Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 18/03/2021

Par : **Damien Loureiro**

Titre : **Utilisation de la DSN et de l'open data pour élaborer et expliquer un zonier incapacité**

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

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

Membres présents du jury de la filière
Caroline Hillairet

Entreprise : Malakoff Humanis
Nom : Solange Hamel
Signature :

Membres présents du jury de l'Institut des Actuaires

Directeur du mémoire en entreprise :

Nom : Julie Séguéla & Solange Hamel
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

Secrétariat :

Signature du candidat

Bibliothèque :

Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE)
5, avenue Henry Le Chatelier - 91120 PALAISEAU, FRANCE

Résumé

Dans un contexte de dégradation continue du risque arrêt de travail depuis plusieurs années en France [1], la généralisation de la Déclaration Sociale Nominative (DSN) offre de nouvelles opportunités pour mieux anticiper et comprendre ce risque. Ainsi, dans un environnement de plus en plus concurrentiel, l'utilisation des données de masse fournies dans la DSN peut permettre de proposer des tarifs plus segmentés. Ce mémoire traite de l'élaboration d'un zonier incapacité dans le but de prendre en compte la localisation de l'entreprise dans la tarification prévoyance collective. Il est construit en utilisant une classification ascendante hiérarchique avec contraintes de proximité géographique (Chavent et al., 2018)[2]. Cette méthode présente l'avantage de tenir compte du voisinage dans l'attribution des classes de risques à chaque territoire. Cette classification est basée sur les résidus agrégés d'une régression logistique multinomiale prenant en compte les critères usuels de tarification. Ces résidus contiennent donc la part non expliquée de la sinistralité après prise en compte des critères de tarification déjà utilisés. Une analyse du zonier obtenu est réalisée à l'aide de données démographiques et socio-économiques de l'INSEE et des SHAP [2] values, permettant ainsi de mieux appréhender les caractéristiques des différentes zones. Ces open data sont aussi utilisées pour estimer le risque porté par les territoires n'ayant que très peu d'affiliés chez Malakoff Humanis. Enfin, l'apport du zonier dans la connaissance du risque arrêt de travail est validé en comparant les modélisations avec/sans zonier, ainsi qu'en ayant recours à une base de données "test".

Mots-clés : zonier, classification ascendante hiérarchique, contraintes de proximité géographique, data science, structure de voisinage, SHAP, régression multinomiale, arrêt de travail, déclaration sociale nominative (DSN), open data

1. (Malakoff Humanis, 2020) [25] et (DREES, 2020) [26].
2. SHapley Additive exPlanations.

Abstract

Against a backdrop of a steady deterioration in the sick leave risk over the past several years in France [1], the nominative social declaration (DSN) [2] offers new opportunities to better anticipate and understand this risk. Thus, in an increasingly competitive environment, the use of the data provided in the DSN may make it possible to offer a more segmented pricing. This dissertation deals with the creation of a zoning in order to take into account the company location in the pricing system of group insurance contracts. It is constructed using a hierarchical clustering with spatial constraints (Chavent et al., 2018)[2]. This method has the advantage to take into account the neighbourhood when assigning risk classes to each territory. This clustering is based on the aggregated residuals of a multinomial logistic regression taking into account the usual pricing criteria. These residuals therefore contain the unexplained part of the occurrence of sick leaves, after taking into account the pricing criteria already used. An analysis of the zoning obtained is carried out using demographic and socio-economic data from INSEE and SHAP [3] values, thus providing a better understanding of the characteristics of the different zones. These public data are also used to estimate the risk borne by territories with few or no Malakoff Humanis policyholders. Finally, the contribution of the zoning in the knowledge of the sick leave risk is validated by comparing the models with and without zoning and by using a "test" dataset.

Keywords: zoning, hierarchical clustering, spatial constraints, data science, neighbourhood structure, SHAP, multinomial regression, sick leave, nominative social declaration, open data

1. (Malakoff Humanis, 2020) [25] and (DREES, 2020) [26].
2. DSN for Déclaration Sociale Nominative in French.
3. SHapley Additive exPlanations.

Remerciements

Avant tout développement, je souhaite remercier tout particulièrement mes deux tutrices en entreprise, Julie Séguéla, docteur en statistique et Data-Scientist, et, Solange Hamel, actuaire et responsable du pilotage technique chez Malakoff Humanis. Je les remercie pour leurs précieux conseils, leur apport scientifique et leur suivi sans faille tout au long de la réalisation de ma mission. J'ai grandement apprécié travailler avec elles et j'espère que mes futures collaborations dans ma vie professionnelle seront tout aussi enrichissantes.

Je suis extrêmement reconnaissant envers Raphaël Soulignac, responsable de l'équipe Data-Science et Stéphane Barde, directeur Data chez Malakoff Humanis pour m'avoir fait confiance en me donnant l'opportunité de travailler sur ce sujet plein de challenges et en l'abordant avec des méthodes innovantes.

Je tiens également à remercier Ao Sun, Data-Scientist pour avoir partagé avec moi ses connaissances les plus fines sur la DSN.

J'adresse aussi mes sincères remerciements à toutes les équipes de la direction Data et de la direction Technique pour leurs constantes bonne humeur, ce qui m'a permis de travailler dans un environnement serein.

Je remercie aussi Caroline Hillairet, ma tutrice académique pour son encadrement et sa relecture. Plus généralement, je remercie les équipes pédagogiques de l'ENSAE Paris et de la Toulouse School of Economics pour la qualité de leurs enseignements.

Enfin, je tiens à exprimer ma profonde gratitude à mes parents, ma soeur et mes amis, qui m'ont toujours soutenu tout au long de mes études.

Sommaire

Introduction **5**

I Contexte de l'étude 8

1 La garantie incapacité de travail dans les contrats d'assurance prévoyance 9
1.1 La prévoyance . . . 9
1.2 La garantie incapacité de travail temporaire . . . 10
1.3 Tarification de la garantie incapacité de travail temporaire . . . 10

2 Données et méthodologie 12
2.1 Présentation et traitements des données . . . 12
2.1.1 La Déclaration Sociale Nominative . . . 12
2.1.2 Catégorisation des variables continues . . . 14
2.1.3 Analyse descriptive des données . . . 17
2.2 Méthodologie mise en place pour élaborer le zonier . . . 19
2.2.1 Déontologie actuarielle . . . 19
2.2.2 Différentes étapes pour l'élaboration du zonier incapacité . . . 20

II Modélisation du risque incapacité avec les critères classiques de tarification 23

3 Modélisation du risque incapacité avec une régression multinomiale 24
3.1 Modèles linéaires généralisés . . . 24
3.1.1 Généralités autour des GLM . . . 24
3.1.2 Régression logistique multinomiale . . . 25
3.2 Application de la régression logistique multinomiale . . . 26
3.2.1 Interprétation des coefficients . . . 26
3.2.2 Impact des différentes variables dans l'estimation du risque incapacité . . . 30

4 Traitements des résidus de la régression multinomiale 33
4.1 Agrégation des résidus à la maille code postal . . . 33
4.2 Sélection des indicateurs . . . 36

III Construction du zonier incapacité 40

5 Structure de voisinage et intérêts d'un zonier incapacité 41
5.1 Différents types de structures de voisinage . . . 41
5.2 Intérêts d'un zonier incapacité . . . 44

5.2.1 Premières cartes... 44
5.2.2 Autocorrélation spatiale 46

6 Construction du zonier avec une classification hiérarchique spatiale 50
6.1 Théorie autour de la classification hiérarchique spatiale 50
6.1.1 Généralités autour de la classification ascendante hiérarchique 50
6.1.2 Ajout des contraintes spatiales 52
6.2 Application au portefeuille 53
6.2.1 Nombre de zones et poids de la contrainte spatiale 53
6.2.2 Résultats de la classification 56

IV Modélisation du zonier à l'aide de données démographiques et socio-économiques de l'INSEE 61

7 Construction d'indicateurs socio-économiques et démographiques à la maille code postal 62
7.1 Agrégation des données INSEE à la maille code postal 63
7.1.1 Données socio-économiques et démographiques initialement à la maille IRIS 63
7.1.2 Données socio-économiques et démographiques initialement à la maille commune 64
7.2 Indicateurs socio-économiques et démographiques à la maille code postal 66

8 Interprétation du zonier et traitement des codes postaux avec peu ou pas d'affiliés 68
8.1 Analyse descriptive des données démographiques et socio-économiques selon les classes de risques 68
8.2 Construction du modèle de prédiction des classes de risques à partir des données INSEE 70
8.3 Interprétation du zonier via les SHAP values 73
8.3.1 Présentation de la valeur de Shapley dans le cadre de la théorie des jeux 73
8.3.2 Utilisation de la valeur de Shapley pour interpréter le modèle de prédiction des classes de risques 75
8.3.3 Résultats de l'application de SHAP sur le modèle de prédiction des classes de risques 77
8.4 Affectation des classes de risques aux codes postaux avec peu ou pas d'affiliés . . 82

V Validation de l'apport du zonier dans la connaissance du risque incapacité 85

9 Apports du zonier dans la connaissance du risque incapacité 86
9.1 V de Cramer 86
9.2 Significativité des coefficients de régression 87
9.3 Comparaison des modélisations du risque incapacité avec et sans zonier 89
9.3.1 Test statistique 89
9.3.2 Ajustement du modèle 90
9.4 Application du zonier sur une base de données "test" : les DSN 2019 91

10 Avantages et limites d'un zonier incapacité **95**
10.1 Avantages d'utiliser la localisation de l'entreprise dans la tarification prévoyance collective . . . 95
10.2 Limites de l'utilisation de la localisation de l'entreprise dans la tarification prévoyance collective . . . 96

Conclusion **98**

Bibliographie **101**

Annexes **103**
A Choix de la classification pour l'âge . . . 103
B Preuve sur la régression multinomiale : équivalence des équations 3.1 et 3.2 . . . 104
C Choix du poids de la contrainte spatiale dans la classification hiérarchique spatiale 105
D Contributions des variables (SHAP values) dans le cas du 9e arrondissement de Paris . . . 106
E SHAP values pour les zones 3 et 4 ("summary plot") . . . 107

Note de synthèse **108**

Executive summary **114**

Introduction

La Déclaration Sociale Nominative (DSN) est un système permettant à tout employeur, de déclarer de façon unique, dématérialisée et mensuelle, un ensemble d'informations liées à la protection sociale de ses salariés. Elle est devenue obligatoire au 1er janvier 2017 [1] et regroupe des renseignements sur le contrat de travail, les arrêts de travail, les rémunérations et primes des salariés. Ce dispositif vise à simplifier les démarches des employeurs puisqu'il remplace plusieurs autres déclarations destinées à différents acteurs (Urssaf, Pôle emploi, Caisse Primaire d'Assurance Maladie (CPAM), Organismes complémentaires, Centre des finances publiques, AGIRC-ARRCO et autres caisses de retraite des régimes spéciaux, etc.). Au-delà de la simplification administrative, la DSN a également permis une amélioration de la qualité de la déclaration puisqu'elle est basée sur l'acte de paie et illustre donc la situation des salariés à cet instant-là. Le fait qu'elle soit transmise chaque mois permet d'avoir accès à des informations récentes qui ont un impact sur la paie, comme la maladie, la maternité/paternité, la fin de contrat de travail, un changement de rémunération, etc.

Les données issues de la DSN permettent aux assureurs prévoyance d'avoir une connaissance plus fine de leurs affiliés sous contrat collectif et du risque qu'ils portent, ce qui pourrait se concrétiser par l'ajout d'un (ou plusieurs) degré(s) de segmentation dans leur tarification prévoyance collective. Par exemple, la localisation de l'entreprise peut apparaître comme un nouveau critère de tarification de cette assurance, et ce, par le biais d'un zonier. Un zonier incapacité pouvait tout à fait être construit avant l'arrivée de la DSN. Cependant, cette "nouvelle" base de données apporte des informations supplémentaires et permet notamment d'avoir une vision globale de l'arrêt de travail. En effet, la DSN offre l'opportunité de disposer des arrêts de travail dès leur déclaration, ce qui permet d'avoir accès à des arrêts dont la durée est inférieure au délai de franchise.

Bien que les zoniers soient très présents dans les tarifications des assurances multirisque habitation (MRH) et automobile, ils sont peu utilisés en assurance prévoyance collective. L'objectif principal de ce mémoire est donc de réaliser un zonier afin de démontrer l'intérêt d'un tel critère dans la connaissance du risque incapacité de travail. Ce zonier est construit à la maille code postal car il s'agit d'une donnée disponible dans la DSN et pouvant être facilement demandée aux clients au moment de la tarification. Une analyse est aussi menée afin d'identifier les caractéristiques démographiques et socio-économiques des zones ayant différents niveaux de risque.

Les zoniers (MRH et auto) sont en général construits par lissage spatial d'un indicateur, qui est le plus souvent le résidu d'un modèle linéaire généralisé. Cette méthode est, par exemple, implémentée dans le mémoire d'actuariat (Pariente, 2017)[9] où l'auteur utilise notamment la méthode de pondération par l'inverse à la distance et le krigeage comme méthodes d'interpola-

1. Certaines exceptions demeurent sur le caractère obligatoire de cette déclaration, notamment pour les employeurs de la fonction publique pour lesquels cet exercice sera progressivement imposé d'ici le 1er janvier 2022.

tion spatiale. Une autre démarche est d'élaborer un modèle de Machine Learning pour prédire les résidus. Dans ce cas, la notion de "proximité géographique" peut se faire par l'ajout dans le modèle de variables basées sur le voisinage de chaque observation. (Beraud-Sudreau, 2017)[8] applique cette méthode pour l'assurance MRH dans son mémoire d'actuariat.

Dans notre étude, nous allons, comme dans les références citées ci-dessus, élaborer un zonier à partir des résidus d'une régression qui prend en compte les variables déjà utilisées dans la tarification. Néanmoins, les modèles implémentés sont différents de ceux évoqués précédemment et innovants dans le domaine de l'actuariat. Ainsi, la première régression est une régression logistique multinomiale[19] et le zonier est réalisé à l'aide d'une classification ascendante hiérarchique avec contraintes de proximité géographique. Cette classification non supervisée, présentée dans (Chavent et al., 2018)[2] prend en compte la proximité géographique en lui accordant un certain poids par rapport aux résidus. Elle permet aussi d'élaborer les clusters à partir d'un ensemble d'indicateurs contrairement aux méthodes de lissage mentionnées ci-dessus, qui elles, ne traitent que d'un seul indicateur.

En résumé, l'idée de ce mémoire est de mêler plusieurs opportunités. Premièrement, nous commençons à avoir un peu de recul sur les données de la DSN (plus de 3 ans) ce qui rend possible une amplification de leurs utilisations par les assureurs de personnes. Ces données peuvent offrir de nouvelles perspectives sur une multitude de sujets ou même remplacer d'autres bases de données internes pour faire certaines analyses. Ensuite, la localisation de l'entreprise n'est pas une dimension aujourd'hui exploitée dans la tarification prévoyance collective de Malakoff Humanis. L'élaboration d'un zonier permet donc d'évaluer la pertinence d'un tel critère de tarification. Enfin, les méthodes appliquées, et plus particulièrement la classification ascendante hiérarchique avec contraintes de proximité géographique, n'avaient jamais été testées chez Malakoff Humanis.

Afin de rendre compte de l'étude réalisée, le présent mémoire est divisé en 5 parties. Dans un premier temps, le contexte de l'étude est présenté, en décrivant plus particulièrement la garantie incapacité de travail, les données utilisées ainsi qu'une vue d'ensemble de la méthodologie mise en place.

La deuxième partie expose la modélisation du risque incapacité par un GLM [2] utilisant les critères classiques de tarification. La théorie autour des modèles linéaires généralisés et en particulier celle de la régression logistique multinomiale est détaillée. L'impact des différentes variables sur le risque incapacité de travail est aussi explicité tout comme l'agrégation des résidus à la maille code postal.

La troisième partie de ce mémoire traite de la construction du zonier incapacité. Le choix de la structure de voisinage, notion clé pour des données spatiales, est notamment précisé avant de décrire la théorie et l'application de la classification ascendante hiérarchique avec contraintes de proximité géographique.

Dans la quatrième partie, l'interprétation du zonier est réalisée à l'aide de données démographiques et socio-économiques de l'INSEE et des SHAP [3] values (ou valeurs de Shapley). Ces travaux permettent aussi d'affecter une classe de risque aux territoires dans lesquels Malakoff Humanis n'a que très peu (ou pas du tout) d'affiliés, et ce, en se basant sur les caractéristiques démographiques et socio-économiques de ces territoires en question. Par ailleurs, le travail

2. Generalized Linear Model.
3. SHapley Additive exPlanations.

d'agrégation des données de l'INSEE à la maille géographique souhaitée est développé.

Enfin, la cinquième et dernière partie valide l'apport du zonier dans la connaissance du risque incapacité de travail en comparant les modèles avec/sans zonier, ainsi qu'en utilisant une base de données "test". Les avantages et limites d'incorporer la localisation de l'entreprise dans la tarification prévoyance collective sont aussi présentés en guise de conclusion de cette partie.

Première partie

Contexte de l'étude

Chapitre 1

La garantie incapacité de travail dans les contrats d'assurance prévoyance

1.1 La prévoyance

La prévoyance est le produit d'assurance étudié dans le cadre de ce mémoire. La loi n° 89-1009 du 31 décembre 1989, dite loi Evin, indique que la prévoyance rassemble les « opérations ayant pour objet la prévention et la couverture du risque décès, des risques portant atteinte à l'intégrité physique de la personne ou liés à la maternité ou des risques d'incapacité de travail ou d'invalidité ou du risque chômage »[15]. Ce type d'assurance vise donc à couvrir les aléas de la vie en diminuant les conséquences financières de ces évènements.

Les prestations au titre d'un contrat prévoyance peuvent être versées à l'assuré (notamment en cas d'invalidité ou d'incapacité temporaire de travail [1]) ou à ses bénéficiaires dans le cas du décès de l'assuré. Par exemple, si l'assuré décède, les bénéficiaires peuvent avoir droit à un capital décès, une rente "conjoint" permettant de compenser la perte de revenu et/ou une rente éducation visant à subvenir aux dépenses courantes et au financement des études des enfants bénéficiaires.

De la même façon qu'une complémentaire santé, un contrat d'assurance prévoyance peut être collectif ou individuel. Une entreprise peut donc souscrire une assurance prévoyance collective pour ses salariés. Cependant, le caractère obligatoire de ces deux couvertures diffère entre la complémentaire santé et la prévoyance. La loi oblige tout employeur du secteur privé à proposer une complémentaire santé collective à l'ensemble de ses salariés alors que la prévoyance n'est en général obligatoire que pour les cadres et assimilés. En effet, ces derniers bénéficient au minimum d'une garantie décès souscrite par leur employeur, et ce, grâce à la convention collective nationale de retraite et de prévoyance des cadres du 14 mars 1947[16], perpétuée par l'Accord National Interprofessionnel (ANI) du 17 novembre 2017 relatif à la prévoyance des cadres[17], étendu et élargi par arrêté ministériel[18]. Néanmoins, certaines conventions collectives ou accords de branche ont aussi rendu obligatoire la souscription d'un contrat prévoyance pour les salariés non-cadres. C'est notamment le cas des CCN [2] sport et restauration rapide. Les contrats individuels sont quant à eux destinés aux salariés non couverts ou ayant des garanties insuffisantes dans leur contrat collectif mais aussi aux indépendants et professions libérales. Dans ce mémoire, seulement l'assurance collective est considérée.

1. Tout au long de ce mémoire, le terme "incapacité" fera référence à "l'incapacité temporaire de travail" et le terme "invalidité" à "l'incapacité permanente de travail".
2. CCN : Convention Collective Nationale.

1.2 La garantie incapacité de travail temporaire

En France, il existe un régime de base de prévoyance géré par les organismes de Sécurité Sociale. Par exemple, en ce qui concerne l'incapacité de travail, garantie étudiée dans le présent mémoire ; si un salarié est en arrêt maladie, il est indemnisé par l'Assurance Maladie à hauteur de 50% de son salaire de base. Néanmoins, le salaire de base pris en compte dans le calcul de l'indemnisation est plafonné à 1.8 fois le montant du SMIC en vigueur, ce qui limite les indemnités journalières au 1er janvier 2020, à 45,55€ bruts[24]. Cette prestation financière protège les salariés en cas d'arrêt de travail mais elle n'est que partielle et en général insuffisante pour maintenir le niveau de vie. C'est la raison pour laquelle les garanties incapacité de travail sont présentes dans la plupart des contrats prévoyance des assureurs, mutuelles et institutions de prévoyance. Ces garanties permettent de compenser, au moins partiellement et de manière complémentaire aux indemnités journalières versées par le régime de base, la perte de salaire des assurés qui seraient dans l'impossibilité temporaire d'exercer leur activité professionnelle.

Différentes couvertures concernant l'incapacité de travail existent. Par exemple, les contrats d'assurance peuvent se différencier sur le taux de maintien du salaire ou du délai de franchise. Le maintien de salaire peut être partiel (par exemple 80%) ou bien total et dans ce cas, le porteur du risque compense l'intégralité de la perte de salaire. Le délai de franchise correspond, en cas de sinistre, à la durée pendant laquelle l'organisme d'assurance n'indemnise pas son assuré. Ainsi, dans le cas d'une franchise de 90 jours, l'assuré ne recevra une indemnisation de la part de sa mutuelle, institution de prévoyance ou assureur qu'à partir du 91e jour d'arrêt [3].

Les indemnités journalières versées par l'organisme d'assurance ne peuvent excéder une durée de trois ans [4]. Au-delà de 36 mois d'arrêt maladie, soit le salarié reprend le travail (à temps partiel ou temps complet), soit il est mis en invalidité (partielle ou totale) et dans ce cas, il perçoit une pension d'invalidité de la Sécurité sociale et éventuellement une rente (ou capital) invalidité de son organisme de prévoyance complémentaire.

1.3 Tarification de la garantie incapacité de travail temporaire

De manière simplifié, le fonctionnement général de l'assurance est le suivant : en échange des garanties offertes, les assurés payent une cotisation/prime à leur assureur. Ces cotisations peuvent être en partie (ou intégralement) financées par l'employeur des assurés s'il s'agit d'un contrat collectif. Comme tout contrat d'assurance, les contrats prévoyance font l'objet d'une tarification adaptée en fonction des risques portés par les assurés. Néanmoins, dans le cas de contrat collectif, les cotisations sont mutualisées parmi les salariés d'une même entreprise [5]. Ces salariés paient donc tous le même taux de cotisation pour les mêmes garanties. La tarification de ces contrats se fait donc par l'utilisation de critères à la maille de l'entreprise, comme l'âge moyen, le nombre de salariés, la répartition homme/femme, le secteur d'activité, la réparti-

3. Différents types de franchises existent, notamment les franchises continue et discontinue. Dans le premier cas, les durées des arrêts de travail sont comptabilisées distinctement tandis que dans le deuxième, le calcul se fait en sommant les durées des précédents arrêts sur une période donnée.

4. Cette durée de trois ans est aussi celle retenue par la Sécurité Sociale pour les affections de longue durée (ALD).

5. Un contrat prévoyance collectif peut concerner l'ensemble des salariés d'une même entreprise mais il est possible qu'il ne soit proposé qu'à une certaine catégorie d'entre eux (seulement les cadres par exemple). Cependant, cette distinction ne peut se faire par des critères de revenu, d'âge ou d'état de santé.

tion des différentes catégories socio-professionnelles (CSP), etc. Ces précédents critères sont les principaux utilisés sur le marché et le fait d'ajouter de nouveaux critères apporte un degré de segmentation supplémentaire et permet de proposer des tarifs plus proches du risque porté par les assurés, ce qui peut être un avantage concurrentiel. En effet, si un assureur a une tarification moins segmentée que sur le marché, il aura tendance à attirer les assurés qui portent un risque important. Par exemple, si un assureur propose le même tarif à deux groupes d'assurés ne portant pas le même risque alors le groupe le moins risqué souscrira sûrement chez la concurrence (qui propose un tarif plus segmenté et donc moins cher) et l'assureur se contentera du groupe le plus risqué, qui lui, payera une prime inférieure à son risque. C'est le phénomène d'antisélection.

Il est évident que le tarif ne dépend pas seulement de critères de tarification mais aussi des garanties choisies par le client. Un client souhaitant disposer de franchises plus courtes et/ou d'un taux de maintien de salaire plus élevé payera une cotisation plus importante.

Chapitre 2

Données et méthodologie

2.1 Présentation et traitements des données

2.1.1 La Déclaration Sociale Nominative

Présentation de la DSN

La Déclaration Sociale Nominative (DSN) est un système permettant à tout employeur, de déclarer de façon unique, dématérialisée et mensuelle, un ensemble d'informations liées à la protection sociale de ses salariés. La DSN est considérée comme la dernière étape de la paie et reprend donc des données comme le salaire, les cotisations payées aux différents organismes, le NIR, le SIRET de l'établissement, les numéros de contrats de prévoyance et santé complémentaire. D'ailleurs, c'est le logiciel de paie qui génère la DSN et il doit donc être compatible avec la norme en vigueur. Il est important de noter que la DSN est effectuée par établissement (SIRET), c'est-à-dire qu'une entreprise enverra autant de DSN qu'elle compte d'établissements. L'établissement (SIRET) sera donc la maille retenue dans notre étude. En outre, la DSN est le canal utilisé dans le cadre du prélèvement à la source mis en place depuis le 1er janvier 2019. Elle sert notamment à indiquer les taux à appliquer aux salariés et à transmettre les paiements à la Direction Générale des Finances Publiques.

La DSN est très riche en information puisqu'elle remplace une multitude d'autres déclarations (Urssaf, Pôle emploi, CPAM, AGIRC-ARRCO, Organismes de prévoyance, etc.). Toutes ces données sont réparties en blocs qui correspondent chacun à un thème précis. "Déclaration", "Entreprise", "Individu" et "Fin de contrat" sont des exemples de ces blocs. En tant qu'organisme de prévoyance, Malakoff Humanis reçoit les blocs liés à son activité d'assurance, et ce, uniquement pour ses entreprises clientes. Ainsi, Malakoff Humanis a, par exemple, accès aux blocs "Adhésion Prévoyance", "Affiliation Prévoyance" et "Arrêt de travail" alors que les blocs "Actions gratuites" ou "Options sur titres (stock options)" ne sont, eux, pas disponibles. Ceci se justifie par le fait qu'il est inutile pour les institutions de prévoyance de disposer de ces informations. Malgré tout, la quantité d'information disponible est considérable et uniquement certaines données sont requises pour notre étude. Il a donc fallu faire une sélection des blocs afin de construire un datamart [1] unifié qui contient les informations les plus importantes des différents blocs. Cette préparation de données a été faite par d'autres personnes de l'équipe et antérieurement aux travaux sur le zonier. Ce datamart, nommé "DM_DSN" tout au long de ce rapport est notamment utilisé pour d'autres sujets que le zonier incapacité.

1. Un datamart est un ensemble de données relatives au même thème. L'idée d'un datamart est de présenter les données de manière organisée et souvent agrégée pour répondre à un besoin spécifique, ce qui permet de faciliter l'usage de la donnée en entreprise.

Le canal "DSN" est également utilisé pour des signalements d'évènements. Par exemple, les arrêts maladie, accidents de travail, maternité, paternité et adoption sont des évènements à déclarer dans les 5 jours suivant la connaissance de l'évènement [2] afin d'indemniser rapidement les salariés concernés, excepté si l'employeur pratique la subrogation où dans ce cas, il n'est pas soumis à cette obligation et peut déclarer l'ensemble des arrêts de travail lors de la DSN mensuelle. Ces informations concernant les arrêts de travail sont celles qui seront les données principales de notre étude.

Avantages et inconvénients de la DSN

La Déclaration Sociale Nominative présente de nombreux bénéfices pour différents acteurs.

- Elle permet une simplification très importante des démarches administratives des employeurs, en supprimant une multitude d'autres déclarations aux différents organismes (CPAM, Urssaf, etc.).
- Le fait que la DSN soit destinée à plusieurs organismes assure un certain niveau de qualité des données. Par exemple, la date de fin d'un contrat de travail est une information essentielle pour Pôle emploi afin de calculer les droits au chômage des individus perdant leur travail, et, tout écart peut avoir des conséquences financières importantes.
- La mensualisation de la déclaration réduit le risque d'erreur/omission puisque les évènements et renseignements sont déclarés au fur et à mesure.
- La mensualisation permet aussi de disposer d'informations actualisées et notamment sur les sinistres, ce qui n'était pas le cas auparavant en prévoyance puisqu'il était difficile de disposer de données récentes fiables. En effet, les arrêts de travail sont vus dans la DSN avant d'être présents dans nos bases de prestation à cause du délai de franchise et de gestion. Cette propriété pourrait légitimer l'utilisation de la DSN pour de nombreux nouveaux usages. Par exemple, une étude visant à estimer la durée d'un arrêt de travail dès qu'il est déclaré dans la DSN pourrait être très utile, notamment pour prédire s'il va dépasser ou non la franchise et ainsi être indemnisé ou non. Ceci permettrait d'estimer les futures prestations à verser.
- Enfin, la DSN facilite l'accès à une immense quantité de données. Elle centralise beaucoup d'informations via un même canal (salaires, arrêts de travail, affiliations, ...) alors qu'auparavant, les données étaient plus dispersées, ce qui les rendaient plus difficilement disponibles. Cette accessibilité des données est un argument de plus en faveur de l'utilisation de la DSN pour de nouveaux usages.

L'utilisation de la DSN affiche aussi quelques inconvénients. Premièrement, le fait de disposer de cette masse d'informations est une réelle chance mais aussi un challenge pour les assureurs de personnes. En effet, rendre exploitable et accessible toutes ces données n'est pas si simple. Les données de la DSN sont réparties en différents blocs et pas forcément à une maille intéressante pour les actuaires et data scientists. C'est pourquoi l'utilisation du datamart "DM_DSN" facilite considérablement l'utilisation des données de la DSN. Ensuite, bien que la DSN soit basée sur l'acte de paie, des erreurs de déclaration peuvent être présentes car c'est un nouveau processus non encore maîtrisé par tous les employeurs. Néanmoins, afin d'assurer une certaine qualité de données, des contrôles existent et il est possible pour les employeurs de corriger leur déclaration par une DSN "annule et remplace" ou dans la DSN suivante.

2. Des signalements sont donc émis tout au long du mois et non uniquement lors de la DSN mensuelle. Néanmoins, la DSN mensuelle inclut un récapitulatif des signalements déclarés dans le mois.

Traitements des données pour les rendre exploitables

Comme évoqué précédemment, un datamart unifié contenant les principales informations de la DSN, nommé "DM_DSN" a été créé lors de précédents travaux. À partir de ce dernier, un deuxième datamart centré sur les arrêts de travail a aussi été construit pour des projets antérieurs au zonier incapacité. Cette base est une des principales sources de données de notre étude. L'objectif de cette table est que chaque ligne représente un arrêt de travail. La construction de ce datamart n'est pas détaillée de manière exhaustive dans le présent mémoire mais certains retraitements sont explicités. Par exemple, il a fallu gérer les annulations d'arrêts de travail précédemment déclarés, ainsi que les prolongations et rechutes. Des incohérences de date, sûrement causées par des erreurs humaines, ont aussi été corrigées.

Le zonier incapacité détaillé dans le présent mémoire n'est pas réalisé sur l'ensemble du portefeuille de Malakoff Humanis. Tout d'abord, comme évoqué précédemment, la principale source de données utilisée dans ces travaux est la DSN, ce qui limite notre étude aux affiliés collectifs. Cette contrainte n'est pas très forte dans le sens où l'essentiel du portefeuille en prévoyance est du collectif. Ensuite, certaines DSN n'étaient pas disponibles au format exploitable au moment de la rédaction de ce mémoire, ce qui a réduit notre périmètre d'étude à environ la moitié du portefeuille. Une idée a été d'essayer de remplacer les DSN manquantes par des données de gestion mais la qualité et l'exhaustivité des données n'étaient pas jugées assez satisfaisantes (valeurs manquantes qui entraînaient des biais, absence du motif d'arrêt de travail, présence uniquement des arrêts de travail ayant dépassé le délai de franchise, etc.). Néanmoins, les DSN manquantes seront bientôt disponibles et l'étude réalisée pourra donc être étendue à l'ensemble du portefeuille. Enfin, les congés maternité, paternité et adoption ont été exclus du périmètre d'étude car ils sont pris en charge dans le cadre de garanties spécifiques qui ne sont pas l'objet de ce zonier.

2.1.2 Catégorisation des variables continues

Catégorisation de la variable d'intérêt

L'objectif des travaux présentés dans ce mémoire est de construire un zonier sur le risque incapacité temporaire de travail. Le sujet n'est pas abordé d'un point de vue coût, mais plutôt avec les dimensions fréquence et durée. Plus précisément, la variable cible reflétant la sinistralité est l'occurrence de l'évènement "avoir un arrêt de travail d'une certaine durée" par individu sur une année. Ce n'est donc pas un modèle de durée, dans le sens où le but n'est pas de prédire la durée d'un arrêt de travail. Ce n'est pas non plus un classique modèle de fréquence cherchant juste à estimer le nombre de sinistres par individu. La variable à expliquer prend en compte ces deux dimensions. Ce choix de variable est lié à notre ambition d'avoir une meilleure connaissance globale du risque incapacité selon la localisation de l'entreprise ; et disposer d'une seule variable qui reflète à la fois les notions de fréquence et durée est un atout important.

En effet, lorsque l'on regarde simplement la fréquence, les analyses sont du type "Toutes choses égales par ailleurs, le territoire A connaît plus d'arrêts de travail que le territoire B". Hors la durée est importante dans la tarification des garanties puisque des franchises sont en général présentes pour l'incapacité temporaire de travail et l'assureur n'intervient qu'après le délai de franchise révolu. Ainsi, plus la durée de la franchise est longue, moins la garantie est chère.

Inversement, lorsque l'on regarde simplement la durée des sinistres, il est, par exemple, possible d'identifier les zones qui ont en moyenne les arrêts de travail les plus longs. Néanmoins, si cette zone n'a que très peu de sinistres, elle est très certainement moins à risque qu'une zone avec une durée moyenne d'arrêt légèrement inférieure mais avec une fréquence beaucoup plus

élevée. Les notions fréquence et durée sont donc toutes les deux essentielles, ce qui justifie le fait de prédire les deux dimensions en modélisant la survenance d'un arrêt d'une certaine durée.

Afin de pouvoir mener cette modélisation, il faut, au préalable, catégoriser les arrêts de travail en différentes classes. La classification choisie est la suivante :

— Absence d'arrêt de travail ("Absence d'AT"),
— Arrêt de travail de 0-15 jours ("]0,15]"),
— Arrêt de travail de 15-30 jours ("]15,30]"),
— Arrêt de travail de 30-90 jours ("]30,90]")
— Arrêt de travail de plus de 90 jours ("+90")

Ces seuils ont été choisis en fonction des différentes franchises pratiquées dans le groupe. Ainsi, en fonction du poids des différentes franchises dans le portefeuille, il est possible d'accorder plus ou moins d'intérêt à certaines modalités. De même, le fait de disposer de ces différentes modalités sera utilisé lors de l'élaboration du zonier pour sur-pondérer les modalités qui portent le plus de risques ("]30,90]", "+90") et sous-pondérer les autres ("]0,15]", "]15,30]").

La catégorisation de la variable cible a aussi l'avantage de résoudre partiellement la problématique liée aux données censurées, qui sont omniprésentes lorsque des durées sont étudiées. La figure 2.1 détaille différents types d'arrêts de travail selon leur date de survenance et montre comment la sélection du périmètre d'étude permet de limiter les effets de cette censure.

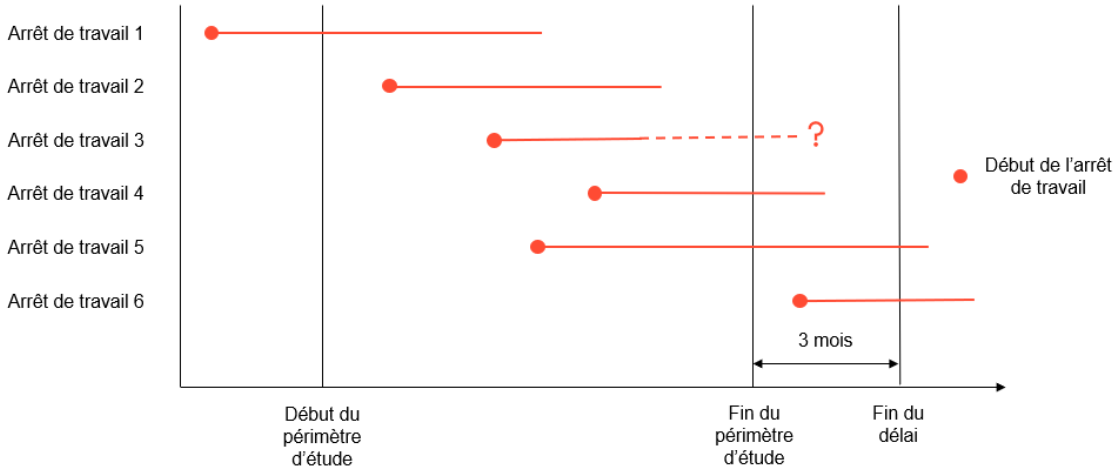

FIGURE 2.1 – Périmètre d'étude et différents types d'arrêts de travail

L'arrêt de travail 1 est un exemple de censure à gauche, puisqu'il est survenu avant le début de l'observation et s'est terminé pendant la période d'observation. Cet arrêt de travail n'est pas inclus dans le périmètre d'étude. La raison vient du fait que dans le cas de l'incapacité de travail, la garantie est activée si la date du début de l'arrêt de travail est comprise entre la date de prise d'effet et la date de résiliation de la garantie. Ainsi, dans le cas d'une entreprise changeant d'assureur prévoyance, les salariés qui sont en arrêt de travail précédemment garanti par un autre assureur sont toujours majoritairement indemnisés par cet ancien assureur [3]. Il est possible de faire l'analogie avec les deux fonctionnements de déclenchement des garanties dans le cas des assurances responsabilité civile : base "fait dommageable" et "réclamation". En base "fait dommageable", la garantie s'enclenche dès lors que le fait à l'origine des dommages est survenu entre la date de prise d'effet et la date de résiliation de la garantie alors qu'en base "réclamation", la garantie est activée dès lors qu'il y a une déclaration de sinistre et même si le

3. Toutefois, ces salariés peuvent aussi bénéficier d'indemnités journalières complémentaires de la part du nouvel assureur, notamment si les garanties sont plus généreuses chez ce dernier.

fait générateur est antérieur à la souscription de la garantie [4]. Le fonctionnement en base "réclamation" existe notamment pour certaines assurances de responsabilité civile professionnelle.

