes de réflexion.

Nous serons également en mesure de faire le lien entre la mortalité modélisée et les éléments abordés au début de ce mémoire. Cette section sera l'occasion de confirmer nos intuitions et la cohérence de nos travaux.

Enfin, nous évoquerons les -nombreuses- perspectives de développement que nous envisageons pour les travaux présentés ici.

Cette partie sera, en définitive, l'occasion de conclure la partie théorique de ce mémoire par un tour d'horizon des résultats obtenus.

1 Modélisation finale et analyse du territoire

Revenons rapidement sur la méthodologie déployée.

Nous avons estimé des ratios de mortalité pour l'ensemble des cantons métropolitains, sur la population de 30 à 62 ans en 2018 et 2019. Nous avons utilisé, d'une part, le recensement comme base d'exposition et, d'autre part, des taux de mortalité nationaux, améliorés et différenciés par catégories socioprofessionnelles. Nous avons alors pu constater que ces ratios n'étaient pas exploitables directement, car extrêmement variables.

Nous avons par la suite introduit un certain nombre de données prédictives. Par elles, nous avons cherché à modéliser les déviations de mortalité observées sur des critères interprétables et cohérents. Les résultats que nous proposons sont l'agrégation de quatre niveaux de modélisation :

— Une modélisation excluant les grandes villes.

— Une modélisation spécifique aux grandes villes.

— Une modélisation spécifique à Paris.

— Une modélisation spécifique au sud-ouest de l'Ile-de-France [23].

Nous avons finalement évalué la robustesse et les performances de notre modèle. Au terme de nos travaux, nous disposons maintenant d'une modélisation fine, fiabilisée et contextualisée.

Les ratios modélisés ici sont donc le reflet d'une déviation de mortalité résiduelle, une fois l'âge, le sexe et la csp des populations retraités.

22. En comptant les arrondissements de Paris, Marseille et Lyon
23. Cluster *Moyennes villes aisées*

1.1 Modélisation à l'échelle du canton

Nous obtenons finalement la modélisation suivante :

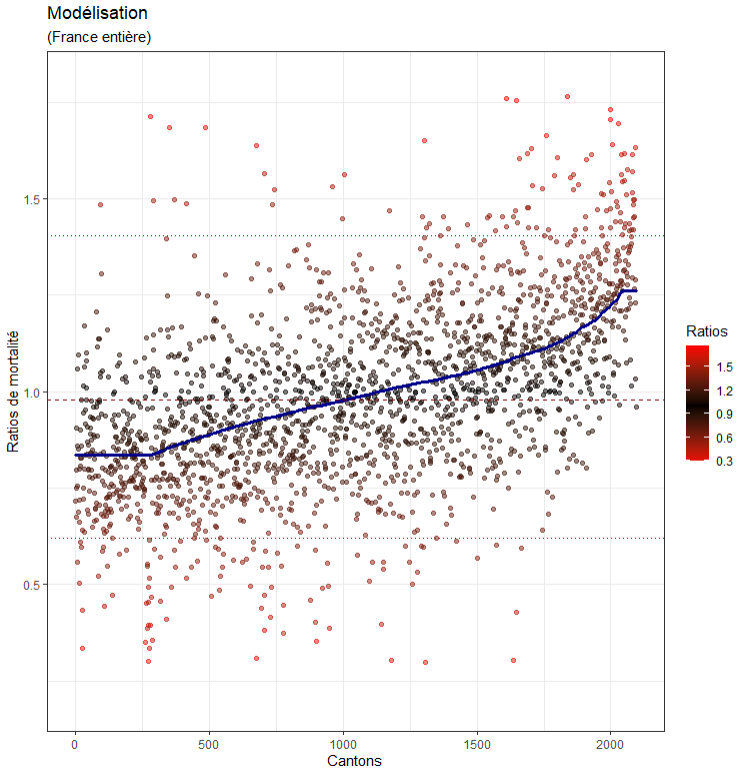


FIGURE 48 – Modèle linéaire final - Représentation

Cette représentation permet de tirer plusieurs constats :

— L'allure de notre modèle est satisfaisante. Nos ratios modélisés sont correctement distribués autour de 1. Les valeurs sont bien échelonnées et l'on n'observe pas de cassure. Le plafonnement des ratios appliqué apparaît également cohérent.

— Nous avons globalement réussi à traiter la grande variabilité de nos données. Le nuage de points suit bien la dynamique de notre modèle. Comme nous l'avons vu, le modèle n'introduit pas de biais et évite la sur-interprétation en expliquant simplement un tiers des variations.

Il faut noter que les cantons présentant un trop faible nombre de décès ont été réintroduits dans cette représentation, ce qui explique les quelques valeurs aberrantes que nous pouvons constater.

Une fois nos ratios mis en carte, nous obtenons la représentation suivante :

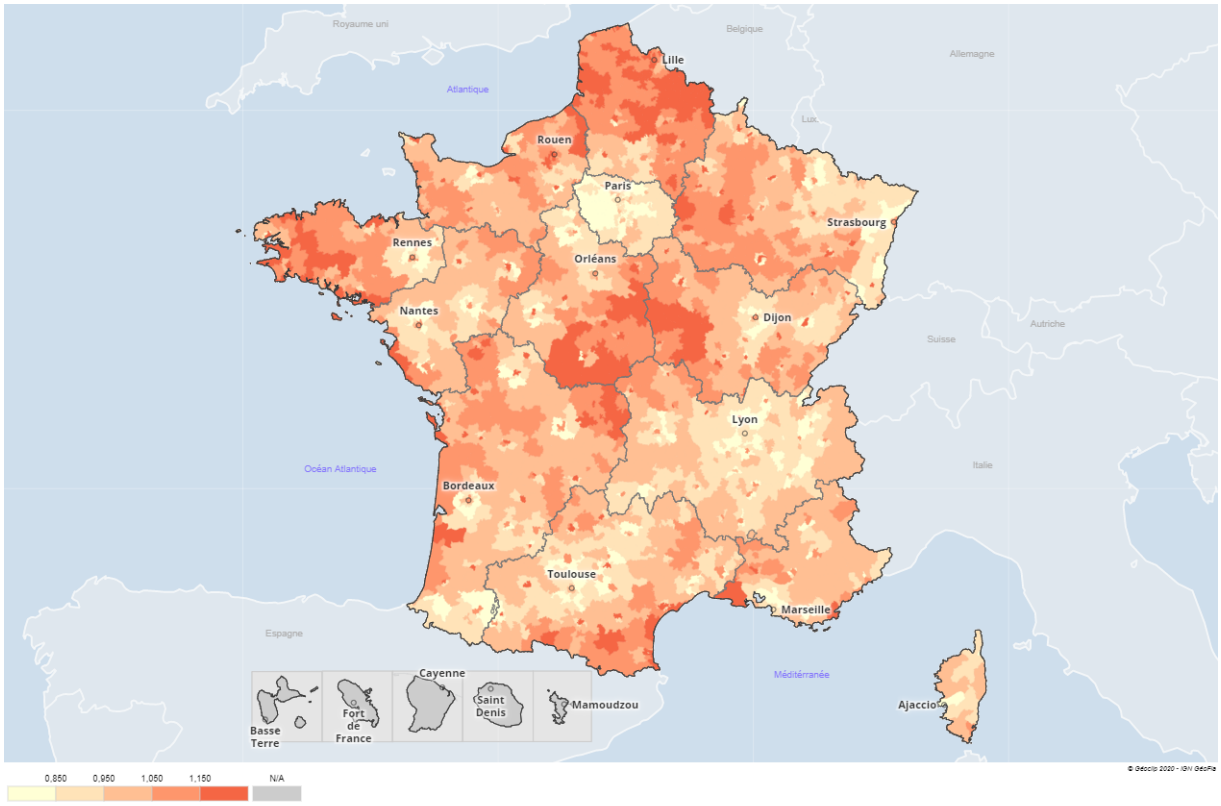


FIGURE 49 – Modélisation des ratios de mortalité
Obtenu à partir de www.france-decouverte.geoclip.fr - Données importées [20]

Nous pouvons voir que le rendu est très cohérent. On observe une certaine homogénéité dans les territoires et l'on peut, d'ores et déjà, distinguer des zones présentant une sur ou une sous-mortalité claire.

Mais, avant de nous tourner vers des considérations plus macroscopiques, il est intéressant de tirer un certains nombre de conclusions sur les dynamiques mises en évidences.

1.2 L'impact du cadre de vie sur la mortalité

1.2.1 Métropoles et périurbanisation

Sur notre représentation, on remarque un schéma récurrent : un effet de tâche qui permet d'identifier facilement les villes (sur toute la partie ouest et sud-ouest de la France cet effet est très visible).

Cet effet est en fait symptomatique de l'étalement urbain et a déjà été mentionné lorsque nous évoquions l'impact de la mobilité sur la mortalité.

Les villes semblent en effet avoir une influence positive sur les territoires qui les entourent. Cela se voit particulièrement autour de Rennes, Nantes et Bordeaux par exemple. On observe toujours le même phénomène :

- Dans la ville : une mortalité supérieure à la moyenne nationale.
- Autour de la ville : une mortalité très inférieure à la moyenne nationale (une partie de la banlieue étant parfois exclue de cet effet positif ; voir présentant une surmortalité.)

Il semble également que, plus les villes sont grandes, plus cet effet est étendu.

On comprend en fait que les périphéries des villes attirent (et donc concentrent) des populations moins à risque. Il s'agit essentiellement d'individus étant sortis de la ville pour profiter d'un meilleur

cadre de vie (logement en maison ou appartement plus spacieux) tout en profitant du dynamisme de la ville (d'où l'usage de la voiture).

Nous avons mentionné l'importance de la taille moyenne des ménages dans notre modélisation. Ce point peut à nouveau être abordé ici. On peut supposer que le fait d'avoir des enfants pousse à s'éloigner des centres urbains et serait corrélé avec un faible risque de mortalité. Et à l'inverse, que la proportion d'individus isolés serait plutôt révélateur du risque de surmortalité.

En résumé, l'effet positif affiché par la périurbanisation s'explique. Notre modélisation démontre que la mortalité de ces populations est inférieure à la moyenne nationale de 10% à 15% en moyenne [24].

Il peut paraître surprenant, en revanche, que le centre des villes présente une mortalité supérieure à la moyenne nationale. Il faut en fait savoir que la pauvreté se concentre dans le centre des grandes villes. L'Insee nous apprend à ce titre que [53] :

"Au sein des grandes aires urbaines, le taux de pauvreté est presque toujours plus élevé dans les villes-centres. Il atteint parfois deux à trois fois celui des banlieues et plus de quatre fois celui des couronnes péri-urbaines.".

Ce sujet soulève en lui même certaines interrogations et sera abordé plus en profondeur dans la section dédiée à l'applicabilité de nos travaux. Nous y évoquerons l'hétérogénéité des villes comme une occasion pour prendre du recul sur nos travaux.

En somme, notre modèle met en évidence un lien clair entre périurbanisation et faible risque de mortalité. Si nous avons pu proposer des explications, il apparaît surtout que cet effet est modélisable, et donc exploitable d'un point de vue actuariel.

1.2.2 Corrélation des résultats avec l'impact des csp

Il est intéressant montrer ici comment l'introduction des csp a impacté nos estimations de mortalité (avant toute modélisation).

Nous définissons le ratio suivant : [25] :

$$I_{cv} = \frac{D_{cv}^{csp}}{D_{cv}} - 1$$

Il s'agit de l'impact relatif qu'a l'introduction de la csp sur nos estimations. Par exemple, si le canton x présente un impact $I_x = 0.1$, nous estimons qu'il faut augmenter l'estimation de décès de ce canton de 10% pour tenir compte des proportions de csp.

Une fois ces impacts calculés, nous obtenons la représentation suivante :

24. Il n'est pas si rare que les zones péri-urbaines des grandes villes atteignent la borne inférieure que nous nous sommes fixée.

25. Multiplié par 100 pour être exprimé en pourcentages

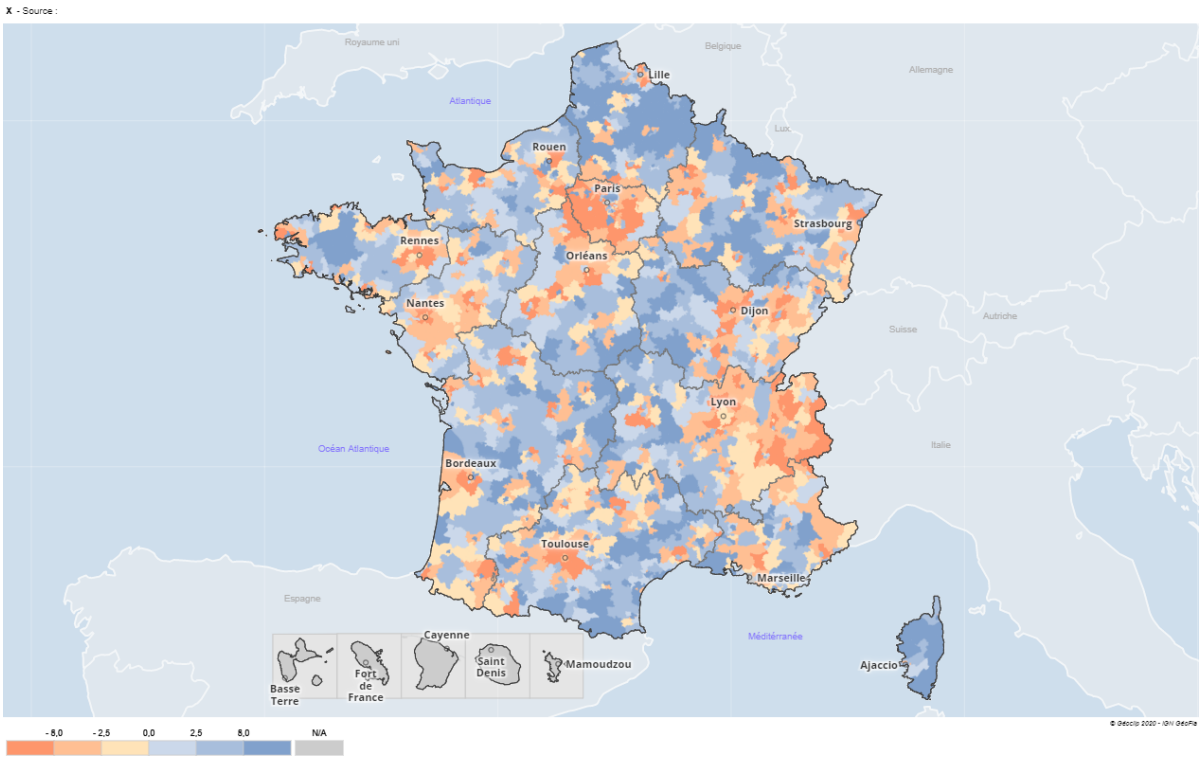


FIGURE 50 – Impact relatif de l'introduction des csp sur les estimations de mortalité
Obtenu à partir de www.france-decouverte.geoclip.fr - Données importées [20]

Il est intéressant de constater une corrélation entre notre modélisation et ces impacts.

Cet lien s'explique et est très cohérent.

Il faut comprendre ici que notre modèle met essentiellement en avant des effets de richesse, comme en témoigne l'importance du salaire dans nos modélisations.

Dans la partie précédente, nous avons notamment montré la sous-mortalité des espaces périurbains. Généralement, dans ces zones, les individus ont eu les moyens de s'éloigner de la ville pour disposer de meilleures conditions de vie. Ils bénéficient également d'un marché du travail dynamique et d'un salaire avantageux.

Il se trouve que la richesse économique, le niveau d'étude et l'exercice d'une profession avantageuse sont corrélés. Les zones aisées concentrent donc logiquement des individus de classe socioprofessionnelle avantageuse ; d'où la corrélation entre la répartition des csp sur le territoire et les effets que nous mettons en évidence.

On notera également que l'introduction de la différentiation par csp n'est pas suffisante pour gommer ces différences territoriales.

Ainsi, il est clair que les catégories socioprofessionnelles sont en fait un indicateur parmi d'autres. On peut citer :

— Le niveau d'études.

— La capacité à accéder à des soins et l'attention portée à la santé.

— Le sport, l'alimentation, une faible consommation d'alcool et de tabac.

— Le salaire et l'accès au logement.

— ...

Tous ces aspects sont corrélés. Ils sont révélateurs d'une même réalité que l'on peut, en définitive, résumer par le terme de réalité "socio-économique" des individus. Cette notion porte en elle-même

tous les aspects qui composent le niveau et les habitudes de vie des individus. Ces aspects, comme nous venons de le montrer, sont révélateurs du risque de mortalité.