Le deuxième arrêt de travail est lui non censuré et inclus dans le périmètre d'étude. Quant à l'arrêt de travail 3, la date de survenance est bien comprise dans la période d'observation mais la date de fin d'arrêt n'est pas précisée. Il peut, par exemple, s'agir d'un salarié dont l'entreprise a résilié son contrat de prévoyance auprès de Malakoff Humanis et dans ce cas-là, l'assureur ne dispose plus des DSN. Cet individu n'est pas retiré de l'étude et la date prévisionnelle de fin d'arrêt de travail est utilisée pour calculer la durée de l'arrêt.

Le quatrième et cinquième arrêts de travail sont survenus pendant la période d'observation mais se sont terminés après la fin du périmètre d'étude. Il s'agit donc de données censurées à droite. Afin de pouvoir calculer la durée réelle de ce type d'arrêt, un délai de trois mois a été appliqué. Autrement dit, pour ces arrêts-là, nous regardons s'ils se sont terminés dans les trois mois suivant la fin du périmètre d'étude. C'est le cas dans la situation 4 et la véritable durée de cet arrêt est donc connue. La date de fin de l'arrêt de travail 5 dépasse, quant à elle, ce délai et la durée réelle de cet arrêt n'est donc pas établie. Néanmoins, puisque le délai de 3 mois est dépassé, cet arrêt dure a minima 90 jours et appartient donc à la classe des arrêts de plus de 90 jours. Ainsi, le fait de catégoriser la variable permet de limiter les effets de la censure puisqu'il n'est pas nécessaire de connaître la durée exacte de tous les arrêts de travail. Enfin, le sixième arrêt de travail n'est quant à lui pas retenu dans l'étude puisqu'il est survenu après la fin du périmètre considéré.

Catégorisation de l'âge

L'âge est la seule variable continue parmi nos (futures) variables explicatives (âge, CSP, secteur d'activité, genre). Ces quatre variables correspondent aux critères de tarification déjà utilisés chez Malakoff Humanis en prévoyance collective [5]. L'impact de l'âge sur la sinistralité en arrêt de travail n'est pas le même en fonction de l'âge. Ainsi, les populations jeunes sont plus concernées par les arrêts de travail de courte durée alors que les personnes les plus âgées ont plus de risque de subir de longs arrêts de travail. C'est la raison pour laquelle, la variable âge est utilisée de manière catégorisée et non continue car cela permettra de capturer les différents impacts décrits ci-dessus. La première classification testée pour l'âge est : [15,20],]20,25],]25,30],]30,35],]35,40],]40,45],]45,50],]50,55],]55,60] et +60 ans. Le tableau A.1, en annexe A reporte les coefficients de la régression multinomiale modélisant le risque incapacité avec cette classification de l'âge [6]. Les coefficients associés aux modalités]30,35] et]35,40] d'une part et ceux liés aux modalités]40,45] et]45,50] d'autre part affichent une certaine proximité, ce qui signifie que leur sinistralité sont proches. C'est la raison pour laquelle il a été décidé de regrouper ces modalités afin de réduire le nombre de paramètres à estimer dans le modèle.

Par ailleurs, les coefficients associés à la modalité "+60" ans sont inférieurs à ceux estimés pour les modalités comprises entre 35 et 60 ans. Toutes choses égales par ailleurs, la sinistralité des individus de plus de 60 ans serait donc inférieure à celle des individus âgés de 35 à 60 ans. Plusieurs justifications peuvent expliquer ce phénomène.

— Des individus qui ne se rendent plus à leur travail mais qui ne sont pas encore en retraite

4. Cependant, en base "réclamation", la garantie ne s'enclenche pas si l'assuré avait connaissance du sinistre lors de la souscription du contrat.

5. Dans la tarification prévoyance collective, ce n'est pas le genre qui est utilisé mais la répartition homme/femme. En effet, un tarif différencié n'est pas proposé aux femmes et hommes d'une même entreprise mais deux entreprises avec une répartition homme/femme différentes peuvent ne pas avoir le même tarif. C'est le genre qui est ici utilisé car les données sont dans un premier temps à la maille individuelle.

6. Les détails autour de la régression multinomiale sont donnés dans le chapitre 3. Ici, nous comparons juste des coefficients pour potentiellement regrouper des modalités.

appartiennent à la classe des plus de 60 ans. Par exemple, les salariés qui liquident leurs congés avant leur retraite sont comptés dans les effectifs de l'entreprise sans toutefois y travailler (et ne sont donc pas en arrêt de travail pendant cette période).

— Il y a une sur-représentation des cadres parmi les individus âgés de plus de 60 ans (44% de plus par rapport à l'ensemble du portefeuille étudié) et une sous-représentation d'ouvriers dans cette tranche d'âge (43% de moins par rapport à l'ensemble du portefeuille). Les ouvriers sont une population sur-sinistrée en termes d'arrêts de travail alors que l'inverse est constaté pour les cadres [7]. Bien que la CSP soit incluse dans la modélisation, cela n'est peut être pas suffisant pour contrôler cet effet.

— La classe des "+60" ans comporte des personnes travaillant au-delà de leur date de départ légal en retraite. Bien que certaines personnes le fassent pour des raisons financières, d'autres individus continuent à travailler car ils ont une bonne qualité de vie professionnelle. Il est donc légitime de penser que ces derniers aient moins de risque d'avoir un arrêt de travail d'origine professionnelle (moins enclin à être en burn-out, etc.), ce qui expliquerait en partie la moindre sinistralité des plus de 60 ans.

Néanmoins, les individus de plus de 60 ans représentent moins de 1.5% du portefeuille et ce résultat sur la sinistralité de cette tranche d'âge doit être consolidé. Ainsi, par mesure de prudence dans la modélisation, il a été décidé de mutualiser le risque sur les tranches]55,60] et +60. La classification finale choisie pour l'âge est donc : [15,20],]20,25],]25,30],]30,40],]40,50],]50,55] et +55 ans.

Les étapes de préparation de données nécessaires à nos premiers travaux ont été détaillées. La section suivante présente donc des statistiques descriptives de nos données.

2.1.3 Analyse descriptive des données

Quelques statistiques descriptives sont fournies dans cette section. Elles sont limitées pour des raisons de confidentialité.

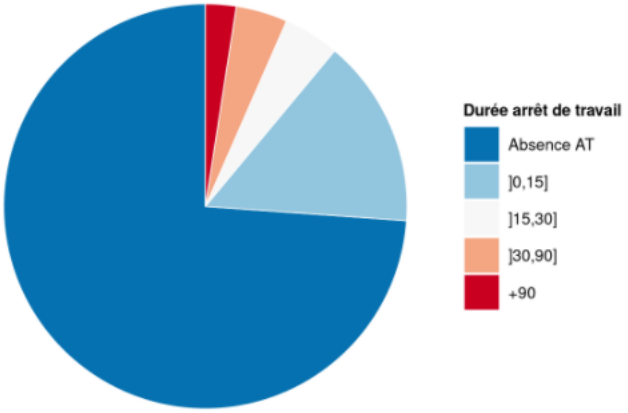

FIGURE 2.2 – Répartition des durées d'arrêts de travail sur tout le périmètre

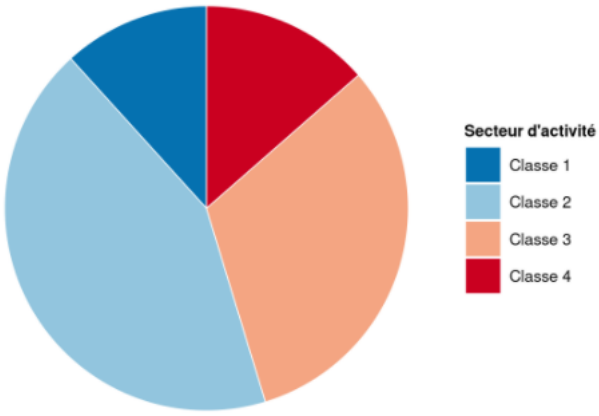

FIGURE 2.3 – Répartition des affiliés parmi les classes de risques "secteur d'activité"

Presque trois quarts des assurés Malakoff Humanis n'ont pas d'arrêt de travail au cours d'une année (modalité "Absence d'AT"). Environ 15% des individus étudiés subissent un arrêt de travail de moins de 15 jours et 10% un arrêt de plus de 15 jours. Les arrêts de travail supérieurs à 30 ou 90 jours, ceux qui nous intéressent le plus car correspondent aux franchises les plus fréquemment prévues dans les garanties des contrats, sont donc des évènements peu

7. Ce résultat sera abordé dans la section suivante.

fréquents. Les statistiques présentées ci-dessus sont basées sur tout le périmètre considéré [8]. Il faut garder en tête que ces fréquences varient énormément en fonction du profil des assurés. Ce sujet est évoqué plus loin.

La figure 2.3 représente la répartition des affiliés parmi les différentes classes de risques portées par le secteur d'activité. En effet, les modalités de "Classe 1" à "Classe 4" sont des classes de risques associées au code NAF [9] de l'entreprise. Cette classification est le résultat d'un travail précédent, développé en interne. Elle est exprimée par ordre croissant du risque, c'est-à-dire que la classe 1 contient les codes NAF les moins risqués et la classe 4 les plus risqués. Plus de 70% des assurés travaillent dans un secteur d'activité ayant un risque classé 2 ou 3.

Les figures 2.4, 2.5, 2.6 et 2.7 représentent la fréquence de l'évènement "avoir un arrêt de travail d'une certaine durée" en fonction de la CSP, du secteur d'activité, de l'âge et du genre. Il s'agit d'une première preuve que ces variables permettent de mieux segmenter le risque incapacité de travail temporaire.

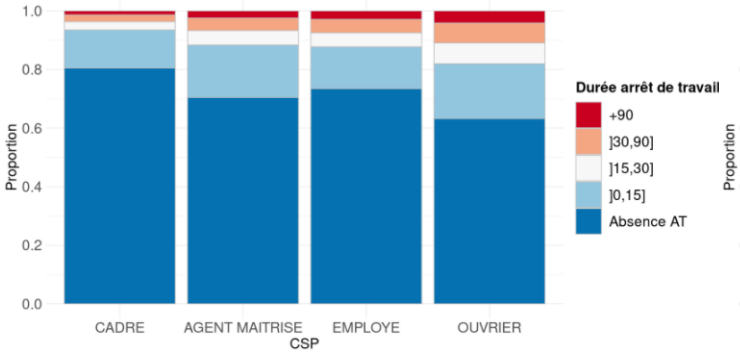

FIGURE 2.4 – Répartition des durées d'arrêts de travail selon la CSP

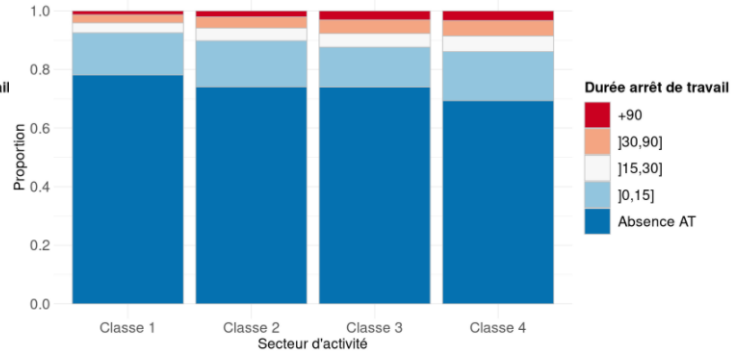

FIGURE 2.5 – Répartition des durées d'arrêts de travail selon le secteur d'activité

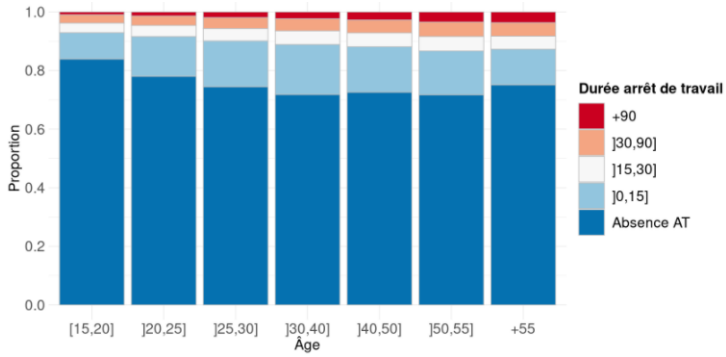

FIGURE 2.6 – Répartition des durées d'arrêts de travail selon l'âge

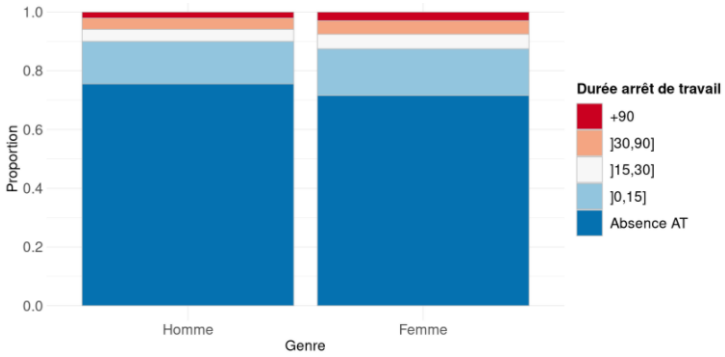

FIGURE 2.7 – Répartition des durées d'arrêts de travail selon le genre

Avant d'interpréter ces graphiques, il faut bien garder en tête qu'il s'agit d'une analyse bivariée qui ne tient donc pas compte des potentielles corrélations entre les variables. Par exemple, les hommes ont tendance à être sur-représentés parmi les cadres. Ainsi, bien que ces graphiques donnent une première idée de la sinistralité en fonction de plusieurs critères, il faut faire attention de ne pas en tirer de conclusions trop hâtives.

8. Tous les travaux de ce mémoire n'utilisent pas les données de 2020. Étant donné la pandémie de COVID-19 et les confinements, il est certain que ces statistiques soient très différentes pour cette année-là.
9. NAF : Nomenclature d'Activité Française.

Catégories socio-professionnelles Au vu des différences de sinistralité observées, la catégorie socio-professionnelle semble être un critère très discriminant. Les cadres seraient les moins touchés par l'arrêt de travail, et ce, quelle que soit la durée. 80% d'entre eux n'ont pas eu d'arrêt de travail au cours d'une année contre moins de 65% pour les ouvriers. Les ouvriers seraient la population la plus à risque, peu importe la durée considérée. Les agents de maîtrise et employés sont des profils intermédiaires en termes de sinistralité. Les agents de maîtrise présentent plus d'arrêts courts que les employés mais cela s'inverse pour les arrêts plus longs.

Secteurs d'activité Le secteur d'activité semble être un facteur moins discriminant que la CSP. Néanmoins, la classification des secteurs d'activité en classes de risques semble validée puisque les classes considérées comme les plus risquées sont aussi les plus sinistrées. Ceci est particulièrement vrai pour les arrêts de plus de 15 jours où la sinistralité augmente continuellement de la classe 1 à la classe 4. Ce lien est plus ambigu pour les arrêts de moins de 15 jours et pour la modalité associée au fait de ne pas avoir d'arrêt de travail ("Absence d'AT"). Cependant, pour cette dernière, là où presque 80% des assurés travaillant dans un secteur classé 1 ne subissent pas d'arrêt de travail, ils représentent moins de 70% dans les secteurs classés 4.

Âge Comme évoqué dans la section précédente, l'évolution de la sinistralité en fonction de l'âge n'est pas linéaire. Les populations jeunes sont plus touchées par les arrêts de travail de courte durée et les personnes plus âgées par les arrêts de longue durée. Ceci est confirmé par le graphique 2.6 où il est observé une hausse continue des arrêts de travail de plus de 30 jours avec l'âge, ce qui n'est pas le cas pour les arrêts plus courts. En effet, la fréquence d'arrêts de travail de moins de 15 jours est la plus élevée pour les populations âgées de 25 à 50 ans. Les personnes de plus de 50 ans présentent certes davantage d'arrêts longs mais moins d'arrêts courts. Quelle que soit la durée des arrêts, les moins de 20 ans sont clairement les moins sinistrés. Près de 85% d'entre eux ne subissent aucun arrêt de travail au cours d'une année.

Genre Les femmes semblent légèrement plus sinistrées que les hommes quelle que soit la durée d'arrêt de travail considérée. Par exemple, 10% des hommes subissent un arrêt de travail de plus de 15 jours au cours d'une année contre environ 14% des femmes [10].

Comme évoqué précédemment, ces graphiques ne tiennent pas compte des possibles corrélations entre les variables. Tous les résultats exposés devront donc être confirmés par un modèle dans un cadre multivarié (cf. chapitre 3). Maintenant que les données ont été présentées, il est possible de détailler la méthodologie mise en place pour construire le zonier. C'est l'objet de la section suivante.

2.2 Méthodologie mise en place pour élaborer le zonier

2.2.1 Déontologie actuarielle

Depuis le 25 mai 2018, le Règlement Général sur la Protection des Données (RGPD) [11] est en application dans l'Union Européenne. Il s'agit du texte de référence sur la protection des données à caractère personnel (DCP). Les principales DCP sont le nom, prénom et numéro de Sécurité Sociale mais le numéro de téléphone et l'adresse mail en font aussi partie. Tous ces renseignements ne sont pas nécessaires pour réaliser notre étude. La source de données mise à

10. Pour rappel, les congés maternité, paternité et adoption ont été retirés du périmètre d'étude et ne sont donc pas considérés pour ces statistiques.
11. En anglais, GDPR : General Data Protection Regulation.

notre disposition ne contient donc pas ces informations et un identifiant anonyme a été créé pour remplacer le nom-prénom.

Par ailleurs, il faut être vigilant de ne pas inclure de biais dans les données. Comme évoqué précédemment, les travaux présentés dans ce mémoire sont réalisés sur une partie importante et représentative du portefeuille de Malakoff Humanis.

Aussi, nous avons accès dans la DSN aux arrêts de travail dont la durée est inférieure au délai de franchise et qui sont donc non indemnisés par l'assureur. Les données de la DSN ne reflètent donc pas une vision prestation. Ainsi, grâce à l'exhaustivité des données de la DSN, aucun biais n'est introduit en ce qui concerne la durée des arrêts de travail. Toutefois, cette étude devra être étendue à l'ensemble du portefeuille pour améliorer la robustesse des résultats.

En résumé, les données utilisées dans le cadre de cette étude respectent globalement les 3 critères suivants :

— **Pertinence**. La DSN répertorie les arrêts de travail et c'est le risque étudié dans le présent mémoire.
— **Exhaustivité**. Par le caractère obligatoire de la DSN, les données sont complètes et l'accès aux arrêts de travail d'une durée inférieure au délai de franchise est un réel plus.
— **Exactitude**. Grâce aux contrôles effectués au moment de la déclaration de l'entreprise, et aux retraitements mis en place en interne, la qualité des données est jugée bonne.

2.2.2 Différentes étapes pour l'élaboration du zonier incapacité

Pour rappel, l'objectif des travaux présentés dans ce mémoire est de construire un zonier sur le risque incapacité temporaire de travail. Cette section détaille les différentes étapes mises en œuvre pour élaborer ce zonier. Un schéma récapitulatif de cette méthodologie est présenté en Figure 2.8.

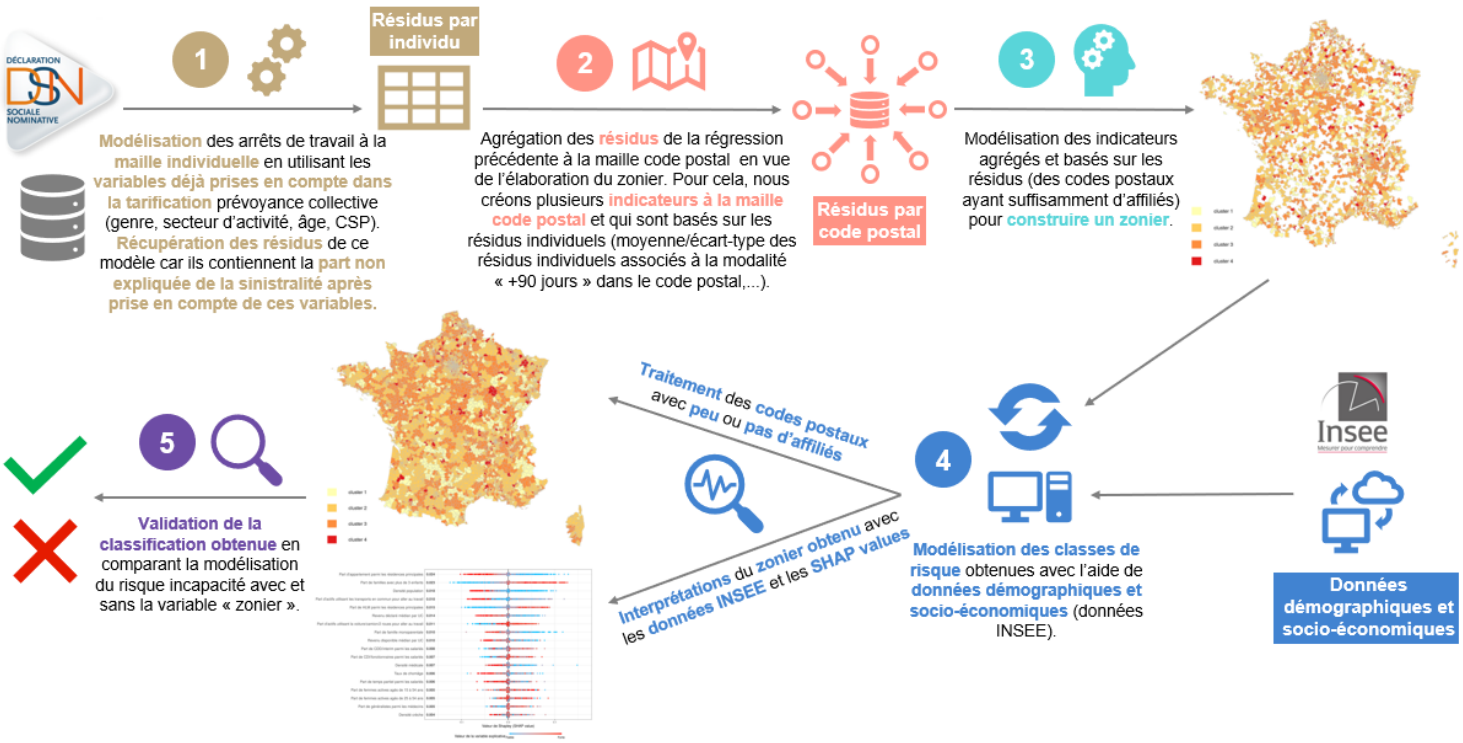

FIGURE 2.8 – Méthodologie d'élaboration du zonier

De la même façon que (Pariente, 2017)[9] et (Beraud-Sudreau, 2017)[8], le zonier incapacité est élaboré à partir des résidus d'une régression qui prend en compte les variables déjà utilisées dans la tarification (âge, CSP, genre et secteur d'activité) [12]. Cette modélisation du risque incapacité se fait à la maille individuelle et le modèle utilisé est une régression multinomiale puisque la variable à expliquer est catégorisée en 5 modalités ("Absence d'AT", "]0,15]", "]15,30]", "]30,90]" et "+90"). Les résidus de cette régression contiennent la part non expliquée de la sinistralité après prise en compte de l'âge, de la CSP, du genre et du secteur d'activité. L'effet, s'il existe, de la localisation de l'entreprise sur la sinistralité est donc présent dans ces résidus et c'est la raison pour laquelle ils sont utilisés. Par exemple, un résidu positif sur la modalité «+90» signifie que la sinistralité pour les arrêts longs est sous-estimée pour cet individu.

La deuxième étape consiste à agréger les résidus à la maille code postal en vue de l'élaboration du zonier. En effet, l'unité géographique retenue est le code postal de l'établissement de l'entreprise car c'est une donnée présente dans la DSN et pouvant être facilement demandée aux clients au moment de la tarification. Plusieurs indicateurs basés sur les résidus individuels sont donc créés à la maille "code postal de l'entreprise". Par exemple, la moyenne et l'écart-type des résidus individuels associés à la modalité "+90" jours dans le code postal sont deux de ces indicateurs. Au vu du nombre de variables créées, une sélection est réalisée avant de modéliser le zonier. Ainsi, davantage d'indicateurs basés sur les modalités associées aux arrêts les plus longs ont été retenus puisque c'est la survenance d'arrêts de travail longs (ceux qui dépassent le délai de franchise) qui représente le risque le plus important dans les garanties incapacité temporaire de travail.

La troisième partie de la méthodologie traite de la construction du zonier incapacité à partir des indicateurs sélectionnés précédemment. Au lieu d'avoir recours à un classique lissage spatial d'un seul indicateur, la méthode utilisée dans cette étude est une classification ascendante hiérarchique avec contraintes de proximité géographique, basée sur plusieurs indicateurs. Cette modélisation n'est réalisée que sur les codes postaux ayant suffisamment d'affiliés et l'objectif de cette méthode est de tenir compte de la proximité géographique dans la constitution des clusters, qui deviendront les futures classes de risques associées à la localisation de l'entreprise.

La quatrième phase a pour but de conduire une analyse du zonier et d'en dégager des interprétations. Pour cela, une modélisation des classes de risques obtenues précédemment est réalisée avec l'aide de données démographiques et socio-économiques de l'INSEE. L'objectif de ce modèle est double :

- Interpréter, avec l'aide des SHAP [13] values, le clustering spatial obtenu précédemment. Cela permettra de comprendre si l'appartenance d'un code postal à une certaine classe de risque est corrélée avec son niveau de richesse, ses équipements, etc.
- Prédire la classe de risque pour les codes postaux avec peu ou pas d'affiliés, en utilisant leurs caractéristiques démographiques et socio-économiques, tout en tenant compte de la proximité géographique.

Pour finir, une étape de validation de la classification obtenue est nécessaire. À ce titre, une comparaison de la modélisation du risque incapacité avec et sans la variable «zonier» est effectuée.

12. Dans la tarification prévoyance collective, ce n'est pas le genre qui est utilisé mais la répartition homme/femme. En effet, un tarif différencié n'est pas proposé aux femmes et hommes d'une même entreprise. Par contre, deux entreprises avec une répartition homme/femme différentes peuvent ne pas avoir le même tarif.

13. SHapley Additive exPlanations.

Toutes les applications, ainsi que les résultats présentés dans ce mémoire ont été faits en utilisant le langage de programmation R [14]. En effet, plusieurs packages R facilitant l'utilisation de données spatiales en ayant recours à des fonctions spécialement conçues pour ce type de données, sont disponibles en libre-accès. Ceci facilite la mise en pratique des travaux théoriques dans ce domaine, ce qui constitue la raison principale de ce choix de langage.

Cette première partie a permis de détailler le contexte de l'étude présentée dans le cadre de ce mémoire. La partie suivante se concentre sur les deux premières étapes de la méthodologie, à savoir la modélisation du risque incapacité avec les critères usuels de tarification.

14. `https://www.r-project.org/`.

Deuxième partie

Modélisation du risque incapacité avec les critères classiques de tarification

Chapitre 3

Modélisation du risque incapacité avec une régression multinomiale

Dans le cadre de ce mémoire, la modélisation du risque incapacité avec les critères classiques de tarification se fait par le biais d'un modèle linéaire généralisé, et plus particulièrement par une régression logistique multinomiale. Les modèles linéaires généralisés (GLM) dont la régression logistique multinomiale sont donc introduit d'un point de vue théorique dans une première section, avant de détailler l'application sur nos données, qui sera elle présentée dans un second temps.

3.1 Modèles linéaires généralisés

3.1.1 Généralités autour des GLM

Les modèles linéaires généralisés, plus connus sous l'acronyme anglais GLM [1], sont une généralisation de la régression linéaire. Ils permettent notamment une relation non linéaire entre les variables explicatives et l'espérance conditionnelle et relâchent l'hypothèse de normalité de la variable à expliquer. Ils sont particulièrement répandus en actuariat et plus spécifiquement en tarification grâce à leur facilité d'interprétation.

Quelques notations sont introduites ci-après. Soient Y le vecteur des valeurs de la variable à expliquer pour les n individus et X une matrice de taille $n \times p$ correspondant aux valeurs des $X_1, ..., X_p$ variables explicatives pour chaque individu.
Un GLM se définit par deux hypothèses :

- $(Y_i)_i$ sont indépendants et Y_i suit une distribution appartenant à la famille exponentielle.
- Une fonction de lien g différentiable, monotone et inversible qui décrit comment la moyenne de la variable cible $\mu = \mathbb{E}[Y|X]$ est reliée à une combinaison linéaire des prédicteurs $\eta = X'\beta$: $\eta = g(\mu)$. Autrement dit, $\forall i \in \{1, ..., n\}, \mu_i = \mathbb{E}[Y_i|X_i] = g^{-1}(\eta_i) = g^{-1}(X_i'\beta)$.

Pour rappel, la famille exponentielle rassemble les lois de probabilité dont la densité peut s'écrire [2] :

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

θ est appelé paramètre naturel (ou canonique), $a(.), b(.)$ et $c(.)$ sont des fonctions et ϕ peut être qualifié de paramètre de nuisance.

1. GLM : Generalized Linear Models.
2. Une définition plus générale existe mais n'est pas nécessaire dans le cadre des GLM.

La majorité des lois usuelles (Normale, Bernoulli, Binomiale, Poisson, Gamma, etc.) appartiennent à cette famille. Par exemple, la loi Bernoulli(p) correspond au cas $\theta = \log(\frac{p}{1-p}), \phi = 1, a(\phi) = 1, b(\theta) = \log(1 + \exp(\theta)), c(y, \phi) = 0$ et est associée au modèle logistique.

L'estimation des paramètres dans un GLM se fait par la méthode du maximum de vraisemblance. La (log)-vraisemblance des modèles de la famille exponentielle est donnée par :

$$\log \mathcal{L}(\theta_1, ..., \theta_n, \phi, y_1, ..., y_n) = \sum_{i=1}^{n} \Big[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \Big].$$

Pour déterminer les paramètres β, les équations du score sont nécessaires. Elles s'obtiennent en dérivant la log-vraisemblance par rapport à β. Néanmoins, ces équations ne se résolvent pas analytiquement et il faut avoir recours à un algorithme appelé "Iterative Weighted Least Square" (IWLS) pour obtenir les paramètres. Pour davantage de détails sur la théorie générale autour des GLM, le lecteur pourra se reporter à (McCullagh et Nelder, 1989)[1].

Bien que très utilisé dans le domaine de l'actuariat, les modèles linéaires généralisés présentent le défaut d'imposer la forme de la loi conditionnelle de la variable à expliquer en fonction des variables explicatives. C'est une hypothèse forte et un risque de modèle est pris lorsque les GLM sont utilisés.

Comme évoqué précédemment, dans notre étude, la variable à expliquer est catégorisée en 5 modalités, d'où le recours à la régression logistique multinomiale. La théorie autour de ce type de régression est détaillée dans la sous-section suivante.

3.1.2 Régression logistique multinomiale

Les notations instaurées précédemment sont conservées et de nouvelles sont introduites ci-dessous.

Soit m le nombre de modalités de la variable cible ($m = 5$ dans cette étude). Les modalités de Y sont nommées $a_1, ..., a_m$. L'objectif d'une régression logistique multinomiale est d'estimer pour tout $k \in \{1, ..., m\}$ la probabilité :

$$p_k(x) = \mathbb{P}(\{Y = a_k\}|\{(X_1, ..., X_p) = x\}), x = (x_1, ..., x_p).$$

L'hypothèse fondamentale de ce modèle est que pour tout $k \in \{2, ..., m\}$,

$$\ln\left(\frac{p_k(x)}{p_1(x)}\right) = \beta_0^{(k)} + \beta_1^{(k)}x_1 + ... + \beta_p^{(k)}x_p \tag{3.1}$$

avec $\beta_0^{(k)}, ..., \beta_p^{(k)}$ des coefficients inconnus que l'on va chercher à estimer.
a_1 joue le rôle de la modalité de référence et chaque coefficient s'interprète par rapport à cette modalité. D'ailleurs, il est important de noter que dans ce modèle, nous obtenons un coefficient par variable et par modalité.

Cette hypothèse fondamentale (équation 3.1) peut aussi s'écrire : $\forall k \in \{2, ..., m\}$

$$p_k(x) = \frac{\exp\left(\beta_0^{(k)} + \beta_1^{(k)}x_1 + ... + \beta_p^{(k)}x_p\right)}{1 + \sum_{k=2}^{m} \exp\left(\beta_0^{(k)} + \beta_1^{(k)}x_1 + ... + \beta_p^{(k)}x_p\right)} \tag{3.2}$$

L'équivalence des équations 3.1 et 3.2 est démontrée en Annexe B.

Remarques :

1. La somme des probabilités étant égale à 1, on a : $p_1(x) = 1 - \sum_{k=2}^{m} p_k(x)$.
2. Le cas $m = 2$ revient à la régression logistique classique (binaire).

Pour estimer les paramètres de ce modèle, il faut utiliser la méthode du maximum de vraisemblance. La vraisemblance de ce modèle s'écrit :

$$\mathcal{L}(\beta_0^{(2)}, ..., \beta_p^{(2)}, ..., \beta_0^{(m)}, ..., \beta_p^{(m)}, y_1, ..., y_n) = \prod_{i=1}^{n} \prod_{k=1}^{m} p_k(x_i)^{\mathbb{1}_{y_i = a_k}}$$

La régression logistique multinomiale appartient à la famille des modèles linéaires généralisés et il n'existe pas de solution analytique pour obtenir une estimation des coefficients. Il faut donc utiliser l'algorithme IWLS pour estimer les paramètres.

Ce modèle suppose que les modalités de la variable à expliquer ne sont pas ordonnées. Dans notre étude, un certain ordre se dégage puisque la variable cible est catégorisée selon la durée des arrêts de travail. L'utilisation de ce modèle dans notre contexte peut donc être considérée comme une limite. Une piste pour améliorer la modélisation serait de tester un modèle faisant l'hypothèse que les modalités de Y sont ordonnées, comme la régression logistique ordinale (ou régression polytomique ordonnée) et de comparer les résultats avec ceux obtenus par la régression logistique multinomiale.

Cette section a permis d'expliciter quelques éléments théoriques autour des GLM et plus particulièrement de la régression logistique multinomiale. La section suivante est consacrée à l'application de ce modèle dans le cadre de l'incapacité temporaire de travail.

3.2 Application de la régression logistique multinomiale

La section 3.2 présente les résultats et interprétations de la régression logistique multinomiale appliquée à nos données.

3.2.1 Interprétation des coefficients

Ce premier modèle est une régression de la sinistralité incapacité sur les variables déjà utilisées en tarification prévoyance collective (âge, CSP, genre et secteur d'activité). Étant donné que cette modélisation se fait à la maille individuelle, la variable "genre" est utilisée au lieu de la variable "répartition homme/femme". En tarification collective, il est autorisé de tarifer selon cette répartition alors qu'il est interdit de le faire en assurance individuelle. La variable explicative "genre" est conservée dans notre modélisation afin que l'effet de cette variable sur la sinistralité ne soit pas présent dans les résidus.

Pour rappel, toutes les variables explicatives de ce modèle sont des variables catégorielles. La liste des modalités de chaque variable est donnée ci-dessous, avec en gras la modalité de référence.