A ce titre, nous pouvons dire que les catégories socioprofessionnelles sont en fait une façon sous-optimale de capter les effets dont nous souhaitons, au fond, tenir compte lorsque nous estimons un risque de décès. Néanmoins, les csp sont simples à manipuler et répondent à un besoin concret de segmenter les individus, et ce à un niveau plus fin que l'échelle d'une ville. Cette notion (ou une analogue) reste indispensable en ce sens.

Pour résumer, et bien que cela ne soit pas nécessairement une surprise, notre modèle capte avant tout des effets socio-économiques au travers de la géographie.

1.2.3 Facteurs environnementaux et risque de mortalité

Il pourrait être surprenant de ne pas voir apparaître des effets environnementaux comme la pollution et l'usage des pesticides par exemple. Au regard de l'importance des aspects socio-économiques que nous venons d'aborder, nous estimons que ces effets sont en fait imperceptibles sur la population étudiée (30-62 ans) étant donné la méthodologie employée.

Afin de préciser ce point, il est intéressant de revenir sur le modèle de Cox présenté dans la première partie de ce mémoire.

Le modèle proposé par l'Insee indique que, *toutes choses égales par ailleurs*, l'Ile de France a un impact plutôt négatif sur la mortalité. Cela pourrait s'expliquer, par exemple, par un facteur environnemental, à savoir un niveau de pollution élevé. Or, notre modélisation fait ressortir cette région comme présentant une sous-mortalité.

Ces conclusions pourraient sembler contradictoires. Ce n'est pas le cas.

La différence réside dans le terme *toutes choses égales par ailleurs*. Si nous étions capables de distinguer parfaitement l'impact de chaque paramètre utilisé dans l'étude de l'Insee (sexe, âge, niveau de vie, csp et diplôme) pour les répercuter dans nos évaluations de mortalité, nous devrions retrouver les résultats de l'Insee.

Notre modélisation capte bien des effets jusque-là ignorés. Par exemple :

— Si nous démontrons, comme l'Insee, que les habitants du nord de la France sont sujets à une surmortalité, c'est que d'importants effets existent.

— Si nous ne démontrons pas que l'Ile-de-France a un effet négatif sur la mortalité mais que, au contraire, il faut la considérer comme ayant un effet positif, c'est que les effets socio-économiques (niveau de vie et diplôme en l'occurrence) sont en fait plus importants que les effets de la "pollution" [26].

Les effets que nous modélisons étant ignorés par les techniques de tarification classique, il est pertinent de les prendre en compte. Ils sont en revanche dépendants de l'estimation a priori que nous faisons des décès.

S'il est certain que la pollution a des effets sur la mortalité, ils sont en tous cas trop diffus pour permettre d'identifier d'éventuelles zones à risques. Sur notre population, les aspects environnementaux sont donc peu impactant au regard des aspects socio-économiques mis en évidence.

En revanche, sur une population de retraités par exemple, si nous pouvons supposer que les aspects socio-économiques restent déterminants, nous ne pouvons pas présupposer de l'importance des facteurs environnementaux.

Ces divers éléments nous permettent de nous faire une idée de la hiérarchie des paramètres déterminant le risque de mortalité. Avoir conscience de ces effets oriente les choix qui doivent être faits dans la récolte de données. Si, demain, un assureur choisit de demander à ses clients s'ils font

26. Ou tout autre facteur expliquant cette surmortalité.

usage de la voiture en présupposant un risque aggravé (qui existe bien entendu), il pourrait très bien s'apercevoir que les assurés ayant répondu oui présentent un risque de décès plus faible que les autres. L'assureur serait plus avisé de s'intéresser, dans un premier temps, au métier, au niveau d'étude et, par exemple, au capital assuré pour estimer le niveau de vie de ses assurés. [27]

Le risque de mortalité peut être approché par divers pans mais, en vérité, les effets s'entrecroisent et cela complexifie l'analyse. Nous avons mentionné l'impact de la pollution. Il apparaît impossible, de par la méthode employée et les données récoltées, de raffiner nos travaux jusqu'à un tel niveau de détail. Les sujets mentionnés en introduction trouvent ainsi leur écho ici : pour atteindre cette finesse dans l'analyse, les assureurs et réassureurs doivent disposer -encore et toujours- de données de plus en plus précises.

Ces considérations posées, nous pouvons maintenant passer à une analyse plus générale des effets observés. Pour étudier les zones en sur ou sous-mortalité, il sera plus efficace de nous tourner vers notre modélisation lissée, quitte à revenir vers cette première carte pour apporter des nuances.

1.3 Cartographie des déviations de mortalité en métropole

Nous proposons ici d'étudier une cartographie de notre modélisation lissée.

Les ratios que nous affichons ici sont le résultat du lissage le plus fort que nous avons réalisé (bande passante de 400) et n'ont pas vocation a être utilisés en conditions réelles.

La représentation suivante est utilisée ici comme support pour la présentation de nos résultats et doit essentiellement être vue comme une carte d'aide à la décision :

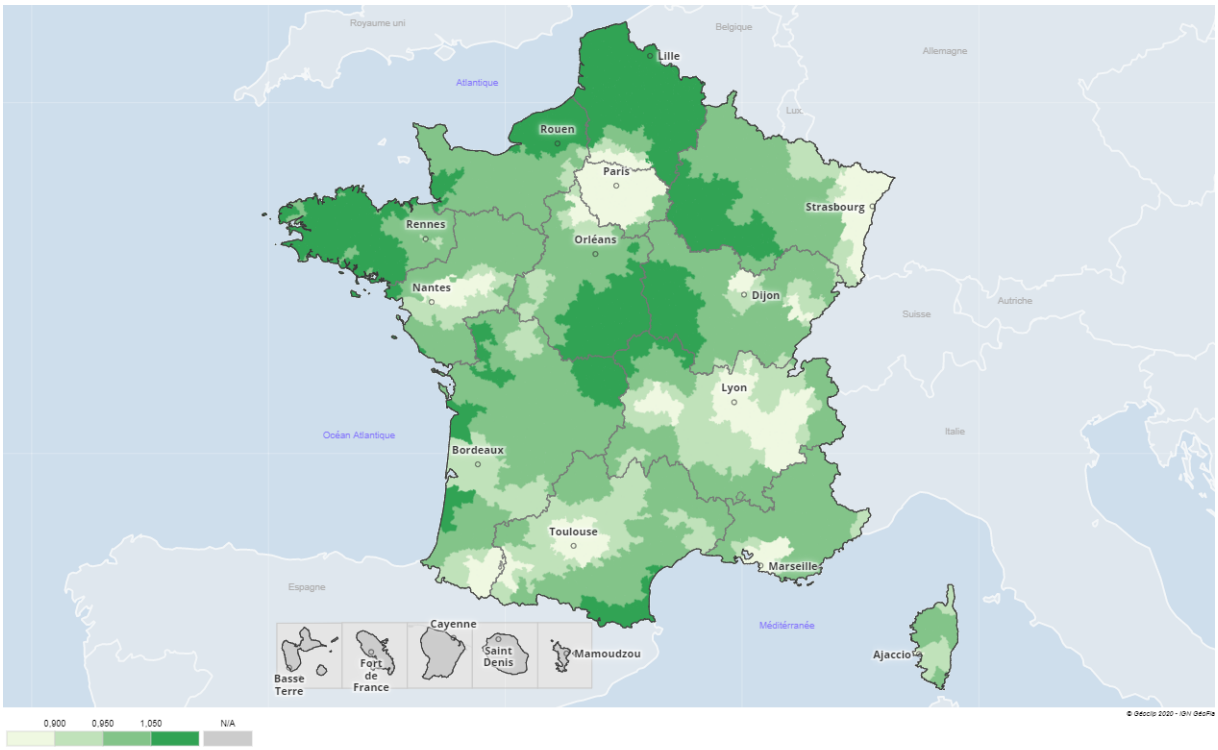


FIGURE 51 – Modélisation des ratios de mortalité - Lissage maximal

Obtenu à partir de www.france-decouverte.geoclip.fr - Données importées [20]

27. Ces propos sont bien évidemment à nuancer, dans le sens où un portefeuille assuré pourrait tout à fait être suffisamment homogène pour identifier des effets comme le risque d'accident de voiture (qui représente 2% des décès chez les moins de 65 ans [11]).

Grâce à cette carte, nous pouvons maintenant identifier clairement des zones en situation de sur ou de sous-mortalité, que nous allons nous attacher à détailler ici. Nous prendrons également soin de nous référer :

— à notre première modélisation pour apporter des nuances lorsque cela est nécessaire.

— à notre modélisation des grandes villes pour comparer les villes entre elles.

Zones en sous-mortalité

Nous pouvons observer une sous mortalité dans les zones suivantes :

— la région Auvergne-Rhône-Alpes :

Il s'agit clairement d'une zone très favorisée qui apparaissait déjà sur notre première modélisation. Globalement, l'ensemble des territoires autour de Lyon sont en sous-mortalité. Ce constat semble pouvoir s'étendre à l'ouest de la région (vers Clermont-Ferrand).
On dispose d'une modélisation au niveau des arrondissements de Lyon, qui bénéficient tous d'une sous-mortalité importante (à l'exception du septième qui présente une légère surmortalité).

— la région Ile de France

L'île de France apparaît également de façon très positive dans nos représentations. Cela s'explique par un haut niveau de vie et la proportion d'emplois avantageux dans la région.

L'uniformité affichée est néanmoins exagérée. En Île-de-France comme partout, des problèmes sociaux existent. Mais, si une surmortalité ponctuelle peut exister, elle est en tous cas marginale et éclipsée par la prépondérance de bons risques. Il faut rappeler -à ce titre- la prise en compte des csp dans notre analyse. Si nous reprenons l'exemple de Saint-Denis, où le taux de pauvreté est très fort, l'impact positif que nous lui attribuons n'en est pas moins réel ; une ville possédant une telle proportion d'inactifs devrait compter plus de décès que ce qui est constaté.
Notons également que les communes de la mégalopole sont en moyenne sept fois plus denses (8 600 habitants au km^2) que les onze plus grandes métropoles de province. Les considérations sur la densité et la pertinence d'une étude infra-communale se pose donc particulièrement sur une région aussi complexe.

Nous constatons également que cette zone à très faible niveau de mortalité s'étend au sud-ouest de la région Île-de-France (vers Chartes).

— les départements Bas Rhin et Haut Rhin :

Nous constatons une sous-mortalité claire. Il est possible que la proximité avec l'Allemagne puisse avoir un rôle, les frontaliers bénéficiant potentiellement d'un marché du travail dynamique.

Plusieurs grandes villes apparaissent également sur cette représentation. De manière marquée :

— Toulouse :

Pour cette dernière on semble pouvoir faire une analogie avec Lyon. Le centre de l'Occitanie est ainsi relativement uniforme.
Sur notre modélisation non lissée, on note néanmoins que les villes de taille intermédiaire gravitant autour de Toulouse sont en situation de légère surmortalité, alors que les territoires entourant Lyon sont déjà particulièrement homogènes et en sous-mortalité.

— Marseille :

La seconde ville de France fait figure d'exception et présente, malgré sa taille, une périurbanisation peu étendue. D'une façon générale, les ville portuaires concentrent des emplois ouvriers ; on peut donc supposer que leur dynamique est assez différente de celle d'une grande ville plus "classique".
Au niveau infra-communale, si la plupart des arrondissements présentent une très forte sous-mortalité, le constat est plus nuancé que pour Lyon. Les 2 et 5ème arrondissements présentent en effet une très légère surmortalité, tandis que le 16ème arrondissement présente une surmortalité plutôt importante.

Les trois villes suivantes sont, globalement, assez similaires. Elles se caractérisent par la structure particulière de leur espace péri-urbain et sont donc moins interprétables dans notre représentation simplifiée :

— Nantes : dont la zone en sous-mortalité s'étend ici jusqu'à Angers de façon quelque peu exagérée.

— Rennes : qui, du fait d'une banlieue proche assez vaste et en surmortalité, ne ressort pas dans cette représentation.

— Bordeaux : qui présente une sous-mortalité légère.

La périphérie de ces villes présente donc une sous-mortalité apparente. Les villes en elles-mêmes se voient attribuer un ratio à peu près équivalent par notre modèle, c'est à dire de 1.1 environ.

Il faut aussi noter la présence de deux zones moins peuplées et qui présentent une sous-mortalité assez nette :

— Dijon : pour laquelle notre analyse sur l'influence positive des grandes villes s'applique, bien que Dijon soit de taille moyenne.

— L'est des Pyrénées-Atlantiques : qui présente un effet ponctuel ; cette zone est une des rares disposant des caractéristiques d'une zone en sous-mortalité sans pour autant être proche d'une grande ville.

Zones nuancées

Parmi les dix plus grandes villes de France, seules deux ne présentent pas d'effet clair :

— Montpellier :

Cette ville ressort peu dans nos représentations pour une ville aussi grande.
Néanmoins, à l'inverse de la majorité des grandes villes de France, elle ne présente presque pas de surmortalité. On peut donc, tout de même, lui attribuer une influence positive.

— Nice :

Cinquième ville de France, elle mérite d'être citée mais est quasiment absente de notre représentation. Elle apparaît relativement neutre du point de vue de la mortalité et reste peu prise en compte du fait de sa forte proportion de personnes âgées.

Certains effets de périurbanisation sont atténués dans notre représentation. Les villes de moyenne taille, si elles disposent aussi d'un péri-urbain moins à risque, sont généralement entourées de zones moins urbaines et présentant une légère surmortalité. Notre représentation lissée les présente donc comme ayant un niveau de mortalité national. Parmi les villes concernées, nous pouvons citer Tours, Poitiers et le Mans par exemple, qui sont plutôt pénalisées dans notre modélisation des grandes villes.

Zones en surmortalité

Sur cette représentation, on remarque que certaines zones comportent des populations présentant un risque de mortalité supérieur au niveau national. On en distingue trois :

— la diagonale des faibles densités

Cet ensemble est parfois appelé de façon abusive "Diagonale du vide". Il englobe le centre de la France et également l'ouest de la région Grand-Est (où l'on trouve notamment Reims). Cet ensemble se caractérise par une absence de grandes villes et une densité relativement réduite.

— le nord de la France

Cette dénomination regroupe le département de Seine maritime et la région Hauts de France. Parmi les grandes villes concernées il faut essentiellement citer Lille, Amiens, Rouen et le Havre.

— l'ouest de la Bretagne

Cette zone comporte notamment Brest, très pénalisée dans notre modèle grande ville et qui possède d'ailleurs le ratio brut le plus élevé.

Comme nous l'avons mentionné, la surmortalité observée résulte d'aspects socio-économiques.

Il semble ici pertinent d'aborder le rôle de l'alcool et du tabac, qui représentent une part importante de la mortalité prématurée (i.e. avant 65 ans). Le nord de la France et la Bretagne sont notamment concernées par un fort impact de l'alcool sur la mortalité prématurée. Le Nord de la France et la région Grand-Est sont quant à elles concernées par un fort impact du tabac.

En annexe, nous proposons une analyse succincte de cartographies par régions (obtenues dans le rapport DREES 2017 sur l'état de santé de la population française [11]) présentant l'impact du tabac et de l'alcool sur la mortalité prématurée.

En résumé, les effets les plus marqués restent les suivants :

— les régions Île-de-France et l'est de la région Rhône-Alpes sont clairement en situation de sous mortalité. Toulouse peut également être citée.

— la diagonale de faible densité, l'ouest de la Bretagne et le nord de la France sont en situation de surmortalité.

Sur cette carte simplifiée que nous proposons, des effets sont logiquement masqués ou atténués. Cette représentation a également tendance à valoriser l'étalement urbain.

Néanmoins, si cette carte demande de l'attention pour être bien interprétée, elle permet de se faire une idée rapide des différents de niveaux de mortalité que l'on peut trouver sur le territoire métropolitain.

2 Applicabilité, cohérence et perspectives

Dans cette partie, nous allons maintenant prendre du recul sur nos travaux.

Pour ce faire, nous nous poserons un certain nombre de questions critiques et tâcherons d'y apporter des réponses. Nous nuancerons ainsi la portée de nos travaux qui, s'ils présentent un intérêt assurantiel certain, ne sauraient être appliqués à tout type de produit sans être challengés en amont. Nous identifierons à ce moment des pistes d'ajustements.