— Genre : **Homme**, Femme
— Âge : **[15,20]**,]20,25],]25,30],]30,40],]40,50],]50,55] et +55 ans
— CSP : **Cadre**, Agent de maîtrise, Employé, Ouvrier
— Secteur d'activité : **Classe 1**, Classe 2, Classe 3, Classe 4

L'hypothèse fondamentale de la régression multinomiale (équation 3.1) définit les différentes équations du modèle. Par exemple, en choisissant "Absence d'AT" comme modalité de référence pour la variable cible ($k = 1$), l'équation associée à la modalité "+90" ($k = 5$) est :

$$\begin{aligned}\ln\left(\frac{p_5(x)}{p_1(x)}\right) &= \beta_0^{(5)} + \beta_1^{(5)}\mathbb{1}_{\text{Femme}} + \beta_2^{(5)}\mathbb{1}_{\text{Âge }]20,25]} + \beta_3^{(5)}\mathbb{1}_{\text{Âge }]25,30]} + \beta_4^{(5)}\mathbb{1}_{\text{Âge }]30,40]} \\ &+ \beta_5^{(5)}\mathbb{1}_{\text{Âge }]40,50]} + \beta_6^{(5)}\mathbb{1}_{\text{Âge }]50,55]} + \beta_7^{(5)}\mathbb{1}_{\text{Âge }+55} + \beta_8^{(5)}\mathbb{1}_{\text{CSP Agent maîtrise}} \\ &+ \beta_9^{(5)}\mathbb{1}_{\text{CSP Employé}} + \beta_{10}^{(5)}\mathbb{1}_{\text{CSP Ouvrier}} + \beta_{11}^{(5)}\mathbb{1}_{\text{Secteur d'activité Classe 2}} \\ &+ \beta_{12}^{(5)}\mathbb{1}_{\text{Secteur d'activité Classe 3}} + \beta_{13}^{(5)}\mathbb{1}_{\text{Secteur d'activité Classe 4}} \end{aligned} \tag{3.3}$$

Chaque coefficient associé à la modalité "+90" s'interprète donc en fonction de la modalité "Absence d'AT". Le lien entre ces coefficients et les odds ratios est détaillé ci-dessous dans le cas d'une variable explicative binaire X. Soient $Z = (Z_1, ..., Z_{p-1})$ les autres variables explicatives du modèle. Dans le cas de la modalité de référence "Absence d'AT", l'odds (cote) est définie par :

$$\text{Odds}(Y = +90|X, Z_1, ..., Z_{p-1}) = \frac{\mathbb{P}(Y = +90|X, Z_1, ..., Z_{p-1})}{\mathbb{P}(Y = \text{Absence d'AT}|X, Z_1, ..., Z_{p-1})} = \frac{p_5(x)}{p_1(x)}$$

L'odds ratio (rapport des cotes) de la variable X ajustée par les covariables $Z_1, ..., Z_{p-1}$ se définit par :

$$\begin{aligned}\text{Odds Ratio}_{\text{Femme/Homme}}(Y = +90|X, Z_1, ..., Z_{p-1}) &= \frac{\text{Odds}(Y = +90|X = 1, Z_1, ..., Z_{p-1})}{\text{Odds}(Y = +90|X = 0, Z_1, ..., Z_{p-1})} \\ &= \frac{\dfrac{\mathbb{P}(Y = +90|X = 1, Z_1, ..., Z_{p-1})}{\mathbb{P}(Y = \text{Absence d'AT}|X = 1, Z_1, ..., Z_{p-1})}}{\dfrac{\mathbb{P}(Y = +90|X = 0, Z_1, ..., Z_{p-1})}{\mathbb{P}(Y = \text{Absence d'AT}|X = 0, Z_1, ..., Z_{p-1})}}\end{aligned}$$

Le rapport des cotes est une mesure d'association entre X et Y. Pour cela, il suffit de le comparer à 1. S'il est supérieur (respectivement inférieur) à 1, il y a une dépendance positive (respectivement négative) entre X et Y. Dans le cas où il vaut 1, les deux variables ne sont pas liées.
En prenant l'exemple de la variable "genre" pour X et puisque $\text{Odds}(Y = +90|X, Z) = \dfrac{p_5(x)}{p_1(x)}$, l'odds ratio précédent peut se réécrire :

$$\begin{aligned}\frac{\text{Odds}(Y = +90|X = 1, Z)}{\text{Odds}(Y = +90|X = 0, Z)} &= \frac{\exp(\beta_0^{(5)} + \beta_1^{(5)} + \beta_2^{(5)}\mathbb{1}_{\text{Âge }]20,25]} + ... + \beta_{13}^{(5)}\mathbb{1}_{\text{Secteur d'activité Classe 4}})}{\exp(\beta_0^{(5)} + \beta_1^{(5)} \times 0 + \beta_2^{(5)}\mathbb{1}_{\text{Âge }]20,25]} + ... + \beta_{13}^{(5)}\mathbb{1}_{\text{Secteur d'activité Classe 4}})} \\ &= \exp(\beta_1^{(5)})\end{aligned}$$

En résumé, après passage à l'exponentielle, les coefficients de la régression logistique multinomiale peuvent s'interpréter comme des odds ratios. Ainsi, pour tout $l \in \{1, ..., p\}$, lorsque $\exp(\beta_l^{(5)}) > 1$, le risque d'avoir un arrêt de travail de plus de 90 jours est plus important pour la modalité de CSP, d'âge, de genre ou de secteur d'activité associée à $\beta_l^{(5)}$ par rapport à la modalité de référence correspondante. Inversement si $\exp(\beta_l^{(5)}) < 1$. Si $\exp(\beta_l^{(5)}) = 1$, le risque d'avoir une longue incapacité de travail est le même entre la modalité associée au coefficient et la modalité de référence de la variable. Il est aussi possible d'interpréter les coefficients sans

passage à l'exponentielle. Dans ce cas, il faut faire le même raisonnement que précédemment en comparant avec 0 au lieu de 1. Des exemples d'interprétations des coefficients sont donnés plus loin.

La manière d'interpréter les coefficients du modèle a été explicitée pour les arrêts les plus longs (modalité "+90" jours) afin de faciliter la compréhension mais elle peut tout à fait se généraliser sur l'ensemble des modalités de la variable à expliquer. D'ailleurs, des interprétations sont données ci-après sur les arrêts de travail ayant une durée inférieure à 90 jours.

Le tableau 3.1 reporte les coefficients obtenus par la régression multinomiale décrite plus haut.

	Dependent variable :			
	]0,15]	]15,30]	]30,90]	+90
Femme	0.217***	0.282***	0.271***	0.398***
	(0.004)	(0.007)	(0.007)	(0.009)
Âge]20,25]	0.509***	0.324***	0.345***	0.604***
	(0.012)	(0.020)	(0.022)	(0.037)
Âge]25,30]	0.739***	0.581***	0.655***	1.130***
	(0.012)	(0.020)	(0.021)	(0.035)
Âge]30,40]	0.858***	0.731***	0.813***	1.390***
	(0.012)	(0.019)	(0.020)	(0.034)
Âge]40,50]	0.746***	0.712***	0.847***	1.533***
	(0.012)	(0.019)	(0.020)	(0.033)
Âge]50,55]	0.707***	0.744***	0.940***	1.738***
	(0.012)	(0.020)	(0.021)	(0.034)
Âge +55	0.448***	0.583***	0.819***	1.735***
	(0.013)	(0.020)	(0.021)	(0.034)
CSP Agent maîtrise	0.488***	0.704***	0.731***	0.708***
	(0.006)	(0.010)	(0.011)	(0.015)
CSP Employé	0.236***	0.642***	0.793***	0.888***
	(0.005)	(0.009)	(0.010)	(0.012)
CSP Ouvrier	0.705***	1.236***	1.357***	1.409***
	(0.006)	(0.010)	(0.011)	(0.014)
Secteur d'activité Classe 2	0.015**	0.047***	0.079***	0.152***
	(0.006)	(0.012)	(0.013)	(0.018)
Secteur d'activité Classe 3	−0.161***	0.034***	0.201***	0.454***
	(0.007)	(0.012)	(0.013)	(0.018)
Secteur d'activité Classe 4	0.103***	0.239***	0.367***	0.547***
	(0.007)	(0.013)	(0.014)	(0.020)
Constante	−2.618***	−4.188***	−4.566***	−5.984***
	(0.013)	(0.021)	(0.023)	(0.037)
Akaike Inf. Crit.	4,020,985.000	4,020,985.000	4,020,985.000	4,020,985.000
Note :				*p<0.1 ; **p<0.05 ; ***p<0.01

TABLE 3.1 – Coefficients obtenus avec la régression multinomiale

Au vu du nombre important de coefficients estimés, il serait fastidieux de tous les analyser. Afin de simplifier l'interprétation du modèle, des représentations graphiques sont présentées dans la section suivante. Toutefois, des exemples d'interprétations de coefficients sont donnés

ci-dessous.

De manière générale, la majorité des coefficients sont positifs. Ceci vient du fait que pour chaque variable explicative, la modalité de référence est celle la moins risquée. De plus, les coefficients sont tous significatifs à 5%, signe de la pertinence des critères de tarification déjà utilisés. Par ailleurs, le test d'adéquation de la déviance sur ce modèle conclut que cette modélisation est adaptée.

Genre Les coefficients associés à la modalité "Femme" sont positifs, signe que les femmes sont une population plus à risque que les hommes sur l'incapacité temporaire de travail, et ce, quelle que soit la durée des arrêts.

Âge L'âge est plus difficile à interpréter. La tranche d'âge la plus risquée pour les arrêts courts est $]30, 40]$ alors que ce sont les personnes les plus âgées qui subissent le plus d'arrêts longs.

Catégories socio-professionnelles L'ensemble des coefficients associés à la CSP sont positifs, ce qui signifie que les cadres sont la population la moins à risque quelle que soit la durée des arrêts. De plus, les agents de maîtrise ont des coefficients supérieurs à ceux des employés pour les modalités d'arrêts courts alors que l'inverse est estimé pour les arrêts longs. Bien que l'objectif de ce mémoire ne soit pas d'étudier précisément la durée des arrêts de travail selon les CSP, ce résultat demeure très intéressant et pas forcément attendu. Les ouvriers sont quant à eux la population la plus sinistrée sur toutes les durées d'arrêts de travail.

Par exemple, toutes choses égales par ailleurs, la probabilité pour un ouvrier d'être plus de 90 jours en arrêt de travail par rapport à celle de ne pas être en arrêt de travail est 4.0919 ($e^{1.409}$) fois (soit 309,19% plus élevé) l'odds (côte) pour un cadre. En résumé, si un certain profil de cadres a un odds de 0.03 alors cet odds pour le même profil, sauf pour un ouvrier est de $0.03 \times 4.0919 \approx 0.12$. Ainsi, pour ce cadre, la probabilité d'être en arrêt plus de 90 jours représente 3% de celle d'absence d'arrêt de travail alors qu'elle représente 12% pour cet ouvrier. Autrement dit, ce cadre a une probabilité de ne pas être en arrêt de travail 33.33 (1/0.03) fois supérieure à celle d'avoir un arrêt de plus de 90 jours alors que pour l'ouvrier, elle n'est que 8.33 fois supérieure.

Secteurs d'activité Pour rappel, les modalités de 1 à 4 pour cette variable sont des classes de risques associées au code NAF de l'entreprise. La classification est exprimée par ordre croissant du risque, c'est-à-dire que la classe 1 contient les codes NAF les moins risqués et la classe 4 les plus risqués. Les coefficients confirment ce résultat puisque les classes les plus risquées ont globalement des coefficients supérieurs à celles moins risquées. Par exemple, toutes choses égales par ailleurs, la probabilité pour un individu dont le secteur d'activité est classé 4 d'être plus de 90 jours en arrêt de travail par rapport à celle de ne pas être en arrêt est 1.728 ($e^{0.547}$) fois (soit 72,8% plus élevé) l'odds (côte) pour un individu dont le secteur d'activité est classé 1.

Jusque-là, les coefficients ont été interprétés par rapport aux modalités de référence. Des exemples hors modalités de référence sont présentés ci-dessous.

Interprétation de coefficients en n'utilisant pas la modalité de référence de la variable explicative Au lieu de comparer les employés avec les cadres (modalité de référence), il peut être intéressant de mesurer la différence de sinistralité sur les arrêts de 30-90 jours entre les employés et les ouvriers. Toutes choses égales par ailleurs, la probabilité pour un ouvrier

d'avoir un arrêt de travail de 30-90 jours par rapport à celle de ne pas être en arrêt de travail est $e^{1.357-0.793} = e^{0.564} = 1.7577$ fois (soit 75.77% plus élevé) l'odds d'un employé.

Interprétation de coefficients en n'utilisant pas la modalité de référence de la variable à expliquer "Absence d'AT" Enfin, il est possible de faire une interprétation sans comparer les coefficients à la modalité de référence de la variable à expliquer. Par exemple, la probabilité pour un ouvrier d'être plus de 90 jours en arrêt de travail par rapport à celle d'être entre 0 et 15 jours en arrêt est $e^{1.409-0.705} = e^{0.704} = 2.0218$ fois la cote (soit 102.18% plus élevé) d'un cadre, ceteris paribus.

3.2.2 Impact des différentes variables dans l'estimation du risque incapacité

Cette partie expose plusieurs graphiques qui permettent de mieux visualiser les impacts des différentes variables explicatives sur la sinistralité. L'idée de chaque figure est de figer toutes les variables explicatives excepté une que l'on fait varier et de prédire la sinistralité sur chacune des modalités de cette variable. Cela permet d'observer, pour un profil d'individu donné [3], l'évolution de la sinistralité prédite en fonction de la variable étudiée.

Âge Les deux premiers profils étudiés (Figure 3.1 et 3.2) sont ceux d'une femme et d'un homme ouvriers travaillant dans un secteur de classe 4, classe la plus risquée. Ces profils font partie des populations les plus à risque concernant l'arrêt de travail. Les figures 3.3 et 3.4 sont associées à un profil peu risqué et les graphiques 3.5 et 3.6 sont plutôt des profils à niveau de risque intermédiaire. Quels que soient les profils, la même dynamique de dérive de la sinistralité est observée en fonction de l'âge. En effet, une augmentation des arrêts courts est prédite entre les âges 20 et 40 ans, pour ensuite re-diminuer après 40 ans. Cette évolution ne se retrouve pas pour les arrêts longs où la prédiction des arrêts de plus de 30 jours augmente de façon continue avec l'âge. Bien que présent dans tous les profils, l'amplitude de ces effets diffère en fonction du risque porté par chaque profil. Dans les profils les plus risqués (Figures 3.1 et 3.2), cette dynamique est très marquée alors qu'elle l'est moins pour les profils intermédiaires (Figures 3.5 et 3.6), et encore moins pour le profil peu risqué (Figures 3.3 et 3.4).

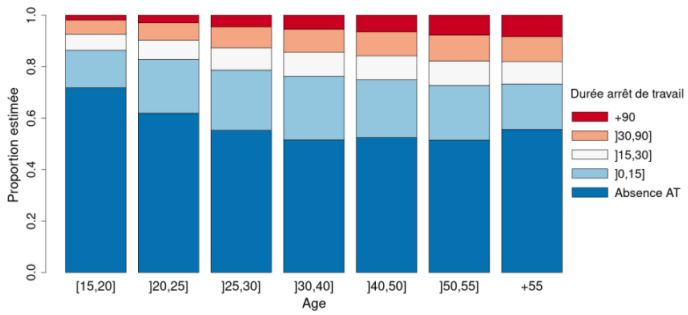

FIGURE 3.1 – Évolution de la sinistralité prédite en fonction de l'âge pour une femme ouvrière travaillant dans un secteur très risqué

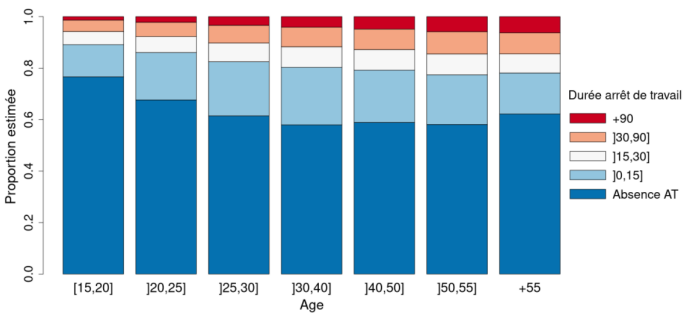

FIGURE 3.2 – Évolution de la sinistralité prédite en fonction de l'âge pour un homme ouvrier travaillant dans un secteur très risqué

3. Ce profil correspond aux variables fixées au préalable.

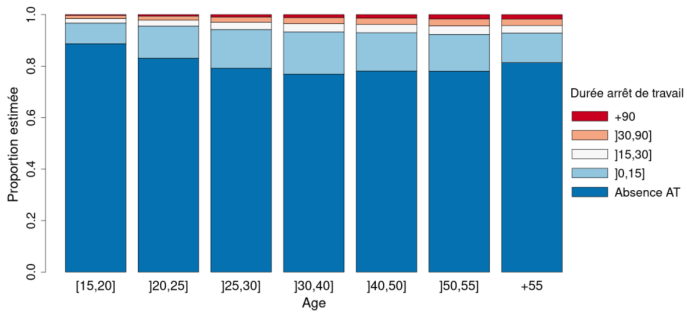

FIGURE 3.3 – Évolution de la sinistralité prédite en fonction de l'âge pour une femme cadre travaillant dans un secteur très peu risqué (classe 1)

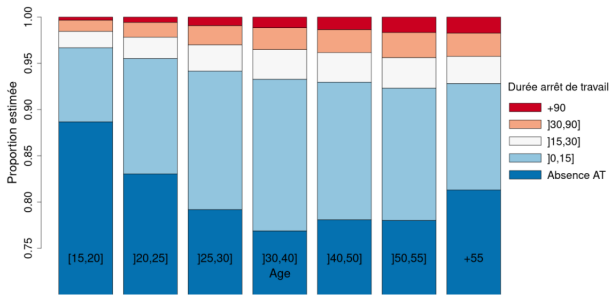

FIGURE 3.4 – Évolution de la sinistralité prédite en fonction de l'âge pour une femme cadre travaillant dans un secteur très peu risqué (classe 1) (zoom)

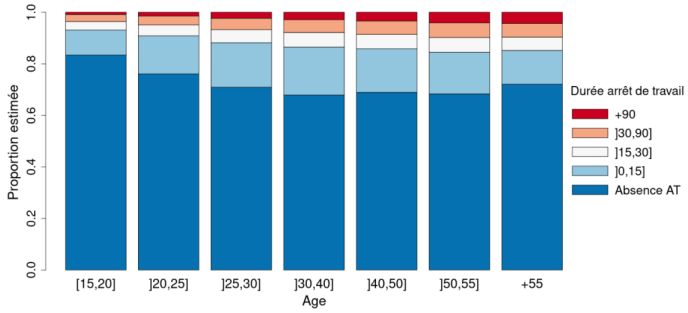

FIGURE 3.5 – Évolution de la sinistralité prédite en fonction de l'âge pour une femme employée travaillant dans un secteur peu risqué (classe 2)

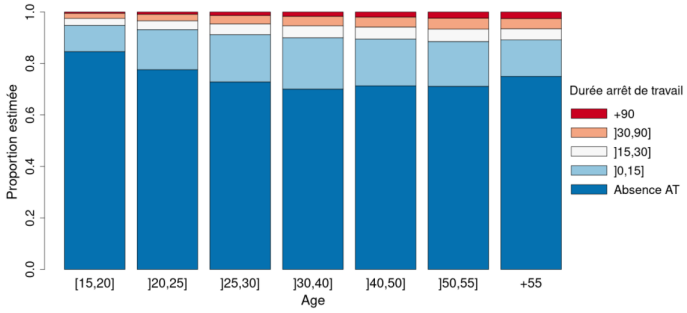

FIGURE 3.6 – Évolution de la sinistralité prédite en fonction de l'âge pour un homme agent de maîtrise travaillant dans un secteur peu risqué (classe 2)

Catégories socio-professionnelles Quelle que soit la durée des arrêts de travail considérée, les cadres sont la catégorie socio-professionnelle la moins risquée et les ouvriers représentent la plus risquée. Comme évoqué dans la section précédente, la situation est plus ambiguë pour hiérarchiser la sinistralité prédite entre les employés et les agents de maîtrise. Ces derniers porteraient plus de risque que les employés sur les arrêts courts (moins de 30 jours) mais moins sur les arrêts longs (plus de 30 jours).

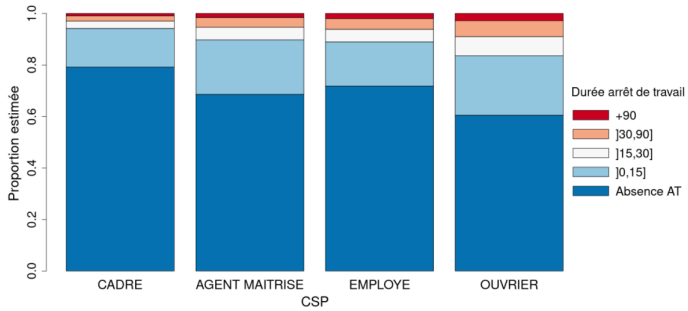

FIGURE 3.7 – Évolution de la sinistralité prédite en fonction de la CSP pour une femme âgée de 30 ans travaillant dans un secteur très peu risqué (classe 1)

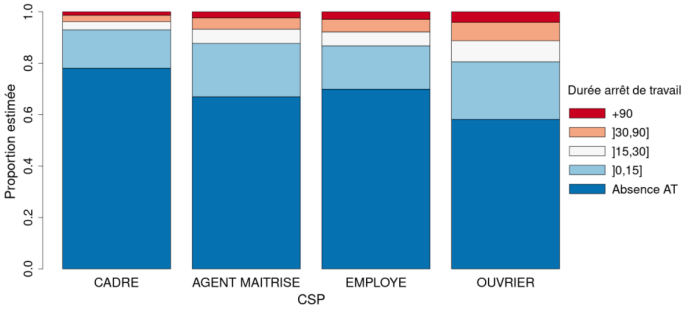

FIGURE 3.8 – Évolution de la sinistralité prédite en fonction de la CSP pour une femme âgée de 50 ans travaillant dans un secteur très peu risqué (classe 1)

Secteur d'activité Le secteur d'activité semble être un critère moins discriminant que les autres, notamment pour les classes de risques 1 et 2. Cependant, le fait de travailler dans un secteur d'activité classé 3 ou 4 augmente la probabilité d'avoir un arrêt de plus de 30 jours. De la même façon que pour les variables précédentes, l'amplitude de cet effet est plus important pour les profils plus à risque (figure 3.10) que dans le cas de profils peu risqués (figure 3.9).

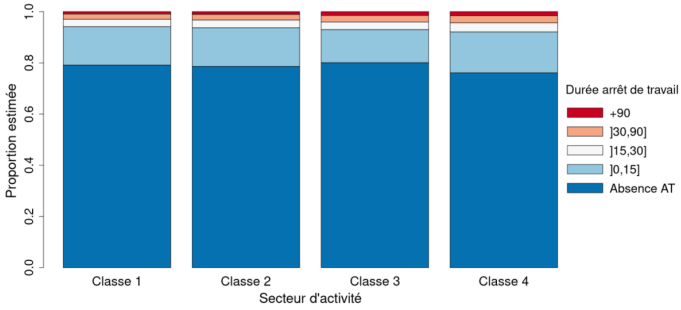

FIGURE 3.9 – Évolution de la sinistralité prédite en fonction du secteur d'activité pour une femme cadre âgée de 30 ans

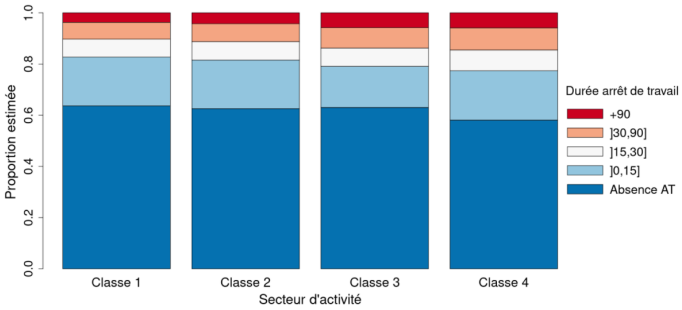

FIGURE 3.10 – Évolution de la sinistralité prédite en fonction du secteur d'activité pour un homme ouvrier âgé de 55 ans

Genre Les femmes sont une population plus à risque que les hommes en ce qui concerne le risque incapacité de travail temporaire. Cet effet est plus marqué dans la figure 3.12 (profil risqué) que dans la figure 3.11 (profil peu risqué).

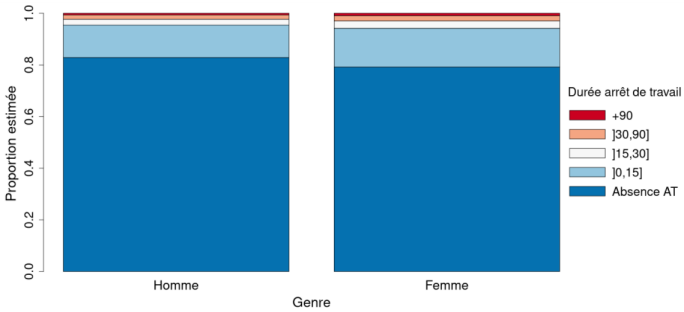

FIGURE 3.11 – Évolution de la sinistralité prédite en fonction du genre pour une personne cadre âgée de 30 ans travaillant dans un secteur peu risqué (classe 1)

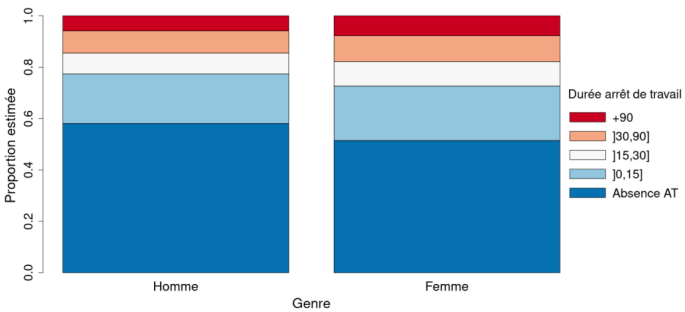

FIGURE 3.12 – Évolution de la sinistralité prédite en fonction du genre pour une personne ouvrière âgée de 55 ans travaillant dans un secteur très risqué (classe 4)

Tous ces graphiques permettent de se rendre compte de la pertinence des critères usuels de tarification en prévoyance collective. En effet, les niveaux de risques sont très différents en fonction des profils étudiés. Par ailleurs, la régression multinomiale a permis d'isoler la part non expliquée de la sinistralité après prise en compte des variables âge, genre, secteur d'activité et CSP. Cette part correspond aux résidus de la régression. Les traitements qui leur sont appliqués en vue de l'élaboration du zonier sont détaillés dans le chapitre suivant.

Chapitre 4

Traitements des résidus de la régression multinomiale

Ce chapitre expose la manière dont les résidus de la régression logistique multinomiale vont être utilisés pour élaborer le zonier. Plus précisément, il s'agit d'agréger les résidus individuels au code postal, unité géographique retenue pour le zonier. Pour cela, plusieurs indicateurs basés sur les résidus individuels sont donc créés à la maille code postal. Au vu du nombre important d'indicateurs construits, une sélection, détaillée dans la deuxième section de ce chapitre, est réalisée avant de modéliser le zonier.

4.1 Agrégation des résidus à la maille code postal

La régression logistique multinomiale présentée dans le chapitre précédent avait pour objectif de contrôler les effets des variables déjà utilisées en tarification (âge, CSP, répartition homme/femme, secteur d'activité). Les résidus de cette régression contiennent donc la part non expliquée de la sinistralité après prise en compte de ces variables. L'effet, s'il existe, de la localisation de l'entreprise sur la sinistralité est donc présent dans ces résidus. Il faut noter que puisque le modèle utilisé est une régression multinomiale, il n'y a pas qu'un seul résidu par individu mais un résidu par individu et modalité, soit 5 au total. Par exemple, un résidu positif sur la modalité «+90» signifie que nous sous-estimons la sinistralité en arrêt de travail de plus de 90 jours pour cet individu. A l'inverse, un résidu négatif indiquerait une surestimation de cette sinistralité.

Par mesure de simplicité, il a été décidé d'utiliser les résidus les plus courants dans notre étude, c'est-à-dire $\epsilon_{i,k} = y_{i,k} - \widehat{y_{i,k}}$ [1]. Cependant, ce type de résidus présente le défaut d'avoir une variance qui dépend de la variable à expliquer. Un axe d'amélioration de cette étude serait donc de tester d'autres résidus ayant potentiellement de meilleures propriétés, comme les résidus de Pearson ($\epsilon_{i,k} = \frac{y_{i,k} - \widehat{y_{i,k}}}{\sqrt{V(\widehat{y_{i,k}})}}$) ou les résidus de déviance qui représentent la contribution de l'observation à la déviance du modèle par rapport au modèle saturé. Il est aussi possible d'avoir recours à des résidus définis plus spécifiquement pour la régression multinomiale comme dans (Romeo et al., 2015)[4] et dans (Seber et Nyangoma, 2000)[5]. Néanmoins, ces méthodes ont le défaut d'être beaucoup plus difficiles à implémenter d'un point de vue pratique.

La régression logistique multinomiale est réalisée à la maille individuelle. En vue d'élaborer un zonier, il nous a donc fallu agréger les résidus à la maille code postal qui est l'unité

1. Il y a deux indices dans cette définition : i pour désigner l'individu et k pour désigner la modalité considérée.

géographique choisie dans notre étude. Ce choix se justifie pour différentes raisons.

— Le code postal de l'entreprise [2] est une donnée disponible dans la DSN.
— Il s'agit d'une information pouvant être facilement demandée aux clients lors de la tarification.
— Choisir le code postal comme unité géographique est un bon compromis. D'une part, agréger au niveau du département n'aurait pas été aussi précis au vu des fortes disparités socio-démographiques présentes au sein d'un même département. D'autre part, nous n'avions pas à notre disposition une donnée plus fine que le code postal. Quand bien même, une maille comme l'IRIS aurait sûrement été trop petite, ne permettant pas une mutualisation du risque satisfaisante. De même, la maille commune n'est pas appropriée car elle est à la fois trop fine pour les petites communes et trop large pour les métropoles. Le code postal permet de mieux prendre en compte les distorsions socio-démographiques au sein d'une commune puisque les grandes métropoles sont découpées en plusieurs codes postaux.

Pour faire cette agrégation, nous créons plusieurs indicateurs par modalité, à la maille code postal et basés sur les résidus individuels. Par exemple, la moyenne des résidus individuels associés à la modalité "+90" dans le code postal est un de ces indicateurs. La liste exhaustive est donnée ci-dessous.

— Moyenne des résidus par modalité ("Absence d'AT", "]0,15]", "]15,30]", "]30,90]", "+90" jours) dans le code postal. Cette variable est un indicateur de position permettant de résumer le différentiel entre la sinistralité prédite et observée dans chaque code postal.
— Moyenne des résidus positifs (respectivement négatifs) par modalité dans le code postal.
— Écart-type des résidus par modalité dans le code postal. Cet indicateur permet de résumer la variabilité des résidus au sein du code postal.
— Écart-type des résidus positifs (respectivement négatifs) par modalité dans le code postal.
— Quantiles de résidus par modalité dans le code postal. Ces variables sont des indicateurs de position permettant de capturer les parties de la distribution des résidus qui nous intéressent le plus.

Le but de créer plusieurs indicateurs (et non, un seul comme la moyenne) est d'essayer de résumer la distribution des résidus de chaque modalité par code postal. Ceci se justifie d'autant plus que ces distributions sont particulièrement spécifiques. En guise d'exemple, les distributions des résidus individuels associés aux modalités "Absence d'AT" et "+90" sont présentées en figure 4.1 et 4.2. Elles permettent de mieux comprendre les raisons de la construction de certains indicateurs comme les quantiles ou ceux distinguant résidus positifs/négatifs.

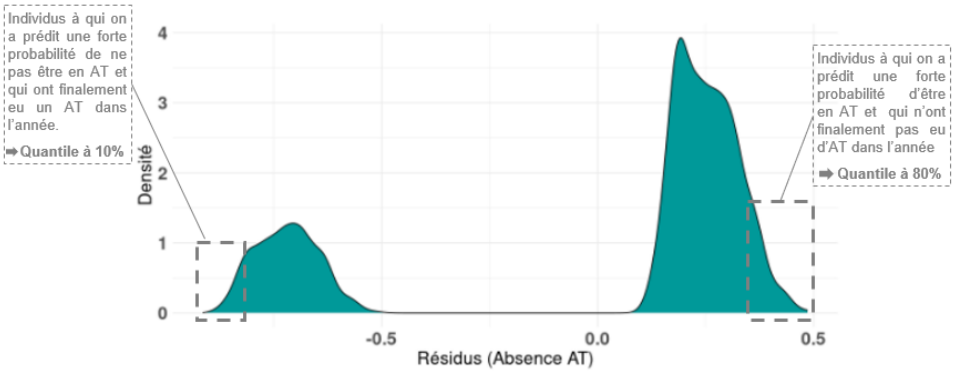

FIGURE 4.1 – Distribution des résidus individuels (modalité "Absence d'AT")

2. Tout au long de ce mémoire, le code postal de l'entreprise fait référence au code postal de l'établissement concerné et non au code postal du siège social de l'entreprise.

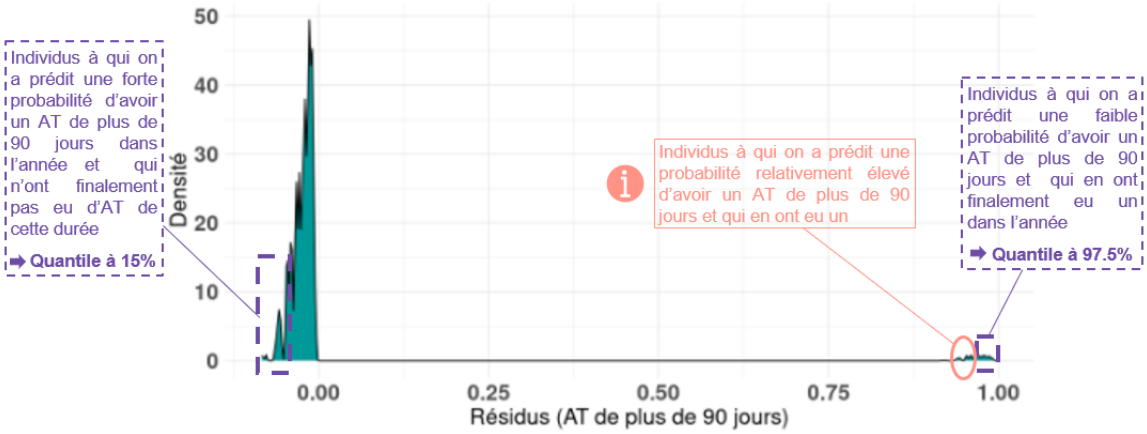

FIGURE 4.2 – Distribution des résidus individuels (modalité "+90" jours)

Les distributions des résidus sont particulièrement atypiques par rapport à ce qui est habituellement observé lors d'une régression linéaire classique par exemple. En effet, deux "pics" sont présents aux deux extrémités de chaque graphique. Ce phénomène s'explique par le fait que la variable à expliquer soit une variable catégorielle. La régression logistique multinomiale renvoie pour chaque individu, sa probabilité d'être dans chacune des classes ("Absence d'AT", "]0,15]", "]15,30]", "]30,90]", "+90" jours).

Par exemple, si un individu i n'a pas eu d'arrêt de travail dans l'année, il est associé à la modalité "Absence d'AT" et $y_{i,k=1} = 1$. S'il a une probabilité de ne pas avoir d'arrêt $\widehat{y_{i,k=1}} = 0.7$, son résidu pour cette modalité est de 0.3. Si cet individu a finalement eu une incapacité temporaire de travail dans l'année alors $y_{i,k=1} = 0$ et son résidu sur cette modalité serait de -0.7. En fait, dans la figure 4.1, le pic de droite (résidus positifs) correspond aux individus qui n'ont pas eu d'arrêt de travail alors que le pic de gauche (résidus négatifs) est associé aux personnes ayant eu un arrêt de travail. Dans la figure 4.2, les individus n'ayant pas eu d'incapacité de travail de plus de 90 jours sont présents dans le pic de gauche alors que ceux qui ont eu un tel arrêt sont dans le pic de droite. En résumé, la présence de ces deux pics justifie la création d'indicateurs distinguant résidus positifs/négatifs afin de mieux résumer la distribution des résidus.

L'objectif de construire des indicateurs basés sur les quantiles est de capturer les parties de la distribution qui ont le plus d'intérêt, à savoir les queues de distribution. Elles sont particulièrement intéressantes car elles correspondent aux individus pour lesquels la sinistralité est la moins bien prédite et où la marge d'amélioration de la connaissance du risque est la plus grande. Il peut s'agir d'individus prédits comme ayant un fort risque et qui n'ont pourtant pas eu d'arrêt de travail ou inversement des personnes considérées comme peu risquées et qui ont pourtant des arrêts de longue durée.

Des exemples de quantiles sont donnés dans les figures 4.1 et 4.2. Ainsi, dans le cas des incapacités de travail de plus de 90 jours, le quantile à 97.5% a été choisi comme seuil pour considérer que les observations sont mal prédites. Il est intéressant de remarquer que les résidus encerclés en couleur rose ne font pas partie des individus mal prédits. En effet, la faible densité dans le pic de droite montre que subir un arrêt de travail de plus de 90 jours est plutôt rare et avoir une probabilité égale à 6-7% est déjà considérable pour cette modalité. Enfin, un certain quantile dont le seuil n'est pas dévoilé par souci de confidentialité, est créé pour chaque modalité. Ce dernier est choisi de façon à ne pas prendre le maximum des résidus observés dans le code postal, même pour les codes postaux avec peu d'affiliés. L'idée est d'obtenir un compromis entre le fait de capturer du signal sur les résidus les plus forts tout en ne prenant pas le maximum des résidus dans le code postal.

Au total, 45 indicateurs sont créés. Ce nombre étant important, il a été décidé de procéder à une sélection de ces indicateurs à l'aide de la méthode LASSO (Least Absolute Shrinkage and Selection Operator).

4.2 Sélection des indicateurs

La sélection des indicateurs en amont de l'élaboration d'un zonier se justifient pour plusieurs raisons.

— Elle aide à identifier les variables qui expliquent la variable cible.
— La variance des estimateurs est réduite si le nombre de paramètres du modèle diminue, permettant ainsi d'obtenir des estimateurs plus précis.
— Elle permet de choisir davantage d'indicateurs sur les modalités d'arrêts de travail les plus importantes pour un organisme assureur, à savoir les modalités associées aux arrêts les plus longs et à l'absence d'arrêt de travail. Ainsi, les modalités ("Absence d'AT", "]30,90]" et "+90") seront sur-pondérées et les modalités ("]0,15]", "]15,30]") seront sous-pondérées. Cette pondération se justifie par le fait que la majorité des contrats du portefeuille étudié ont des franchises supérieures à 30 jours.

Cette sélection de variables est faite en utilisant la pénalisation lasso. L'estimateur lasso est défini à l'aide du problème d'optimisation suivant :

$$\hat{\beta}_{LASSO} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} ||Y - X\beta||_2^2 + \lambda||\beta||_1 \quad \text{avec } \lambda > 0 \tag{4.1}$$

Il est possible que la solution du problème de minimisation ne soit pas unique alors que les prédictions $X\hat{\beta}_{LASSO}$ le sont toujours. Par ailleurs, le paramètre de régularisation λ est le plus souvent déterminé par validation croisée. Plus λ est grand, plus la pénalisation est forte.

Le lasso présente l'avantage de faciliter l'interprétation d'un modèle en sélectionnant les variables les plus influentes. Néanmoins, il peut être moins performant en présence de fortes corrélations entre des variables explicatives. En effet, si des variables sont très corrélées et influencent beaucoup la prédiction, le lasso aura tendance à privilégier une des variables au détriment des autres.

Pour étudier "toutes choses égales par ailleurs" l'influence de chaque indicateur sur la sinistralité, les régressions lasso sont effectuées à la maille individuelle. Autrement dit, chaque individu est associé aux indicateurs correspondant au code postal de son entreprise. Ensuite, il faut régresser la sinistralité sur les indicateurs et les variables déjà utilisées en tarification (âge, CSP, genre et secteur d'activité). Il a été décidé de ne pas faire une régression lasso multinomiale mais plutôt une régression lasso logistique par modalité de durée d'arrêt de travail. En effet, une régression lasso multinomiale chercherait à estimer les coefficients de tous les indicateurs pour chaque modalité alors que chaque indicateur n'est construit qu'à partir d'une seule modalité d'arrêt de travail. Par exemple, le coefficient associé à l'indicateur "résidu moyen de la modalité "+90" dans le code postal" aurait aussi été estimé pour la modalité "]0,15]".

En d'autres termes, en effectuant la régression lasso multinomiale, nous ferions l'hypothèse que les indicateurs correspondant à une modalité d'arrêt de travail porteraient aussi du signal sur les autres modalités. Bien que ce soit potentiellement le cas, ce signal est certainement faible et il est préférable de se concentrer sur l'impact de chaque indicateur sur la sinistralité correspondante à sa propre modalité. Ainsi, 5 régressions lasso logistiques sont mises en place, soit une par tranche de durée d'arrêt de travail. Pour chacune de ces régressions, nous incorporons seulement les indicateurs issus de la modalité étudiée, et ce, en complément des variables déjà

utilisées en tarification. Par exemple, dans la régression logistique associée aux arrêts de travail les plus longs (au moins 90 jours), les variables incluses dans le modèle sont :

— âge, CSP, secteur d'activité (en classe), genre ;
— résidu moyen associé à la modalité "+90" dans le code postal ;
— écart-type des résidus associés à la modalité "+90" dans le code postal [3] ;
— quantiles des résidus associés à la modalité "+90" dans le code postal.

Les résultats de cette régression lasso logistique sont présentés en figure 4.3 et 4.4.

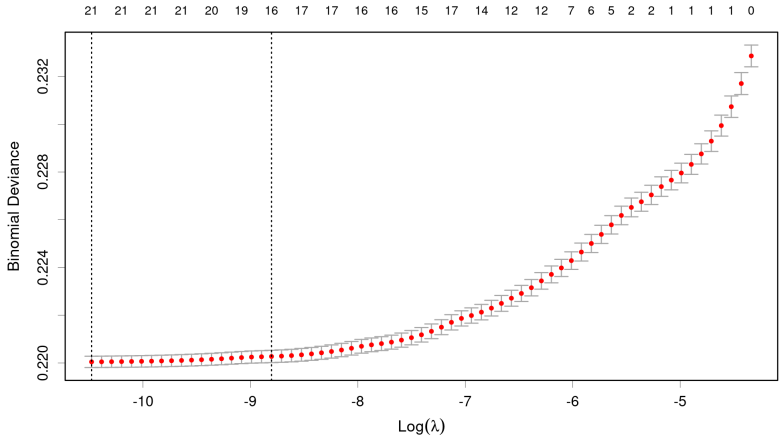

FIGURE 4.3 – Choix du paramètre de régularisation λ (arrêts de travail de plus de 90 jours)

La figure 4.3 permet de mesurer la qualité du modèle en fonction de différentes valeurs du paramètre de régularisation λ. Dans notre exemple, il est obtenu en considérant une validation croisée à 5 blocs (5-fold cross-validation). Le graphique fournit également le nombre de paramètres estimés différents de 0 en fonction de λ (chiffres en haut de la figure). Ce graphique apporte donc une aide à la décision du paramètre de régularisation. Il existe principalement deux règles pour choisir le λ.

— Choisir le λ qui minimise l'erreur moyenne de validation croisée, appelée "lambda.min" dans R et qui correspond à la droite en pointillé $x = \ln(2.815.10^{-5}) = -10.4780$.
— Choisir la plus grande valeur de λ de telle sorte que l'erreur soit à moins d'un écart-type de l'erreur minimale, nommée "lambda.1se" dans R et qui correspond à la droite en pointillé $x = \ln(1.502.10^{-4}) = -8.8035$.

La deuxième règle pénalise plus que la première et permet donc d'obtenir un plus petit nombre de coefficients différents de 0. C'est la raison pour laquelle cette règle a été privilégiée à la première dans nos travaux. La figure 4.4 représente les principaux coefficients estimés en fonction du paramètre de régularisation. La droite verticale correspond au λ choisi en ayant recours à la deuxième règle.

3. Pour rappel, les moyennes et écarts-types sont aussi calculés en distinguant résidus positifs/négatifs.

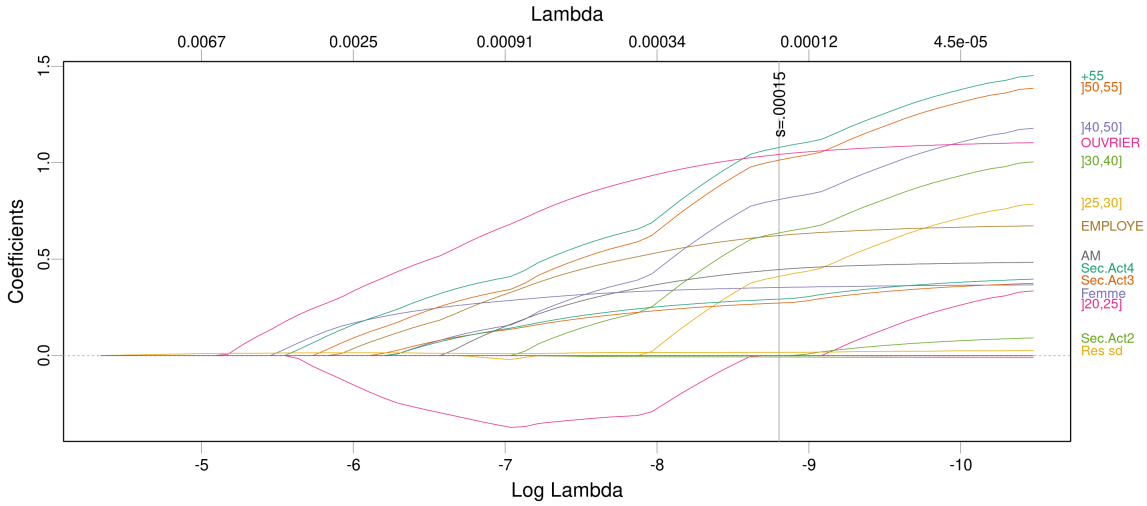

FIGURE 4.4 – Évolution des principaux coefficients de la régression lasso logistique pour les arrêts de travail de plus de 90 jours en fonction de λ

Dans cette figure, l'ordonnée correspond à la valeur des coefficients et l'abscisse au logarithme népérien du paramètre de régularisation. Ce graphique permet donc de montrer l'évolution de la valeur des coefficients en fonction de la pénalisation choisie. Il faut remarquer que l'axe des abscisses est "inversé". Autrement dit, les "grandes" valeurs de λ se situe à gauche du graphique et les "petites" à droite. Ainsi, plus on se décale vers la droite, moins la pénalisation est forte. Ce graphique vient donc confirmer les résultats obtenus avec la régression logistique multinomiale. En effet, lorsque la pénalisation est faible, les coefficients présentés (sur la partie droite du graphique) sont tous positifs, comme c'est le cas pour ceux associés à la modalité "+90" dans la régression multinomiale (cf. tableau 3.1). Aussi, la hiérarchie des coefficients est très proche entre les deux modèles. La régression lasso logistique estime également que les personnes de plus de 50 ans ont un fort risque d'avoir un arrêt de travail de plus de 90 jours, tout comme les ouvriers.

La plupart des coefficients sont assez stables lorsque la pénalisation varie. Ce graphique est donc plutôt satisfaisant par rapport à ce qu'il est possible d'obtenir avec d'autres applications de lasso. Néanmoins, il n'est pas parfait. Le coefficient associé à la tranche d'âge]20,25] est notamment difficilement interprétable. En effet, lorsque la pénalisation est forte (courbe rose sur la partie basse du graphique), ce coefficient est négatif et devient positif quand la pénalisation est plus faible (courbe rose sur la partie droite du graphique).

Bien que la confirmation des résultats de la régression multinomiale soit une bonne chose, l'objectif initial de ces régressions lasso logistique était de sélectionner les indicateurs pertinents pour chaque tranche d'arrêt de travail. En ce qui concerne les arrêts de plus de 90 jours, l'utilisation de la règle "lambda.1se" a permis de sélectionner 5 indicateurs basés sur les résidus. Étant donné que les arrêts de longue durée sont ceux qui intéressent le plus les organismes assureurs, le nombre d'indicateurs retenus par lasso a été borné à 5 pour les autres modalités d'arrêts de travail. Le but de cette limite est de ne pas accorder trop de poids aux arrêts de travail qui ont un coût faible, voire nul (franchise) pour l'assureur. Ainsi, il a été décidé de fixer au préalable un nombre d'indicateurs à sélectionner par tranche d'arrêt de travail :

— 5 indicateurs relatifs à la modalité "Absence d'AT"
— 2 indicateurs relatifs à la modalité "]0,15]" jours
— 2 indicateurs relatifs à la modalité "]15,30]" jours

— 4 indicateurs relatifs à la modalité "]30,90]" jours.

La majorité des franchises du portefeuille étudié sont supérieures à 30 jours, d'où le fait de sous-pondérer les modalités "]0,15]" et "]15,30]". L'absence d'arrêt de travail est ici surpondérée car savoir que les individus travaillant dans un certain territoire ne sont pas souvent en arrêt de travail est très intéressant pour l'assureur, bien plus que de savoir qu'un autre territoire est associé à une fréquence élevée d'arrêts de moins de 15 jours. En effet, s'il n'y a pas d'arrêt de travail, le coût sera nul pour l'organisme de prévoyance alors qu'un arrêt de 15 jours peut être indemnisé par l'assureur puisque Malakoff Humanis a de manière minoritaire, des contrats à franchise courte dans son portefeuille. Cependant, dans son ensemble, les garanties proposées par Malakoff Humanis ont globalement des délais de franchise supérieurs à 30 jours, d'où le fait d'être plutôt intéressé par l'absence d'arrêt de travail et les arrêts de plus de 30 jours.

Cette pondération a été choisie de façon arbitraire pour mettre en relief les arrêts de travail (et absence d'arrêt) qui nous intéressent le plus. Si nous avions eu accès à d'autres données, une autre solution aurait pu être de choisir le nombre d'indicateurs par modalité en fonction de l'impact financier de chacune d'entre elles chez Malakoff Humanis. Pour cela, il faudrait regarder, en tenant compte des différentes franchises pratiquées dans le groupe, combien coûtent à l'assureur les différents types d'arrêts de travail ("Absence d'AT", "]0,15]", "]15,30]", "]30,90]", "+90").

Maintenant que le nombre d'indicateurs est figé par modalité, il suffit de lancer les régressions lasso logistiques et de jouer sur le λ pour s'assurer de conserver le nombre prédéfini d'indicateurs par modalité. Ce travail a bien été mené mais les graphiques ne sont pas présentés afin de ne pas alourdir le mémoire.

Après avoir sélectionné un ensemble d'indicateurs, l'étude se poursuit avec la construction du zonier incapacité à l'aide d'une classification ascendante hiérarchique avec contraintes de proximité géographique. Tous les concepts nécessaires pour implémenter cette méthode, ainsi que les résultats sont présentés dans la troisième partie.

Troisième partie

Construction du zonier incapacité

Chapitre 5

Structure de voisinage et intérêts d'un zonier incapacité

Le chapitre 5 introduit dans une première section le concept essentiel de structure de voisinage en explicitant différentes définitions de "voisin". Dans un deuxième temps, l'élaboration d'un zonier incapacité est justifiée grâce notamment aux notions de structure de voisinage et d'autocorrélation spatiale.

5.1 Différents types de structures de voisinage

L'unité géographique retenue dans cette étude est le code postal de l'entreprise. Autrement dit, l'objectif final est de disposer pour chaque code postal d'une classe de risque. La figure 5.1 expose les contours des codes postaux de France métropolitaine [1]. On en dénombre plus de 6 000.

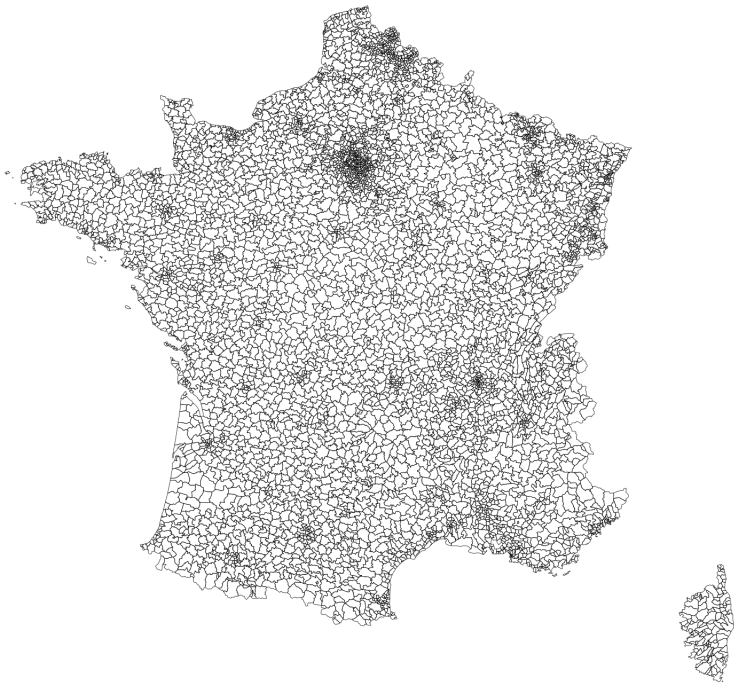

FIGURE 5.1 – Contour des codes postaux de France métropolitaine

1. Les cartes présentées dans ce mémoire ont pu être tracées grâce à la mise à disposition du fond de carte des codes postaux au format "Shapefile" par Géoclip[20]. Le format "Shapefile" est un des principaux formats de fichier pour stocker des données spatiales.

En regardant cette carte, il est très facile d'identifier les plus grandes métropoles françaises comme Paris, Lyon, Marseille, Toulouse, etc. En effet, les codes postaux sont beaucoup plus petits (en superficie) dans les aires urbaines et c'est la raison pour laquelle il est plus difficile de les distinguer sur une carte de France métropolitaine. Il est bien évidemment possible de faire un zoom comme le montre la figure 5.2 dans le cas de Paris. Les zones rurales sont quant à elles aisément remarquables par leurs codes postaux à grandes superficies.

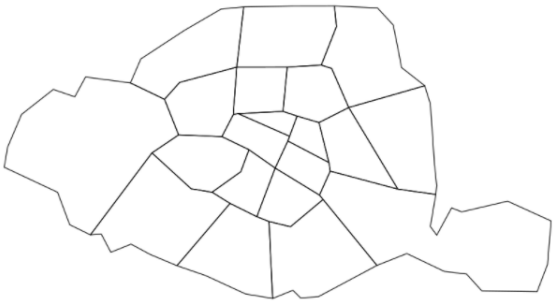

FIGURE 5.2 – Contour des arrondissements (codes postaux) de Paris

Afin de pouvoir démontrer l'intérêt de réaliser un zonier incapacité, il nous faut caractériser les voisins de chaque code postal. Cela autorisera la comparaison des codes postaux avec leurs voisins pour déterminer si une certaine corrélation se dégage (cf. indice de Moran dans la section suivante). La prise en compte de cette structure de voisinage dans la modélisation du zonier servira aussi à faire du lissage entre les codes postaux. En effet, il est légitime de penser que, globalement, deux codes postaux proches géographiquement devraient avoir des risques similaires (autocorrélation spatiale positive), ce qui justifierait de les mettre dans une même classe de risque. Par ailleurs, le fait de proposer un tarif très différent pour deux codes postaux voisins est difficilement compréhensible pour les entreprises. Bien qu'il y aura forcément des codes postaux limitrophes associés à deux classes de risques différentes (et donc deux tarifs), il faudra essayer de limiter ces cas et le lissage est une méthode qui permet d'atteindre cet objectif.

Plusieurs définitions existent pour déterminer les voisins d'un code postal donné et en fonction de la définition choisie, les résultats diffèrent. Chaque définition entraîne donc une structure de voisinage distincte. Les figures 5.3, 5.4 et 5.5 schématisent les trois principales définitions qui sont contiguïté, distance et k plus proches voisins.

Dans les 3 figures ci-dessous :

— Chaque carré correspond à un code postal.
— Le point bleu turquoise correspond au centroïde de chaque code postal.
— Le code postal d'intérêt est représenté en orange foncé.
— Les carrés colorés en rose saumon sont les voisins du code postal d'intérêt.

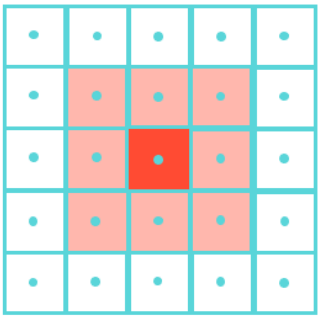

FIGURE 5.3 – Structure de voisinage avec la contiguïté

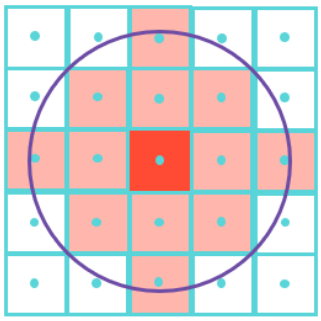

FIGURE 5.4 – Structure de voisinage avec les distances

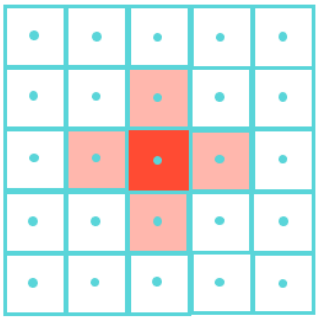

FIGURE 5.5 – Structure de voisinage avec les 4-NN

Contiguïté La définition utilisant la contiguïté pour définir les voisins est la plus intuitive. En effet, si deux codes postaux partagent au moins une "frontière" commune ou un même sommet, ils sont considérés comme voisins. Le problème de cette méthode est que si le code postal n'a aucune frontière ou sommet commun avec un autre code postal, il n'a pas de voisins, ce qui est le cas pour des îles n'ayant qu'un seul code postal. Cet exemple arrive pour certaines îles de Bretagne, Nouvelle-Aquitaine et Pays de la Loire. Par ailleurs, des variantes de cette définition existent. Il est possible de ne considérer que les frontières communes et non les sommets ou inversement.

Distance La définition basée sur les distances consiste à fixer un seuil maximal de distance d et à considérer comme voisin tout code postal dont le centroïde est à l'intérieur du cercle de rayon d. Il est aussi possible d'imaginer d'autres figures géométriques que le cercle. Cette définition est clairement la moins adaptée à notre contexte. Comme évoqué précédemment, les codes postaux d'aires urbaines sont plus petits que ceux de zones rurales. Ainsi, en utilisant cette définition, les codes postaux ruraux auraient beaucoup moins de voisins que les codes postaux urbains alors que les disparités sont beaucoup plus fortes d'un kilomètre à l'autre en zones urbaines. Une solution serait d'utiliser des seuils de distance différents en fonction de la densité de population mais cela complexifierait la caractérisation de la structure de voisinage.

k plus proches voisins/k-nearest neighbours (k-NN) La troisième définition est basée sur les k plus proches voisins. Autrement dit, les voisins d'un code postal donné sont ceux pour lesquels les centroïdes font partie des k plus proches en termes de distance par rapport au centroïde du code postal considéré. Ainsi, tous les codes postaux présentent le même nombre de voisins, ce qui résout le problème rencontré avec la définition basée sur les distances. C'est une des raisons pour lesquelles la méthode "k-NN" a été privilégiée dans notre étude. La limite de cette définition est qu'il faut choisir le nombre de voisins à considérer k. Ce nombre a été fixé à 4 dans nos travaux, ce choix sera justifié dans la section suivante [2] car il fait intervenir des concepts non encore présentés.

Le portefeuille étudié dans ces travaux n'est pas uniformément réparti sur la France métropolitaine. En effet, des zones comme le centre de la France ne sont que très peu représentées dans notre base de données, c'est-à-dire que peu d'assurés du portefeuille travaillent dans ces codes postaux. Par conséquent, peu de données de sinistralité sont associées à ces codes postaux, ce qui ne les rend pas assez robustes pour les incorporer dans notre modélisation. Ils sont donc exclus dans un premier temps. Pour cela, il faut aussi que ces codes postaux avec peu ou pas d'affiliés soient écartés de la construction de la structure de voisinage. Les figures 5.6 et 5.7 montrent les structures de voisinage obtenues avec la définition "4-NN" en tenant compte de tous les codes postaux (figure 5.6) et en ne considérant que ceux ayant un nombre d'affiliés suffisant (figure 5.7).

2. Plus précisément, le choix du nombre de voisins est détaillé à la fin de la sous-section 5.2.2.

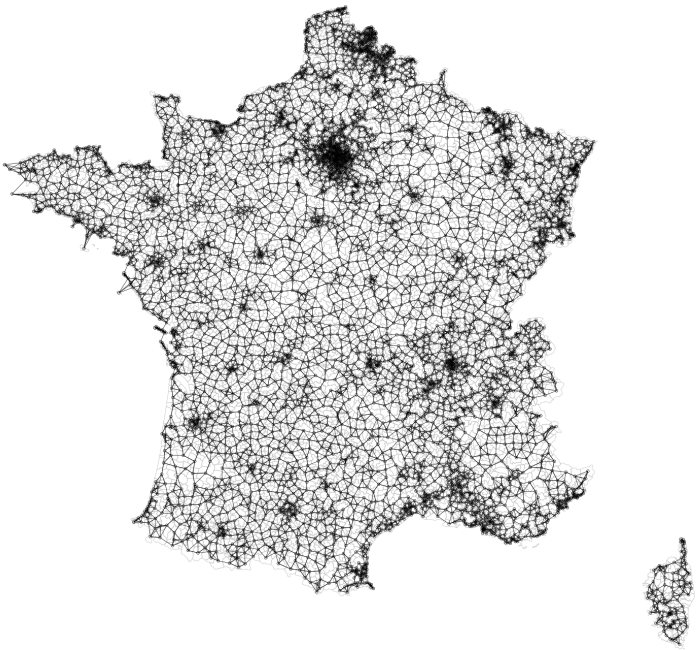

FIGURE 5.6 – Structure de voisinage en tenant compte de tous les codes postaux de France métropolitaine

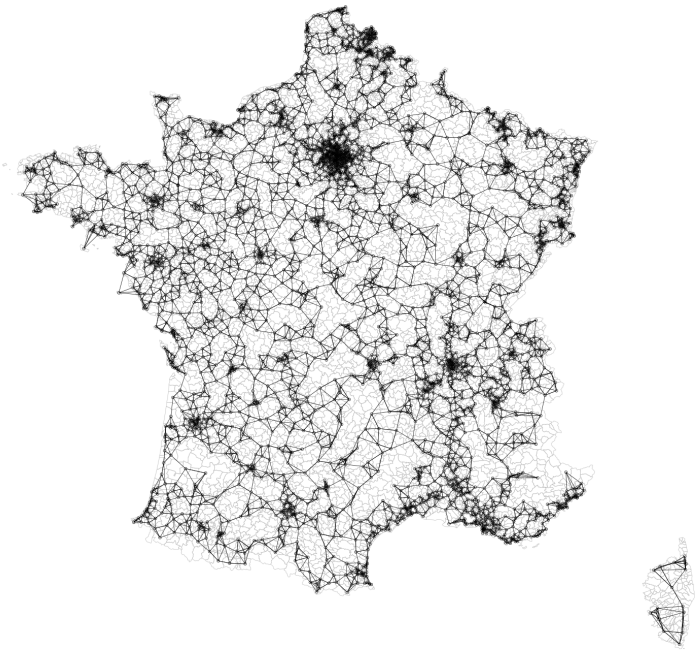

FIGURE 5.7 – Structure de voisinage basée uniquement sur les codes postaux avec suffisamment d'affiliés

Dans les deux figures plus haut, le contour des codes postaux est dessiné en gris clair et les codes postaux sont reliés à leurs voisins (en noir). Dans la figure 5.7, les codes postaux avec peu ou pas d'assurés sont facilement remarquables puisque ce sont ceux qui ne sont pas reliés à d'autres codes postaux. Ces zones se concentrent essentiellement dans les territoires montagneux (Alpes et Massif central) et ne seront pas étudiées dans un premier temps. Le fait de ne pas tenir compte de ces codes postaux entraîne un changement significatif de la structure de voisinage, essentiellement dans les zones rurales. Ce phénomène se constate facilement en comparant les figures 5.6 et 5.7. Certains codes postaux ont des voisins plus éloignés dans la figure 5.7, ce qui représente le principal désavantage de ce choix.

5.2 Intérêts d'un zonier incapacité

5.2.1 Premières cartes...

La section 5.2 a pour but de justifier l'intérêt de réaliser un zonier incapacité. Dans un premier temps, les figures 5.8, 5.9, 5.10 et 5.11 exposeront les résidus moyens par code postal pour les arrêts de travail les plus longs ("]30,90]" et "+90"), ce qui montrera qu'il est difficile de construire un zonier uniquement basé sur ces graphiques et sans modélisation, ni lissage. Ensuite, le concept d'autocorrélation spatiale est introduit et servira à démontrer que réaliser un zonier incapacité a du sens.

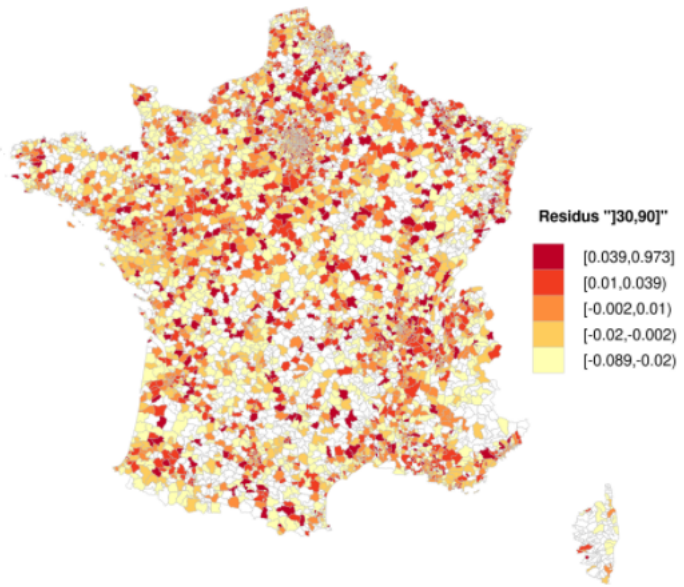

FIGURE 5.8 – Carte choroplèthe des résidus "]30,90]" moyens dans le code postal (France métropolitaine)

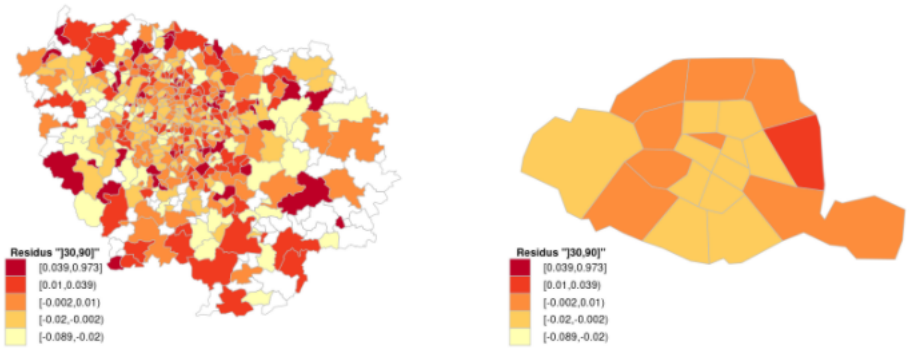

FIGURE 5.9 – Carte choroplèthe des résidus "]30,90]" moyens dans le code postal (Île de France et Paris)

Les figures 5.8 et 5.9 s'intéressent aux arrêts de travail de 30 à 90 jours. Pour rappel, plus un code postal a un résidu moyen important, plus sa sinistralité prédite en ne tenant compte que de l'âge, CSP, genre et secteur d'activité est sous-estimée. Autrement dit, plus un code postal a un résidu moyen important, plus il vire au rouge et plus il est risqué en ce qui concerne les arrêts de 30 à 90 jours. Des territoires se démarquent comme plus risqués sur cette durée d'arrêt de travail, comme le nord de l'Isère (sud de Lyon). Néanmoins, il est très difficile, en regardant simplement cette carte de pouvoir délimiter des zones reflétant des niveaux de risques différents sur cette modalité. Le même exercice est réalisé pour les arrêts de plus de 90 jours (figures 5.10 et 5.11) mais il faut garder en tête que chacune de ces cartes n'est basée que sur un seul type d'arrêts de travail. Ainsi, déterminer un zonier à partir d'une de ces cartes ne permettrait pas de considérer l'ensemble des durées d'arrêts de travail.

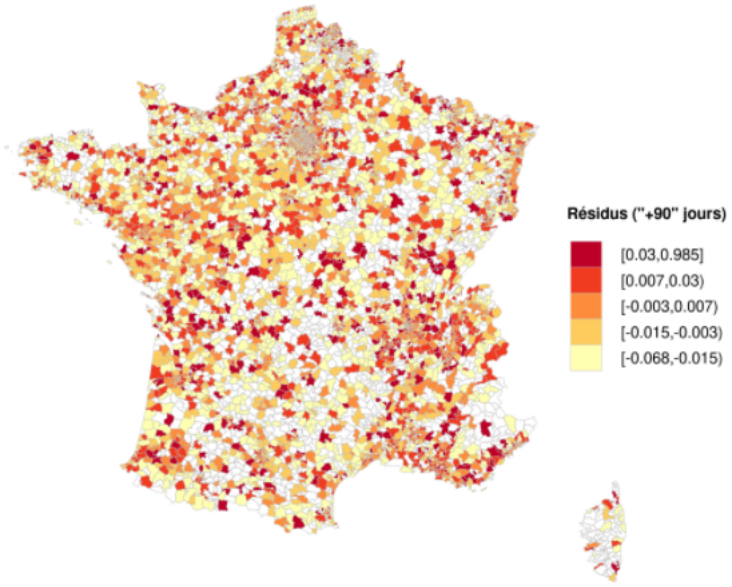

FIGURE 5.10 – Carte choroplèthe des résidus "+90" jours moyens dans le code postal (France métropolitaine)

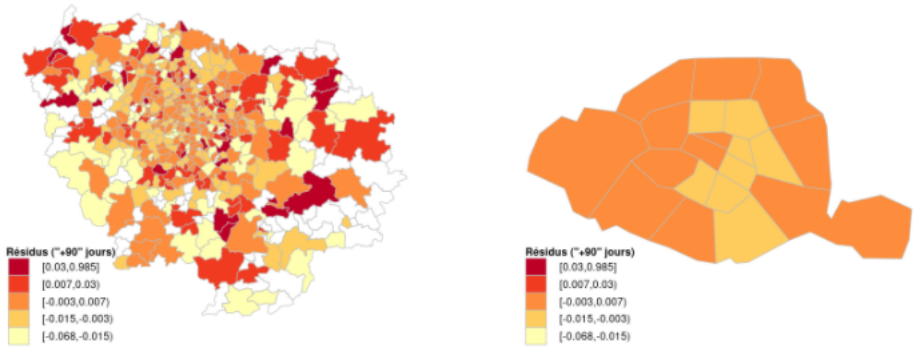

FIGURE 5.11 – Carte choroplèthe des résidus "+90" jours moyens dans le code postal (Île de France et Paris)

A nouveau, il est compliqué de définir des zones reflétant des niveaux de risques différents à partir des deux figures ci-dessus. Néanmoins, quelques zones semblent partager un risque élevé sur les arrêts de travail de 30 à 90 jours mais aussi sur les arrêts de plus de 90 jours. C'est notamment le cas de l'Isère et de la Côte d'Azur.

Ces cartes rappellent aussi le nombre important de codes postaux dont la sinistralité est considérée comme n'étant pas assez robuste. Ils sont représentés en blanc. Le fait que nous ne disposions pas de données complètes sur tous les codes postaux est aussi un argument en faveur de la réalisation d'un modèle pour construire un zonier.

5.2.2 Autocorrélation spatiale

Afin d'introduire le concept d'autocorrélation spatiale, il est possible de faire une analogie avec les séries temporelles. En effet, les modèles de séries temporelles prennent en compte la potentielle corrélation entre les valeurs associées à des périodes proches. Par exemple, si l'on souhaite modéliser les cours de bourse d'une action (ou son rendement), sa valeur d'aujourd'hui est plus corrélée à celle d'hier plutôt que celle de deux ans auparavant. Ce phénomène

est aussi présent pour des données spatiales sauf que la dimension temporelle est remplacée par la dimension spatiale. Là où de la corrélation entre périodes proches est recherchée en série temporelle, il s'agit de corrélation entre territoires voisins dans le cas de données spatiales. Par ailleurs, il n'y a pas d'ordre naturel dans l'espace a contrario du temps où il y a un passé et un futur.

L'autocorrélation spatiale se définit comme étant la corrélation entre les valeurs d'une même variable dans différents endroits de l'espace. Trois types d'autocorrélation existent d'un point de vue qualitatif.

- Dans le cas d'**autocorrélation spatiale positive**, la variable d'intérêt a des valeurs similaires pour des zones géographiques voisines. Des regroupements géographiques de valeurs similaires de la variable d'intérêt sont donc constatés.
- Dans le cas d'**autocorrélation spatiale négative**, la variable d'intérêt a des valeurs disparates pour des zones géographiques voisines.
- Dans le cas d'**absence d'autocorrélation spatiale**, la proximité géographique n'aurait pas de lien avec le degré de similarité des valeurs de la variable.

Afin d'illustrer plus simplement ces différences, les figures 5.12, 5.13 et 5.14 schématisent ces trois types d'autocorrélation sur un exemple simplifié. Chaque carré correspond à une unité géographique (code postal dans notre étude) et une variable d'intérêt est étudiée. Si la zone est associée à une forte (respectivement faible) valeur de cette variable, le carré est coloré en rouge (respectivement bleu).

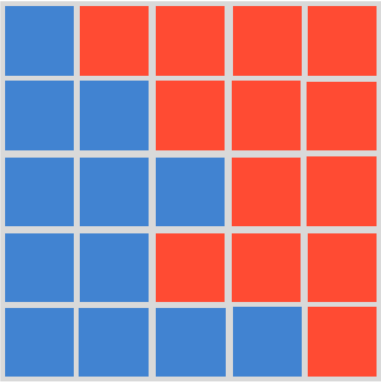

FIGURE 5.12 – Autocorrélation spatiale positive

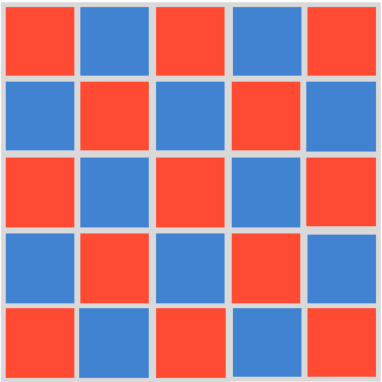

FIGURE 5.13 – Autocorrélation spatiale négative

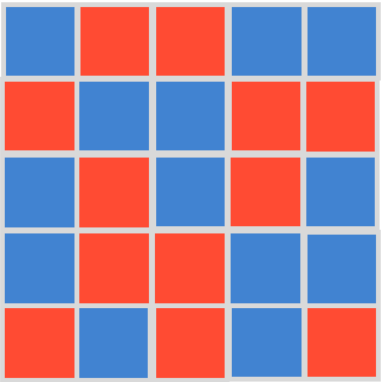

FIGURE 5.14 – Absence d'autocorrélation spatiale

Les deux regroupements de carrés rouges et bleus sont facilement observables dans la figure 5.12, signe d'une autocorrélation spatiale positive. Dans la figure 5.13, chaque zone géographique a une valeur de sa variable d'intérêt opposée à celles de ses voisins, synonyme d'autocorrélation spatiale négative. Dans la figure 5.14, les carrés rouges et bleus sont répartis de façon aléatoire et il est plus difficile de voir un lien entre proximité géographique et similarité entre valeurs de la variable. Il faut maintenant détailler comment quantifier cette autocorrélation spatiale de manière théorique et l'appliquer dans nos travaux.

L'indice de Moran est une mesure de l'autocorrélation spatiale. Afin de pouvoir le calculer, il faut définir une matrice de poids W. Cette matrice W sert à refléter la structure de voisinage choisie. Chaque élément de cette matrice indique le degré de proximité entre deux unités géographiques. Dans le cas de structure de voisinage définie par k-NN, pour une zone i, les indices

j tels que $w_{ij} = 1/k$ sont associés aux zones voisines de i (au sens de k-NN)[3]. Pour toutes les autres zones j qui ne sont pas voisines de i, $w_{ij} = 0$. En considérant n zones et Y la variable d'intérêt, le terme l du vecteur WY correspond à la moyenne de Y dans le voisinage de la zone l.