Dans un premier temps, nous poserons le problème de l'hétérogénéité des populations qui peut remettre en question les ratios proposés dans les zones denses. Nous aborderons alors le lissage minimal construit précédemment comme une piste d'amélioration. Dans un second temps, nous discuterons du lien entre la population assurée et la population générale, qui n'est en fait pas immédiat. Finalement, nous ramènerons nos taux à l'échelle de la région et du département, ce qui permettra de mettre en avant la cohérence des résultats proposés.

2.1 Le cas des mégalopoles

Nous avons évoqué plus avant l'impact de la pauvreté en ville sur les ratios de mortalité qui y sont proposés. Cela amène donc la question suivante : à quel point nos ratios sont-il applicables à des zones hétérogènes comme le centre des villes ?

Il est assez évident que les zones denses sont plus complexes à cerner. Les inégalités y sont plus fortes et il n'est pas évident qu'un ratio de mortalité général soit idéal.

Pour le centre des grandes aires urbaines, il faudrait dans l'absolu pouvoir traiter une maille infra-communale (comme l'IRIS). Néanmoins, quand bien même nous arriverions à atteindre un tel niveau de modélisation, l'application en condition réelle nécessiterait de traiter l'adresse des assurés. Or :

— Il faudrait appliquer notre méthodologie à une maille très fine et donc, soit disposer de suffisamment d'années d'expositions pour rendre les résultats robustes (approche grandes villes), soit disposer de nombreuses zones infra-communales (approche générale).

— Au delà de la modélisation des ratios de mortalité, il faudrait disposer de données extrêmement fines pour caractériser ces zones.

— Une fois les résultats obtenus, il faudrait potentiellement réaliser d'importants traitements pour rattacher l'adresse des assurés à une zone infra-communale.

— En tant que réassureur, nous ne disposons que rarement -à l'heure actuelle- d'un tel niveau de détails dans les données qui nous sont transmises.

On comprend donc que de nombreux paramètres viennent freiner la mise en place de ce type d'analyse.

En revanche, nous estimons que les ratios calculés pour les *Grandes* et *Très grandes villes*, sans être parfaitement adaptés, restent pertinents. En effet :

— Dans le cadre de produits emprunteurs, il possible que les ratios proposés soient trop forts. Dans ces zones denses, il doit vraisemblablement exister une proportion importante d'individus à risque accédant peu à la propriété et pénalisant donc notre estimation à tort.
Néanmoins, à défaut de plus amples informations, garder des ratios élevés reste conservateur.

— Dans le cadre de produits de prévoyance individuelle, la cible de chaque produit est déterminante. Sauf à déterminer un ajustement pour chacune d'elles, ce taux général reste un a priori cohérent.

— Dans le cadre de produits groupes, la population sous-jacente est bien plus hétérogène que celle d'un portefeuille emprunteur. Ces ratios sont donc vraisemblablement adaptés.

Dans certains cas extrêmes, il est possible que certains quartiers du centre ou de la banlieue soient effectivement assimilables à des zones péri-urbaines aisées. A l'inverse, il est aussi possible que des assurés logent dans des quartiers défavorisés. A défaut de disposer d'une maille extrêmement fine, et sans prétendre qu'ils sont parfaitement ajustés, nous estimons que ces ratios doivent pouvoir s'appliquer ; d'autant plus que la question est de savoir s'ils doivent être nivelés à la baisse.

A ce titre, nous verrons qu'appliquer un lissage à notre modélisation a tendance à niveler les ratios des villes vers le bas. Cela pourrait donc permettre, à terme, de corriger notre modélisation si une sur-pénalisation des villes venait à se manifester.

2.2 Le lissage des ratios comme piste d'ajustement

Nous présentons ici les deux lissages modérés présentés plus avant, à savoir le lissage modéré (bande passante de 121.2) et le lissage minimal (bande passante de 53.6). Nous discutons ici de la pertinence de ces deux niveaux d'homogénéisation.

Comme nous l'avons mentionné, il pourrait se révéler pertinent de diminuer les ratios de mortalité proposés dans les villes. En effet, le lissage minimal que nous proposons homogénéise les zones hétérogènes. Dans le cas d'une ville dont la périphérie est en situation de sous-mortalité, cela revient donc essentiellement (pour notre lissage minimal) à diminuer le ratio de mortalité de la ville. Nous pouvons représenter l'impact du lissage minimal sur notre modélisation. Sur la cartographie suivante, on affiche en bleu l'amélioration des ratios et en rouge leur détérioration. :

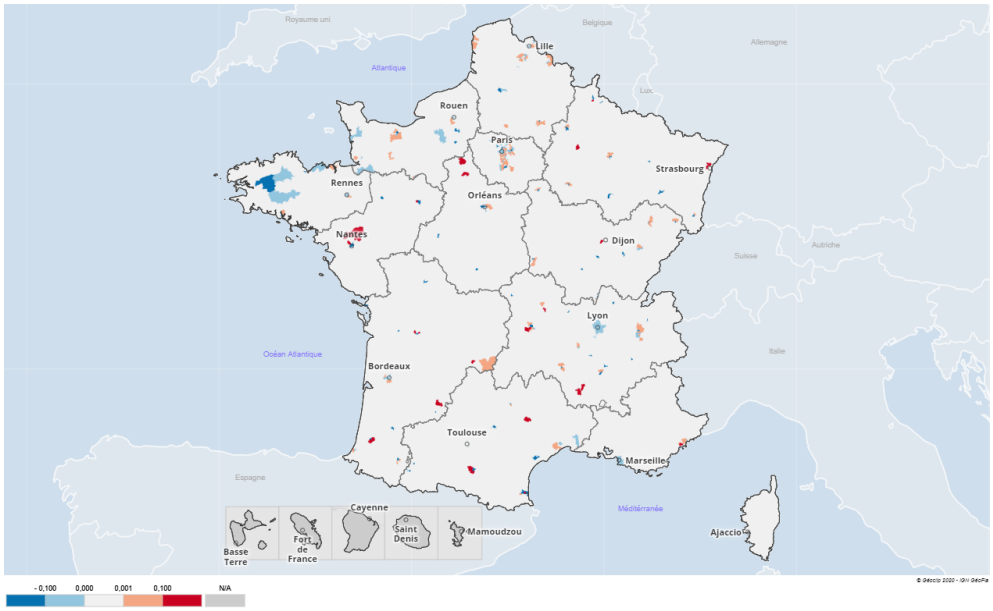


FIGURE 52 – Impact du lissage minimal sur les ratios modélisés
Obtenu à partir de www.france-decouverte.geoclip.fr - Données importées [20]

Au delà d'une diminution des ratios des villes, nous pouvons voir que certaines zones sont modifiées assez fortement. Or, il apparaît complexe de savoir si ces modifications sont le fruit d'une quelconque réalité. Si cette modélisation lissée a vocation à être exploitée, il serait envisageable de limiter son impact aux grandes villes uniquement.

L'effet du lissage minimal reste néanmoins assez circonscrit. [28]

Observons maintenant l'impact de notre lissage modéré :

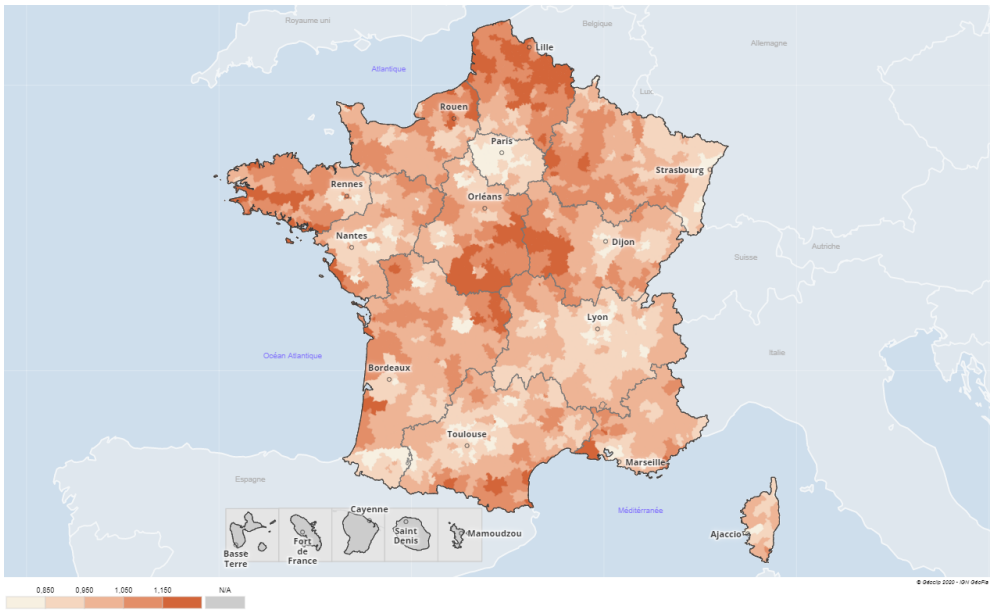


FIGURE 53 – Lissage modéré des ratios modélisés
Obtenu à partir de www.france-decouverte.geoclip.fr - Données importées [20]

28. Une représentation est disponible en annexe (figure 61). Elle est très proche de notre modélisation initiale.

Nous pouvons voir que ce lissage est bien plus impactant et a tendance à pénaliser les banlieues. Encore une fois, on remarque des modifications dont il est difficile de contrôler la pertinence. A ce titre, on peut remarquer des incohérences entre les deux niveaux de lissage ce qui, même si cet effet est plutôt minime, pose question.

Nous pouvons voir sur cette représentation que l'effet du lissage est cette fois relativement fort :

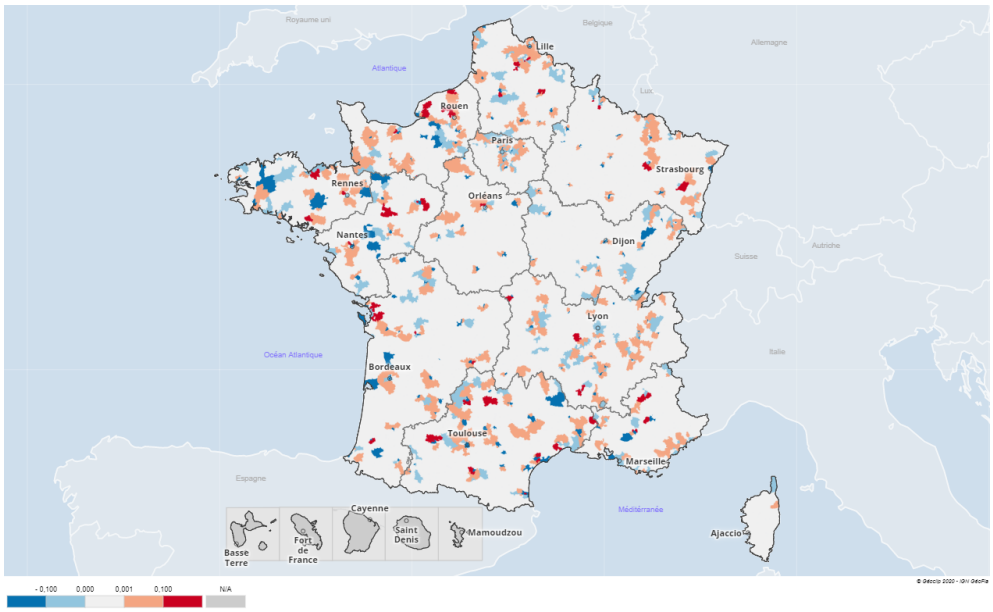


FIGURE 54 – Impact du lissage modéré sur les ratios modélisés
Obtenu à partir de www.france-decouverte.geoclip.fr - Données importées [20]

La carte proposée gagne en lisibilité. Néanmoins, en proposant des zones homogènes, ce lissage limite l'importante de la distance à la ville et gomme une part importante des effets mis en évidence ; il est donc légitime de se demander si ce type de lissage est applicable.

Bien que nous ayons insisté sur la nécessité de proposer une analyse fine, on peut imaginer que, sur des zones proches, un relatif brassage des populations existe et il n'est pas absurde de chercher à capter cet effet.

De plus, il peut sembler raisonnable de se dire que deux cantons proches devraient avoir un niveau de mortalité similaire. Est-il raisonnable qu'un canton présentant, par exemple, un ratio de 1.10 soit adjacent à un canton présentant un ratio de 0.90 ? Notre perception du risque mérite-elle d'être autant modifiée, sachant que seuls quelques kilomètres les séparent ?

Comme nous l'avons déjà mentionné, et malgré toutes les nuances que nous aimerions idéalement apporter, nous estimons que la réponse à ces questions est oui. La ségrégation spatiale des individus se faisant à des échelles très fines, lisser nos ratios pourrait conduire à perdre de l'information et, à ce titre, l'une des forces de notre analyse reste la finesse géographique proposée.

Dans l'immédiat, nous préférons donc appliquer nos ratios non lissés.

Nous verrons dans l'application que même un léger lissage a tendance à niveler la prime vers le bas et que le second lissage ne compense pas cette diminution en augmentant les ratios du péri-urbain.

Ces lissages pourront éventuellement permettre de répondre à un problème potentiel que nous avons identifié. Ils doivent donc, pour le moment, être considérés comme une piste de développement.

2.3 Population assurée et population générale

Si nos taux modélisés reflètent une réalité certaine, nous avons mentionné qu'ils pourraient trouver des limites dans le traitement des villes.

Si nous étendons les interrogations formulées, le sujet est le suivant : la population assurée étant différente de la population générale, est-il raisonnable de lui appliquer un ratio calculé sur la population nationale ?

Il faut garder à l'esprit qu'il s'agit bien de ratios et non pas de valeurs absolues. La question n'est pas de savoir si un assuré habitant en centre ville a un niveau de mortalité inférieur au niveau national. Il est probable que ce soit le cas. La question est de savoir si, relativement à une estimation sur une population assurée, son niveau de mortalité est plus élevé.

Un assuré habitant en centre ville est un assuré qui n'est pas parti vivre dans un pavillon en zone péri-urbaine. Ce qui a été mentionné précédemment est que les niveaux *relatifs* proposés sont peut-être exagérés dans le cadre des centres-villes.

Dans les zones très denses (i.e. les mégalopoles), nous avons mentionné que nous percevions :

— Des limites dans le cas de portefeuilles emprunteurs.

— Des interrogations dans le cas de produits prévoyance.

Nous avons répondu à ces interrogations en considérant les ratios actuels comme conservateurs et en proposant le lissage minimal comme piste d'ajustement.

Pour le reste, nous estimons que les constats généraux que nous tirons sont vraisemblablement cohérents avec ceux que l'on ferait pour une population d'assurés. En dehors des centres urbains, l'hétérogénéité des territoires diminue de plus drastiquement. Le fait que nous étudions des niveaux *relatifs* reste l'argument principal quant à l'applicabilité de ces taux à une population assurée, une fois l'âge, le sexe et les csp traités.

Encore une fois, nos travaux ont vocation à évoluer. Si nous sommes confiants quant à leur applicabilité et leur capacité à améliorer la vision du risque, nous ne prétendrons pas que des décalages subtils existent. A ce titre, RGA France reste une entité récente et, à mesure qu'elle acquerra de l'expérience de portefeuilles, nos travaux pourront être challengés.

En définitive, les taux dont nous disposons sont la meilleure estimation a priori que nous ayons. Ils donnent de vraies indications et permettent, ainsi, une vision du risque accrue. L'application que nous proposerons pour conclure ce mémoire donnera à ce titre un premier exemple illustratif.

2.4 Cohérence des résultats et perspectives

2.4.1 Mise à l'échelle de la région et du département

Dans le cadre d'une application réelle, il est possible que nous ne disposions pas du code postal d'un assuré, mais simplement de son département ou de sa région de résidence. C'est pourquoi nous proposons ici des ratios de mortalité par région et département. Bien que nous ayons apporté beaucoup de nuances sur la finesse de l'analyse, ce type de lissage est plus pertinent dans le sens où il tient bien compte des proportions de populations. En effet, nous posons :

$$Y_{Reg} = \frac{1}{N_{Reg}} \sum_{cv \in Reg} N_{cv} * Y_{cv}$$

Où :

— $Y_{Reg/cv}$ est le ratio de mortalité de la région/du canton.

— $N_{Reg/cv}$ est le nombre d'habitants entre 30 et 62 ans de la région/du canton Reg/cv.

En appliquant cette pondération, nous obtenons donc un ratio de mortalité par région :

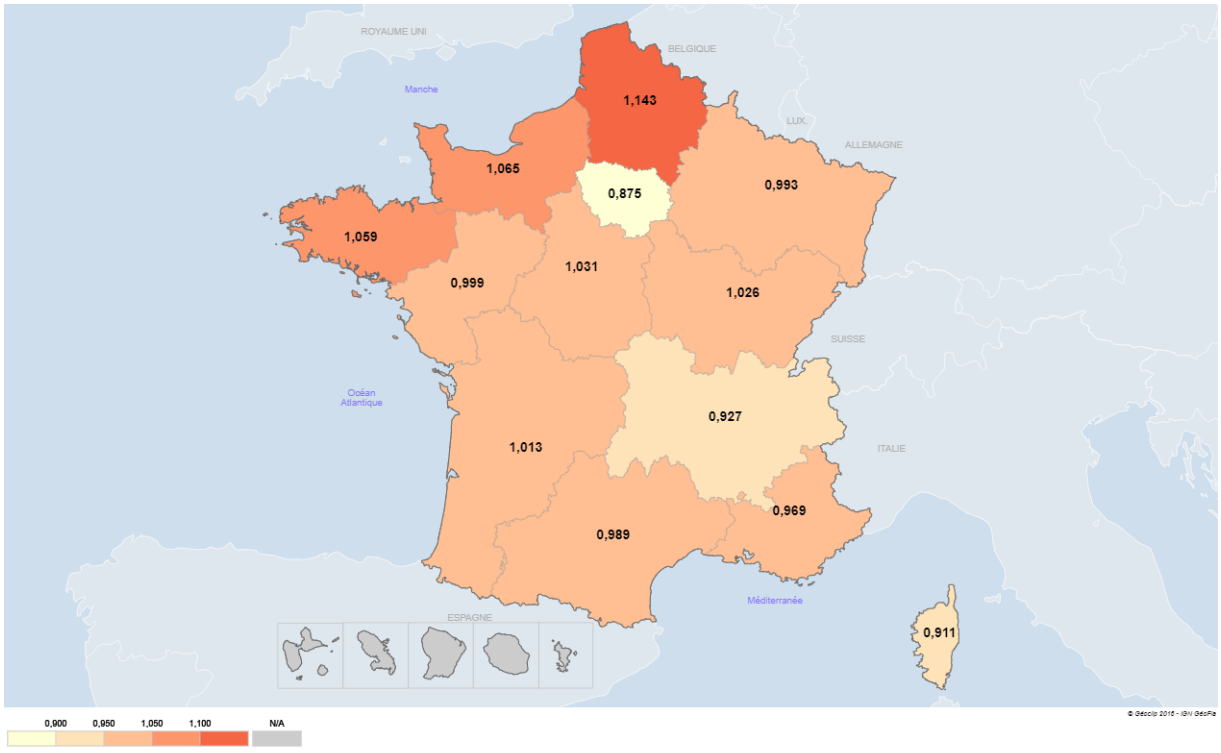


FIGURE 55 – Modélisation des ratios de mortalité régionaux

Obtenu à partir de www.france-decouverte.geoclip.fr - Données importées [20]

En comparaison avec les ratios de taux standardisés proposés au début de ce mémoire, ces taux ont l'avantage de tenir compte des proportions de csp.

Il est intéressant de revenir ici sur cette première proposition. Pour rappel, nous avions défini les ratios suivants :

$$Y_{Reg}^{Insee} = \frac{\tilde{Q_{Reg}}}{\tilde{Q_{France}}}$$

A l'issue de nos travaux, nous sommes maintenant en mesure de comparer nos ratios (indépendants de l'âge, du sexe et des csp) à ces ratios de taux standardisés (indépendants de l'âge, du sexe seulement). Cela nous permet notamment d'évaluer les écarts constatés entre ces deux approches.

Les résultats sont les suivants :

Région	Y_{Reg}	Y_{Reg}^{Insee}	$Y_{Reg} - 1$	$Y_{Reg}^{Insee} - 1$
Île-de-France	0,87	0,83	-0,13	-0,17
Centre-Val de Loire	1,03	1,06	0,03	0,06
Bourgogne-Franche-Comté	1,03	1,11	0,03	0,11
Normandie	1,07	1,17	0,07	0,17
Hauts-de-France	1,14	1,28	0,14	0,28
Grand Est	0,99	1,06	-0,01	0,06
Pays de la Loire	1,00	1,00	0,00	0,00
Bretagne	1,06	1,11	0,06	0,11
Nouvelle-Aquitaine	1,01	1,00	0,01	0,00
Occitanie	0,99	1,00	-0,01	0,00
Auvergne-Rhône-Alpes	0,93	0,89	-0,07	-0,11
Provence-Alpes-Côte d'Azur	0,97	1,00	-0,03	0,00
Corse	0,91	0,89	-0,09	-0,11

TABLE 11 – Étude des ratios de mortalités régionaux

Il est intéressant d'observer la cohérence de nos résultats avec les coefficients proposés au début de ce mémoire.

Tout d'abord, cela conforte la confiance que nous avons dans nos résultats. De plus, deux éléments déjà mentionnés se retrouvent ici confirmés :

— Il y a une corrélation extrêmement forte entre les effets de déviations de la mortalité modélisés et ceux dus à la répartition des classes socioprofessionnelles. Ainsi, les dynamiques observées se maintiennent malgré le traitement des csp.

— Nous avons réussi à retraiter l'effet des classes socioprofessionnelles que portent les ratios de taux standardisés. Nos ratios peuvent donc être appliqués conjointement à un traitement des csp dans le cadre d'une tarification.

L'ensemble des conclusions tirées ici restent valides à l'échelle départementale. Les résultats obtenus à cette échelle sont proposés en annexe (figure 62).

2.4.2 Pistes de développements

Si les travaux présentés dans ce mémoire nous ont permis d'analyser la structure du territoire métropolitain, ils ont également ouvert un certain nombre de perspectives. L'étude proposée ici a vocation à être poursuivie en dehors du cadre de ce mémoire.

Parmi les principales pistes de développement envisagées, nous pouvons évoquer :

— L'actualisation des données.
Lorsque nous disposerons des recensements 2018 et 2019 ainsi que des taux de mortalité correspondants, il sera possible de réaliser une nouvelle estimation. Dès lors, le redressement des estimations ne sera plus nécessaire. Nous ne nous attendons pas à remettre en question nos conclusions. Néanmoins, il est possible que cela ajuste légèrement le niveau des ratios de mortalité modélisés. A ce titre, la mise à jour des coefficients par csp restera probablement l'effet le plus fort.

— L'application éventuelle des techniques de lissage pour moduler l'impact des grandes villes selon le type de portefeuille.

— L'exploitation de la base de données construite.
Il est entendu que les travaux réalisés ici ne sauraient rendre compte de tous les facteurs expliquant les différences territoriales de mortalité. De ce fait, la base de donnée que nous avons construite recèle encore d'effets à mettre en évidence. A l'avenir, nous pourrons également alimenter cette base avec de nouvelles données exogènes (comme la pollution de l'air ou les accidents industriels chimiques par exemple). Cela permettra de raffiner encore notre modèle et d'étendre nos travaux. Nous serons de plus en mesure de proposer de nouvelles études sur, par exemple, d'autres facteurs explicatifs que ceux mis en évidence dans ce mémoire.

— Étudier l'impact de la Covid-19 de façon géographique.
Maintenant que nous disposons d'une première modélisation, il sera plus simple d'étudier les zones où la mortalité a dévié. D'autant plus que les données proposées par l'Insee seront au même format que celles déjà exploitées.
Pour nuancer nos attentes, il faut mentionner que le confinement national et les changements de comportement ont modifié les niveaux de mortalité (baisse des accidents de la route par exemple), ce qui complexifiera l'analyse [42]. Dès lors que les données de décès 2020 seront complétées par l'Insee, cette étude pourra être envisagée.

— Étendre nos travaux aux territoires d'outre-mer.
Plusieurs raisons nous ont poussé à cantonner notre étude à la France métropolitaine. Si nous voulons étendre notre étude aux territoires d'outre-mer, il nous faudra notamment étudier dans quelles mesures nous disposons de suffisament de données pour comprendre leurs spécificités. Dans le cadre de ce mémoire, nous avons d'ores et déjà pu construire des ratios de mortalité pour ces départements/régions. A défaut de disposer de bases de mortalités spécifiques, il s'agit d'une piste prometteuse pour ajuster nos estimations de mortalité sur ces territoires.

— Proposer un modèle Homme et un modèle Femme.
Les femmes représentent seulement un tiers (34,6%) des décès que nous observons chez les 30-62 ans. Or, nous disposons souvent de données différenciées par sexe. Ce type de modèle pourrait ainsi être simple à mettre en place. Au delà d'affiner l'ajustement tarifaire, il serait intéressant de savoir si les facteurs explicatifs sont les mêmes. Nous pourrions de plus comparer les ratios obtenus et, éventuellement, identifier des différences géographiques.

3 Applications illustratives

Nous proposons ici d'appliquer nos ratios à deux portefeuilles réels (un portefeuille de prévoyance individuelle et un portefeuille emprunteur).

L'idée est ici de réaliser une tarification classique, puis de venir comparer la prime pure estimée avec celle calculée en introduisant la pondération par lieu de domicile.

En réalisant ces estimations, nous serons capables de mesurer l'impact des ratios modélisés dans ce mémoire sur une tarification. Nous discuterons également des résultats obtenus.

3.1 Mise en place de l'analyse

On définit l'impact de la répartition géographique des assurés sur le portefeuille par :

$$Y_{global} = \frac{\tilde{P}}{P}$$

Où :

— Y_{global} est l'impact du lieu de domicile des assurés sur notre estimation de prime pure.

— P est la prime pure du portefeuille.

— $\tilde{P}$ est la prime pure du portefeuille où la pondération par lieu de domicile a été introduite.

Le ratio des primes pures calculées ici peuvent également être vu comme le *ratio géographique moyen* du portefeuille ; c'est à dire une moyenne pondérée de nos ratios (la pondération étant, en l'occurrence, la prime pure de chaque assuré).

$$\frac{\tilde{P}}{P} = \frac{1}{P} \sum_{i \in ptf} Y_i p_i$$

Où :

— p_i est la prime pure de l'individu i.

— Y_i est le ratio pondérant le risque de mortalité de l'individu i en fonction de son lieu de domicile.

Pour les deux portefeuilles étudiés, nous disposons du code postal des assurés.

La majorité du temps, les codes postaux sont une maille plus fine que les cantons. Il suffit alors de considérer, pour chaque code postal, le ratio du canton en question. Lorsque plusieurs cantons recouvrent un même code postal, nous réalisons une pondération des ratios en fonction du nombre d'habitants (nous disposons du nombre d'habitants sur chaque intersection ; par exemple, pour 10 000 habitants pour le canton de ratio Y_a et 90 000 habitants pour le canton de ratio Y_b, nous estimons un ratio $Y_c = Y_a/10 + 9Y_b/10$).

Pour appliquer nos ratios à une table de mortalité sans introduire de biais, il est important d'évaluer la distribution géographique de la population de référence (i.e. celle utilisée pour construire la table de mortalité). De cette manière, on ne comptabilisera pas deux fois les mêmes effets de concentration géographique.

Concernant la référence RGA utilisée dans cette partie illustrative, nous savons que la distribution ayant permis de la construire suit bien la densité de la population nationale. Les impacts que nous calculerons seront donc de ce fait directement interprétables.

3.2 Impact de la répartition géographique des lieux de domicile sur le niveau de mortalité d'un portefeuille de prévoyance individuelle

Étudions dans un premier temps un portefeuille de prévoyance individuelle. Il s'agit d'un produit très largement distribué, avec près de 175 000 individus en portefeuille sur l'année considérée. Nous sommes en mesure d'appliquer nos ratios de mortalité à l'essentiel de ces assurés, avec environ 9 500 codes postaux impossibles à évaluer (domiciliés à l'étranger ou en outre-mer), soit 5.5% des assurés, ce qui est satisfaisant.

Nous pouvons, dans un premier temps, observer la répartition de nos assurés sur le territoire national :

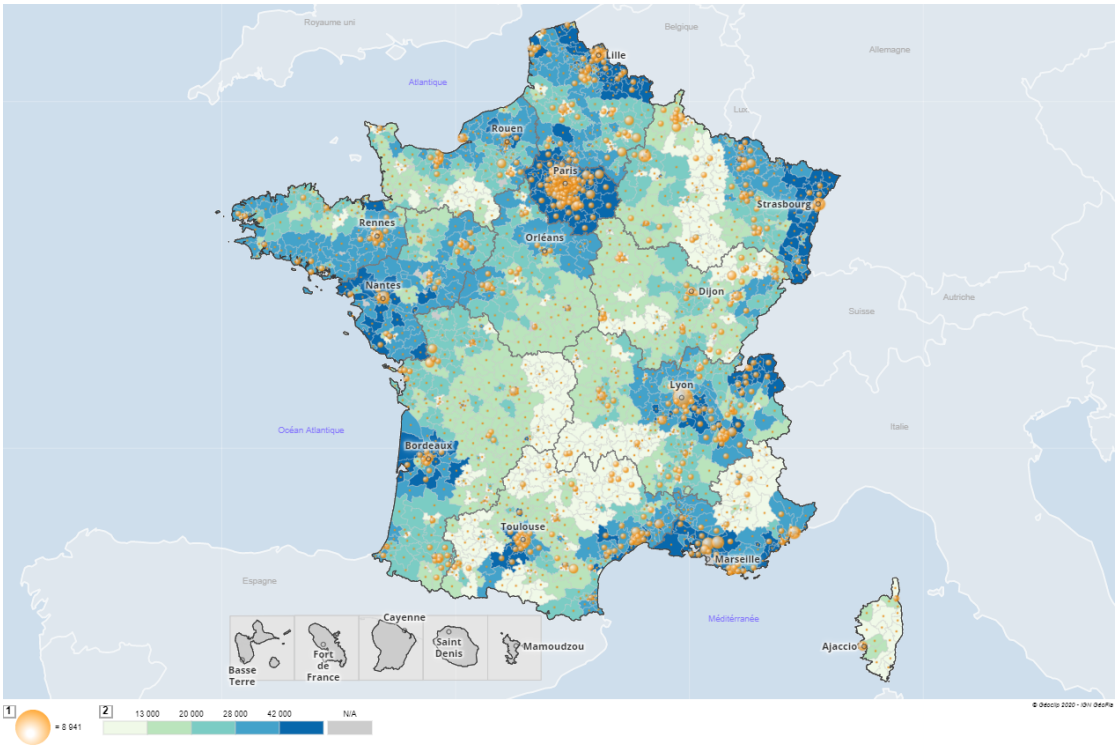


FIGURE 56 – Répartition des assurés - Portefeuille de prévoyance individuelle
Obtenu à partir de www.france-decouverte.geoclip.fr - Données importées et Insee Populations légales (2018) [20]

Sur cette représentation, nous avons coloré les territoires en fonction du nombre d'habitants en population générale. Nous pouvons ainsi voir que ce portefeuille a été, non seulement, largement distribué, mais que sa distribution est nationale et suit de très près la population française. Nous nous attendons donc à ce que le *ratio géographique moyen* de ce portefeuille soit proche de 1.

Dans la base de données utilisée, nous ne disposons pas de la profession de l'individu, mais simplement de l'âge, du sexe et du code postal.

On réalise donc ici l'estimation suivante :

$$P = \sum CA \cdot q_{a,s}^{RGA} \cdot Y_{cp}$$

Où :

— P est la prime pure estimée.

— CA est le capital assuré.

— $q_{(a,s)}^{RGA}$ taux de mortalité pour les personnes d'âge a et de sexe s tiré de la table de mortalité RGA.

— Y_{cp} ratio modélisé pour le code postal *cp*. Dans le cas d'une prime non ajustée par code postaux, ce ratio vaut toujours 1.

Nous obtenons ainsi les résultats suivants :

Prime pure				
Non Ajustée *(i.e. $Y_{cp} = 1$)*	Ajustée	Ajustée *(Lissage minimal)*	Ajustée *(Lissage modéré)*	Ajustée *(Lissage maximal)*
6 626 370	6 517 922	6 410 306	6 284 604	6 279 804

Les chiffres présentés ici ont été modifiés linéairement.

TABLE 12 – Prime pure d'un portefeuille de prévoyance selon l'ajustement géographique appliqué

Impact des ratios			
Ajustée	Ajustée *(Lissage minimal)*	Ajustée *(Lissage modéré)*	Ajustée *(Lissage maximal)*
-1,647%	-3,261%	-5,168%	-5,230%

TABLE 13 – Impact de la répartition géographique des assurés d'un portefeuille de prévoyance

Malgré la ressemblance constatée entre la répartition géographique des assurés et celle de la population française, nous observerons une légère diminution de la prime. Il est possible que, même pour des produits très bien distribués, on observe une concentration des assurés dans les zones en situation de sous-mortalité.