L'indice de Moran se définit par :

$$I = \frac{\dfrac{\sum_{i,j} w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_{i,j} w_{ij}}}{\dfrac{\sum_i (Y_i - \bar{Y})^2}{n}} \tag{5.1}$$

Il peut se mettre sous la forme plus générale suivante : $Corr(Y, WY) = \dfrac{Cov(Y, WY)}{\sqrt{Var(Y)Var(WY)}}$

Il peut donc être vu comme un coefficient de corrélation entre la valeur de la variable dans une unité géographique et celle dans son voisinage.

- Si I est "fortement" positif (respectivement négatif) alors les codes postaux ont globalement des valeurs proches (respectivement différentes) de celles de ses voisins sur la variable étudiée. Cela indique une autocorrélation positive (respectivement négative) (cf. figures 5.12 et 5.13).
- Une valeur de I proche de 0 indique une absence d'autocorrélation spatiale (cf. figure 5.14).

Afin de déterminer si I est suffisamment grand pour conclure à une autocorrélation spatiale (positive ou négative) de la variable étudiée, des tests statistiques existent. Ils permettent de tester l'absence d'autocorrélation spatiale pour une variable Y :

$$H_0 : \text{absence d'autocorrélation spatiale} \quad \text{vs} \quad H_1 : \text{présence d'autocorrélation spatiale} \tag{5.2}$$

Dans le cas où une autocorrélation spatiale positive est anticipée, il est aussi possible d'implémenter le test suivant :

$$H_0 : \text{absence d'autocorrélation spatiale vs } H_1 : \text{présence d'autocorrélation spatiale positive} \tag{5.3}$$

Cette notion d'autocorrélation spatiale va être utilisée sur nos indicateurs basés sur les résidus agrégés au code postal. L'idée est de vérifier que les codes postaux proches géographiquement ont des valeurs similaires de résidus. Ainsi, dans notre étude, il est légitime d'anticiper une autocorrélation spatiale positive des résidus, plutôt que négative. Dans le cas où l'autocorrélation spatiale serait négative, il serait difficilement justifiable de construire un zonier avec du lissage puisque cela impliquerait de mettre dans une même classe de risque des codes postaux voisins ayant des sinistralités très différentes. Par ailleurs, les indices de Moran calculés pour différents indicateurs de résidus sont tous positifs (cf. tableau 5.1). Ainsi, le deuxième test (5.3) sera implémenté plutôt que le premier (5.2).

Afin d'implémenter ce test, il faut faire une hypothèse sur la distribution de la variable Y sous H_0. Deux hypothèses sont couramment utilisées et engendrent deux tests distincts :

- **Test gaussien** : Il teste si $Y_1, ..., Y_n$ est représentatif de la distribution d'un vecteur gaussien de composantes indépendantes et identiquement distribuées.
- **Test de permutation** : Il teste si l'échantillon observé est représentatif d'une allocation aléatoire uniforme des valeurs sur les n zones.

3. Le $1/k$ permet d'avoir une matrice de poids normalisée.

Les deux tests ci-dessus sont implémentés pour 5 indicateurs basés sur les résidus. Les résultats sont présentés dans le tableau 5.1.

	Indice de Moran	Type de test	p-value
Résidu moyen "Absence d'AT"	0.08176	Test gaussien Test de permutation	4.198×10^{-14} 4.178×10^{-14}
Résidu moyen "]0,15]"	0.08356	Test gaussien Test de permutation	1.202×10^{-14} 1.186×10^{-14}
Résidu moyen "]15,30]"	0.02908	Test gaussien Test de permutation	0.003783 0.003769
Résidu moyen "]30,90]"	0.01550	Test gaussien Test de permutation	0.0756 0.07544
Résidu moyen "+90"	0.03098	Test gaussien Test de permutation	0.002227 0.002207

TABLE 5.1 – Indices de Moran et tests statistiques pour différents indicateurs

Les indices de Moran liés aux 5 indicateurs étudiés sont tous positifs. Quel que soit le type de test implémenté (gaussien/permutation), les conclusions sont les mêmes. Toutes les p-values sont inférieures à 1%, excepté celles associées au résidu moyen "]30,90]" qui sont, elles inférieures à 10%. Ainsi, tous les tests sont significatifs à 1% sauf ceux basés sur le résidu moyen "]30,90]" qui sont eux significatifs à 10%. Par conséquent, avec un degré de significativité des tests de 10%, chacune des hypothèses nulles (H_0) est rejetée. Autrement dit, il y a présence d'autocorrélation spatiale positive sur les 5 indicateurs étudiés mais cette autocorrélation est moins significative pour le résidu moyen "]30,90]. Cette conclusion est très importante puisqu'elle justifie le fait de regrouper dans une même classe de risque des codes postaux proches géographiquement (lissage). Elle confirme donc tout l'intérêt de réaliser un zonier incapacité.

La notion d'autocorrélation spatiale est présentée uniquement dans un cadre univarié dans ce mémoire, ce qui constitue une limite. Des travaux ont été menés pour étendre cette notion dans un cadre multivarié comme (Wartenberg, 1985)[7]. Néanmoins, ces concepts ne se sont pas démocratisés dans le domaine de la statistique spatiale et demeurent beaucoup moins utilisés que ceux définis dans un cadre univarié.

Par ailleurs, l'indice de Moran est sensible à la structure de voisinage choisie puisque W apparaît dans la formule 5.1. Nous nous sommes donc servis de cet indice pour déterminer le nombre de voisins le plus approprié pour définir la structure de voisinage. En effet, nous avons choisi les 4 plus proches voisins car c'est cette définition qui maximise l'indice de Moran pour la plupart des indicateurs.

Enfin, bien que l'indice de Moran soit le plus utilisé, d'autres indices d'autocorrélation spatiale existent comme l'indice de Geary. Il est également possible de mener des tests statistiques de présence ou d'absence d'autocorrélation spatiale avec cet indice.

Tous les concepts nécessaires à l'élaboration du zonier ont été présentés. La construction du zonier d'un point de vue théorique et pratique, est donc détaillée dans le chapitre suivant.

Chapitre 6

Construction du zonier avec une classification hiérarchique spatiale

Ce chapitre expose l'élaboration du zonier d'un point de vue théorique dans un premier temps puis pratique. Le choix du nombre de zones est ainsi détaillé avant de présenter les résultats de la classification.

6.1 Théorie autour de la classification hiérarchique spatiale

L'objectif principal de ce mémoire est de construire un zonier incapacité à l'aide d'un clustering spatial. La méthode implémentée est présentée dans (Chavent et al., 2018)[2]. Cette méthode est nommée "classification hiérarchique spatiale" tout au long de ce mémoire pour en faciliter la lecture. Elle consiste à incorporer des contraintes spatiales dans une Classification Ascendante Hiérarchique (CAH). La CAH est une des méthodes de clustering les plus populaires qui est basée sur l'agrégation de groupes d'individus "proches". Plus précisément, la méthode consiste à considérer dans un premier temps autant de clusters que d'individus et de regrouper ensemble les clusters par étapes successives jusqu'à obtenir un seul cluster contenant tous les individus.

Les généralités liées à la classification ascendante hiérarchique (CAH) sont présentées dans une première sous-partie et la manière d'introduire des contraintes spatiales juste après.

6.1.1 Généralités autour de la classification ascendante hiérarchique

Le but d'un clustering est de créer des groupes d'individus tels que :

— chaque groupe (cluster) soit le plus homogène possible (c'est-à-dire que les individus se ressemblent le plus possible au sein du groupe) ;

— chaque groupe soit le plus différent possible des autres groupes.

L'objectif n'est donc pas d'expliquer une variable cible. Toutes les variables jouent le même rôle. On parle de méthode non supervisée.

La notion de "ressemblance" entre les individus se formalise par l'utilisation de distances. Dans nos travaux, seulement la distance euclidienne sera utilisée. En notant $X_1, ..., X_m$ les m individus et $X^1, ..., X^p$ les p variables quantitatives, la distance euclidienne entre les individus i et j se définit par :

$$d(X_i, X_j) = \sqrt{\sum_{k=1}^{p} (X_i^k - X_j^k)^2}$$

Ainsi, plus cette distance est petite, plus les individus ont des caractéristiques (basées sur les p variables quantitatives) proches.

La distance euclidienne est sensible à l'échelle et à la variance des variables. C'est la raison pour laquelle il est fortement recommandé de centrer-réduire les données avant de lancer une procédure de clustering. En effet, si cela n'est pas fait, les variables avec de grandes variances auront plus de poids dans l'élaboration du clustering puisqu'elles impactent davantage le calcul des distances entre les individus.

L'inertie est aussi une notion centrale dans le domaine du clustering. Elle se définit par :

$$I = \sum_{k=1}^{p} Var(X^k) = \frac{1}{m} \sum_{k=1}^{p} \sum_{i=1}^{m} (X_i^k - \overline{X^k})^2 = \frac{1}{m} \sum_{i=1}^{m} d^2(X_i, \overline{X})$$

En notant $C_1, ..., C_L$ les différents groupes de tailles respectives $m_1, ..., m_L$ et $\overline{X}^{(1)}, ..., \overline{X}^{(L)}$ les vecteurs moyens des clusters $C_1, ..., C_L$, l'inertie totale peut se décomposer en :

$$I = \sum_{l=1}^{L} \frac{1}{m} \sum_{i \in C_l} d^2(X_i, \overline{X}^{(l)}) + \sum_{l=1}^{L} \frac{m_l}{m} d^2(\overline{X}, \overline{X}^{(l)})$$

avec $\sum_{l=1}^{L} \frac{1}{m} \sum_{i \in C_l} d^2(X_i, \overline{X}^{(l)})$ l'inertie intra-cluster et $\sum_{l=1}^{L} \frac{m_l}{m} d^2(\overline{X}, \overline{X}^{(l)})$ l'inertie inter-cluster.

Maintenant que le concept d'inertie est introduit, il est possible de reformuler de manière plus précise l'objectif d'un clustering qui est de trouver les clusters qui :

- maximisent l'inertie inter-classe (chaque groupe soit le plus différent possible des autres groupes) ;
- minimisent l'inertie intra-classe (chaque groupe soit le plus homogène possible).

Comme évoqué précédemment, le principe de la CAH est d'agréger par étapes successives des groupes d'individus "similaires". Pour cela, il faut définir une distance entre les clusters et pas seulement entre les observations. La distance de Ward est la plus utilisée dans ce contexte. En notant c_a (respectivement c_b) le centre du cluster a de taille m_a (respectivement b de taille m_b), la distance de Ward entre les clusters a et b se définie par :

$$D^2_{Ward}(a, b) = \frac{m_a m_b}{m_a + m_b} d^2(c_a, c_b)$$

où $d(c_a, c_b)$ est la distance euclidienne entre les centres des clusters.

L'avantage de la méthode de Ward est de minimiser, à chaque itération, la baisse de l'inertie inter-classe. En effet, lorsque deux clusters sont regroupés, l'inertie inter-classe décroît obligatoirement mais nous souhaitons que cette baisse soit la plus petite possible, puisque l'objectif d'un clustering est de maximiser l'inertie inter-classe et de minimiser l'inertie intra-classe.

Le principal inconvénient de la CAH est d'être coûteuse en temps de calcul dans le cas de grosses bases de données. Étant donné que le clustering implémenté dans ces travaux est à la maille code postal, le nombre de lignes de la base de données est plutôt faible. Aussi, la sélection des indicateurs, présentée plus haut a permis de limiter le nombre de variables en entrée du clustering. Ainsi, nous n'avons pas été impactés par cette problématique de temps de calcul et les résultats de la CAH étaient donnés quasi-instantanément.

A contrario d'une méthode comme k-means, la CAH présente l'avantage de ne pas imposer le nombre de clusters a priori. En utilisant une représentation des regroupements successifs, appelé dendrogramme et détaillé plus bas, la CAH permet de choisir le nombre de clusters a posteriori.

6.1.2 Ajout des contraintes spatiales

Ajouter des contraintes géographiques dans la construction d'un clustering peut s'avérer très pertinent avec des données spatiales. Ces contraintes permettent de gagner en termes de cohérence géographique en jouant le même rôle qu'un lissage spatial puisqu'en fonction du poids que nous leur accordons, les codes postaux "voisins" sont plus ou moins regroupés dans une même classe de risque, réduisant ainsi le nombre de codes postaux "isolés".

Cette partie détaille la manière d'incorporer des contraintes spatiales dans une CAH. La majorité des concepts et notations présentés sont repris de (Chavent et al., 2018)[2]. Leur méthode nécessite d'utiliser deux matrices :

— Une matrice de dissimilarité (aussi appelée matrice de distance), notée D_0 et qui correspond aux distances des individus dans l'espace des variables. Il s'agit de la matrice utilisée dans une CAH classique.

— Une deuxième matrice, notée D_1 et qui est associée aux "distances" géographiques entre les codes postaux. D_1 peut contenir des distances en kilomètres entre les centres des codes postaux mais peut aussi être basée sur une structure de voisinage.

Un paramètre de mélange α est introduit pour contrôler le poids de la contrainte spatiale dans la classification. Le critère minimisé à chaque regroupement devient une combinaison convexe d'un critère d'homogénéité calculé avec D_0 et d'un autre calculé avec D_1 alors que dans une CAH classique (sans contrainte spatiale), la minimisation n'est basée que sur un critère d'homogénéité calculé avec D_0. Le paramètre α correspond au poids de cette combinaison convexe. Ainsi, lorsque α augmente, l'homogénéité calculée avec D_1 augmente (meilleure contiguïté spatiale) et celle calculée avec D_0 diminue. Autrement dit, plus α est grand, plus la contrainte spatiale est forte et plus le lissage est important. Les deux cas extrêmes sont :

— Si $\alpha = 0$, la classification hiérarchique spatiale est équivalente à une CAH (sans contrainte spatiale).

— Si $\alpha = 1$, la classification hiérarchique spatiale n'est basée que sur D_1, c'est-à-dire seulement sur la proximité géographique entre les codes postaux.

Le but sera donc de fixer une valeur de $\alpha \in]0, 1[$ qui permet d'augmenter la contiguïté spatiale (les codes postaux voisins sont plus souvent dans la même zone) tout en limitant la dégradation par rapport à une CAH classique (sans contrainte spatiale). Pour davantage de détails sur les équations associées à la classification hiérarchique spatiale, le lecteur pourra se reporter à (Chavent et al., 2018)[2].

La classification hiérarchique spatiale présente plusieurs avantages.

1. Elle prend en compte des contraintes spatiales de manière souple. En effet, il est possible d'envisager des méthodes incorporant des contraintes beaucoup plus strictes. Par exemple, regrouper ensemble seulement des clusters voisins à chaque étape successive est beaucoup plus restrictif que la méthode présentée ci-dessus. Implémenter une telle approche pourrait regrouper des clusters très différents en termes de sinistralité mais proches géographiquement et à l'inverse diviser des clusters similaires concernant la sinistralité mais éloignés géographiquement.
2. Contrairement à un lissage spatial classique, la classification hiérarchique spatiale est une méthode multivariée, permettant de classifier (et de lisser) en utilisant plusieurs indicateurs.

Il faut toutefois garder en tête que le fait de rajouter des contraintes spatiales dégrade la qualité de la classification, comme c'est aussi le cas pour un lissage spatial.

6.2 Application au portefeuille

L'application de cette méthode avec nos données a pour objectif d'affecter à chaque code postal une classe de risque, tout en tenant compte de la proximité géographique.

Le zonier incapacité est construit à partir des indicateurs basés sur les résidus et agrégés au code postal. Pour rappel, ces indicateurs reflètent la part non expliquée de la sinistralité après prise en compte de l'âge, de la CSP, du genre et du secteur d'activité. Étant donné que ces indicateurs sont des moyennes, écarts-types et quantiles, toutes les variables sont quantitatives. Dans notre étude, la matrice D_0 contient donc la distance euclidienne (basée sur les indicateurs centrés-réduits) entre les codes postaux.

La structure de voisinage, décrite dans une matrice A, a été construite en utilisant la règle des 4-NN et seulement en tenant compte des codes postaux ayant suffisamment d'affiliés (cf. figure 5.7). C'est la raison pour laquelle, seulement ces codes postaux seront affectés à un cluster dans un premier temps. Pour être plus précis, chaque élément de la matrice A indique le degré de proximité entre deux unités géographiques. Autrement dit, $a_{ij} = 1$ si i et j sont des voisins et $a_{ij} = 0$ sinon. Par convention, $a_{ii} = 1$. Comme évoqué précédemment, D_1 porte la notion de proximité géographique. Ainsi, une forte (respectivement faible) valeur de $D_{1,ij}$ indique que les codes postaux i et j sont éloignés (respectivement proches) géographiquement. Pour garder cette cohérence, D_1 se définit par $D_1 = \mathbb{1}_n - A$ dans le cas de contraintes spatiales basées sur une structure de voisinage et non sur des distances.

6.2.1 Nombre de zones et poids de la contrainte spatiale

En plus du choix de la structure de voisinage et de la distance entre les variables, la classification hiérarchique spatiale nécessite de déterminer un nombre de zones, ainsi que le poids de la contrainte spatiale.

(Chavent et al., 2018)[2] préconise de choisir le nombre de zones avant de déterminer le paramètre α associé au poids de la contrainte spatiale. En effet, les auteurs présentent une procédure pour sélectionner une valeur adaptée de α pour un nombre donné de clusters. Pour choisir le nombre de groupes, ils suggèrent d'utiliser le dendrogramme obtenu avec la CAH classique (c'est-à-dire sans contrainte spatiale) et de déterminer le nombre de groupes qui semble le plus adéquat.

La figure 6.1 correspond au dendrogramme associé à la classification ascendante hiérarchique réalisée sur nos données. Ce diagramme permet de représenter les regroupements successifs obtenus avec la CAH. La hauteur de chaque "branche" correspond à la baisse de l'inertie interclasse lorsque deux clusters sont regroupés (le diagramme se lit du bas vers le haut). Ainsi, quand la hauteur est importante lors d'un regroupement de clusters, la perte d'inertie interclasse est grande et on regroupe des clusters qui sont différents. Inversement, quand la hauteur est petite, on regroupe des clusters similaires. Par conséquent, cette figure est un outil d'aide à la décision pour déterminer le nombre de clusters puisqu'elle permet de couper le dendrogramme lorsque nous considérons que trop d'inertie est perdue en prenant moins de clusters.

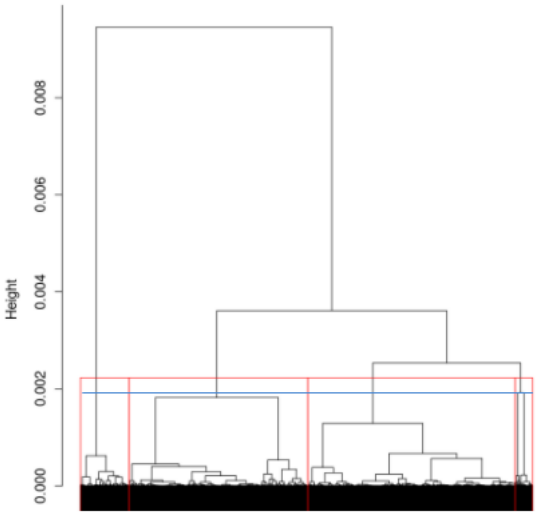

FIGURE 6.1 – Dendrogramme de la CAH (sans contrainte spatiale)

Afin de faciliter l'application de ces travaux en tarification, Malakoff Humanis souhaitait disposer d'un zonier composé au maximum de 6 zones. La figure 6.1 suggère de choisir un nombre de groupes égal à 4. En effet, si un cinquième cluster était ajouté (correspond à la ligne bleu), il serait issu d'un découpage au sein du cluster le plus petit (celui le plus à droite dans le dendrogramme). Étant donné que ce groupe est déjà celui contenant le moins de codes postaux, il paraît difficilement envisageable de le diviser une fois de plus.

Maintenant que le nombre de groupes est fixé, il faut déterminer la valeur du paramètre α. Ce paramètre peut être vu comme un paramètre de lissage puisqu'il correspond au poids accordé à la contrainte spatiale. Afin de mieux visualiser l'effet du paramètre α, les figures 6.2 et 6.3 représentent les classifications obtenues respectivement avec $\alpha = 0$ (sans contrainte spatiale) et $\alpha = 0.4$.

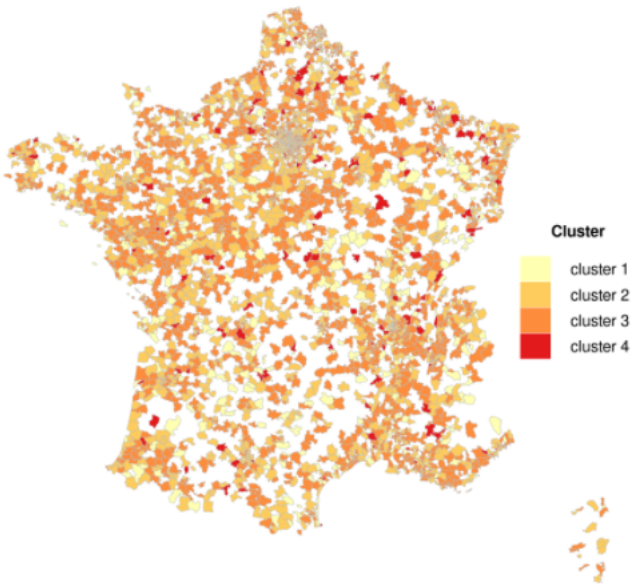

FIGURE 6.2 – Répartition des codes postaux parmi les différents clusters (CAH, $\alpha = 0$)

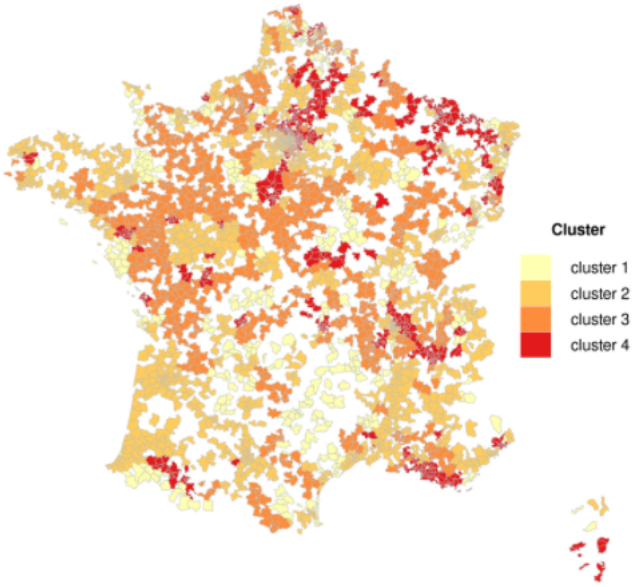

FIGURE 6.3 – Répartition des codes postaux parmi les différents clusters ($\alpha = 0.4$)

La classification obtenue avec $\alpha = 0.4$ est beaucoup plus lissée que dans le cas d'une CAH (sans contrainte spatiale). Elle pourrait donc apparaître plus intuitive à utiliser que celle obtenue

dans la figure 6.2. Néanmoins, il faut aussi prendre en compte la perte d'homogénéité en termes de sinistralité, qui, comme explicité plus bas, est beaucoup trop importante avec $\alpha = 0.4$.

L'idée est donc de choisir une valeur de α qui améliore la contiguïté spatiale (augmentation de l'homogénéité géographique) sans trop dégrader l'homogénéité des clusters obtenus avec une CAH classique. Ces homogénéités peuvent être mesurées par les inerties intra-classe [1]. En notant $W_0(P_1)$ l'inertie totale basée sur D_0 et $W_0(P_4^\alpha)$ l'inertie intra-classe basée sur D_0, il est possible d'obtenir la part d'inertie expliquée par la partition P_4^α :

$$Q_0(P_4^\alpha) = 1 - \frac{W_0(P_4^\alpha)}{W_0(P_1)}$$

$Q_0(P_4^\alpha)$ est un indicateur de la qualité de la classification uniquement du point de vue de la sinistralité (D_0). En effet, plus $Q_0(P_4^\alpha)$ est élevé, plus les clusters sont homogènes (plus les individus se "ressemblent" dans chaque cluster) pour D_0 (sans tenir compte de la proximité géographique). Puisque la classification du point de vue de la sinistralité se dégrade au fur et à mesure que la contrainte spatiale se renforce, $Q_0(P_4^\alpha)$ aura tendance à décroître lorsque α augmente. $Q_0(P_4^\alpha)$ atteint donc son maximum pour $\alpha = 0$.

De la même manière, il est possible de définir le même critère $Q_1(P_4^\alpha)$ pour évaluer la qualité de la classification d'un point de vue géographique (D_1). Dans ce cas-là, $Q_1(P_4^\alpha)$ aura tendance à croître lorsque α augmente (meilleure contiguïté géographique). $Q_1(P_4^\alpha)$ atteint donc son maximum pour $\alpha = 1$.

L'objectif de (Chavent et al., 2018)[2] consiste à comparer ces deux indicateurs pour déterminer un α adéquat. $Q_0(P_4^\alpha)$ et $Q_1(P_4^\alpha)$ ne sont pas forcément à la même échelle. Afin de faciliter leur comparaison, ces indicateurs sont donc renormalisés en divisant par leur valeurs maximales respectives, ce qui donne :

$$Q_0^*(P_4^\alpha) = \frac{Q_0(P_4^\alpha)}{Q_0(P_4^0)} \text{ et } Q_1^*(P_4^\alpha) = \frac{Q_1(P_4^\alpha)}{Q_1(P_4^1)}$$

La figure 6.4 reporte les valeurs de $Q_0^*(P_4^\alpha)$ (en noir) et $Q_1^*(P_4^\alpha)$ (en rouge) en fonction de α.

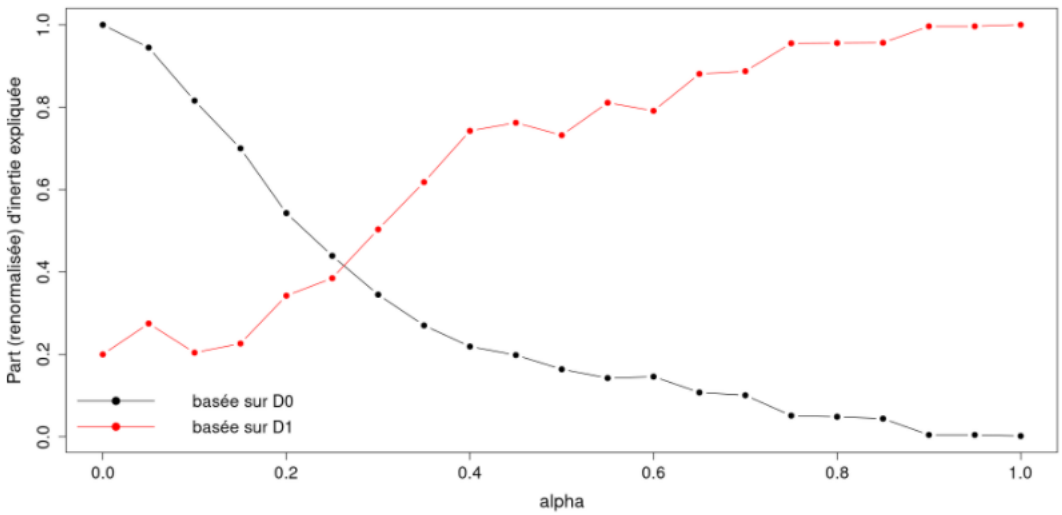

FIGURE 6.4 – Choix du paramètre α pour un nombre de clusters égal à 4

La figure 6.4 suggère de choisir un $\alpha = 0.05$. Cela correspond à une baisse d'environ 5.5 points de l'homogénéité basée sur la sinistralité et à une hausse de l'homogénéité géographique

1. Les formules détaillées sont données dans (Chavent et al., 2018)[2].

de 7.5 points en comparant avec la CAH (sans contrainte spatiale) [2]. Choisir une valeur de α supérieure à 0.05 ne paraît pas justifié. En effet, il faudrait prendre au minimum $\alpha = 0.2$ pour obtenir une amélioration significative de la cohérence géographique. Néanmoins, cette valeur de α engendrerait une perte de 45 points de l'homogénéité basée sur les indicateurs. Cette baisse est jugée trop importante pour s'assurer d'une cohérence suffisante en termes de sinistralité, des codes postaux au sein d'un même cluster. Autrement dit, avec une telle valeur de α, les codes postaux de chaque cluster ne seraient pas assez "proches" en termes de sinistralité.

En résumé, le poids de la contrainte spatiale imposée dans notre zonier est plutôt faible. Le lissage spatial, sera donc limité. Il est légitime de penser que si davantage de codes postaux étaient retenus dans l'élaboration du zonier, le lissage spatial serait sûrement plus important puisque les codes postaux auraient globalement des voisins plus proches, ce qui serait certainement synonyme de sinistralité plus similaire.

L'approche détaillée ci-dessus présente le défaut de fixer le nombre de clusters a priori et de déterminer la valeur de α conditionnellement à ce nombre de clusters. Une méthode définissant un critère global pour choisir le couple optimal (nombre de clusters, paramètre α) pourrait potentiellement être envisagée mais cela apporterait sûrement un degré de difficulté supplémentaire.

6.2.2 Résultats de la classification

Cette sous-partie présente les résultats de l'application de la classification hiérarchique spatiale sur nos données. Les figures 6.5 et 6.6 représentent le zonier obtenu en imposant un poids pour la contrainte spatiale α de 0.05.

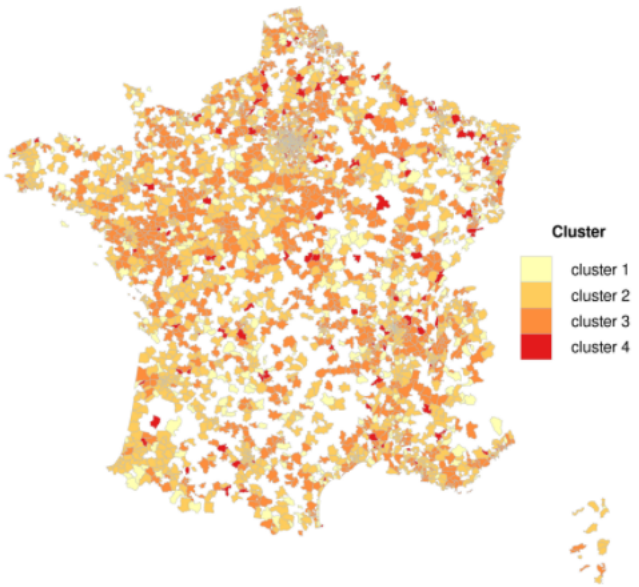

FIGURE 6.5 – Zonier avant traitement des codes postaux avec peu ou pas d'affiliés (France métropolitaine)

2. Ces résultats, plus précis que ceux visualisés dans la figure 6.4 sont disponibles en annexe C.

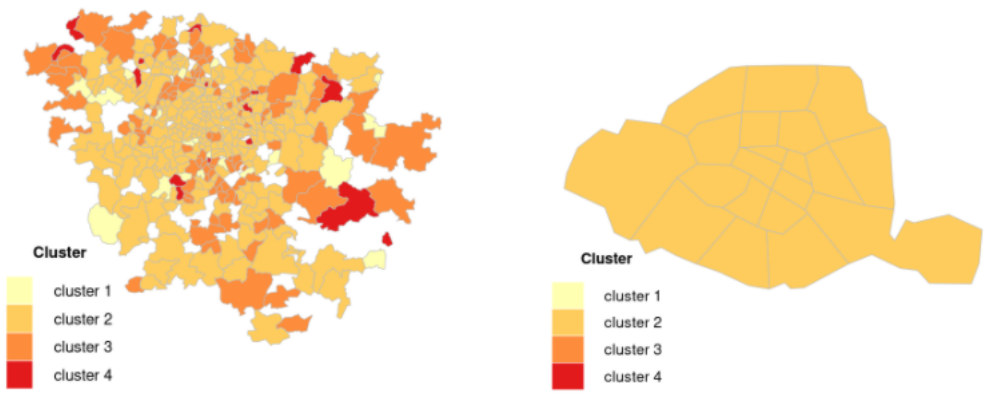

FIGURE 6.6 – Zonier avant traitement des codes postaux avec peu ou pas d'affiliés (Île de France et Paris)

Comme évoqué précédemment, seulement les codes postaux considérés comme ayant suffisamment d'affiliés sont affectés à une classe de risque. Ainsi, les codes postaux représentés en blanc sur les cartes sont ceux dont le calcul de la sinistralité est jugé insuffisamment robuste. Comme attendu, le lissage entre les codes postaux voisins est plutôt limité. Les codes postaux voisins ne sont pas forcément affectés à la même classe de risque. Néanmoins, il y a certaines zones où cela se produit. C'est notamment le cas pour les régions Pays de la Loire et Centre-Val de Loire où les codes postaux sont dans la grande majorité classés dans le groupe 3.

Les codes postaux n'étant pas encore associés à une classe de risque devront l'être dans un second temps puisque Malakoff Humanis se doit de pouvoir proposer un tarif partout en France. Cette affectation, présentée dans le chapitre 8, tiendra compte des classes de risques du voisinage et permettra ainsi d'obtenir un zonier davantage lissé.

Les clusters ont été triés dans l'ordre croissant du risque avant de les représenter dans les figures 6.5 et 6.6. Le cluster 1 de ces figures est donc celui ayant le niveau de risque le plus faible tandis que le cluster 4 est associé au risque le plus élevé. Le numéro du cluster obtenu initialement avec la classification hiérarchique spatiale (sans triage) n'établissait aucune relation d'ordre. Autrement dit, le groupe "1" initial ne voulait rien dire de particulier et c'est la raison pour laquelle il a fallu trier les groupes par ordre croissant du risque avant de les afficher sur une carte. Cette étape de triage a été faite en analysant la sinistralité dans les différents clusters. Les boîtes à moustache 6.7 à 6.11 ont notamment aidé à classer les clusters du moins au plus risqué.

Le tableau 6.1 affiche la répartition des codes postaux au sein des différents clusters. Les classes de risques obtenues sont déséquilibrées. Les clusters 2 et 3 rassemblent environ 85% des codes postaux alors que les clusters 1 et 4 contiennent respectivement 11.03% et 3.81% des codes postaux étudiés. Il paraît intuitif que le cluster le plus risqué soit le plus petit puisque associé à une sinistralité potentiellement atypique. Inversement pour les clusters 2 et 3 qui sont plutôt associés à une sinistralité proche de la moyenne. Cependant, le nombre de codes postaux dans le cluster 4 est tout de même faible et il aurait été préférable d'obtenir un cluster un peu plus conséquent. Le fait de ne pas pouvoir imposer la taille des clusters (ou au moins une taille minimale) est notamment une des limites de la méthode implémentée.

Cluster	Nombre de codes postaux	Proportion
Cluster 1	397	11.03%
Cluster 2	1 816	50.47%
Cluster 3	1 248	34.69%
Cluster 4	137	3.81%

TABLE 6.1 – Répartition des codes postaux au sein des différents clusters

Afin d'effectuer une première validation de la classification obtenue, les figures 6.8, 6.9, 6.10 et 6.11 reportent la répartition des fréquences des arrêts de travail (par code postal) en fonction de la durée et du cluster. La figure 6.7 affiche, selon le cluster, la répartition de la part d'individus (par code postal) n'ayant pas eu d'arrêt de travail dans l'année.

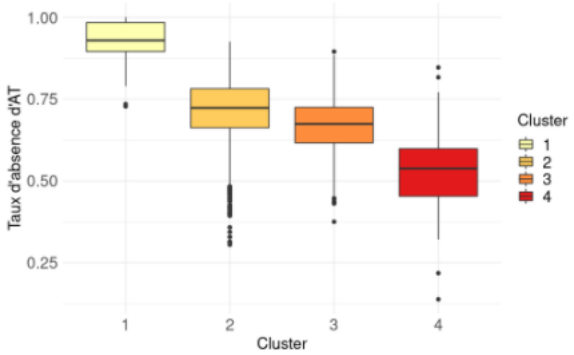

FIGURE 6.7 – Part d'individus n'ayant pas eu d'arrêt de travail par cluster

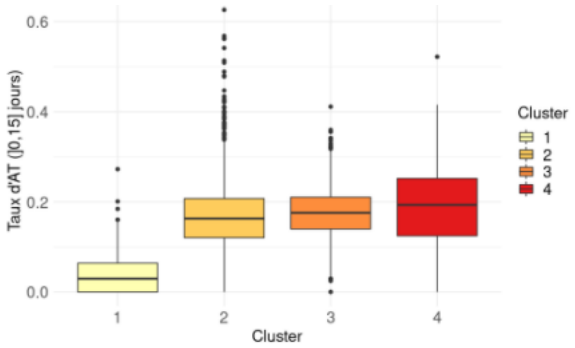

FIGURE 6.8 – Fréquence des arrêts de travail de 0 à 15 jours par cluster

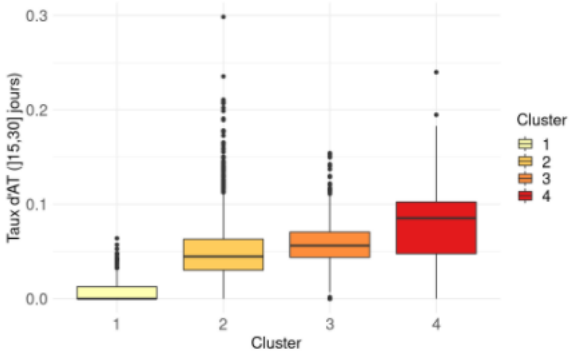

FIGURE 6.9 – Fréquence des arrêts de travail de 15 à 30 jours par cluster

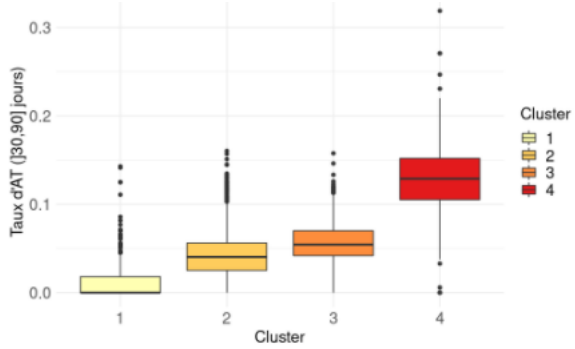

FIGURE 6.10 – Fréquence des arrêts de travail de 30 à 90 jours par cluster

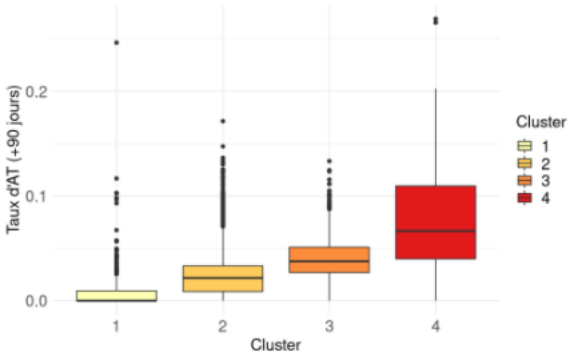

FIGURE 6.11 – Fréquence des arrêts de travail de plus de 90 jours par cluster

Avant d'interpréter ces boîtes à moustaches, il faut bien garder en tête qu'il s'agit d'une analyse bivariée qui ne tient donc pas compte des potentielles corrélations avec d'autres variables (âge, CSP, genre et secteur d'activité). Des résultats prenant en compte ces corrélations seront présentés dans le chapitre 9.

Quelle que soit la durée d'arrêt de travail, la sinistralité croît avec les classes de risques obtenues via la classification hiérarchique spatiale (les boîtes à moustaches sont globalement plus élevées au fur et à mesure que les classes de risques augmentent). Les clusters 3 et 4 ont bien une sinistralité plus importante que les clusters 1 et 2. Cependant, les différences de sinistralité entre les clusters sont plus ou moins importantes selon la durée des arrêts de travail considérée. En effet, les écarts de fréquence des arrêts de travail de 0 à 15 jours sont moindres par rapport à ceux observés pour les arrêts de travail de plus de 30 jours. Ce résultat est cohérent puisqu'un poids moins important a été accordé aux arrêts de travail de moins de 30 jours. Pour rappel, moins d'indicateurs basés sur les modalités "]0,15]" et "]15,30]" avaient été inclus dans la classification.

Concernant la part d'individus (par code postal) n'ayant pas eu d'arrêt de travail dans l'année, les écarts sont notables entre les clusters. En effet, la part d'individus n'ayant pas eu d'arrêt de travail est supérieure à 89% dans plus de trois-quarts des codes postaux du cluster 1 [3]. À l'inverse, cette statistique ne dépasse pas les 60% pour trois quarts des codes postaux de la classe 4 [4]. L'écart est moins important entre les classes 2 et 3. La part médiane d'individus n'ayant pas eu d'arrêt de travail est de 72% dans le cluster 2 contre 67% dans le cluster 3.

Comme dans tout travail statistique, l'application de la classification hiérarchique spatiale sur nos données présente des limites.

1. Le cluster le plus risqué contient peu de codes postaux. Imposer un nombre de codes postaux minimum dans chaque cluster pourrait être une solution mais complexifierait encore plus la méthode utilisée.
2. Le poids de la contrainte spatiale imposée dans notre zonier est plutôt faible ($\alpha = 0.05$). Il est légitime de penser que si davantage de codes postaux étaient retenus dans l'élaboration du zonier, le lissage spatial serait sûrement plus important puisque les codes postaux auraient globalement des voisins plus proches, synonyme de sinistralité plus similaire. Ceci sera rendu possible par l'ajout de données supplémentaires. En effet, l'historique de nos données est peu profond puisque la DSN est devenue obligatoire seulement en 2017. Ainsi, le fait de rajouter les données les plus récentes dans l'étude pourrait permettre de consolider la sinistralité calculée sur les codes postaux ayant peu d'affiliés, les permettant d'être inclus dans la structure de voisinage et de leur affecter une classe de risque.

3. Cela correspond au premier quartile.
4. Cela correspond au troisième quartile.

3. Les territoires ultramarins (Guadeloupe, Guyane, Martinique, La Réunion, Mayotte, Nouvelle-Calédonie, Polynésie française, Saint-Barthélemy, Saint-Martin, Saint-Pierre-et-Miquelon, les Terres Australes et Antarctiques Françaises et les îles de Wallis-et-Futuna) ne rentrent pas dans le cadre de cette étude. D'ailleurs une telle méthode serait difficilement applicable pour ces territoires puisqu'ils sont souvent isolés (île). Il serait donc difficile d'imaginer une structure de voisinage satisfaisante. Pour ces territoires, il serait préférable d'établir une règle plus simple, basée notamment sur une analyse descriptive de la sinistralité qui tient compte des variables âge, CSP, secteur d'activité, genre. Par exemple, une idée serait de faire cette analyse sur les indicateurs basés sur les résidus de la régression multinomiale et agrégés à la maille code postal.

La classification hiérarchique spatiale a permis d'affecter une classe de risque à tous les codes postaux ayant suffisamment d'affiliés. Pour ce qui est des codes postaux restants, cette affectation est faite en ayant recours à de l'open data. Plus précisément, l'idée est de modéliser les classes de risques à partir de données démographiques et socio-économiques de l'INSEE. D'une part, ce modèle permettra d'interpréter le zonier en appréhendant mieux les caractéristiques des différentes classes de risques. D'autre part, il rendra possible l'affectation d'une classe de risque aux codes postaux ayant peu ou pas d'affiliés, et ce, en utilisant leurs caractéristiques démographiques et socio-économiques. C'est tout l'objet de la quatrième partie.

Quatrième partie

Modélisation du zonier à l'aide de données démographiques et socio-économiques de l'INSEE

Chapitre 7

Construction d'indicateurs socio-économiques et démographiques à la maille code postal

L'institut national de la statistique et des études économiques (INSEE) donne accès librement à une quantité importante de données dans différents domaines. Une liste non exhaustive des principaux thèmes étudiés par l'INSEE est donnée ci-dessous :

- l'économie (taux de croissance, dette publique, etc.)
- la démographie (recensement des naissances et décès, recensement de la population, etc.)
- les revenus et la consommation (inégalités de revenus, etc.)
- les conditions de vie des Français (équipements des logements, égalité femmes-hommes, éducation, etc.)
- le marché du travail (taux de chômage, salaires, etc.)
- les entreprises (nombre de créations/défaillances d'entreprises, etc.)

Dans le cadre de cette étude, nous avons sélectionné les données qui permettraient potentiellement de mieux comprendre les caractéristiques démographiques et socio-économiques des différentes classes de risques du zonier. Ces données sont principalement reliées aux thèmes suivants et des exemples d'indicateurs sont aussi explicités ci-dessous.

- **Activité :** nombre de chômeurs de 15 à 64 ans, nombre de personnes actives de 15 à 64 ans, nombre de personnes salariées de 15 ans ou plus à temps partiel, nombre de personnes salariées de 15 ans ou plus, etc.
- **Revenus :** revenu disponible médian par unité de consommation (UC) [1], revenu déclaré médian par UC
- **Équipements et fonction médicale :** nombre de médecins généralistes, nombre de spécialistes en cardiologie, etc.
- **Population/Famille :** nombre de familles monoparentales, nombre de familles, etc.
- **Logement :** nombre de résidences principales HLM louées vides, nombre de résidences principales de type appartement, nombre de résidences principales.

Cette sélection est quasi-exclusivement basée sur des données de comptage (nombre de chômeurs de 15 à 64 ans, nombre de personnes actives de 15 à 64 ans, etc.) et non d'indicateurs

1. En reprenant la définition de l'INSEE, l'unité de consommation (UC) est un "système de pondération attribuant un coefficient à chaque membre du ménage et permettant de comparer les niveaux de vie de ménages de tailles ou de compositions différentes". Avec ce système, il est donc possible d'obtenir un nombre d'unités de consommation en fonction du ménage. Actuellement, la pondération utilisée par l'INSEE est : 1 UC pour le premier adulte du ménage, 0.5 UC pour les autres personnes de 14 ans ou plus, 0.3 UC pour les enfants de moins de 14 ans.

(taux de chômage, etc.). La construction des indicateurs souhaités sera donc réalisée par nos soins dans le cadre de ces travaux. Ce choix se justifie pour deux raisons. Premièrement, cela rend possible une certaine liberté dans la création d'indicateurs, nous permettant de ne pas se limiter aux indicateurs usuels (taux de chômage, etc.). Ensuite, les données de l'INSEE ne sont pas fournies à la maille code postal qui est la maille du zonier. Elles sont produites à la maille IRIS et/ou commune. Étant donné qu'une commune et un IRIS peuvent avoir plusieurs codes postaux, il est impossible dans un premier temps de relier les classes de risques obtenues par code postal avec les données INSEE. Ainsi, une étape indispensable de préparation de données a dû être mise en place pour agréger les différentes informations à la maille code postal.

7.1 Agrégation des données INSEE à la maille code postal

Avant de détailler l'agrégation des données à la maille code postal, un premier retraitement a été effectué concernant la mise en cohérence des différentes géographies issues de ces données. Nous avons retenu pour chaque variable le rafraîchissement le plus récent mis à disposition au moment de la réalisation de l'étude. Toutes les variables ne sont donc pas issues de la même année. Par exemple, les données de population utilisées sont basées sur l'année 2016 alors que celles liées aux équipements/fonction médicale sont de 2018. Le fait que les données utilisées ne soient pas de la même année n'est pas problématique en soi, ce qui l'est davantage est la différence de géographie. En effet, les contours des unités géographiques utilisées (IRIS et communes) peuvent varier d'une année à l'autre via par exemple, des fusions ou scissions de communes. Il a donc fallu figer une géographie de référence pour assurer une certaine cohérence aux indicateurs. Avec l'aide de tables de passage mises à disposition par l'INSEE, les données ont pu être mises à la géographie de 2019.

L'approche pour agréger les données de la maille IRIS à code postal est différente de celle utilisée pour passer de la commune au code postal. C'est la raison pour laquelle ces deux méthodes font l'objet de deux sous-parties distinctes.

7.1.1 Données socio-économiques et démographiques initialement à la maille IRIS

La majorité des données démographiques et socio-économiques sélectionnées sont fournies par l'INSEE à la maille IRIS. Les îlots regroupés pour l'information statistique (IRIS) sont des découpages infra-communaux, créés par l'INSEE dans le cadre du recensement de la population de 1999. Les communes peu peuplées (moins de 5 000 habitants) ne sont pas découpées en IRIS. Seulement les communes de plus de 10 000 habitants et la majorité des communes de 5 000 à 10 000 habitants le sont. En incluant les IRIS des petites communes, la France compte environ 50 000 IRIS. L'IRIS est donc la maille de référence pour toute statistique infra-communale.

Afin de rendre possible le passage de la maille IRIS à la maille code postal, nous avons eu recours à la Base Adresse Nationale (BAN). Cette base de données, produite par Etalab, répertorie, à la date d'écriture de ce mémoire, plus de 24.9 millions d'adresses françaises. Bien que très riche, elle n'est pas encore tout à fait complète et est mise à jour de façon hebdomadaire. Cette base de données ne contient pas l'IRIS. Antérieurement à cette étude, un travail a été mené en interne par d'autres personnes de l'équipe pour relier chaque adresse (avec le code postal) à l'IRIS correspondant. Ce travail est essentiel pour pouvoir agréger les données de la maille IRIS à code postal.

La France métropolitaine compte environ 50 000 IRIS pour plus de 6 000 codes postaux. Les codes postaux sont donc composés de plusieurs IRIS. Cependant, il arrive qu'un IRIS soit à cheval sur plusieurs codes postaux. L'objectif est donc de construire, à partir de la BAN, une table de correspondance dans laquelle chaque IRIS n'est associé qu'à un seul code postal. Un schéma récapitulatif de l'approche utilisée pour agréger les données de la maille IRIS au code postal est présentée en figure 7.1.

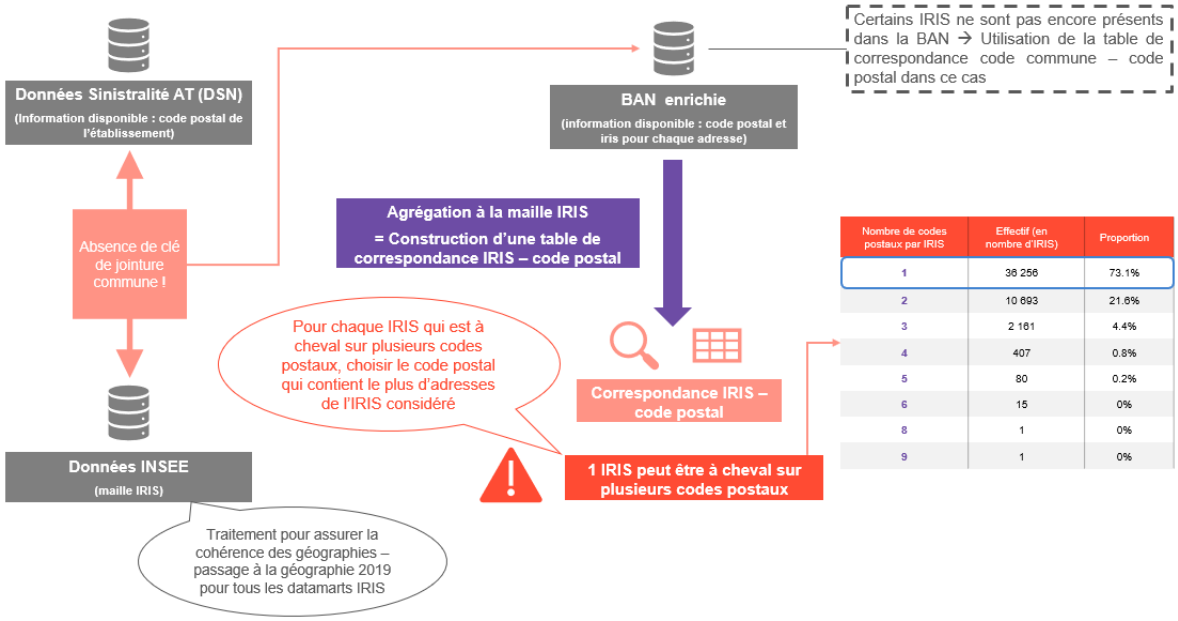

FIGURE 7.1 – Méthodologie pour agréger les données INSEE de la maille IRIS à code postal

Pour construire la table de correspondance "IRIS - code postal", l'idée est de calculer le nombre d'adresses par (IRIS, code postal). Seulement 73% des IRIS n'ont des adresses que dans un seul code postal (cf. tableau en bas à droite de la figure 7.1). Cela signifierait que près de 27% des IRIS sont à cheval sur plusieurs codes postaux. Toutefois, ce dernier chiffre est potentiellement sur-estimé à cause d'éventuelles erreurs dans la BAN enrichie de l'IRIS. En effet, certains IRIS qui seraient à cheval sur plusieurs codes postaux ont plus de 2 000 adresses dans un code postal pour une seule dans un autre. Ainsi, pour aboutir à une correspondance où chaque IRIS n'est relié qu'à un seul code postal, nous ne retenons que le code postal qui contient le plus d'adresses de l'IRIS considéré.

Par ailleurs, il est à noter que certains IRIS n'étaient pas encore présents dans la BAN au moment de la réalisation de cette étude. Dans ce cas, la table de correspondance "commune - code postal" a été utilisée. Sa construction est détaillée dans la sous-partie suivante.

7.1.2 Données socio-économiques et démographiques initialement à la maille commune

Bien que la majorité des données INSEE sélectionnées pour nos travaux soient fournies à la maille IRIS, certaines ont dû être utilisées à la maille commune pour des raisons de complétude (notamment pour les petites communes).

La France métropolitaine compte près de 35 000 communes pour plus de 6 000 codes postaux. En général, un code postal contient plusieurs communes, mais des cas plus complexes se présentent également. De la même manière que pour les données à la maille IRIS, l'objectif est de construire une table de correspondance qui associe chaque commune à un code postal. Un

schéma récapitulatif de l'approche utilisée pour agréger les données de la maille commune au code postal est présentée en figure 7.2.

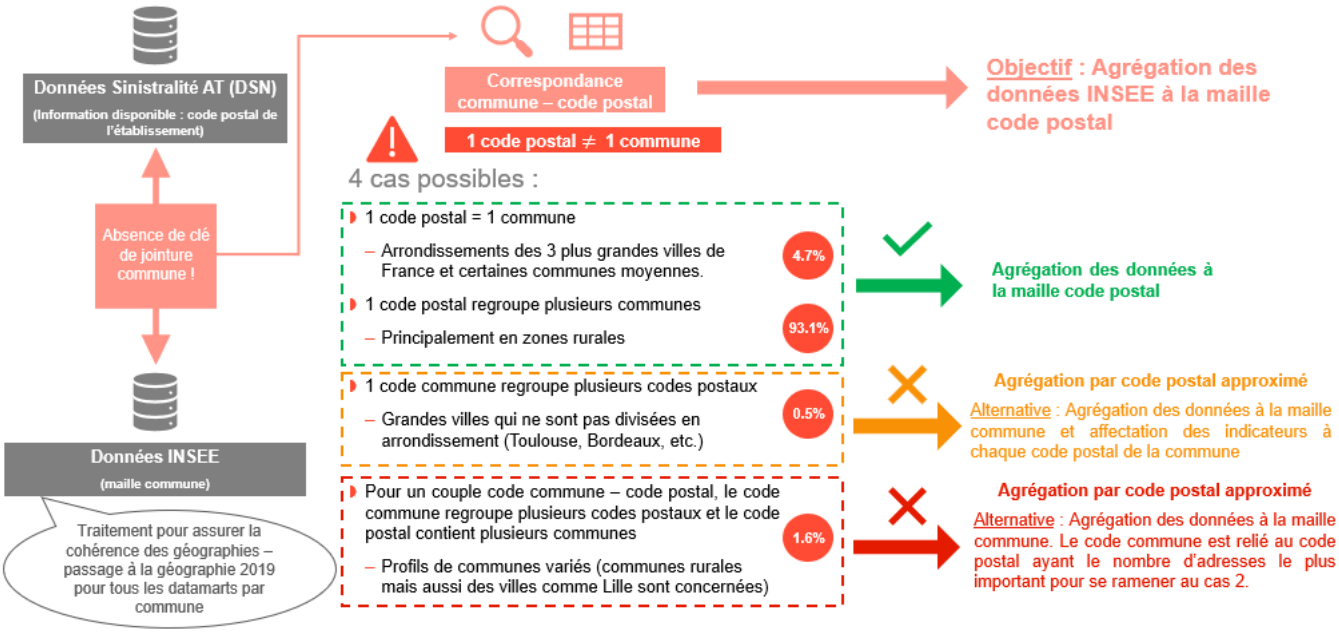

FIGURE 7.2 – Méthodologie pour agréger les données INSEE de la maille commune à code postal

Comme évoqué précédemment, il n'y a pas forcément qu'une seule commune dans un code postal et inversement. Plus précisément, quatre cas sont à considérer pour construire la table de correspondance "commune - code postal".

1. **Un code postal = une commune**, c'est-à-dire que les contours du code postal sont les mêmes que ceux de la commune. Cela représente près de 5% des cas et concerne les arrondissements des 3 plus grandes villes de France (Paris, Marseille et Lyon) et des villes moyennes. La correspondance "commune - code postal" est évidente dans ce cas-là.
2. **Un code postal regroupe plusieurs communes**. C'est la situation la plus fréquente avec 93% des cas. Elle concerne essentiellement les zones rurales. La correspondance "commune - code postal" ne pose pas de problème dans ce cas-là puisqu'il suffit de relier chaque commune à son unique code postal.
3. **Un code commune regroupe plusieurs codes postaux**. Cela concerne généralement les grandes villes non divisées en arrondissement (Toulouse, Bordeaux, etc.). Les données de l'INSEE ne pourront pas être disponibles pour ces codes postaux-là. Il faudra affecter à chaque code postal de ces communes, les indicateurs calculés sur toute la commune en question.
4. **Pour un couple code commune – code postal, le code commune regroupe plusieurs codes postaux et le code postal regroupe plusieurs communes**. Cela concerne des profils variés de communes (communes rurales mais aussi Lille par exemple). Ce cas est clairement le plus difficile à retraiter. L'idée est de construire une table qui référence le code postal principal (celui contenant le plus d'adresses) de chaque code commune. Cette table permet de se ramener au deuxième cas et permet de finaliser la correspondance "commune - code postal".

7.2 Indicateurs socio-économiques et démographiques à la maille code postal

Les tables de correspondances "IRIS-code postal" et "commune-code postal" étant établies, il est possible de construire des indicateurs INSEE à la maille code postal. La majorité des variables de l'INSEE sont des données de comptage. La création des indicateurs consiste principalement à établir des ratios de ces variables pour obtenir des taux. La liste de ces indicateurs triés par thème est donnée en figure 7.3.

Activités
Taux de chômage
Part de femmes actives dans la tranche d'âge 15-54 ans
Part de CDI/titulaires de la fonction publique parmi les « salariés »
Part de CDD ou intérim parmi les « salariés »
Part de temps partiel parmi les salariés
Part d'actifs qui vont au travail en transports en commun
Part d'actifs qui vont au travail en voiture/camion ou en deux-roues

Equipements
Nombre de crèches pour 1000 habitants (âgés de 18 à 39 ans)
Densité médicale = Nombre de médecins pour 1000 habitants
Part de généralistes parmi les médecins

Population/Famille
Densité de population
Part de familles monoparentales
Part de familles ayant 3 enfants ou plus de moins de 25 ans

Revenus
Revenu déclaré médian approximé par UC dans le code postal
Revenu disponible médian approximé par UC dans le code postal

Logement
Part de HLM parmi les résidences principales
Part d'appartements parmi les résidences principales

Les revenus médians sont des informations fournies à la maille commune. Ils sont qualifiés d'approximés car ils correspondent à la moyenne pondérée (par la population) des médianes des communes du code postal.
UC : unité de consommation

FIGURE 7.3 – Indicateurs socio-économiques et démographiques

Presque la moitié des indicateurs créés sont liés à l'activité des habitants dans chaque code postal. Par exemple, le taux de chômage ou la part d'actifs allant au travail en transports en commun pourrait être des informations utiles pour interpréter le zonier incapacité. De même, les indicateurs liés aux revenus, logement et population pourrait aider à mieux comprendre le lien entre richesse/précarité et absentéisme. Enfin, les indicateurs basés sur les équipements reflètent le taux de présence de services dans différents territoires, nous permettant par exemple de savoir si un lien existe entre désert médical et arrêt de travail.

Pour finir, un dernier retraitement a été implémenté pour les codes postaux où toutes les données INSEE n'étaient pas disponibles. Il s'agit de compléter les valeurs manquantes par la médiane non pondérée dans le département. L'idée de le faire par département permettrait de s'approcher de la vraie valeur du code postal en question. Cela ajoute aussi de la variabilité dans les indicateurs puisque si la médiane à l'échelle nationale était utilisée, il y aurait un pic d'observations "artificiel" au niveau de cette médiane. Le choix de faire une médiane non pondérée se justifie car si une pondération par la population était effectuée pour calculer les indicateurs à l'échelle du département, les codes postaux les plus peuplés auraient plus de poids, alors que les données manquantes sont essentiellement présentes sur les codes postaux les moins peuplés. Utiliser la médiane non pondérée permettrait donc d'être plus proche de la vraie valeur du code postal en question.

Tous les retraitements de données effectués pour aboutir à des indicateurs démographiques et socio-économiques ont été présentés dans ce chapitre. L'élaboration d'un modèle prédictif des classes de risques (issues du zonier) à partir de ces indicateurs est donc rendue possible. L'interprétation de ce modèle via l'utilisation des SHAP [2] values permettra de mieux appréhender les caractéristiques socio-économiques et démographiques des différentes classes de risques. Ce modèle servira également à prédire la classe de risque pour les codes postaux n'ayant pas suffisamment d'affiliés. Ces deux objectifs sont détaillés dans le chapitre suivant.

2. SHapley Additive exPlanations.

Chapitre 8

Interprétation du zonier et traitement des codes postaux avec peu ou pas d'affiliés

Les objectifs de ce chapitre sont multiples :

— Interpréter, avec l'aide des valeurs de Shapley et des données de l'INSEE, le zonier obtenu précédemment. Cela permettrait de mieux comprendre les différentes zones. Par exemple, nous pourrions voir si l'appartenance d'un code postal à une certaine classe de risque est corrélée avec ses équipements, sa densité de population, etc.

— Déterminer la classe de risque pour les codes postaux avec peu ou pas d'affiliés, en utilisant leurs caractéristiques démographiques et socio-économiques. Pour rappel, ces codes postaux n'étaient pas associés à une classe de risque puisqu'ils n'étaient pas inclus dans la classification hiérarchique spatiale.

Pour cela, un modèle de prédiction de la classe de risque à partir des indicateurs démographiques et socio-économiques est réalisé. Avant de présenter cette modélisation basée sur les forêts aléatoires (random forest), quelques statistiques descriptives sont analysées dans une première section.

8.1 Analyse descriptive des données démographiques et socio-économiques selon les classes de risques

L'étude descriptive détaillée ci-dessous consiste principalement à présenter une analyse bivariée. Les figures 8.1, 8.2, 8.3 et 8.4 représentent la répartition de quatre indicateurs socio-économiques dans chaque zone/cluster [1].

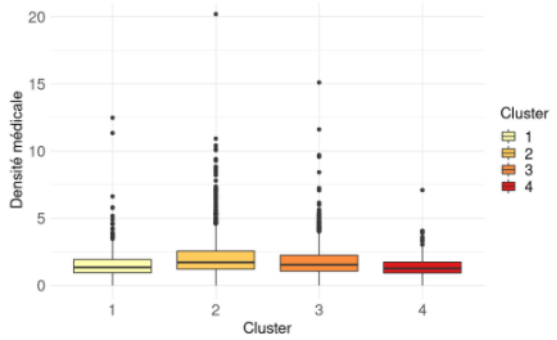

FIGURE 8.1 – Densité médicale par cluster/zone

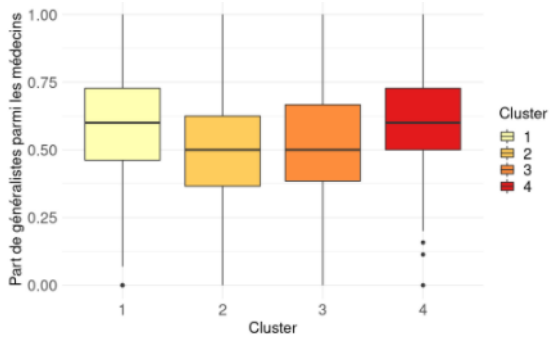

FIGURE 8.2 – Part de généralistes parmi les médecins par cluster/zone

1. Tout au long de ce mémoire, les termes "zone", "cluster" et "groupe" font à chaque fois référence aux classes de risques associées à la localisation de l'entreprise.

Les deux premiers indicateurs étudiés sont la densité médicale (nombre de médecins pour 1000 habitants) et le pourcentage de généralistes parmi l'ensemble des médecins. La densité médicale peut varier énormément d'un code postal à l'autre. En effet, là où certains codes postaux ne disposent d'aucun médecin, le 8e arrondissement de Paris recense plus de 20 médecins pour 1 000 habitants. Les zones 2 et 3 présentent globalement une densité médicale supérieure aux zones 1 et 4. Plus précisément, la densité médicale médiane dans la zone 4 est de 1.26 médecin pour 1 000 habitants contre :

— 1.34 dans la zone 1, zone la moins risquée (soit 6.3% de plus que la zone 4, zone la plus risquée) ;
— 1.55 dans la zone 3 (soit 23% de plus que la zone 4) ;
— 1.72 dans la zone 2 (soit 36% de plus que la zone 4).

La part de généralistes parmi l'ensemble des médecins est plus élevée dans les zones peu équipées (zones 1 et 4) par rapport aux autres. Ainsi, les inégalités sur la répartition des médecins entre territoires semblent encore plus fortes pour les spécialistes que pour les généralistes.

Il est difficile d'établir un lien entre densité médicale et absentéisme dans le code postal avec ces graphiques. D'une part, la zone la plus dépourvue de médecins est aussi la zone associée au risque le plus élevé (zone 4). D'autre part, la zone la moins risquée (zone 1) subit aussi une faible densité médicale. Néanmoins, les zones 1 et 4 sont celles qui contiennent le moins de codes postaux et ont potentiellement des sinistralités très atypiques. Au vu des densités médicales de la zone 2 et 3, il paraît très difficile d'affirmer qu'une forte densité médicale serait associée à un risque plus important d'arrêts de travail puisque la zone 2 (moins risquée) a un nombre de médecins pour 1 000 habitants supérieur à celui de la zone 3. Ainsi, le fait d'avoir accès plus facilement à un médecin ne serait pas signe d'un plus fort taux d'absentéisme mais plutôt l'inverse. Nous pourrions par exemple supposer que la difficulté à voir un médecin (désert médical) pourrait retarder des diagnostics de maladies et in fine allonger les (futurs) arrêts de travail. Cependant, cela reste qu'une hypothèse et une analyse plus approfondie (hors du cadre de ce mémoire) doit être menée pour confirmer ou non cette constatation.

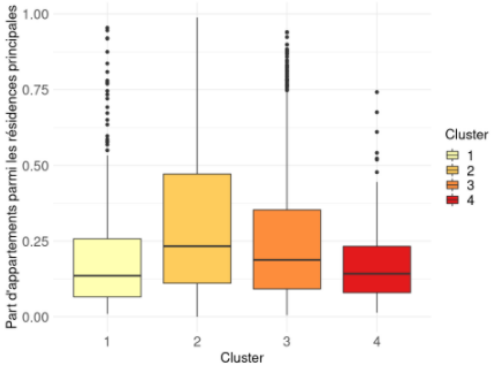

FIGURE 8.3 – Part d'appartements parmi les résidences principales par cluster/zone

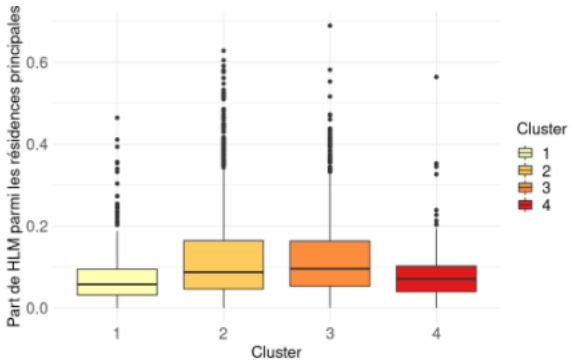

FIGURE 8.4 – Part de HLM parmi les résidences principales par cluster/zone

Les boîtes à moustaches ci-dessus traitent de la part d'appartements et de HLM [2] parmi les résidences principales. De la même manière que pour les deux figures précédentes, il est difficile d'établir un lien entre ces indicateurs et absentéisme. Les codes postaux des zones 1 et 4 contiennent globalement moins d'appartements (en proportion) que ceux des zones 2 et 3. Il s'agirait donc de codes postaux généralement plus ruraux que ceux des zones 2 et 3. Ceci est confirmé par les densités de population dans les différentes zones (cf. tableau 8.1). Néanmoins, les zones 1 et 4 se distinguent sur la présence de HLM. En effet, la part médiane de HLM parmi

2. HLM : Habitation à loyer modéré.

les résidences principales est de 5.74% dans la zone 1 contre 7.08% dans la zone 4 (soit 23% de plus). De la même façon, il y a plus de HLM (en proportion des résidences principales) dans la zone 3 que dans la zone 2 alors que la densité de population y est globalement supérieure dans la zone 2, ce qui serait potentiellement un signe que les habitants des zones 1 et 2 sont plus riches que ceux des zones 3 et 4. Ceci est confirmé avec les statistiques présentées dans le tableau 8.1. Le revenu déclaré médian par UC [3] est plus élevé dans les zones 1 et 2 par rapport aux zones 3 et 4.

	Zone	Min	Q_1	Q_2	Moyenne	Q_3	Max
Densité de population (habitants au km^2)	1	7.57	39.65	84.46	418.71	192.48	30481
	2	1.48	79.20	231.28	1518.42	920.02	44651
	3	5.69	63.54	139.83	617.12	493.94	10514
	4	5.83	45.89	100.89	367.18	260.04	6145
Revenu déclaré médian par UC	1	14520	19210	20607	21382	22918	41860
	2	7590	19331	21350	22344	24406	50280
	3	11230	18778	20277	20843	22455	40280
	4	12520	18827	20461	20972	22910	29420

Note : Q_1 et Q_3 correspondent au premier et troisième quartile. Q_2 est la médiane.
UC : unité de consommation

TABLE 8.1 – Densité de population et revenu déclaré médian par UC selon les zones

En résumé, cette analyse descriptive indiquerait que :
- les codes postaux de la zone 1 seraient globalement des territoires ruraux et plus riches que la zone 4 ;
- les codes postaux de la zone 2 seraient généralement des territoires urbains et plutôt riches ;
- les codes postaux de la zone 3 seraient plutôt des territoires pauvres, moins urbains que la zone 2 mais plus que les zones 1 et 4 ;
- les codes postaux de la zone 4 seraient généralement des territoires ruraux et plus pauvres que la zone 1.

Cette analyse descriptive est loin d'être suffisante pour interpréter le zonier incapacité. Un modèle de prédiction de la zone à partir des indicateurs de l'INSEE est donc réalisé. Cette modélisation est présentée dans la section suivante.

8.2 Construction du modèle de prédiction des classes de risques à partir des données INSEE

Cette section détaille l'élaboration du modèle de prédiction des classes de risques à partir des données INSEE. Cette modélisation est basée sur les forêts aléatoires car elles permettent en général d'obtenir de bonnes performances prédictives et ont tendance à moins sur-apprendre que d'autres modèles comme l'arbre de décision.

Les forêts aléatoires ou random forest correspondent à une combinaison d'arbres de décision dans lesquels chacun d'eux est construit à partir d'un échantillon aléatoire de la base d'apprentissage. Au vu du déséquilibre des classes de risques obtenues (cf. tableau 6.1), une

3. UC : unité de consommation, cf. définition en note de bas de page au chapitre 7.

attention particulière a été portée sur l'élaboration de cet échantillon. Ceci sera détaillé dans le paragraphe suivant. Par ailleurs, dans un random forest, chaque division est précédée d'un tirage aléatoire des variables et la division est construite en utilisant uniquement l'une de ces variables. Le nombre de variables qui sont sélectionnées aléatoirement comme candidates à chaque division fait donc l'objet d'un premier hyperparamètre du modèle. La convention pour la classification est d'utiliser la racine carrée du nombre de variables. Étant donné qu'il y a 17 variables explicatives [4] dans le modèle, ce premier paramètre est fixé à 4. Les deux autres paramètres à déterminer sont le nombre d'arbres et la taille minimale des nœuds terminaux. Pour choisir ces valeurs de manière optimale, une fonction Grid-search dans R est utilisée. Elle permet de comparer différentes combinaisons possibles d'hyperparamètres et choisit celle qui minimise l'erreur. Grâce au caractère aléatoire, un random forest est souvent plus efficace qu'un arbre de décision.

Comme évoqué précédemment, des précisions sont apportées concernant la définition de l'échantillon à considérer pour la construction de chacun des arbres de décision. Par défaut, le random forest fait pour chaque arbre, un tirage aléatoire avec remise de taille 3 598 (correspondant au nombre de codes postaux avec suffisamment d'affiliés). Afin de prendre en compte le déséquilibre des classes, un échantillon stratifié par rapport à la variable «zonier» est tiré à la place, pour chaque arbre. L'idée est de rééquilibrer un peu la distribution des codes postaux selon les zones. Ce ré-échantillonnage se justifie par le fait qu'un random forest a pour objectif de réduire l'erreur globale et a donc des difficultés à gérer les classes déséquilibrées. Il va avoir tendance à être performant pour identifier les codes postaux des zones les plus grandes (2 et 3) et beaucoup moins pour les zones plus petites (1 et 4). En pratique, sans ré-échantillonnage, il ne va quasiment jamais prédire la zone 4 car n'étant pas assez fréquente par rapport aux autres [5]. Ce ré-échantillonage va donc donner plus de poids aux zones les plus petites. Pour être plus précis, le tableau 8.2 compare la répartition des codes postaux selon les zones et celle des échantillons stratifiés.

Zone	Nombre de codes postaux	Proportion
Zone 1	397	**11.03%**
Zone 2	1816	**50.47%**
Zone 3	1248	**34.69%**
Zone 4	**137**	**3.81%**
Total	**3 598**	100%

Répartition des codes postaux selon les zones

Zone	Nombre de codes postaux	Proportion
Zone 1	262	**23%**
Zone 2	400	**35%**
Zone 3	343	**30%**
Zone 4	**137**	**12%**
Total	**1 142**	100%

Echantillon imposé pour la construction des arbres

TABLE 8.2 – Ré-échantillonnage mis en place pour le random forest

Les échantillons stratifiés (tableau de droite) ont une répartition beaucoup plus équilibrée selon les zones par rapport à la distribution initiale. Pour chaque arbre, il a été décidé de sélectionner 137 codes postaux de la zone 4, soit le nombre total de codes postaux de cette zone-là. Néanmoins, ce tirage est avec remise et tous les codes postaux de cette zone ne sont donc pas forcément sélectionnés pour l'élaboration de chaque arbre. Plus généralement, ces échantillons stratifiés permettent de diminuer le poids des zones les plus grandes (les proportions des zones 2 et 3 ont diminué par rapport à la répartition des codes postaux selon les zones) et d'accentuer

4. Ces variables explicatives correspondent aux 17 indicateurs démographiques et socio-économiques listés dans la figure 7.3.