Nous notons également que le lissage spatial a tendance à niveler le niveau de prime vers le bas. Nous verrons que cet effet est également présent lorsque nous tariferons le second portefeuille. Cela confirme que ces taux lissés ne doivent pas être utilisés en l'état. Nous nous serions idéalement attendus à observer une stabilité relative, où à minima des effets légers et/ou des effets différents d'un portefeuille à l'autre. Cela confirme que ces ratios lissés ont tendance à limiter la variance mais à introduire un biais.

3.3 Impact de la répartition géographique des lieux de domicile sur le niveau de mortalité d'un portefeuille emprunteur

Étudions maintenant un portefeuille emprunteur. Ce produit a été distribué de façon plus restreinte, avec environ 17 000 individus sur l'année considérée. Comme pour le précèdent portefeuille, nous sommes en mesure d'appliquer nos ratios de mortalité à l'essentiel de ces assurés, avec environ 500 codes postaux impossibles à évaluer (domiciliés à étrangers ou en outre-mer), soit 2.9% des assurés environ, ce qui est très satisfaisant.

Nous pouvons également observer la répartition de nos assurés sur le territoire national :

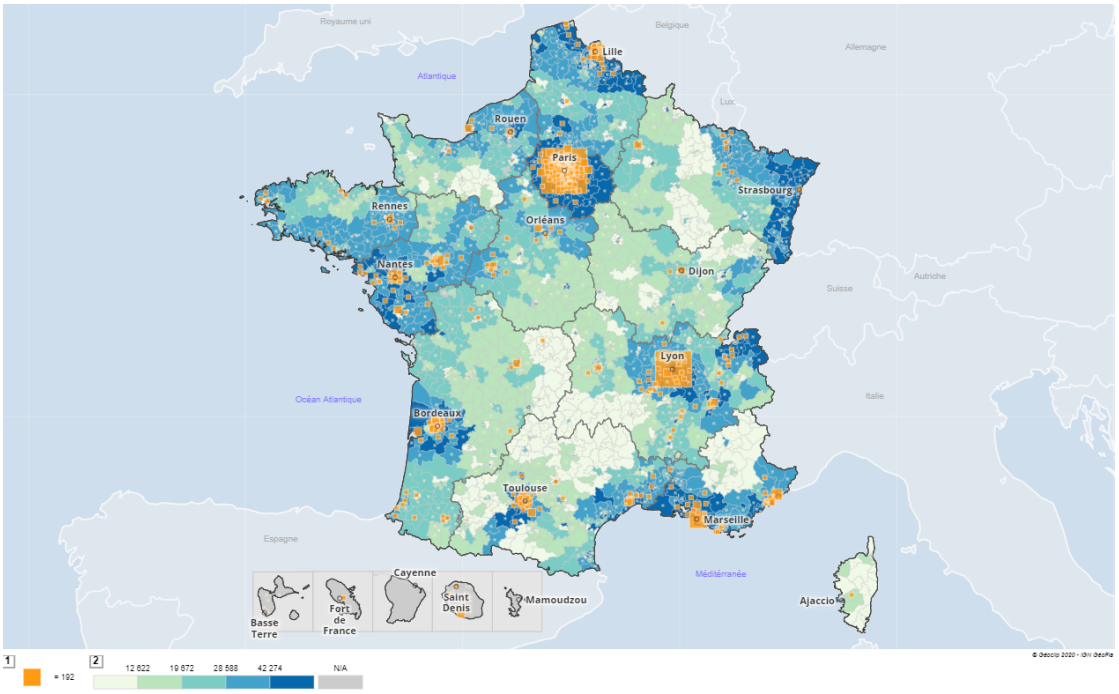


FIGURE 57 – Répartition des assurés - Portefeuille emprunteur

Obtenu à partir de www.france-decouverte.geoclip.fr - Données importées et Insee Populations légales (2018) [20]

On observe ici que les assurés sont moins bien repartis que notre précédent portefeuille. On note une forte concentration des assurés à Paris, Lyon et, dans une moindre mesure, une sur-représentation des villes du sud. Il faut aussi noter que la "diagonale des faibles densités" est presque totalement absente de cette représentation, tout comme l'ouest de la Bretagne et la Normandie. Nous nous attendons donc à ce que le *ratio géographique moyen* de ce portefeuille soit inférieur au précédent.

Nous disposons ici de la catégorisation par classe de risque utilisée dans les tarifications RGA. Nous disposons également de libellés décrivant la profession de l'individu et permettant de nous ramener à une catégorisation par csp.

Nous allons donc réaliser deux estimations plutôt qu'une. A savoir, une tarification classique et une tarification dans les conditions du mémoire. Nous serons alors en mesure d'évaluer la sensibilité du *ratio géographique moyen* à la méthode de tarification utilisées.

On réalise ici les estimations suivantes :

$$P_1 = \sum CRD \cdot q_{a,s}^{RGA} \cdot C_c \cdot Y_{cp}$$

Où :

— P_1 est la prime pure estimée.

— CRD est le capital restant dû de l'individu.

— $q_{(a,s)}^{RGA}$ taux de mortalité pour les personnes d'âge a et de sexe s tiré de la table de mortalité rga.

— C_c le coefficient de la classe de risque c [29].

— Y_{cp} ratio modélisé pour le code postal cp. Dans le cas d'une prime non ajustée par codes postaux, ce ratio vaut toujours 1.

29. Une classe de risque est un regroupement de professions auxquelles on attribue une même prime d'assurance de façon à satisfaire la cible commerciale du produit. L'hypothèse de risque de mortalité est travaillée en parallèle pour tenir compte des diverses professions constituant la classe. Ici, C_c fait référence à l'hypothèse de risque agrégée sur la classe c.

Et on a également :

$$P_2 = \sum CRD \cdot q_{a,s} \cdot C_{csp,a,s} \cdot Y_{cp}$$

Où :

— $q_{(a,s)}$ est tiré de la table de mortalité nationale.

— $C_{csp,a,s}$ est le coefficient pondérant le taux de mortalité $q_{a,s}$ selon a, s et la classe socioprofessionnelle csp.

Il est implicite que les sommes ci-dessus s'appliquent à l'ensemble du portefeuille et que les formules proposées ici sont celles appliquées à chaque assuré.

Une fois ces estimations réalisées nous obtenons les résultats suivants :

Prime pure P_1 - Taux de mortalité RGA				
Non Ajustée *(i.e. $Y_{cp} = 1$)*	Ajustée	Ajustée *(Lissage minimal)*	Ajustée *(Lissage modéré)*	Ajustée *(Lissage maximal)*
253 110	237 149	235 083	230 396	229 540

Les chiffres présentés ici ont été modifiés linéairement.

TABLE 14 – Prime pure d'un portefeuille emprunteur selon l'ajustement géographique appliqué - P_1

Impact des ratios géographiques - Taux de mortalité RGA			
Ajustée	Ajustée *(Lissage minimal)*	Ajustée *(Lissage modéré)*	Ajustée *(Lissage maximal)*
-6,306%	-7,122%	-8,974%	-9,312%

TABLE 15 – Impact de la répartition géographique des assurés d'un portefeuille emprunteur - P_1

Prime pure P_2 - Taux de mortalité nationaux				
Non Ajustée *(i.e. $Y_{cp} = 1$)*	Ajustée	Ajustée *(Lissage minimal)*	Ajustée *(Lissage modéré)*	Ajustée *(Lissage maximal)*
249 894	235 476	233 954	230 318	229 454

Les chiffres présentés ici ont été modifiés linéairement.

TABLE 16 – Prime pure d'un portefeuille emprunteur selon l'ajustement géographique appliqué - P_2

Impact des ratios géographiques - Taux de mortalité nationaux			
Ajustée	Ajustée *(Lissage minimal)*	Ajustée *(Lissage modéré)*	Ajustée *(Lissage maximal)*
-6,557%	-7,249%	-8,903%	-9,295%

TABLE 17 – Impact de la répartition géographique des assurés d'un portefeuille emprunteur - P_2

On constate une forte proximité entre les ajustements géographiques obtenus par l'une et l'autre des tarifications (c'est à dire -6,306% pour la tarification RGA et -6,557% pour la tarifications dans les conditions du mémoire). Cela signifie que notre mesure est moins impactée par les nuances entre les méthodes que par les caractéristiques du portefeuille, c'est à dire l'âge, le sexe, la csp et le capital assuré des individus.

Ici, les effets que nous observons sont beaucoup plus forts que ceux calculés sur le portefeuille précédent. Ce produit bénéficie visiblement d'une distribution géographique avantageuse.

Nous sommes finalement en mesure de tenir compte de la distribution géographique des produits dans nos tarifications. Le potentiel des ratios modélisés pour le calibrage tarifaire se voit particulièrement bien sur le second portefeuille, avec une diminution de la prime pure de plus de 6%.

Les assurés de ces portefeuilles sont donc répartis dans des zones où la population présente globalement un risque de mortalité inférieur à la moyenne. Si la prime pure des produits étudiés ici peut de ce fait être revue à la baisse, nos ratios pourront également permettre de mieux calibrer le risque sur des populations sujettes à une sur-mortalité. Il nous faudra recueillir plus d'expérience pour estimer à quel point les effets géographiques sont différents d'un portefeuille à l'autre.

Les résultats que nous obtenons sont ainsi très encourageants.

La présence d'un paramètre géographique dans les données clients (que ce soit le département, la région ou le code postal) reste néanmoins une nécessité. Ainsi :

— Lorsque les données de portefeuilles le permettront, l'impact géographique sera mesuré de façon systématique dans les études de portefeuille et des comparaisons entre différentes typologies de produits pourront être menées.
Nous enrichissons donc par nos travaux les outils d'analyse de risque de l'entreprise.

— Lorsque nous disposerons de données de sinistres décès suffisantes, nous serons également en mesure de challenger nos travaux plus en profondeur en confrontant nos estimations à une sinistralité réelle.

Les travaux proposés ici apparaissent en somme comme un atout conséquent dans un marché concurrentiel et offrent de nouvelles perspectives d'analyse.

Conclusion

Dans ce mémoire, nous avons approché le risque de décès par l'angle de la géographie.

Pour ce faire, nous avons dans un premier temps construit des ratios de mortalité sur tout le territoire métropolitain et à la maille la plus fine possible. Les deux années d'exposition dont nous disposions se sont néanmoins révélées insuffisantes pour exploiter immédiatement ces résultats. Après avoir réuni et mis en forme un nombre conséquent de données explicatives, nous avons mobilisé plusieurs techniques numériques et statistiques (ACP, Clustering et GLM notamment). Nous avons alors montré qu'il était possible de traiter cette variabilité.

Notre méthodologie a été calibrée pour rendre nos résultats applicables dans les tarifications. Les coefficients que nous avons modélisés sont ainsi le reflet des écarts de mortalité que nous pouvons constater dans la populations française, une fois l'âge, le sexe et la catégorie socioprofessionnelle des individus traités. Nous avons alors pu confirmer que ces trois paramètres -utilisés habituellement dans les tarifications- ne suffisaient pas à capter toutes les différences de mortalité sur le territoire métropolitain.

Nos travaux ont ainsi été l'occasion de tirer un certain nombre de constats sur les facteurs déterminant les écarts observés. Notre modélisation, une fois interprétée, nous a conduit à mettre en évidence plusieurs facteurs explicatifs comme le salaire, la taille moyenne des ménages et la mobilité des actifs. Les informations de ce type -que les assureurs se refusent généralement à demander- se sont ainsi révélées exploitables au travers de la géographie. Nous avons pu observer le rôle majeur des effets socio-économiques sur la mortalité, effets qui ne sauraient se résumer aux seules catégories socioprofessionnelles. Les effets de concentration comme la péri-urbanisation se sont avérés révélateurs du niveau de vie des individus. De fait, nous avons mis en évidence le lien entre la mortalité et la structure des espaces urbains.

Les travaux proposés ici ont ouvert de nombreuses pistes de développement. Nous pouvons notamment citer la construction de ratios différenciés par sexe ou une étude d'impact de la Covid-19 sur la mortalité. Nous pourrons également mobiliser les techniques de lissage que nous avons présentées si nécessaire. A l'avenir, la base de données construite pourra être exploitée et alimentée pour raffiner et étendre nos travaux. Nous serons de plus en mesure de proposer de nouvelles études.

Les ratios modélisés ici entendent s'intégrer aux outils actuariels de l'entreprise de réassurance où ce mémoire a été réalisé.

Ils trouveront notamment leur application dans des études de portefeuille. La notion de *ratio géographique moyen* permettra de comparer la distribution des portefeuilles entre eux. Elle permettra également un suivi de la distribution des produits dans le temps. Identifier une dégradation ou une amélioration de ce ratio pourrait ainsi permettre d'anticiper des déviations de sinistralité. Les cartes que nous avons proposées et interprétées donnent quant à elles des a priori utiles aux décisions commerciales. Les travaux proposés ici sont en somme un atout conséquent en terme de pilotage du risque.

Enfin, l'objectif premier de nos résultats reste leur application dans les tarifications. Nous avons pu observer l'impact des coefficients construits sur un portefeuille de prévoyance individuelle et un portefeuille emprunteur. Les résultats que nous obtenons sont très encourageants et permettront, à l'avenir, d'intégrer la distribution des portefeuille d'assurance dans notre estimation du risque. Pouvoir tenir compte du lieu de domicile des assurés est, en définitive, un atout technique et commercial dans un marché concurrentiel.

Références

[1] Akaike, H. "A new look at the statistical model identification". Dans : *IEEE Transactions on Automatic Control* 19.6 (1974), p. 716-723. doi : 10.1109/TAC.1974.1100705.

[2] Bailly, P. et Carrère, C. *Statistiques descriptives : Théorie et applications*. PUG, 2015, p. 165-167.

[3] Breusch, T. S. et Pagan, A. R. "A Simple Test for Heteroskedasticity and Random Coefficient Variation". Dans : *Econometrica* 47.5 (1979), p. 1287-1294. doi : 10.2307/1911963.

[4] BRUNSDON, C. et al. *Geographically weighted summary statistics : a framework for localised exploratory data analysis*. Computers, Environment et Urban Systems, 2002.

[5] Calinski, T. et Harabasz, J. "A dendrite method for cluster analysis". Dans : *Communications in Statistics* C.3 (1974), p. 1-27.

[6] Chauvel C. "Processus empiriques pour l'inférence dans le modèle de survie à risques non proportionnels". Dans : *Mathématiques générales* (2014), p. 22-26.

[7] Cook, R. D. "Detection of Influential Observations in Linear Regression". Dans : *Technometrics* 19.1 (1977), p. 15-18. doi : 10.2307/1268249.

[8] Cour de justice de l'union européenne. *Primes et des prestations unisexes pour contrats d'assurance*.
https://curia.europa.eu/jcms/upload/docs/application/pdf/2011-03/cp110012fr.pdf. 2011.

[9] Cox, D. R. "Regression models and life–tables (with discussion)". Dans : Series B 34.2 (1972), p. 187-220.

[10] Data.gouv. *Communes Geolocalisees*.
https://www.data.gouv.fr/en/datasets/communes-geolocalisees/.

[11] DREES. *L'état de santé de la population en France*.
https://drees.solidarites-sante.gouv.fr/publications-documents-de-reference/rapports/letat-de-sante-de-la-population-en-france-rapport-2017. 2017.

[12] DREES. *Données*.
http://www.data.drees.sante.gouv.fr/.

[13] Dutang, C. "Some explanations about the IWLS algorithm to fit generalized linear models". Dans : *hal* (2017). doi : 01577698.

[14] EIDER. *Données*.
http://www.stats.environnement.developpement-durable.gouv.fr/Eider/.

[15] Fédération Françaises de l'Assurance. *Les contrats de l'assurance en cas de décès*.
https://www.ffa-assurance.fr/infos-assures/les-contrats-assurance-en-cas-de-deces. 2015.

[16] Fédération Françaises de l'Assurance. *Les contrats d'assurance emprunteur*.
https://www.ffa-assurance.fr/etudes-et-chiffres-cles/les-contrats-assurance-emprunteur-en-2017. 2017.

[17] Fédération Francaises de l'Assurance. *Progression du marché de la prévoyance*.
https://www.ffa-assurance.fr/etudes-et-chiffres-cles/le-marche-de-la-sante-et-de-la-prevoyance-progresse-de-28-en-2018. 2018.

[18] Fox, J et Monette, G. "Generalized Collinearity Diagnostics". Dans : *Journal of the American Statistical Association* 87.417 (1992), p. 178-183. doi : 10.1080/01621459.1992.10475190.

[19] Gareth, J. et al. "An Introduction to Statistical Learning". Dans : *Springer Science+Business Media New York* (2017).

[20] Geoclip. *Carte interactive.* https://france-decouverte.geoclip.fr/.

[21] Hartigan, J. A. et Wong, M. A. "Algorithm AS 136 : A k-Means Clustering Algorithm". Dans : *Journal of the Royal Statistical Society* C.28 (1) (1979), p. 100-108.

[22] INSEE. *Emploi et revenus des indépendants.* https://www.insee.fr/fr/statistiques/4470890. 2020.

[23] INSEE. *Analyse de la mortalité par csp - Méthode et principaux résultats.* https://www.insee.fr/fr/statistiques/2022138.

[24] INSEE. *Base de comparateur de territoire.* https://www.insee.fr/fr/statistiques/2521169.

[25] INSEE. *Caractéristiques de l'emploi.* https://www.insee.fr/fr/statistiques/4171449.

[26] INSEE. *Caractéristiques des entreprises.* https://www.insee.fr/fr/statistiques/2021289.

[27] INSEE. *Carte des taux standardisée de mortalité par département.* https://www.insee.fr/fr/statistiques/2012741.

[28] INSEE. *Conseils d'utilisation du recensement.* https://www.insee.fr/fr/information/2383177.

[29] INSEE. *Décès quotidiens.* https://www.insee.fr/fr/statistiques/4487988?sommaire=4487854.

[30] INSEE. *Définition - Echantillon Démographique Permanent.* https://www.insee.fr/fr/metadonnees/source/serie/s1166.

[31] INSEE. *Définition : Canton-ou-ville.* https://www.insee.fr/fr/metadonnees/definition/c1725.

[32] INSEE. *Définition : Commune.* https://www.insee.fr/fr/metadonnees/definition/c1468.

[33] INSEE. *Définitions.* https://www.insee.fr/fr/metadonnees/definitions.

[34] INSEE. *Diplômes et formation.* https://www.insee.fr/fr/statistiques/4171395.

[35] INSEE. *Document méthodologique - Indicateurs démographiques.* https://www.insee.fr/fr/metadonnees/source/indicateur/p1667/documentation-methodologique.

[36] INSEE. *Documentation sur les fichiers de détail du recensement.* https://www.insee.fr/fr/information/2383306.

[37] INSEE. *Fichiers des personnes décédées depuis 1970.* https://www.insee.fr/fr/information/4190491.

[38] INSEE. *Logement.* https://www.insee.fr/fr/statistiques/4171415?sommaire=4171436.

[39] INSEE. *Mobilités des diplômés du supérieur.* https://www.insee.fr/fr/statistiques/1288054.

[40] INSEE. *Nathalie Blanpain - Liens entre espérance de vie et niveau de vie.* https://www.insee.fr/fr/statistiques/3319895.

[41] INSEE. *Population active.* https://www.insee.fr/fr/statistiques/4171446.

[42] INSEE. *Première estimation provisoire du nombre de décès en 2020.* https://www.insee.fr/fr/information/5013803.

[43] INSEE. *Présentation.* https://www.insee.fr/fr/information/1302230.

[44] INSEE. *Présentation et documentation complémentaires sur le recensement.* https://www.insee.fr/fr/information/2383410.

[45] INSEE. *Professions et catégories socioprofessionnelles.* https://www.insee.fr/fr/metadonnees/pcs2003/categorieSocioprofessionnelleAgregee/1?champRecherche=true.

[46] INSEE. *Résultats recensement de la population.* https://www.insee.fr/fr/information/2008354.

[47] INSEE. *Revenus et pauvreté.* https://www.insee.fr/fr/statistiques/4190006.

[48] INSEE. *Salaire horaire net moyen.* https://www.insee.fr/fr/statistiques/2021266.

[49] INSEE. *Structure de la population.* https://www.insee.fr/fr/statistiques/4171334?sommaire=4171351.

[50] INSEE. *Table d'appartenance géographique des communes et tables de passage.* https://www.insee.fr/fr/information/2028028.

[51] INSEE. *Tableaux de séries longues des taux de mortalité Francais.* https://www.insee.fr/fr/statistiques/4503155?sommaire=4503178.

[52] INSEE. *Tables de mortalité par csp et diplôme.* https://www.insee.fr/fr/statistiques/1893092?sommaire=1893101.

[53] INSEE. *Une pauvreté très présente dans les villes-centres des grands pôles urbains.* https://www.insee.fr/fr/statistiques/1283639.

[54] INSEE. *Une personne sur deux meurt dans son département de naissance.* https://www.insee.fr/fr/statistiques/4204068.

[55] Jolliffe, I. T. *Principal Component Analysis. Springer Series in Statistics.* 2002. ISBN : 978-0-387-95442-4. DOI : 10.1007/b98835.

[56] L'Argus de l'assurance. *Classement 2020 de l'assurance emprunteur.* https://www.argusdelassurance.com/classements/classement-2020-de-l-assurance-emprunteur.170069. 2020.

[57] Ministère de l'éducation nationale. *Méthodologie des indicateurs de résultats des lycées.* https://www.education.gouv.fr/methodologie-des-indicateurs-de-resultats-des-lycees-11948. 2021.

[58] ODICER. *Données.* https://odicer.ofdt.fr/mobile.php.

[59] Pearson, K. "On Lines and Planes of Closest Fit to Systems of Points in Space". Dans : *Philosophical Magazine* 2.6 (1901), p. 559-572.

[60] Rencher, A. C. et Schaalje, G. B. *Linears models in statistics.* John Wiley Sons, 2008.

[61] Saporta G. *Probabilités, analyses des données et statistique*. Technip - 2 ème édition, 2006.

[62] Schwarz, G. E. "Estimating the dimension of a model". Dans : *Annals of Statistics* 6.2 (1978), p. 461-464. DOI : 10.1214/aos/1176344136.

[63] Shapiro, S. S. ; Wilk, M. B. "An analysis of variance test for normality (complete samples)". Dans : *Biometrika* 52.3-4 (1965), p. 591-611. DOI : 10.1093/biomet/52.3-4.591.

[64] Union européenne. *Directive Européenne sur la distribution d'assurances*. https://eur-lex.europa.eu/legal-content/fr/TXT/?uri=CELEX%3A32016L0097. 2016.

Table des figures

Liste des tableaux

Annexes

Représentations complémentaires

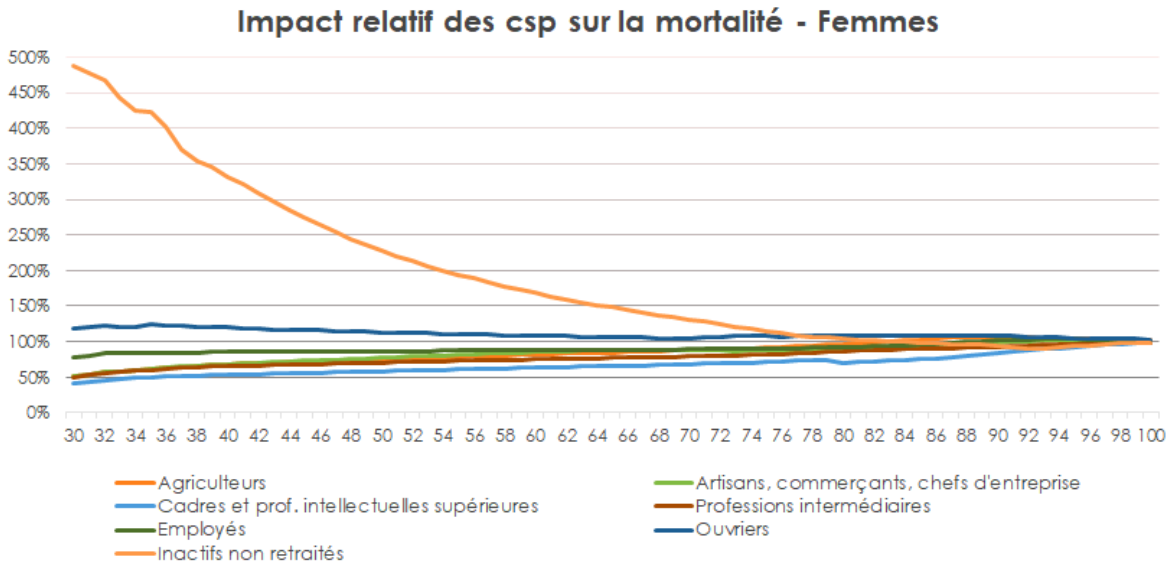


FIGURE 58 – Coefficients appliqués par csp - Femmes

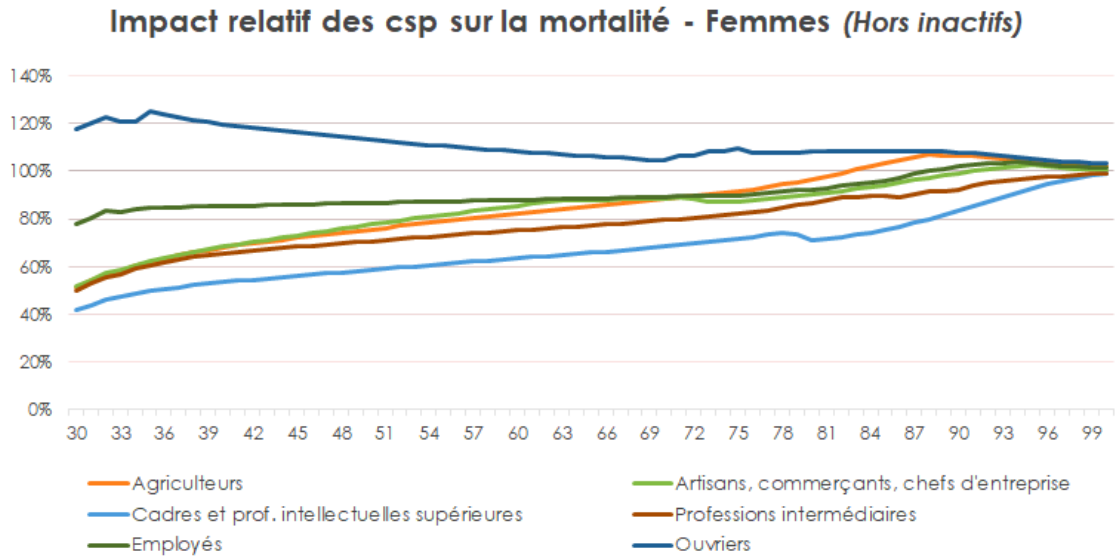


FIGURE 59 – Coefficients appliqués par csp - Femmes *(Hors inactifs)*

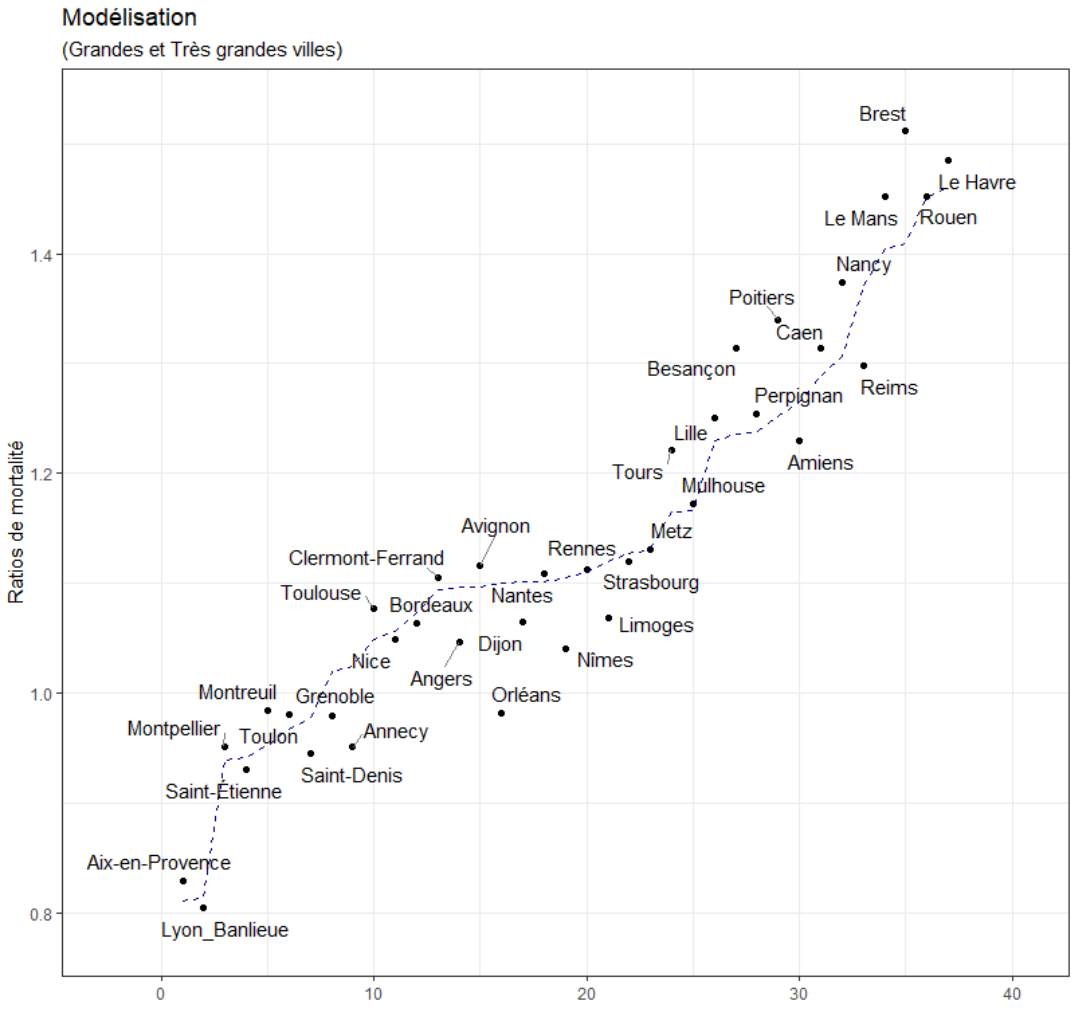


FIGURE 60 – Modèle métropoles *(après pondération)*

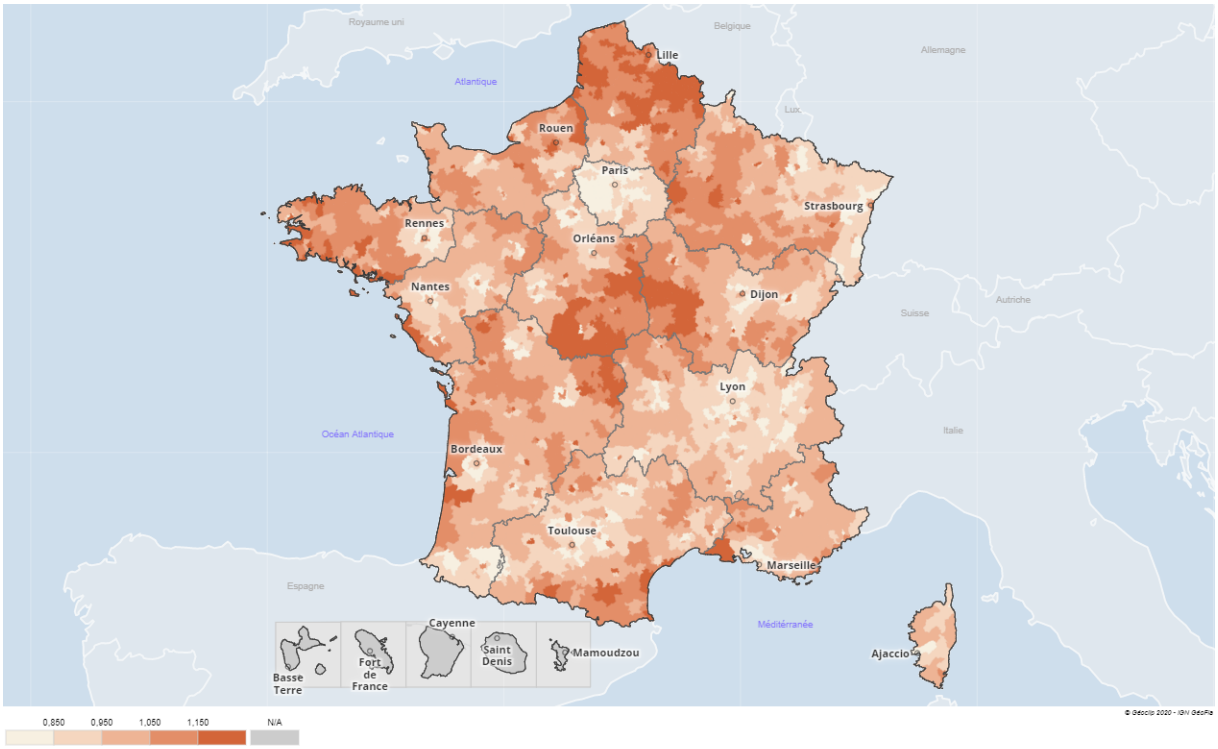


FIGURE 61 – Lissage minimal des ratios modélisés

Obtenu à partir de www.france-decouverte.geoclip.fr - Données importées [20]