5. Dans des exemples encore plus déséquilibrés, le random forest peut même prédire toutes les observations dans une même classe.

ceux des zones les plus petites (les proportions des zones 1 et 4 ont elles augmenté par rapport à la répartition des codes postaux selon les zones).

Les interprétations directes des forêts aléatoires sont plutôt restreintes et se limitent pour la plupart du temps à un calcul d'importance des variables. La figure 8.5 liste les indicateurs de l'INSEE, qui sont les variables explicatives du modèle étudié, par importance.

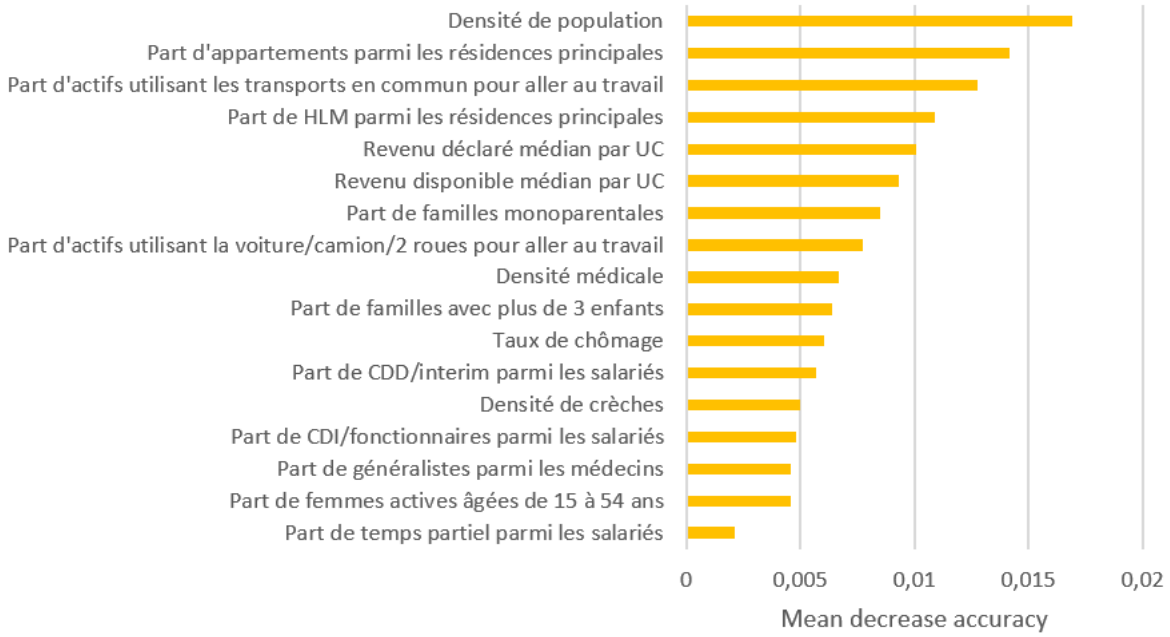

FIGURE 8.5 – Importance des variables (Mean decrease accuracy)

L'importance d'une variable correspond à la diminution moyenne de la précision (ou l'augmentation de l'erreur de prédiction) obtenue lorsque les valeurs de cette variable sont permutées aléatoirement. Plus précisément, si la permutation aléatoire de la variable diminue beaucoup la précision du modèle, la variable est importante. Ainsi, les trois variables qui semblent le plus discriminer les différentes zones sont :

— la densité de population,
— la part d'appartements parmi les résidences principales,
— la part d'actifs utilisant les transports en commun pour aller au travail.

Ces trois variables permettent notamment de distinguer les territoires urbains et ruraux. Il est donc légitime de penser que les différentes classes de risques du zonier ne sont pas équivalentes en termes d'urbanisation mais présentent des caractéristiques différentes à ce niveau-là. Ensuite, les trois variables les plus importantes qui suivent sont :

— la part de HLM parmi les résidences principales,
— le revenu déclaré médian par UC [6],
— le revenu disponible médian par UC.

Ces trois indicateurs se rapportent plutôt au thème de la richesse. Les différentes classes de risques du zonier ne seraient donc pas égalitaires en termes de revenu. Ces deux points confirment ce qui avait été observé avec les statistiques descriptives dans la section précédente 8.1.

Ce graphique est très intéressant puisqu'il aide à avoir une idée des variables les plus importantes. Cependant, il ne permet pas d'avoir le signe de la corrélation en fonction des zones. Par exemple, avec seulement ce graphique à disposition, il n'est pas possible de savoir quelles zones

6. UC : unité de consommation.

sont urbaines/rurales, ni lesquelles sont riches/pauvres. Cette analyse n'est donc pas complète et c'est la raison pour laquelle des interprétations à l'aide des valeurs de Shapley sont réalisées dans la section suivante.

8.3 Interprétation du zonier via les SHAP values

Cette section est divisée en trois sous-parties. Dans un premier temps, le concept de la valeur de Shapley est introduit. Ensuite, l'application de ce concept pour interpréter le modèle de prédiction des classes de risques est présentée. Enfin, une dernière sous-partie est consacrée à la présentation des résultats de cette méthode.

8.3.1 Présentation de la valeur de Shapley dans le cadre de la théorie des jeux

La valeur de Shapley trouve son origine dans la théorie des jeux et a été introduite en 1953 par Lloyd Shapley dans (Shapley, 1953)[6]. Il s'agit d'un concept de solution pour les jeux coopératifs, c'est-à-dire lorsque les joueurs collaborent pour avoir un gain. Plus précisément, l'objectif est de répartir le gain parmi l'ensemble des joueurs selon leur importance. La valeur de Shapley pour le joueur i, notée $\varphi_i(v)$ correspond à la valeur reçue par le joueur i.

Soit un jeu coopératif où $N = \{1, 2, ..., n\}$ est un ensemble fini de joueurs et v une fonction caractéristique qui à tout sous-ensemble de joueurs S associe $v(S)$ le gain de S. L'approche de Shapley est basée sur les quatre axiomes suivants :

- **Efficacité.** La somme des valeurs individuelles doit être égale à ce que peut obtenir la coalition de l'ensemble des joueurs (grande coalition), c'est-à-dire, $\sum_{i \in N} \varphi_i(v) = v(N)$
- **Symétrie.** Si deux joueurs i et j peuvent se substituer quelle que soit la coalition, alors ils doivent avoir le même gain.
  Autrement dit, si $v(S \cup \{i\}) = v(S \cup \{j\}) \ \forall \ S \not\ni i, j$ alors $\varphi_i(v) = \varphi_j(v)$
- **Nullité.** Un joueur est nul si $v(S \cup \{i\}) = v(S) \ \forall \ S \not\ni i$. Si i est nul alors $\varphi_i(v) = 0$.
- **Additivité.** Soit un joueur i qui participe à deux jeux ayant les mêmes joueurs et dont les fonctions caractéristiques sont v et w. On doit avoir $\varphi_i(v + w) = \varphi_i(v) + \varphi_i(w)$.
  Autrement dit, la solution d'un jeu défini comme la somme de deux jeux doit être égale à la somme des solutions de ces deux jeux.

Shapley a montré qu'il existe une unique solution φ qui satisfait l'ensemble de ces quatre axiomes. Il s'agit de la valeur de Shapley et elle se définit pour le joueur i par :

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S)) \tag{8.1}$$

où $|S|$ est le nombre de joueurs de la coalition S et n le nombre total de joueurs.
$\varphi_i(v)$ peut donc être vue comme une somme pondérée des contributions marginales $(v(S \cup \{i\}) - v(S))$ du joueur i.

Elle peut aussi se définir d'une manière plus pratique. Supposons que les joueurs rejoignent dans un certain ordre R la grande coalition [7]. La contribution marginale du joueur i dans l'ordre R correspond à son apport à la coalition composée des joueurs qui l'ont précédé. La valeur de Shapley pour le joueur i peut donc s'écrire :

$$\varphi_i(v) = \frac{1}{n!} \sum_R [v(S_i^R \cup \{i\}) - v(S_i^R)] \tag{8.2}$$

7. La grande coalition correspond à la coalition qui regroupe l'ensemble des joueurs.

où R parcourt toutes les $n!$ permutations de N et $S_i^R \subseteq N$ est la coalition des joueurs qui précèdent i dans l'ordre R.

Plusieurs situations peuvent être associées à des jeux coopératifs avec comme potentielle solution la valeur de Shapley. La formation de cartel d'entreprises pour décider de quotas de production, la coalition entre partis politiques ou la mesure du pouvoir de décision dans une assemblée (détaillée ci-dessous) en sont de bons exemples.

Considérons une assemblée "simplifiée" avec seulement 3 joueurs. L'objectif est de mesurer avec les valeurs de Shapley, le pouvoir de décision de chaque joueur dans différentes situations.

Cas 1 : Un texte est adopté si la majorité de l'assemblée vote en sa faveur
$v(N) = 1$ puisque si l'ensemble des joueurs vote en faveur du texte, il est adopté.
$v(S) = 1$ si $|S| \geq 2$ et $v(S) = 0$ sinon.
D'après l'axiome "efficacité", $\sum_{i \in N} \varphi_i(v) = v(N) = 1$.
Tous les joueurs sont "substituts", ils doivent donc avoir le même gain (axiome "symétrie").
Ainsi, $\varphi_1(v) = \varphi_2(v) = \varphi_3(v) = \frac{1}{3}$

Cas 2 : Le joueur 3 est un dictateur
$v(N) = 1$ puisque si l'ensemble des joueurs vote en faveur du texte, il est adopté.
$v(S) = 1$ si $\{3\} \subseteq S$ et $v(S) = 0$ sinon.
D'après l'axiome "efficacité", $\sum_{i \in N} \varphi_i(v) = v(N) = 1$.
Les joueurs 1 et 2 sont nuls. Ainsi, $\varphi_1(v) = \varphi_2(v) = 0$ d'après l'axiome "nullité". Enfin, $\varphi_3(v) = 1$.

Cas 3 : Le joueur 3 dispose d'un droit de veto. Un texte est adopté si la majorité de l'assemblée vote en sa faveur et si le joueur 3 n'a pas exercé son droit de veto
$v(N) = 1$ puisque si l'ensemble des joueurs vote en faveur du texte, il est adopté.
$v(S) = 1$ si $|S| \geq 2$ et $\{3\} \subseteq S$, $v(S) = 0$ sinon.
D'après les axiomes "efficacité" et "symétrie", $\varphi_1(v) = \varphi_2(v) = (1 - \varphi_3(v))/2$

Afin de pouvoir calculer le pouvoir de décision $\varphi_3(v)$ du joueur 3, la formule 8.2 est utilisée. Les résultats intermédiaires sont présentés dans le tableau 8.3.

Permutation dans l'ordre R	S_i^R	Contribution marginale du joueur 3
123	{1,2}	$v(\{1,2\} \cup \{3\}) - v(\{1,2\}) = 1 - 0 = 1$
132	{1}	$v(\{1\} \cup \{3\}) - v(\{1\}) = 1 - 0 = 1$
213	{2,1}	$v(\{2,1\} \cup \{3\}) - v(\{2,1\}) = 1 - 0 = 1$
231	{2}	$v(\{2\} \cup \{3\}) - v(\{2\}) = 1 - 0 = 1$
312	$\emptyset$	$v(\emptyset \cup \{3\}) - v(\emptyset) = 0 - 0 = 0$
321	$\emptyset$	$v(\emptyset \cup \{3\}) - v(\emptyset) = 0 - 0 = 0$

TABLE 8.3 – Exemple de calcul de valeurs de Shapley (cas 3) : Pouvoir de décision avec droit de veto

À partir du tableau ci-dessus, il est désormais possible d'en déduire la valeur de Shapley pour le joueur 3 : $\varphi_3(v) = \frac{1}{3!}(1 + 1 + 1 + 1) = \frac{4}{6}$.
Enfin, $\varphi_1(v) = \varphi_2(v) = \frac{1-\varphi_3(v)}{2} = \frac{1}{6}$

8.3.2 Utilisation de la valeur de Shapley pour interpréter le modèle de prédiction des classes de risques

La méthode SHAP [8], présentée dans (Lundberg et Lee, 2017)[3] vise à évaluer avec l'aide de la valeur de Shapley, le rôle de chaque variable explicative dans la prédiction d'un modèle. Cette procédure n'est pas totalement explicitée dans ce mémoire, notamment la méthode pour approximer les SHAP values. Pour davantage de détails, le lecteur pourra se reporter à (Lundberg et Lee, 2017)[3].

Dans notre application, le "jeu" correspond au fait de prédire la classe de risque pour un code postal. Les variables explicatives correspondent quant à elles aux "joueurs". Enfin, le "gain" est la différence entre la prédiction de l'observation et la prédiction moyenne sur toute la base. La valeur de Shapley correspond donc ici à la contribution marginale moyenne d'une variable explicative à la prédiction sur l'ensemble des coalitions possibles.

Pour faciliter la compréhension, l'exemple des SHAP values pour le 9e arrondissement de Paris est détaillé ci-dessous. Le tableau 8.4 fournit les prédictions associées à ce code postal ainsi que les prédictions moyennes obtenues sur toute la base.

	Zone 1	Zone 2	Zone 3	Zone 4
Code postal "75009"	0.126	0.806	0.048	0.02
Prédictions moyennes	0.204	0.388	0.313	0.095
Ecart de prédiction	-0.078	0.418	-0.265	-0.075

TABLE 8.4 – Comparaison des prédictions moyennes avec celles du code postal "75009" (9e arrondissement de Paris

La forêt aléatoire prédit la zone 2 (avec un score de 0.806) pour le 9e arrondissement de Paris. Il s'agit bien de la zone associée à ce code postal (cf. figure 6.6). La prédiction moyenne sur cette zone est de 0.388, soit un écart de 0.418 avec le 9e arrondissement de Paris. Étant donné que le modèle fournit une prédiction pour chaque zone, il est possible de calculer pour chacune d'entre elles, un écart par rapport à la prédiction moyenne. L'idée de la méthode SHAP est de quantifier la contribution de chaque variable pour expliquer cet écart. Les figures 8.6 et 8.7 listent ces contributions pour les prédictions des zones 2 et 3 dans le cas du 9e arrondissement de Paris.

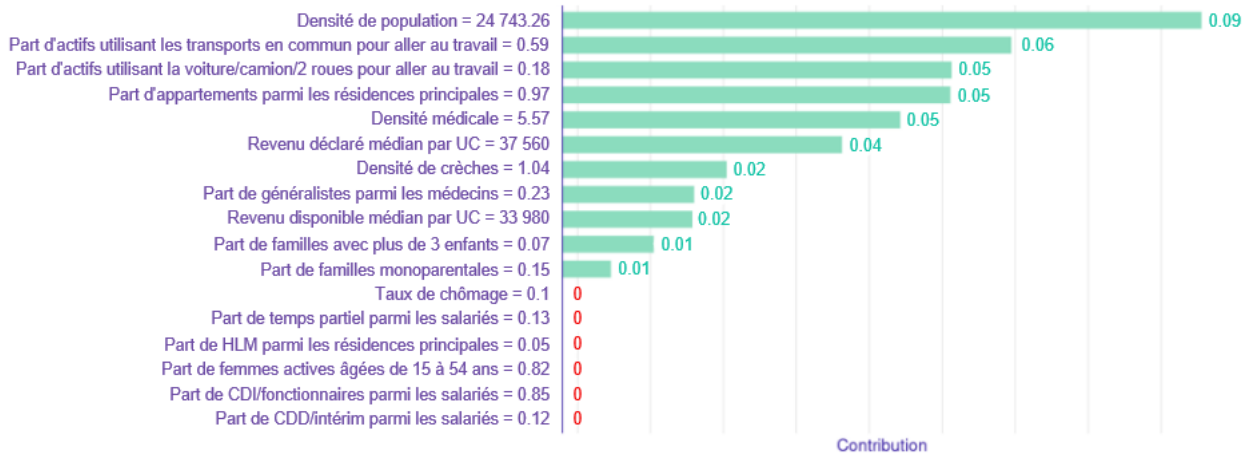

FIGURE 8.6 – Contributions des variables (SHAP values) à la prédiction "Zone 2" du 9e arrondissement de Paris

8. SHapley Additive exPlanations.

Dans ce type de graphique, les variables sont triées par ordre décroissant de contribution. La première variable de la liste est donc celle qui impacte le plus la prédiction pour le 9e arrondissement de Paris. Il s'agit de la densité de population qui contribue à hauteur de 0.09 dans la prédiction pour la zone 2. Autrement dit, pour cette observation et cette zone, la densité de population fait augmenter la prédiction de 0.09. Les autres variables à forte contribution sont :

— la part d'actifs utilisant les transports en commun pour aller au travail,
— la part d'actifs utilisant la voiture, camion ou deux roues pour aller au travail,
— la part d'appartements parmi les résidences principales.

Ces trois variables permettent de distinguer le caractère urbain/rural d'un territoire. Le fait que ces variables soient "importantes" pour la prédiction de la zone 2 est cohérent avec l'analyse descriptive réalisée au début du chapitre. En effet, les codes postaux de la zone 2 semblaient se distinguer de ceux des autres zones sur cette caractéristique [9].

Le graphique 8.6 nous informe directement des contributions de chaque variable mais pas du sens de la corrélation entre la variable et la zone. En effet, c'est la valeur de la variable pour cette observation qui contribue positivement/négativement à la prédiction et non la variable qui est corrélée positivement/négativement à la zone. Pour mieux comprendre, prenons l'exemple des deux variables suivantes.

— La densité de population du 9e arrondissement de Paris est considérable (plus de 24 000 habitants au km^2 alors que la densité de population française est de 105.1 habitants au km^2 en 2017 selon l'INSEE[27]). Ainsi, le fait d'avoir cette forte densité de population contribue positivement à la prédiction de la zone 2. Sur le 9e arrondissement de Paris, la corrélation entre la densité de population et la zone 2 est donc positive.
— 18% des actifs résidant dans le 9e arrondissement de Paris utilisent une voiture, camion ou deux roues pour se rendre au travail. Cette part est très faible par rapport à la moyenne nationale qui est de 74% en 2015 selon l'INSEE[28]. Ainsi, le fait d'avoir cette faible part contribue positivement à la prédiction de la zone 2. Sur le 9e arrondissement de Paris, la corrélation entre la part d'actifs utilisant la voiture, camion ou deux roues pour se rendre au travail et la zone 2 est donc négative.

Enfin, la somme des contributions des variables correspond à la différence entre la prédiction de l'observation et la moyenne. En effet, pour le 9e arrondissement de Paris, cela donne $0.09 + 0.06 + 0.05 \times 3 + 0.04 + 0.02 \times 3 + 0.01 \times 2 = 0.42$, qui correspond (en arrondissant) à l'écart mentionné dans le tableau 8.4.

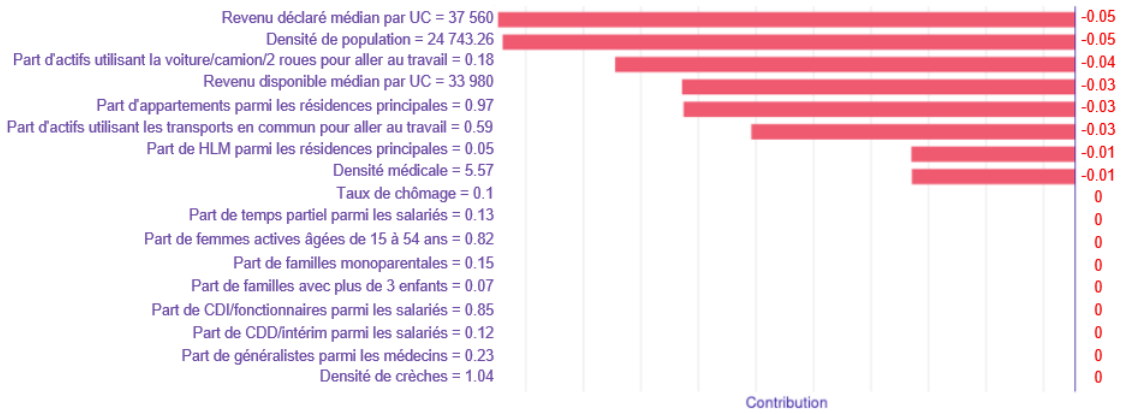

FIGURE 8.7 – Contributions des variables (SHAP values) à la prédiction "Zone 3" du 9e arrondissement de Paris

9. La zone 2 semble globalement plus urbaine que les autres zones.

La figure 8.7 reporte les SHAP values pour la prédiction de la zone 3 dans le cas du 9e arrondissement de Paris. Les contributions sont ici toutes négatives. Pour rappel, la zone 3 est prédite avec un score de 0.048 contre 0.313 en moyenne, soit un écart de -0.265 (cf. tableau 8.4). Le revenu déclaré médian par UC est la variable qui impacte le plus la prédiction de la zone 3 pour le 9e arrondissement de Paris. En effet, les revenus sont très élevés dans cet arrondissement et cela contribuerait à réduire le score associé à la zone 3. Sur le 9e arrondissement de Paris, la corrélation entre les revenus déclarés et la zone 3 est donc négative. Ceci fait écho à l'analyse descriptive réalisée en début de chapitre où de plus faibles revenus avaient été observés en zone 3.

Dans un objectif d'exhaustivité et afin de ne pas alourdir cette partie, les deux mêmes graphiques associés aux zones 1 et 4 sont donnés en annexe D. Bien que ces graphiques soient pertinents pour décrypter les prédictions, ils ne sont associés qu'à un seul code postal donné et il faut avoir une vision plus agrégée pour pouvoir en tirer des conclusions. C'est l'objet de la sous-partie suivante.

8.3.3 Résultats de l'application de SHAP sur le modèle de prédiction des classes de risques

Les graphiques de la sous-partie précédente permettent de décrypter les prédictions du modèle pour un seul code postal. Ceci est un des avantages de cette méthode puisque grâce à elle, il est possible de comprendre les prédictions de chaque observation individuellement, ce qui est particulièrement utile pour les prédictions atypiques par exemple. Néanmoins, dans un objectif d'interprétation globale du modèle, il faut agréger ces informations. Ceci est rendu possible par la représentation d'un "summary plot" qui reporte les SHAP values pour chaque variable et chaque observation. Ce graphique est construit en figure 8.8 pour la zone 1.

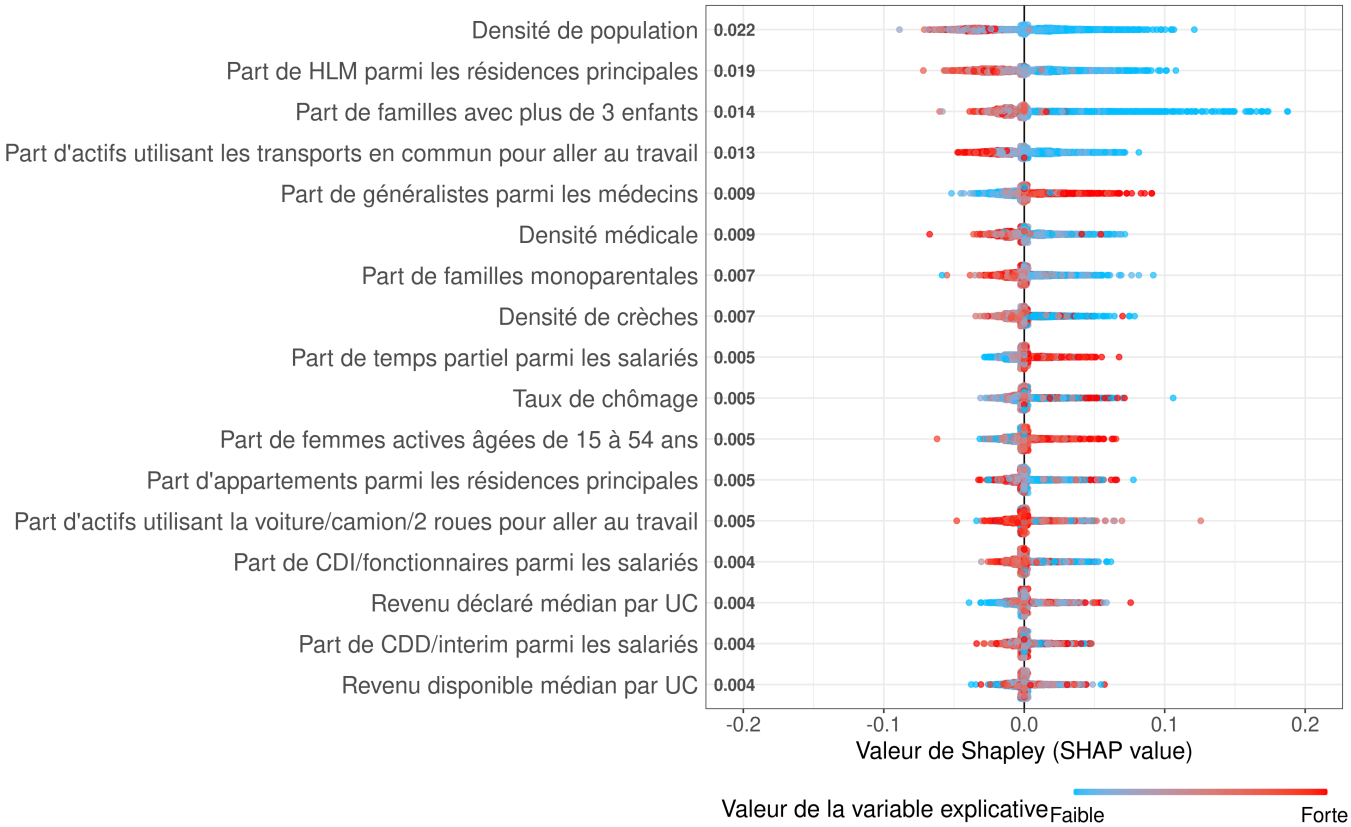

FIGURE 8.8 – SHAP values pour la zone 1 ("summary plot")

Bien que ce type de graphique ne soit pas forcément facile à interpréter dans un premier temps, il est malgré tout très informatif. Quelques détails sont donnés ci-dessous pour faciliter son interprétation. Tout d'abord, les variables explicatives sont triées dans l'ordre décroissant d'importance pour la zone en question. Ensuite, chaque point correspond à la valeur de Shapley pour une variable et un code postal (celles représentées dans les figures 8.6 et 8.7 par exemple). Enfin, ce point est coloré en fonction de la valeur de la variable explicative donnée (forte valeur en rouge et faible valeur en bleu). Ainsi, la densité de population, variable la plus importante pour discriminer la zone 1, zone la moins risquée, serait corrélée négativement avec cette zone. En effet, les codes postaux avec une faible densité de population (les points bleus) ont globalement des SHAP values positives, ce qui signifie qu'une faible densité de population augmente la probabilité d'être en zone 1. De manière complémentaire, les valeurs de Shapley sont généralement négatives pour les codes postaux à forte densité de population (les points rouges), signe qu'une forte densité de population diminue la probabilité d'être en zone 1. Nous retrouvons donc le résultat présenté dans le cadre de l'analyse descriptive qui suggérait que la zone 1 était plutôt composée de codes postaux ruraux.

La figure 8.9 correspond au "summary plot" associé à la zone 2.

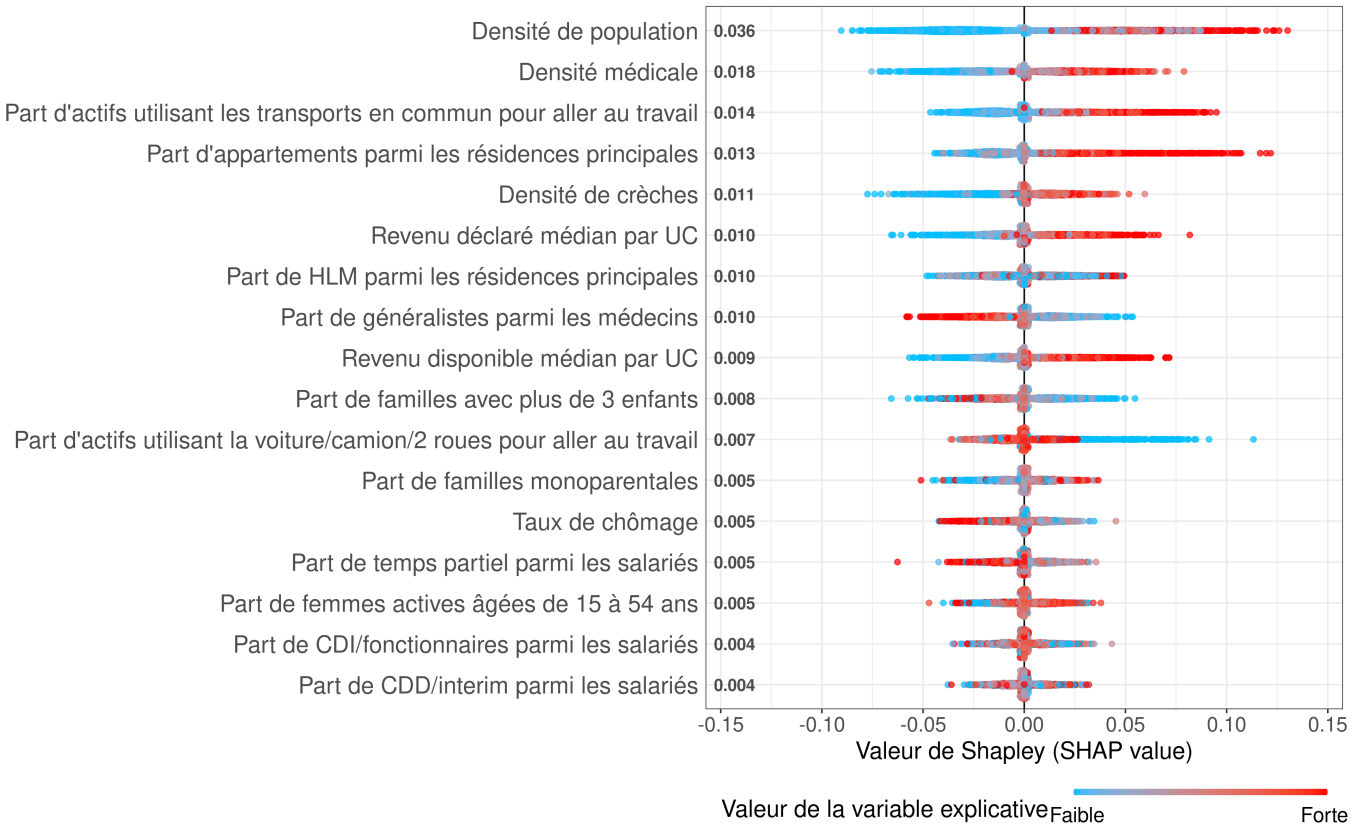

FIGURE 8.9 – SHAP values pour la zone 2 ("summary plot")

Contrairement à la zone 1, la densité de population est corrélée positivement avec la zone 2 puisque les codes postaux densément peuplés (points rouges) ont des valeurs de Shapley positives pour la zone 2. Nous retrouvons donc ici aussi, le résultat présenté dans l'analyse descriptive qui suggérait que la zone 2 était plutôt composée de codes postaux urbains.

Afin de ne pas alourdir cette partie, seulement les "summary plots" de la zone 1 et 2 sont représentés dans le corps du chapitre et ceux relatifs aux zones 3 et 4 sont fournis en annexe E. Néanmoins, le tableau 8.5 synthétise les informations présentes dans ces "summary plots"

et donne un aperçu des corrélations entre les différents indicateurs et les zones. Par exemple, les trois variables les plus importantes pour la zone 1 sont corrélées négativement avec cette dernière, d'où le fait de mentionner "− − −" dans les cases correspondantes.

Variable	Zone 1	Zone 2	Zone 3	Zone 4
Densité de population	− − −	+ + +	−	−
Part de HLM parmi les résidences principales	− − −		+ + +	− − −
Part de familles avec plus de 3 enfants	− − −		+ + +	+
Part d'actifs utilisant les transports en commun pour aller au travail	− −	+ + +	−	
Part de généralistes parmi les médecins	+ +	−	−	+
Densité médicale	− −	+ + +		− −
Part de familles monoparentales	−			
Densité de crèches	−	+ +		−
Part de temps partiel parmi les salariés	+			
Taux de chômage				
Part de femmes actives âgés de 15 à 54 ans				−
Part d'appartements parmi les résidences principales		+ + +		− −
Part d'actifs utilisant la voiture/camion/ 2 roues pour aller au travail			+	
Part de CDI/fonctionnaires parmi les salariés				
Revenu déclaré médian par UC		+ +	− − −	
Part de CDD/intérim parmi les salariés				
Revenu disponible médian par UC		+	− −	

Note : Le nombre de +/- correspond au degré de la corrélation
Note : Les cas où il est difficile d'établir une corrélation entre les variables et les zones (avec les graphiques) sont représentés avec des cases vides

TABLE 8.5 – Récapitulatif des corrélations des variables avec chaque zone

Les variables qui discriminent le plus les différentes zones sont la densité de population et la part de HLM parmi les résidences principales. Par exemple, comme évoqué précédemment, la zone 2 contient globalement des codes postaux plus urbains que les autres zones. De plus, la zone 3 semble se caractériser par une proportion de HLM plus importante que les autres zones. Ce résultat avait déjà été présenté lors de l'analyse descriptive mais la différence entre la zone

2 et 3 semblait moins importante que celle retranscrite dans le tableau 8.5 et issue de l'analyse des SHAP values.

Les corrélations entre les variables et les zones sont en général cohérentes. Par exemple, la zone 2, plus densément peuplée que les autres, se caractérise aussi par :
— une forte part d'actifs allant au travail en transports en commun,
— une forte densité médicale,
— une forte part d'appartements parmi les résidences principales,
— une importante densité de crèches,
— des revenus plus élevés.
Ces caractéristiques sont dans la majorité des cas, vérifiés pour les villes françaises.

Le caractère plus défavorisé de la zone 3 est confirmé dans ce tableau puisque les revenus (déclarés et disponibles) sont liés négativement à la zone 3. La zone 4, plus difficile à interpréter car étant celle avec le moins de codes postaux, se caractérise par des territoires plutôt ruraux et faiblement équipés (faible densité de crèches et de médecins). La proportion de HLM est faible dans cette zone, tout comme dans la zone 1, certainement liée à l'absence d'obligation légale pour les plus petites communes.

En résumé, les SHAP values permettent de dresser le panorama suivant :
— **Zone 1.** La zone la moins risquée en termes de probabilité de tomber en arrêt de travail est aussi la plus rurale de toutes les zones. Elle réunit une majorité des critères associés à une zone rurale : faible densité de population, peu de HLM (en proportion des résidences principales), peu d'actifs utilisant les transports en commun pour aller travailler et des territoires faiblement équipés (peu de médecins et de crèches).
— **Zone 2.** Cette zone est la plus urbaine parmi les quatre. Elle rassemble une grande partie des caractéristiques liées aux zones urbaines : forte densité de population, beaucoup d'actifs utilisant les transports en commun pour se rendre au travail, des résidences principales plutôt composées d'appartements et des territoires plutôt équipés (beaucoup de crèches, de médecins et plus particulièrement de médecins spécialistes). C'est aussi dans cette zone que les habitants sont globalement les plus riches.
— **Zone 3.** Cette zone semble être la plus défavorisée. C'est clairement celle où les habitants ont les revenus les plus faibles et une importante présence de HLM est observée dans cette zone.
— **Zone 4.** La zone la plus risquée en termes de probabilité de tomber en arrêt de travail est aussi une zone plutôt rurale. Elle se caractérise par des faibles parts de HLM et d'appartements parmi les résidences principales, des territoires faiblement équipés (peu de médecins et de crèches) mais une densité de population plus importante que la zone 1 [10].

Il est possible de construire un graphique encore plus agrégé que le "summary plot", en considérant l'ensemble des zones. Plus précisément, il s'agit de reporter par zone, la moyenne (en valeur absolue) des SHAP values pour chaque variable. Cet indicateur qui correspond à une mesure d'importance des variables est représenté en figure 8.10.