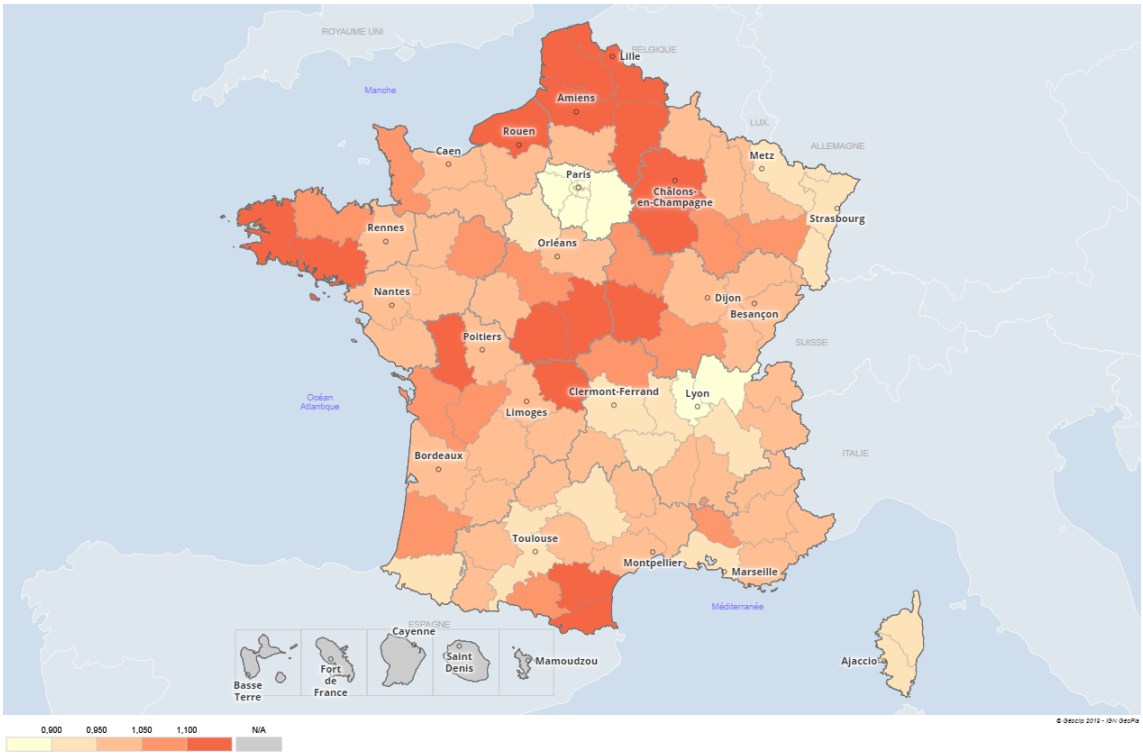


FIGURE 62 – Modélisation des ratios de mortalité départementaux

Obtenu à partir de www.france-decouverte.geoclip.fr - Données importées [20]

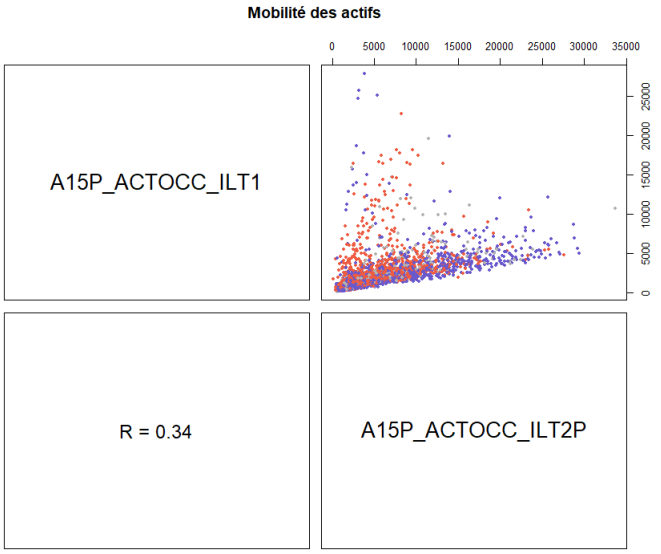
Analyses complémentaires

L'importance de la mobilité domicile-travail

Observons les variables suivantes :

— A15P_ACTOCC_ILT1 : Le nombre d'actifs travaillant dans leur commune de résidence

— A15P_ACTOCC_ILT2P : Le nombre d'actifs travaillant hors de leur commune de résidence.



Sur ce graphique, on remarque deux dynamiques distinctes :

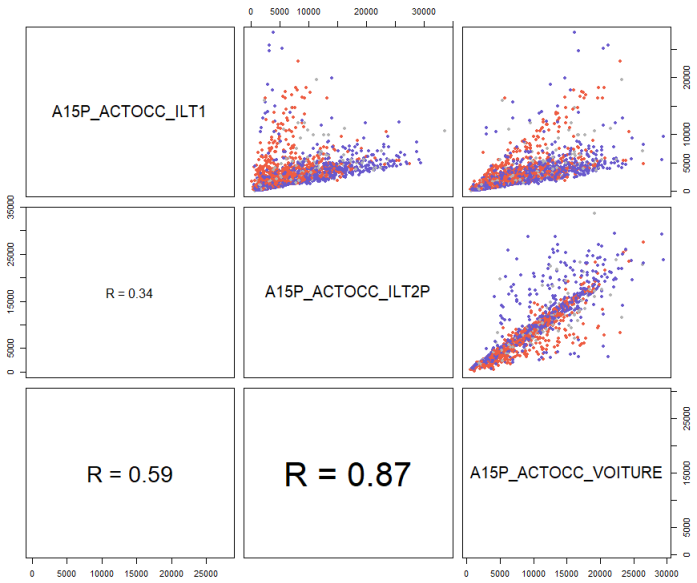
— Les villes où l'on compte beaucoup d'actifs travaillant sur place ont tendance à être en sur-mortalité.

— Les villes où l'on compte beaucoup d'actifs travaillant dans une autre ville ont tendance à être en sous-mortalité.

Aucune ville ne concentre ces deux types de dynamiques, ou alors les trajets vers l'extérieur deviennent prépondérants.

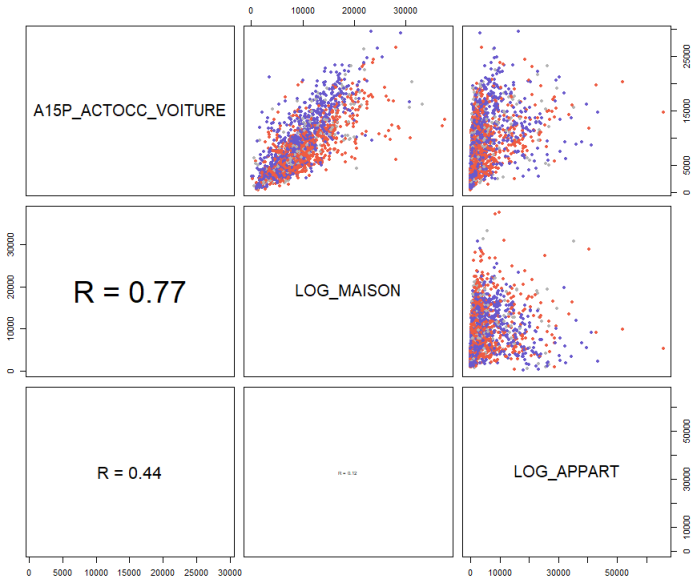
Ajoutons maintenant la variable suivante à notre représentation :

— A15P_ACTOCC_Voiture : Le nombre d'actifs se rendant au travail en voiture.



On constate logiquement que le fait de travailler à l'extérieur du canton est très lié à l'usage de la voiture.
Vers le haut du graphique ILT2P x VOITURE , on notera que les cantons où les individus travaillent à l'extérieur sans recourir de façon importante à la voiture sont en grande majorité en sous-mortalité. Ces cantons disposent très probablement d'un réseau de transport en commun de bonne qualité.

Sur le graphique suivant, on peut voir que l'usage de la voiture pour se rendre au travail est également très corrélé au fait de loger en maison :

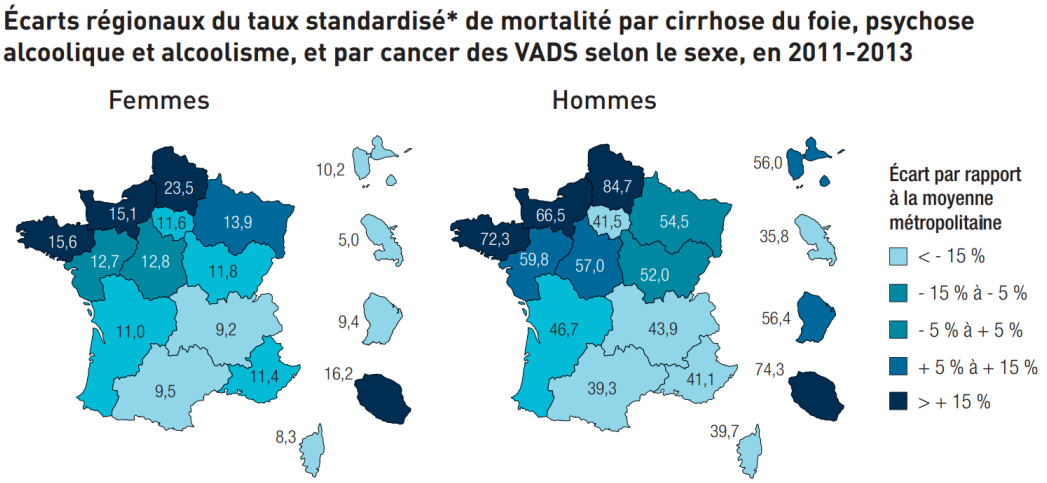


Les éléments que nous venons d'évoquer apparaissent clairement dans notre modélisation. La mobilité des travailleurs est représentative du cadre de vie et, ainsi, du ratio de mortalité des cantons.

L'impact du tabac et de l'alcool sur la mortalité

L'étude ponctuelle de la *Direction de la Recherche, des Études, de l'Évaluation et des Statistiques (DREES)* permet de se faire une idée de l'impact du tabac et de l'alcool sur la mortalité prématurée (ie. avant 65 ans). Ce rapport propose un panorama très détaillé et didactique de l'état de santé des Français. Nous encourageons le lecteur intéressé à se référer au document original [11].

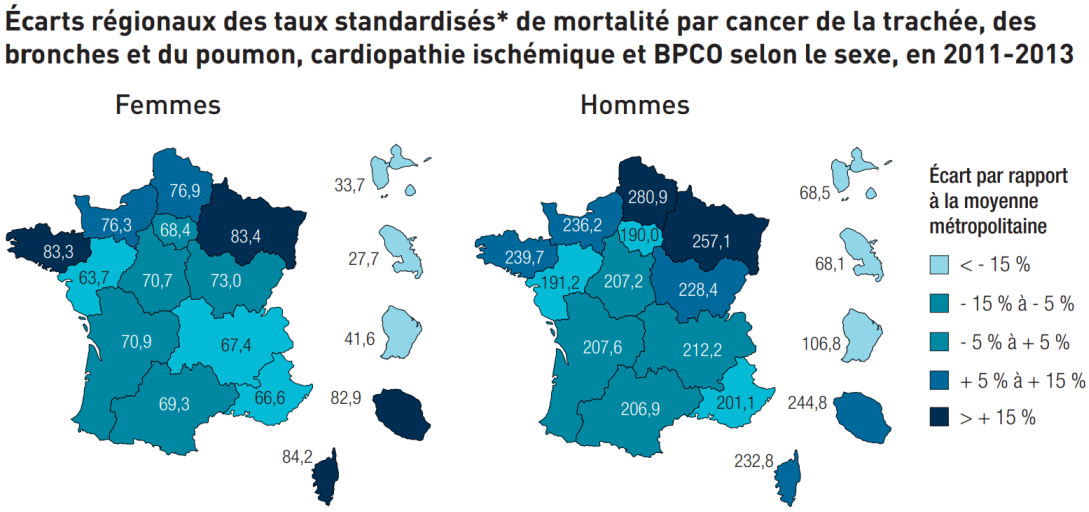
Comme on peut le voir ici, la mortalité prématurée attribuable à l'alcool est très inégalement répartie :



Mortalité attribuée à l'alcool

On constate une tendance lourde dans les Hauts-de-France, Normandie et Bretagne. Plus généralement, à l'exception de Paris, on remarque une graduation du nord ouest au sud est.

Comme pour l'alcool, la mortalité prématurée attribuable au tabac est aussi mal répartie :



Mortalité attribuée au tabac

On remarque une relative corrélation entre ces deux représentations avec, encore une fois, des taux élevés en Bretagne et Hauts-de-France, ainsi qu'une sous mortalité à Paris pour les hommes. On note de plus une sur-mortalité dans la région Grand-Est. Notre modélisation laisse supposer que cet effet est moins marqué dans les départements Bas-Rhin et Haut-Rhin.

LISTE DES VARIABLES

TABLE 18 – Liste des variables utilisées

SOURCE	VARIABLE	DEFINITION
Base comparateur	SUPERF	Superficie
	POP11	Population en 2011
	POP16	Population en 2016
	MEN16	Ménages en 2016
	NAIS1116	Naissances entre 2011 et 2016
	DECE1116	Décès entre 2011 et 2016
	MIGR1116	Migration 2011-2016
	DENS	Densité
	NATAL	Taux de natalité
	MEN_TA	Taille moyenne des ménages
Diplômes	H0217	Hommes 0-18 ans
	H1517	Hommes 15-17 ans
	H1824	Hommes 18-24 ans
	H2529	Hommes 25-29 ans
	F0217	Femmes 0-18 ans
	F1517	Femmes 15-17 ans
	F1824	Femmes 18-24 ans
	F2529	Femmes 25-29 ans
	H0217_SCOL	Hommes scolarisés <18
	H1517_SCOL	Hommes scolarisés 15-17 ans
	H1824_SCOL	Hommes scolarisés 18-24 ans
	H2529_SCOL	Hommes scolarisés 25-29 ans
	F0217_SCOL	Femmes scolarisés <18
	F1517_SCOL	Femmes scolarisées 15-17 ans
	F1824_SCOL	Femmes scolarisées 18-24 ans
	F2529_SCOL	Femmes scolarisées 25-29 ans
	H15P_NSCOL	Hommes 15 ans ou plus non scolarisés
	F15P_NSCOL	Femmes 15 ans ou plus non scolarisées
	H15P_NSCOL_DIPLMIN	Hommes 15 ans ou plus non scol. Sans diplôme ou BEPC, brevet des collèges, DNB
	H15P_NSCOL_CAPBEP	Hommes 15 ans ou plus non scol. CAP-BEP
	H15P_NSCOL_BAC	Hommes 15 ans ou plus non scol. BAC
	H15P_NSCOL_SUP	Hommes 15 ans ou plus non scol. Enseignement sup
	F15P_NSCOL_DIPLMIN	Femmes 15 ans ou plus non scol. Sans diplôme ou BEPC, brevet des collèges, DNB
	F15P_NSCOL_CAPBEP	Femmes 15 ans ou plus non scol. CAP-BEP
	F15P_NSCOL_BAC	Femmes 15 ans ou plus non scol. BAC
	F15P_NSCOL_SUP	Femmes 15 ans ou plus non scol. Enseignement sup

SOURCE	VARIABLE	DEFINITION
Emploi *population active*	H1524	Pop 15-24 ans Hommes en 2011
	H2554	Pop 25-54 ans Hommes en 2011
	H5564	Pop 55-64 ans Hommes en 2011
	F1524	Pop 15-24 ans Femmes en 2011
	F2554	Pop 25-54 ans Femmes en 2011
	F5564	Pop 55-64 ans Femmes en 2011
	H1524_ACTOCC	Actifs occupés 15-24 ans Hommes en 2011
	H2554_ACTOCC	Actifs occupés 25-54 ans Hommes en 2011
	H5564_ACTOCC	Actifs occupés 55-64 ans Hommes en 2011
	F1524_ACTOCC	Actifs occupés 15-24 ans Femmes en 2011
	F2554_ACTOCC	Actifs occupés 25-54 ans Femmes en 2011
	F5564_ACTOCC	Actifs occupés 55-64 ans Femmes en 2011
	H1524_CHOM	Chômeurs 15-24 ans Hommes en 2011
	H2554_CHOM	Chômeurs 25-54 ans Hommes en 2011
	H5564_CHOM	Chômeurs 55-64 ans Hommes en 2011
	F1524_CHOM	Chômeurs 15-24 ans Femmes en 2011
	F2554_CHOM	Chômeurs 25-54 ans Femmes en 2011
	F5564_CHOM	Chômeurs 55-64 ans Femmes en 2011
	INACT1564	Inactifs 15-64 ans en 2011
	ETUD1564	Élèv. Etud. Stag. non rémunérés 15-64 ans en 2011
	RETR1564	Retraités Préretraités 15-64 ans en 2011
	AINACT1564	Autres inactifs 15-64 ans en 2011
	EMPLT	Emplois au LT en 2011
	EMPLT_SLR	Emplois salariés au LT en 2011
	EMPLT_NSLR	Emplois non-salariés au LT en 2011
	EMPLT_AGRI	Emplois au LT Agriculture en 2011
	EMPLT_INDUS	Emplois au LT Industrie en 2011
	EMPLT_CONST	Emplois au LT Construction en 2011
	EMPLT_CTS	Emplois au LT Commerce, Transports, Services divers en 2011
	EMPLT_APESAS	Emplois au LT Adm publique, Enseignement, Santé, Act sociale en 2011
Emploi *caractéristiques*	A15P_ACTOCC	Actifs occupés 15 ans ou plus
	A15P_SLR	Salariés 15 ans ou plus
	A15P_NSLR	Non-salariés 15 ans ou plus
	A15P_ACTOCC_TP	Actifs occ 15 ans ou plus TP
	H15P_ACTOCC	Actifs occupés 15 ans ou plus Hommes
	H15P_SLR	Salariés 15 ans ou plus Hommes
	H15P_SLR_CDI	Salariés 15 ans ou plus Hommes Fonct publ, CDI
	H15P_SLR_CDD	Salariés 15 ans ou plus Hommes CDD
	H15P_SLR_INTERIM	Salariés 15 ans ou plus Hommes Intérim
	H15P_SLR_EMPAID	Salariés 15 ans ou plus Hommes Emplois aidés
	H15P_SLR_APPR	Salariés 15 ans ou plus Hommes Apprentissage - Stage

SOURCE	VARIABLE	DEFINITION
	H15P_NSLR	Non-salariés 15 ans ou plus Hommes
	H15P_NSLR_INDEP	Non-salariés 15 ans ou plus Hommes Indépendants
	H15P_NSLR_EMPLOY	Non-salariés 15 ans ou plus Hommes Employeurs
	H15P_NSLR_AIDFAM	Non-salariés 15 ans ou plus Hommes Aides familiaux
	F15P_ACTOCC	Actifs occupés 15 ans ou plus Femmes
	F15P_SLR	Salariés 15 ans ou plus Femmes
	F15P_SLR_CDI	Salariés 15 ans ou plus Femmes Fonct publ, CDI
	F15P_SLR_CDD	Salariés 15 ans ou plus Femmes CDD
	F15P_SLR_INTERIM	Salariés 15 ans ou plus Femmes Intérim
	F15P_SLR_EMPAID	Salariés 15 ans ou plus Femmes Emplois aidés
	F15P_SLR_APPR	Salariés 15 ans ou plus Femmes Apprentissage - Stage
	F15P_NSLR	Non-salariés 15 ans ou plus Femmes
	F15P_NSLR_INDEP	Non-salariés 15 ans ou plus Femmes Indépendantes
	F15P_NSLR_EMPLOY	Non-salariés 15 ans ou plus Femmes Employeurs
	F15P_NSLR_AIDFAM	Non-salariés 15 ans ou plus Femmes Aides familiales
	H1524_SLR	Salariés 15-24 ans Hommes
	H2554_SLR	Salariés 25-54 ans Hommes
	H5564_SLR	Salariés 55-64 ans Hommes
	F1524_SLR	Salariés 15-24 ans Femmes
	F2554_SLR	Salariés 25-54 ans Femmes
	F5564_SLR	Salariés 55-64 ans Femmes
	A15P_ACTOCC_PASTRANS	Actifs occ 15 ans ou plus pas de transport pour travail
	A15P_ACTOCC_MARCHE	Actifs occ 15 ans ou plus marche à pied pour travail
	A15P_ACTOCC_2ROUES	Actifs occ 15 ans ou plus deux roues
	A15P_ACTOCC_VOITURE	Actifs occ 15 ans ou plus voiture
	A15P_ACTOCC_COMMUN	Actifs occ 15 ans ou plus transport en commun
	A15P_ACTOCC_ILT1	Actifs occ 15 ans ou plus travaillent commune résidence
	A15P_ACTOCC_ILT2P	Actifs occ 15 ans ou plus travaillent autre commune que commune résidence
	A15P_ACTOCC_ILT2	Actifs occ 15 ans ou plus travaillent autre commune même dépt résidence
	A15P_ACTOCC_ILT3	Actifs occ 15 ans ou plus travaillent autre dépt même région résidence
	A15P_ACTOCC_ILT4	Actifs occ 15 ans ou plus travaillent autre région en métropole
	A15P_ACTOCC_ILT5	Actifs occ 15 ans ou plus travaillent autre région hors métropole

SOURCE	VARIABLE	DEFINITION
Entreprises	ENT	Total Ets actifs
	ENT_AGRI	Ets actifs agriculture
	ENT_INDU	Ets actifs industrie
	ENT_CONS	Ets actifs construction
	ENT_COMM	Ets actifs commerce services (hors auto)
	ENT_AUTO	Ets actifs commerce rep auto
	ENT_ADMP	Ets actifs adm publique
	ENT10M	Ets actifs <10 salariés
	ENT10M_AGRI	Ets actifs agriculture <10 salariés
	ENT10M_INDU	Ets actifs industrie <10 salariés
	ENT10M_CONS	Ets actifs construction <10 salariés
	ENT10M_COMM	Ets actifs commerce services (hors auto) <10 salariés
	ENT10M_AUTO	Ets actifs commerce rep auto <10 salariés
	ENT10M_ADMP	Ets actifs adm publique <10 salariés
	ENT50P	Ets actifs de 50 salariés ou plus
	ENT50P_AGRI	Ets actifs agriculture 50 sal ou plus
	ENT50P_INDU	Ets actifs industrie 50 sal ou plus
	ENT50P_CONS	Ets actifs construction 50 sal ou +
	ENT50P_COMM	Ets actifs commerce services (hors auto) 50 sal ou plus
	ENT50P_AUTO	Ets actifs commerce 50 rep auto sal ou plus
	ENT50P_ADMP	Ets actifs adm publique 50 sal ou plus
	POST	Postes des Ets actifs
	POST_AGRI	Postes des Ets actifs agriculture
	POST_INDU	Postes des Ets actifs de l'industrie
	POST_CONS	Postes des Ets actifs de la construction
	POST_COMM	Postes des Ets actifs du commerce services (hors auto)
	POST_AUTO	Postes des Ets actifs du commerce rep auto
	POST_ADMP	Postes des Ets actifs adm publique
	POST10M	Postes des Ets actifs de 1 à 9 salariés
	POST10M_AGRI	Postes des Ets actifs agriculture 1 à 9 salariés
	POST10M_INDU	Postes des Ets actifs industrie 1 à 9 salariés
	POST10M_CONS	Postes des Ets actifs construction 1 à 9 sal
	POST10M_COMM	Postes des Ets actifs commerce services (hors auto) 1 à 9 sal
	POST10M_AUTO	Postes des Ets actifs commerce rep auto 1 à 9 salariés
	POST10M_ADMP	Postes des Ets actifs adm publique 1 à 9 salariés
	POST50P	Postes des Ets actifs > 50 salariés
	POST50P_AGRI	Postes des Ets actifs agriculture > 50 salariés
	POST50P_INDU	Postes des Ets actifs industrie > 50 salariés
	POST50P_CONS	Postes des Ets actifs construction > 50 salariés
	POST50P_CONS	Postes des Ets actifs commerce services (hors auto) > 50 salariés
	POST50P_AUTO	dont Postes des Ets actifs commerce rep auto > 50 salariés
	POST50P_ADMP	Postes des Ets actifs adm publique > 50 salariés

SOURCE	VARIABLE	DEFINITION
Logement	LOG	Logements
	LOG_RP	Résidences principales
	LOG_RS	Rés secondaires et logts occasionnels
	LOG_VAC	Logements vacants
	LOG_MAISON	Maisons
	LOG_APPART	Appartements
	LOG_RP_MAISON	Rés princ type maison
	LOG_RP_APPART	Rés princ type appartement
Revenus	TxPAUV	Taux de pauvreté
	TxMENFIMP	Part des ménages fiscaux imposés (%)
	REV_Q115	1er quartile
	REV_Q315	3e quartile
	Q3_Q1	Écart interquartile
	RD	Rapport interdécile 9e décile/1er decile
	S80S2015	S80/20
	GI	Indice de Gini
	Tx_REV_ACT	Part des revenus dactivité (%)
	Tx_REV_ACT_SAL	dont part des salaires, traitements hors chômage (%)
	Tx_REV_ACT_CHOM	dont part des indemnités chômage (%)
	Tx_REV_ACT_NSLR	dont part des revenus des activités non salariées (%)
	Tx_REV_PRR	Part des pensions, retraites et rentes (%)
	Tx_REV_CAPI	Part des revenus du patrimoine et autres revenus (%)
	Tx_REV_PSOC	Part de l'ensemble des prestations sociales (%)
	Tx_REV_PSOC_FAMI	dont part des prestations familiales (%)
	Tx_REV_PSOC_MINI	dont part des minima sociaux (%)
	Tx_REV_PSOC_APL	dont part des prestations logement (%)
	Tx_REV_PSOC_IMPOTS	Part des impôts (%)
Salaires	SAL	Salaire net horaire moyen
	SAL_H	Salaire net horaire moyen H
	SAL_H_CADR	Salaire net hor. moy. H cadres sup.
	SAL_H_PINT	Salaire net hor. moy. H prof inter.
	SAL_H_EMPL	Salaire net hor. moy. H employés
	SAL_H_OUVR	Salaire net hor. moy. H ouvriers
	SAL_F	Salaire net horaire moyen F
	SAL_F_CARD	Salaire net hor. moy. F cadres sup.
	SAL_F_PINT	Salaire net hor. moy. F prof inter.
	SAL_F_EMPL	Salaire net hor. moy. F employés
	SAL_F_OUVR	Salaire net hor. moy. F ouvriers
	SAL_H_1825	Salaire net horaire moyen H 18 à 25 ans
	SAL_H_2650	Salaire net horaire moyen H 26 à 50 ans
	SAL_H_50P	Salaire net horaire moyen H plus de 50 ans
	SAL_F_1825	Salaire net horaire moyen F 18 à 25 ans
	SAL_F_2650	Salaire net horaire moyen F 26 à 50 ans
	SAL_F_50P	Salaire net horaire moyen F plus de 50 ans

SOURCE	VARIABLE	DÉFINITION
Structure Population	POP16	Population en 2016
	POPH	Pop Hommes
	H0014	Pop Hommes 0-14 ans
	H65P	Pop Hommes 65 ans ou plus
	POPF	Pop Femmes
	F0014	Pop Femmes 0-14 ans
	F65P	Pop Femmes 65 ans ou plus
	POP01P	Pop 1 an ou plus localisée 1 an auparavant
	POP01P_IRAN1	Pop 1 an ou plus habitant 1 an avt même logt
	POP01P_IRAN2	Pop 1 an ou plus habitant 1 an avt autre logt même commune
	POP01P_IRAN3	Pop 1 an ou plus habitant 1 an avt autre commune même dépt
	POP01P_IRAN4	Pop 1 an ou plus habitant 1 an avt autre dépt même région
	POP01P_IRAN5	Pop 1 an ou plus habitant 1 an avt autre région métropole
	POP01P_IRAN6	Pop 1 an ou plus habitant 1 an avt un Dom
	POP01P_IRAN7	Pop 1 an ou plus habitant 1 an avt hors métro ou Dom

TABLE 19 – Variables complémentaires - Indicateurs numériques - Échelle : Commune

Source	VARIABLE	Description	Base(s) utilisé(s)	*Année*
DREES	IND_med	Indice d'accessibilité à un médecin	Indicateur d'accessibilité potentielle localisée (APL) aux médecins généralistes	2018
	IND_med65M	Indice d'accessibilité à un médecin (> 65ans)		
	SU_SMUR_MCS	Temps d'accès aux urgences	Diagnostic d'accès aux soins urgents	2015
	Edu_Rdblt	Part des élèves ayant au moins un an de retard à l'entrée en sixième	C15	2018
	LOG_APL	Part du revenu consacré au loyer pour allocataires d'une aide au logement	C20	2020

TABLE 20 – Variables complémentaires - Indicateurs catégoriels - Échelle : Département

Source	**VARIABLE**	**Description**	**Base(s) utilisé(s)**	*Année*
DREES	LOG_Occ	Niveau d'occupation des logements	C19	2016
	CHOM	Indicateur de l'intensité du chômage	C12	
	TxPauv	Indicateur de pauvreté	C7-C8	
	AgeACC	Age moyen d'accouchement	C24	2017
EIDER	Meteo	Type de climat	TE06	2017
	Eau	Pollution de l'eau	EA42 - EA43	2014
	Equip	Equipement des communes (Nombre de commerces, d'écoles, etc...)	TE08 - TE07	
	Transp	Importance du réseau de transport	QS10	2010
ODICER	AlcTbc	Indicateur d'impact de l'alcool et du tabac sur la mortalité	Taux de décès attribués à l'alcool/au tabac chez les 45-64 ans H/F	2017

TABLE 21 – Description des bases complémentaires - DREES

BASE	Nom	Description	Maille
C07-ISD	Pauvreté monétaire	Part de la population dont le revenu disponible du ménage par unité de consommation est inférieur à 60 % de la médiane nationale *(définition européenne)*.	DEP
C08-ISD	Intensité de la pauvreté monétaire	Ecart relatif, en pourcentage du seuil de pauvreté à 60%, entre le niveau de vie médian des personnes pauvres et ce seuil *(définition européenne)*.	
C12-ISD	Demandeurs d'emploi en fin de mois de catégorie A, B, C dans la population en âge de travailler	Effectifs et taux	
C15-ISD	Part des élèves ayant au moins un an de retard à l'entrée en sixième	Taux pour 100 élèves de 6e.	
C19-ISD	Logements suroccupés	Part des lgoements en sur/sous occupation	
C20-ISD	Taux d'effort net médian des allocataires d'une aide au logement	Ratio entre le coût du logement, déduction faite des allocations logement, et les revenus. (ie. part du revenu des allocataires effectivement consacrée au loyer une fois prises en compte les allocations logement)	
C24-ISD	Répartition des accouchements selon l'âge de la mère	*(Taux pour 100 accouchements)*	
Indicateur d'accessibilité potentielle localisée aux médecins généralistes	Indicateur d'accessibilité aux médecins généralistes	Nombre médian de consultation par an et par habitant. Construit à partir de : - Nombre de consultations et visites effectués dans l'année - Temps de trajet pour accéder à un medecin - Structure d'age *(pondération)*	COM
Diagnostic d'accès aux soins urgents	Diagnostic d'accès aux soins urgents	Temps de trajet pour accéder aux urgences.	

ISD : Indicateurs sociaux départementaux

:

TABLE 22 – Description des bases complémentaires - EIDER

BASE	**Nom**	**Description**	**Maille**
EA42	Qualité physico-chimique générale des eaux superficielles pour les nitrates	Pollution de l'eau au nitrate	DEP
EA43	Qualité physico-chimique générale des eaux superficielles pour l'altération pesticides	Pollution de l'eau aux pesticides	
TE06	Eléments de climatologie	*(Nombre de jours d'ensoleillement, précipitations etc...)*	
TE07 TE08	Equipement des communes	Nombre de commerce, d'écoles, etc... par ville/par habitants	
QS10	La densité des réseaux de communication	Kilomètres d'autoroute, de départementales, de routes nationales et superficie du territoire	