10. Ces résultats peuvent facilement se retrouver à partir du tableau 8.5.

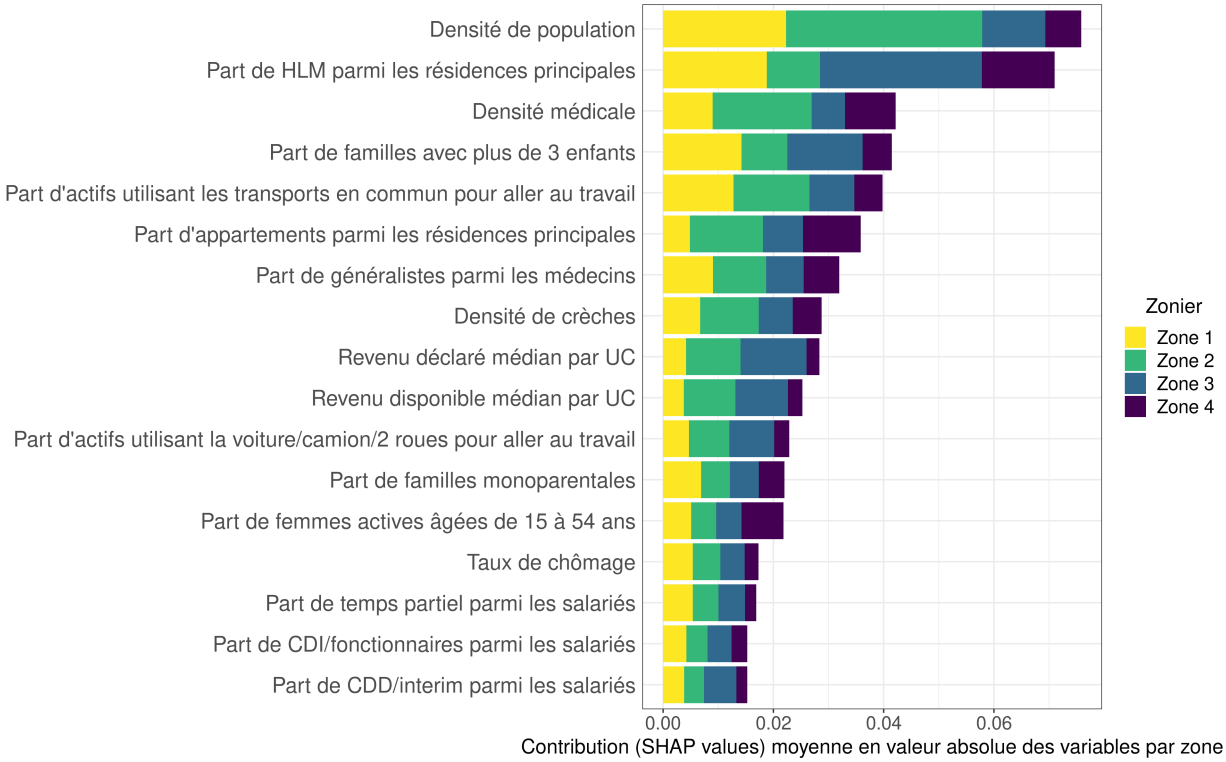

FIGURE 8.10 – Contribution des variables explicatives à la prédiction de chaque classe

Toutes zones confondues, la variable la plus importante selon cette méthode est la densité de population. Cette dernière était aussi la plus importante en utilisant le critère "mean decrease accuracy" (cf. figure 8.5). Dans la suite, le classement des variables par importance est différent pour les deux approches mais demeure globalement proches. Par exemple, les variables liées au marché du travail (taux de chômage, part de CDI/CDD/temps partiel/femmes actives) ne sont pas considérées comme importantes pour discriminer les différentes zones, que ce soit par l'approche SHAP ou par "mean decrease accuracy". Un avantage de l'approche par SHAP est qu'elle permet de quantifier l'importance en fonction des zones, ce qui n'est pas le cas avec l'approche "mean decrease accuracy". Par exemple, les revenus sont des caractéristiques plus importantes pour les zones 2 et 3 que pour les zones 1 et 4.

Ce graphique seul ne permet pas de faire une interprétation complète de chacune des zones. En effet, contrairement aux "summary plots" précédents, ce graphique ne nous informe pas du signe de la corrélation entre la variable et la zone.

La méthode SHAP présente plusieurs avantages.

1. Elle rend possible l'interprétation de modèles de Machine Learning qui sont considérés comme des boîtes noires (Random Forest, XGBoost, etc.)
2. Elle fournit une interprétation globale en indiquant le signe et l'intensité des contributions de chaque variable sur la cible, chose qui n'est pas disponible avec seulement un "classique" graphique d'importance de variables.
3. Elle autorise aussi une interprétation locale puisque chaque code postal a ses propres SHAP values, qui permettent de décomposer individuellement chacune des prédictions.

Cependant, cette méthode admet aussi des inconvénients. En effet, calculer des valeurs de Shapley sur de très grandes bases de données devient très rapidement long, voire impossible. Dans notre cas, étant donné que le code postal était la maille retenue dans le modèle, le nombre de lignes n'était donc pas très élevé et le temps de calcul raisonnable.

Pour rappel, l'élaboration d'un modèle prédictif des classes de risques du zonier a deux objectifs :

— Interpréter le zonier avec l'aide des valeurs de Shapley et des indicateurs démographiques et socio-économiques.
— Déterminer la classe de risque pour les codes postaux avec peu ou pas d'affiliés, en utilisant leurs caractéristiques démographiques et socio-économiques. Pour rappel, ces codes postaux n'étaient pas associés à une classe de risque puisqu'ils n'étaient pas inclus dans la classification hiérarchique spatiale.

Les interprétations du zonier ont été détaillées dans les premières sections de ce chapitre. Les éléments relatifs à l'affectation des classes de risques aux codes postaux avec peu ou pas d'affiliés sont présentés dans la section suivante.

8.4 Affectation des classes de risques aux codes postaux avec peu ou pas d'affiliés

Plus de 2 400 codes postaux de France métropolitaine ont peu ou pas d'affiliés "Malakoff Humanis". Ces derniers ne sont pas encore liés à une classe de risque puisque l'estimation de leur sinistralité est jugée insuffisamment robuste. Néanmoins, afin de pouvoir proposer un tarif partout en France, chaque code postal doit être associé à une zone. La méthodologie pour mener à bien cette affectation est présentée dans cette section.

Le random forest considéré précédemment est un modèle de Machine Learning qui renvoie pour toute nouvelle observation, une probabilité d'appartenance à chacune des zones. Il est donc possible de l'utiliser pour déterminer la classe de risque des codes postaux avec peu ou pas d'affiliés, en leur affectant la zone ayant la probabilité prédite par le modèle la plus élevée. Ainsi, les classes de risques seraient attribuées aux codes postaux concernés en fonction de leurs caractéristiques démographiques et socio-économiques puisque ce sont ces indicateurs qui sont les variables explicatives du random forest. Cette méthode présente le défaut de ne pas prendre en compte la proximité géographique dans l'affectation des classes de risques. En effet, le random forest pourrait attribuer pour un code postal donné, une classe de risque différente de celles de ses voisins, ce qui réduirait le lissage du zonier et créerait des territoires isolés.

Pour résoudre le problème précédent et tenir compte de la proximité géographique, une idée est d'attribuer pour un code postal donné (et non encore affecté à une zone), la classe de risque ayant la plus grande probabilité parmi celles de ses 4 plus proches voisins. Par exemple, si les voisins d'un code postal donné sont associés soit à la classe 1, soit à la classe 2 alors ce code postal sera affecté à la classe qui aura la probabilité prédite par le modèle la plus élevée entre la classe 1 et 2. Cette méthode, qui tient compte des classes de risques du voisinage, a donc pour principal avantage de conserver une cohérence géographique dans le zonier, en ne créant pas de territoires isolés lors de cette étape d'affectation de la classe de risque aux codes postaux avec peu ou pas d'affiliés.

Les figures 8.11 et 8.12 représentent le résultat de cette affectation qui permet d'obtenir un zonier complet. Chaque code postal est désormais associé à une classe de risque.

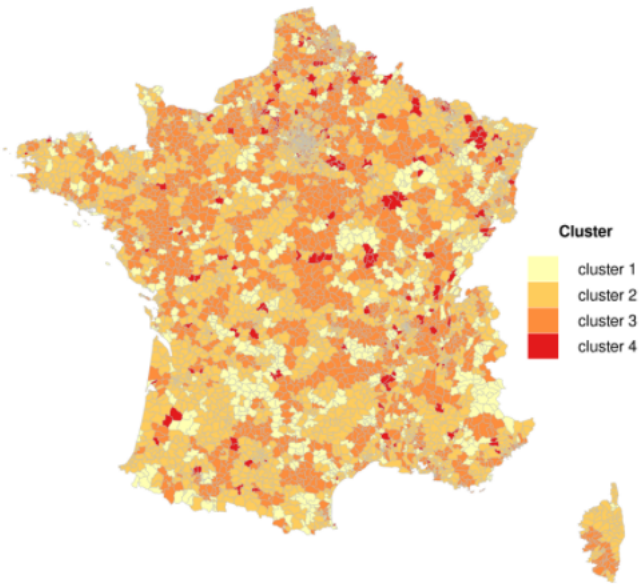

FIGURE 8.11 – Zonier après traitement des codes postaux avec peu ou pas d'affiliés (France métropolitaine)

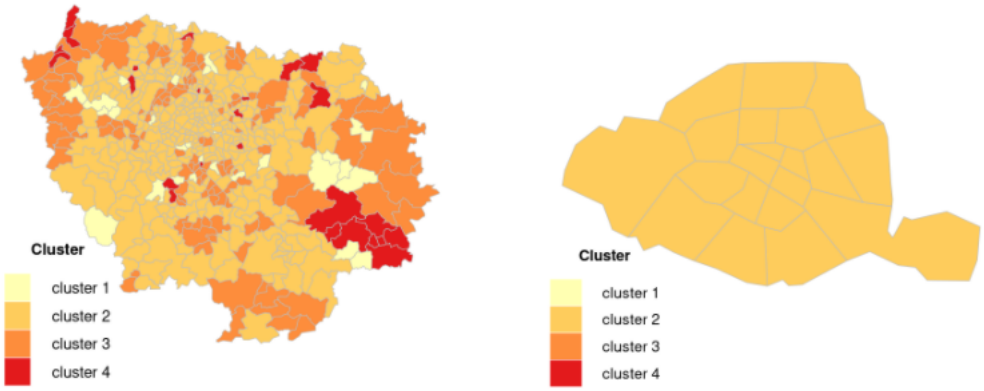

FIGURE 8.12 – Zonier après traitement des codes postaux avec peu ou pas d'affiliés (Île de France et Paris)

Les cartes ci-dessus sont davantage lissées par rapport à celles présentées avant traitement des codes postaux avec peu ou pas d'affiliés (cf. figures 6.5 et 6.6). Ceci est dû au choix de prendre en compte la proximité géographique dans l'attribution des classes de risques aux codes postaux dont le calcul de la sinistralité n'était pas assez robuste.

Le zonier obtenu n'isole pas chaque classe de risque dans une certaine partie de la France métropolitaine. Par exemple, tous les codes postaux de la zone 4 ne sont pas situés qu'au nord/sud, ni à l'est/ouest et ceci est valable pour toutes les autres zones. L'objectif n'était pas d'obtenir une telle séparation mais si ça avait été le cas, elle aurait facilité la communication autour de ce zonier, notamment auprès des personnes moins proches de la fonction actuarielle (commerciaux, assurés, etc.). Quelques regroupements peuvent tout de même s'observer sur certaines zones. C'est notamment le cas pour les régions Pays de la Loire et Centre-Val de Loire où les codes postaux sont dans la grande majorité classés dans la zone 3. Aussi, les codes postaux du Sud-Ouest de la France appartiennent généralement aux classes 1 et 2.

Les classes de risques obtenues après l'étape d'affectation sont toujours déséquilibrées (cf. tableau 8.6). Les zones 2 et 3 demeurent encore majoritaires. Cette disparité était déjà présente

dans le zonier avant cette affectation (cf. tableau 6.1). Néanmoins, la répartition des codes postaux dans les différentes zones sont très proches avant/après l'étape d'attribution des classes de risques aux codes postaux avec peu ou pas d'affiliés. Seule la zone 1 présente une différence un peu plus importante. En effet, elle est légèrement plus représentée dans le zonier final (14.17% contre 11.03% dans le zonier intermédiaire).

Cluster	Nombre de codes postaux	Proportion
Cluster 1	857	14.17%
Cluster 2	2 950	48.78%
Cluster 3	2 045	33.81%
Cluster 4	196	3.24%

Table 8.6 – Répartition de l'ensemble des codes postaux au sein des différents clusters

Avec l'aide des valeurs de Shapley, les open data de l'INSEE ont permis de mieux appréhender les caractéristiques démographiques et socio-économiques de chaque classe de risque du zonier. Elles ont aussi aidé à attribuer une classe de risque aux codes postaux dont le calcul de la sinistralité n'était pas considéré comme robuste. Le zonier est donc dorénavant complet et chaque code postal de France métropolitaine [11] est associé à une classe de risque. Ainsi, il est désormais possible d'évaluer la qualité de ce zonier et plus généralement de valider son apport dans la connaissance du risque incapacité. Tel est l'objectif de la dernière partie de ce mémoire.

11. Pour rappel, cette étude n'est basée que sur la France métropolitaine. La modélisation, via une classification hiérarchique spatiale ne semble pas adaptée aux territoires ultramarins puisqu'il s'agit en général de territoires isolés (peu de voisins, île, etc.).

Cinquième partie

Validation de l'apport du zonier dans la connaissance du risque incapacité

Chapitre 9

Apports du zonier dans la connaissance du risque incapacité

Le chapitre 9 vise à valider l'apport du zonier dans la connaissance du risque incapacité en comparant notamment les modélisations avec et sans zonier. Avant cela, une simple analyse bivariée permettant de quantifier la liaison entre la sinistralité et la variable "zonier" est réalisée avec l'aide du V de Cramer. Dans un deuxième temps, une modélisation du risque incapacité incluant le zonier est construite, permettant d'étudier la significativité des coefficients associés aux différentes zones. Ensuite, d'autres critères comme les critères d'information d'Akaike (AIC) et bayésien (BIC), ainsi qu'un test statistique de comparaison de modèles emboîtés [1] sont présentés. Enfin, le zonier est appliqué sur une base de données "test" pour confirmer sa robustesse et sa pertinence.

9.1 V de Cramer

Le V de Cramer est une mesure d'association entre deux variables qualitatives. Quelques notations sont introduites ci-dessous pour définir ce coefficient. Soit un couple de variables qualitatives (X, Y) où X a I modalités notées $a_1, ..., a_I$ et Y a J modalités notées $b_1, ..., b_J$. n_{ij} représente le nombre de fois où le couple (a_i, b_j) est observé. Aussi, $n_{i.} = \sum_{j=1}^{J} n_{ij}$ et $n_{.j} = \sum_{i=1}^{I} n_{ij}$.

Grâce aux notations introduites, il est possible de définir la distance du χ^2 d'écart à l'indépendance par :

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} \right)$$

Cette distance est utilisée pour définir le V de Cramer :

$$V = \sqrt{\frac{\chi^2}{n \times \min\{I-1, J-1\}}}$$

Le V de Cramer est compris entre 0 et 1 (0 indiquant une absence d'association [2] et 1 une dépendance parfaite [3]). Ce coefficient peut donc être utilisé pour comparer la liaison entre plusieurs couples de variables.

1. Deux modèles sont emboîtés si le plus "grand" modèle contient toutes les variables du plus "petit" modèle.
2. Autrement dit, lorsque $V = 0$, les variables X et Y sont indépendantes.
3. Une dépendance parfaite signifie que chaque variable est complètement déterminée par l'autre.

Le tableau 9.1 reporte les valeurs des V de Cramer entre la sinistralité [4] et chacune des variables suivantes : CSP, âge, genre, secteur d'activité et zonier.

	CSP	Âge	Genre	Secteur d'activité	Zonier
Association (V de Cramer) avec la sinistralité	0.08225	0.04172	0.04867	0.04142	0.06316

TABLE 9.1 – V de Cramer entre la sinistralité et chacune des variables

Les V de Cramer peuvent paraître faibles mais il faut garder en tête que prédire l'arrêt de travail est un sujet assez complexe. En regardant seulement cette statistique, il semblerait que le zonier soit bien corrélé à la sinistralité puisque son V de Cramer est supérieur à ceux de l'âge, du genre et du secteur d'activité.

Cependant, le V de Cramer nous informe seulement du degré d'association entre les deux variables mais pas du signe de cette dépendance par modalité. De plus, il s'agit seulement d'une analyse bivariée qui ne tient pas compte des potentielles corrélations entre les variables. Passer dans un cadre multivarié est donc essentiel pour aboutir à des résultats plus complets et robustes. C'est l'objet de la section suivante.

9.2 Significativité des coefficients de régression

Afin d'évaluer l'impact du zonier sur le risque incapacité, une modélisation via une régression logistique multinomiale est utilisée. Plus précisément, nous reprenons la régression du chapitre 3 à laquelle nous ajoutons la variable "zonier". Ce modèle se fait donc à la maille individuelle et contient en plus de la variable zonier, l'ensemble des critères usuels de tarification, à savoir l'âge, la CSP, le genre [5] et le secteur d'activité.

Le tableau 9.2 reporte les coefficients obtenus par la régression logistique multinomiale décrite plus haut.

4. Pour rappel, la variable de sinistralité correspond à la survenance d'un arrêt de travail d'une certaine durée ("Absence d'AT", "]0,15]", "]15,30]", "]30,90]" et "+90").

5. Étant donné que cette modélisation se fait à la maille individuelle, la variable "genre" est utilisée au lieu de la variable "répartition homme/femme". Pour rappel, la répartition homme/femme est un critère de tarification autorisé en prévoyance collective. Par contre, il est interdit de proposer un tarif différencié entre les femmes et les hommes d'une même entreprise.

	Dependent variable :			
	]0,15]	]15,30]	]30,90]	+90
Femme	0.219***	0.286***	0.275***	0.402***
	(0.004)	(0.007)	(0.007)	(0.009)
Âge]20,25]	0.486***	0.300***	0.323***	0.586***
	(0.012)	(0.020)	(0.022)	(0.037)
Âge]25,30]	0.714***	0.555***	0.629***	1.107***
	(0.012)	(0.020)	(0.021)	(0.035)
Âge]30,40]	0.830***	0.701***	0.782***	1.357***
	(0.012)	(0.019)	(0.020)	(0.034)
Âge]40,50]	0.716***	0.679***	0.812***	1.496***
	(0.012)	(0.019)	(0.020)	(0.034)
Âge]50,55]	0.678***	0.713***	0.905***	1.700***
	(0.013)	(0.020)	(0.021)	(0.034)
Âge +55	0.418***	0.550***	0.784***	1.697***
	(0.013)	(0.020)	(0.022)	(0.034)
CSP Agent maîtrise	0.479***	0.692***	0.712***	0.676***
	(0.006)	(0.011)	(0.011)	(0.015)
CSP Employé	0.239***	0.639***	0.780***	0.864***
	(0.005)	(0.009)	(0.010)	(0.012)
CSP Ouvrier	0.685***	1.207***	1.306***	1.329***
	(0.006)	(0.010)	(0.011)	(0.014)
Secteur d'activité Classe 2	−0.004	0.023**	0.050***	0.116***
	(0.006)	(0.012)	(0.013)	(0.018)
Secteur d'activité Classe 3	−0.186***	0.0005	0.153***	0.388***
	(0.007)	(0.012)	(0.013)	(0.019)
Secteur d'activité Classe 4	0.054***	0.178***	0.289***	0.448***
	(0.008)	(0.014)	(0.015)	(0.020)
Zone 2	1.750***	1.762***	1.547***	1.451***
	(0.026)	(0.045)	(0.042)	(0.053)
Zone 3	1.849***	1.872***	1.693***	1.675***
	(0.026)	(0.045)	(0.042)	(0.053)
Zone 4	2.223***	2.548***	2.674***	2.552***
	(0.032)	(0.052)	(0.048)	(0.060)
Constant	−4.333***	−5.912***	−6.076***	−7.404***
	(0.029)	(0.050)	(0.048)	(0.064)
Akaike Inf. Crit.	4,002,825.000	4,002,825.000	4,002,825.000	4,002,825.000
Note :				*p<0.1 ; **p<0.05 ; ***p<0.01

TABLE 9.2 – Coefficients obtenus avec la régression multinomiale (en incorporant le zonier)

Tout d'abord, les coefficients associés aux variables âge, CSP, genre et secteur d'activité sont globalement stables par rapport à la même régression avant l'ajout du zonier (cf. tableau 3.1). Cela signifie que l'estimation de ces coefficients semble robuste à l'ajout de nouvelles variables.

La modalité de référence pour la variable "zonier" est la zone 1, zone la moins risquée. Il est donc cohérent que tous les coefficients liés à cette variable soient positifs. En effet, cette positivité des coefficients signifie que toutes les zones sont plus risquées que la zone 1. Ces coefficients sont aussi significatifs à 1%, signe d'une certaine pertinence du zonier. De plus, quelle que soit la durée des arrêts de travail considérée ("]0,15]", "]15,30]", "]30,90]" et "+90"), les coefficients augmentent avec la zone. Autrement dit, pour chaque durée d'arrêt de travail, les coefficients liés à la zone 4 sont toujours supérieurs à ceux de la zone 3, qui sont eux plus grands que ceux de la zone 2. Des exemples d'interprétations des coefficients sont donnés ci-dessous. Pour rappel, après passage à l'exponentielle, les coefficients de la régression logistique multinomiale peuvent s'interpréter comme des odds ratios (cf. section 3.2.1, chapitre 3).

Arrêts de travail de plus de 90 jours : Zone 3 vs Zone 1 Toutes choses égales par ailleurs, la probabilité pour un individu travaillant dans la zone 3 d'être plus de 90 jours en arrêt de travail par rapport à celle de ne pas être en arrêt est 5.3388 ($e^{1.675}$) fois (soit 433.88% plus élevé) l'odds (côte) pour un individu travaillant dans la zone 1. En résumé, si un certain profil d'individus de la zone 1 a un odds de 0.03 alors cet odds pour le même profil, sauf pour un individu de la zone 3 est de $0.03 \times 5.3388 \approx 0.16$. Ainsi, pour cet individu de la zone 1, la probabilité d'être en arrêt plus de 90 jours représente 3% de celle d'absence d'arrêt de travail alors qu'elle représente 16% pour cet individu de la zone 3. Autrement dit, cet individu de la zone 1 a une probabilité de ne pas être en arrêt de travail 33.33 (1/0.03) fois supérieure à celle d'avoir un arrêt de plus de 90 jours alors que pour l'individu de la zone 3, elle n'est que 6.25 fois supérieure.

Arrêts de travail de plus de 90 jours : Zone 3 vs Zone 2 Toutes choses égales par ailleurs, la probabilité pour un individu travaillant dans la zone 3 d'avoir un arrêt de travail de plus de 90 jours par rapport à celle de ne pas être en arrêt est $e^{1.675-1.451} = e^{0.224} \approx 1.2511$ fois (soit 25.11% plus élevé) l'odds d'un individu travaillant dans la zone 2.

Les écarts d'odds sont beaucoup plus importants entre les zones 1 et 3 plutôt qu'entre les zones 2 et 3. Ceci est dû au fait que la zone 1 contient des codes postaux faiblement sinistrés et pouvant être considérés comme atypiques. Bien que les écarts de coefficients entre les zones 2 et 3 soient plus réduits, ils demeurent suffisants pour conclure à une différence de sinistralité entre les zones 2 et 3, toutes choses égales par ailleurs.

En résumé, cette section a permis de confirmer la cohérence du zonier puisque "toutes choses égales par ailleurs", les zones considérées comme les plus (respectivement moins) risquées sont celles avec les coefficients les plus (respectivement moins) élevés. De manière plus générale, la section suivante vise à prouver, en comparant les modélisations avec et sans zonier, que le zonier améliore la connaissance du risque incapacité.

9.3 Comparaison des modélisations du risque incapacité avec et sans zonier

Les modélisations avec et sans zonier sont comparées par le biais d'un test statistique et des critères d'information AIC/BIC.

9.3.1 Test statistique

Le test statistique utilisé pour comparer ces deux modélisations est le test du rapport de vraisemblance (likelihood-ratio test en anglais) qui permet de tester des hypothèses sur des modèles emboîtés [6].

Dans notre cas, il s'agit de tester :

$$H_0 : \text{modèle sans zonier} \quad \text{vs} \quad H_1 : \text{modèle avec zonier} \tag{9.1}$$

Pour être plus précis, l'hypothèse nulle (H_0) suppose que l'ensemble des 12 coefficients associés à la variable zonier (cf. coefficients surlignés en orange dans le tableau 9.2) valent 0. Réciproquement, H_1 correspond au cas où au moins l'un de ces 12 coefficients est significativement différent de 0.

6. Deux modèles sont emboîtés si le plus "grand" modèle contient toutes les variables du plus "petit" modèle.

L'application de ce test sur nos données conclut à rejeter H_0 (p-value < 0.01). Autrement dit, le modèle avec zonier est privilégié. Ce résultat n'est pas surprenant puisque, considéré individuellement, chaque coefficient était déjà significatif (cf. tableau 9.2).

9.3.2 Ajustement du modèle

Pour mesurer la qualité d'ajustement d'un modèle linéaire généralisé, la déviance est souvent utilisée. Cependant, comme pour le R^2 dans la régression linéaire qui augmente systématiquement lorsque de nouvelles variables sont incorporées au modèle, la déviance elle, diminue suite à l'ajout de variables explicatives. Afin de juger correctement de la pertinence d'une variable, il faut donc pénaliser cette baisse systématique de déviance par un terme reflétant la complexité du modèle. C'est l'idée des critères d'information d'Akaike (AIC) et bayésien (BIC).

L'AIC (Akaike Information Criterion en anglais) est une mesure de la qualité d'un modèle statistique pénalisant le modèle en fonction du nombre de ses paramètres. Plus précisément, il se définit par $AIC = -2\ln(L) + 2k$ où k est le nombre de paramètres à estimer dans le modèle et L est le maximum de la fonction de vraisemblance du modèle.

Le BIC (Bayesian Information Criterion) est un critère très proche de l'AIC qui permet aussi de sélectionner le modèle le plus pertinent parmi un ensemble fini d'entre eux. Il se définit par $BIC = -2\ln(L) + k\ln n$ avec n le nombre d'observations. La différence avec l'AIC est que la pénalisation dépend aussi de la taille de l'échantillon et pas uniquement du nombre de paramètres. Ainsi, le BIC pénalise plus fortement le nombre de paramètres que l'AIC.

Dans une approche de sélection de modèles, le modèle choisi est celui associé à l'AIC et/ou au BIC le plus faible.

Les tableaux 9.3 et 9.4 reportent les AIC/BIC :

— du modèle construit avec les variables âge, genre, CSP, secteur d'activité et zonier (2e colonne) ;

— du même modèle sauf en enlevant une variable (3e colonne).

La dernière colonne correspond à la différence des deux AIC/BIC calculés dans les 2e et 3e colonnes. L'idée de ces tableaux est de vérifier dans un premier temps que l'AIC et le BIC sont plus petits dans le modèle avec zonier. Ensuite, ils permettent de vérifier que le gain d'information apporté par le zonier est important en le comparant à celui procuré par les autres variables.

Variable	AIC (avec la variable)	AIC (en enlevant la variable)	Différence d'AIC
CSP	4 002 825,03	4 046 914,09	44 089,06
Zonier	4 002 825,03	4 020 984,84	18 159,81
Secteur d'activité	4 002 825,03	4 007 412,58	4 587,55
Âge	4 002 825,03	4 023 259,17	20 434,14
Genre	4 002 825,03	4 009 656,34	6 831,31

TABLE 9.3 – Comparaison des AIC avec/sans chacune des variables

Variable	BIC (avec la variable)	BIC (en enlevant la variable)	Différence de BIC
CSP	4 003 686,80	4 047 623,78	43 936,98
Zonier	4 003 686,80	4 021 694,53	18 007,73
Secteur d'activité	4 003 686,80	4 008 122,27	4 435,47
Âge	4 003 686,80	4 023 816,79	20 129,98
Genre	4 003 686,80	4 010 467,42	6 780,62

TABLE 9.4 – Comparaison des BIC avec/sans chacune des variables

L'AIC (respectivement BIC) associé au modèle contenant l'ensemble des variables (âge, genre, CSP, secteur d'activité et zonier) vaut 4 002 825,03 (respectivement 4 003 686,80) alors que celui obtenu dans le modèle sans zonier est de 4 020 984,84 (respectivement 4 021 694,53), soit une différence de 18 159,81 (respectivement 18 007,73). Puisque l'AIC et le BIC sont plus petits dans le modèle avec zonier, ce dernier est préféré à l'autre. La variable zonier apporte donc de l'information dans la connaissance du risque incapacité. Cependant, il est difficile de se faire une idée de l'importance de ce gain d'information.

Pour cela, le même principe est effectué pour les autres variables. Autrement dit, pour chaque variable, une régression logistique multinomiale est lancée en enlevant la variable donnée. L'AIC et BIC de ce modèle sont alors comparés à ceux du modèle incluant l'ensemble des variables pour calculer un écart d'AIC/BIC. Il est alors possible de comparer ces écarts obtenus pour les différentes variables. Néanmoins, toutes les comparaisons ne sont pas pertinentes. En effet, les variables âge et genre n'ont pas le même nombre de modalités que la variable zonier. C'est pourquoi la pénalisation diffèrent entre ces variables alors que pour la CSP et le secteur d'activité, la pénalisation est la même que pour le zonier puisque chacune de ces variables nécessite le même nombre de paramètres à estimer. Par conséquent, seulement les AIC/BIC liés à l'exclusion des variables CSP et secteur d'activité dans la modélisation sont comparés à ceux obtenus en enlevant la variable zonier du modèle [7].

Ainsi, le fait d'enlever la variable CSP dans la modélisation augmente l'AIC de 44 089 alors que pour le secteur d'activité, cette hausse n'est que de 4 588. L'augmentation de l'AIC est quant à elle de 18 160 suite à la suppression du zonier dans la modélisation du risque incapacité. Le zonier apporterait donc moins d'informations que la CSP mais plus que le secteur d'activité pour expliquer l'incapacité de travail.

Le zonier permet donc une amélioration de la connaissance du risque incapacité. La section suivante vise à confirmer ce résultat en appliquant le zonier à une base de données "test", à savoir les DSN 2019.

9.4 Application du zonier sur une base de données "test" : les DSN 2019

Tous les résultats et modélisations décrits plus haut sont basés sur les DSN de 2017 et 2018. Cette section présente l'application du zonier sur les DSN 2019, qui peuvent être considérées

7. Les variables âge et genre sont donc présentes dans les tableaux 9.3 et 9.4 uniquement pour des raisons d'exhaustivité.

comme des données de test puisqu'elles n'ont pas été utilisées dans la construction du zonier.

Le tableau 9.5 reporte les résidus moyens des codes postaux par zone et par durée d'arrêt de travail pour les années 2017-2018 et 2019. Ces résidus sont obtenus à partir de la régression logistique multinomiale présentée au chapitre 3 et qui permet d'isoler les effets des variables déjà utilisées en tarification (âge, CSP, genre et secteur d'activité). Ces résidus représentent donc la part non expliquée de la sinistralité après prise en compte des variables citées ci-dessus. C'est la raison pour laquelle il est plus intéressant d'utiliser les résidus plutôt que les fréquences "brutes" d'arrêt de travail pour confirmer la pertinence du zonier.

Zonier	Années	Résidu moyen Absence d'AT	Résidu moyen 0-15 jours	Résidu moyen 15-30 jours	Résidu moyen 30-90 jours	Résidu moyen +90 jours
Zone 1	2017-2018	0.141	−0.084	−0.025	−0.019	−0.013
	2019	0.111	−0.066	−0.020	−0.018	−0.007
Zone 2	2017-2018	0.055	−0.034	−0.009	−0.007	−0.005
	2019	0.043	−0.026	−0.008	−0.006	−0.003
Zone 3	2017-2018	0.047	−0.032	−0.008	−0.006	−0.001
	2019	0.035	−0.022	−0.008	−0.004	−0.001
Zone 4	2017-2018	−0.042	−0.011	0.005	0.030	0.018
	2019	−0.023	−0.009	0.001	0.020	0.011
Total	2017-2018	0.062	−0.040	−0.010	−0.007	−0.004
	2019	0.048	−0.030	−0.010	−0.006	−0.002

TABLE 9.5 – Comparaison des résidus moyens par zone en 2017-2018 et 2019

Pour rappel, plus un code postal a un résidu moyen important, plus la prédiction de la modalité concernée ("Absence d'AT", "]0,15]", "]15,30]", "]30,90]", "+90") en ne tenant compte que de l'âge, CSP, genre et secteur d'activité est sous-estimée. Par exemple, un résidu moyen positif sur la modalité "+90" signifie que la sinistralité de ce code postal est sous-estimée en observant uniquement les variables citées ci-dessus. Ce code postal est donc considéré comme risqué. A l'inverse, un résidu moyen positif sur la modalité "Absence d'AT" signifie que le fait de ne pas avoir d'arrêt de travail est sous-estimé, ce qui revient à sur-estimer le risque d'avoir un arrêt de travail sur ce code postal. Ce dernier n'est donc pas un code postal risqué sur cette modalité.

De manière générale, le zonier semble moins performant sur les données 2019 que sur les données 2017-2018. En effet, le résidu moyen "+90" augmente sur les zones 1 et 2 alors qu'il diminue sur la zone 4 entre 2017-2018 et 2019 [8]. A contrario, le résidu moyen "Absence d'AT" diminue sur les trois premières zones alors qu'il augmente sur la quatrième, ce qui signifie que les écarts entre les zones se réduisent entre 2017-2018 et 2019. Ce phénomène est moins marqué sur les autres modalités comme les arrêts de travail de 30 à 90 jours.

Néanmoins, bien que sa performance diminue avec les données de 2019, le zonier semble conserver un certain intérêt puisque le risque croît toujours avec les zones en 2019. En effet, les résidus moyens pour les modalités "]0,15]", "]15,30]", "]30,90]" et "+90" augmentent avec les zones et ils diminuent pour la modalité "Absence d'AT". Ce résultat est un premier élément pour énoncer que le zonier reste pertinent avec les données de 2019. Des tests statistiques sont menés pour conforter ce résultat.

8. Pour rappel, les zones sont triées dans l'ordre croissant du risque : la zone 1 est la moins risquée et la zone 4 est celle associée au risque le plus élevé.

— Une première méthode consiste à étudier la même zone entre 2017-2018 et 2019, en comparant par exemple le résidu moyen de la zone 1 en 2017-2018 et en 2019. Cependant, il est difficile d'implémenter un test de comparaison de moyennes sur les années 2017-2018 contre 2019 car une dérive des arrêts de travail a été observée cette année-là. En effet, la fréquence des arrêts de travail a augmenté entre 2017-2018 et 2019. Par ailleurs, cette dérive a certainement été amplifiée par une amélioration de la qualité des déclarations d'arrêts de travail dans la DSN en 2019 par rapport à 2017-2018 [9].
— Une deuxième méthode vise à comparer les différentes zones entre elles sur la même année, ce qui permettrait de confirmer ou non la significativité des écarts de risque entre les zones et de valider la pertinence du zonier sur plusieurs années. Cette méthode est implémentée ci-dessous.

Afin d'éviter de faire une hypothèse trop forte sur la distribution des données, un test non paramétrique est utilisé, à savoir le test de Wilcoxon [10]. Le test de Wilcoxon est souvent vu comme une alternative au test de comparaison de moyennes de Student dans le cas où les données ne sont pas issues d'une loi normale. Cependant, la philosophie des deux tests est légèrement différente. Dans le test de Student, l'hypothèse nulle correspond à l'égalité des moyennes dans les deux groupes alors que dans le test de Wilcoxon (unilatéral), les hypothèses sont par exemple :

H_0 : les zones A et B portent le même risque sur la modalité Z vs
H_1 : la zone B est plus (ou moins) risquée que la zone A sur la modalité Z

L'idée de ce test est de rassembler les deux groupes et de trier les valeurs par ordre croissant. Ainsi, si les deux groupes suivent la même distribution alors leurs valeurs devraient être régulièrement alternées. A contrario, si les valeurs d'un groupe sont plutôt supérieures (ou inférieures) à l'autre groupe alors l'hypothèse alternative H_1 sera privilégiée.

Ce test est implémenté sur les résidus des 5 modalités ("Absence d'AT", "]0,15]", "]15,30]", "]30,90]", "+90") avec les données 2017-2018 et 2019. Pour chacun de ces résidus, il s'agit d'un test unilatéral qui est réalisé sur les combinaisons les plus pertinentes, à savoir "Zone 1 vs Zone 2", "Zone 2 vs Zone 3" et "Zone 3 vs Zone 4". Par exemple, les hypothèses du test pour la comparaison des résidus "+90" entre les zones 1 et 2 sont les suivantes :

H_0 : les zones 1 et 2 portent le même risque pour les arrêts de travail de plus de 90 jours vs
H_1 : la zone 1 est moins risquée que la zone 2 pour les arrêts de travail de plus de 90 jours.

Les p-values de ces tests sont présentées dans le tableau 9.6.

9. La DSN est devenue obligatoire pour la majorité des employeurs au 1er janvier 2017. Il est donc légitime de penser qu'avec le temps, les déclarations des entreprises soient plus justes et complètes.
10. Des tests de normalité (test de Shapiro-Wilk) ont été réalisés pour confirmer le caractère non gaussien de nos données.

Indicateur	Années	Zone 1 vs Zone 2	Zone 2 vs Zone 3	Zone 3 vs Zone 4
Résidus "Absence d'AT"	2017-2018	$< 2.2 \times 10^{-16}$	7.982×10^{-6}	$< 2.2 \times 10^{-16}$
	2019	$< 2.2 \times 10^{-16}$	5.919×10^{-5}	3.446×10^{-6}
Résidus "]0,15]"	2017-2018	$< 2.2 \times 10^{-16}$	0.0047	2.196×10^{-4}
	2019	$< 2.2 \times 10^{-16}$	0.0224	0.0268
Résidus "]15,30]"	2017-2018	$< 2.2 \times 10^{-16}$	0.0157	7.252×10^{-6}
	2019	$< 2.2 \times 10^{-16}$	0.5457	0.0020
Résidus "]30,90]"	2017-2018	$< 2.2 \times 10^{-16}$	0.0217	$< 2.2 \times 10^{-16}$
	2019	$< 2.2 \times 10^{-16}$	0.0025	8.995×10^{-10}
Résidus "+90"	2017-2018	$< 2.2 \times 10^{-16}$	1.601×10^{-11}	2.638×10^{-13}
	2019	$< 2.2 \times 10^{-16}$	0.0046	3.279×10^{-4}

TABLE 9.6 – p-value des tests de Wilcoxon implémentés

Toutes les p-values sont inférieures à 5%, excepté celle associée à la comparaison des résidus "]15,30]" pour les zones 2 et 3 en 2019. Ainsi, tous les tests sont significatifs à 5% sauf celui cité ci-dessus. Par conséquent, avec un degré de significativité des tests de 5%, chacune des hypothèses nulles (H_0) est rejetée. Autrement dit, quelle que soit la durée des arrêts de travail (excepté les arrêts de 15 à 30 jours pour les zones 2 et 3), la hiérarchie du risque selon les zones est conservée : la zone 4 est plus risquée que la zone 3, qui est elle, plus risquée que la zone 2, etc. Globalement, les p-values ont certes augmenté entre 2017-2018 et 2019 mais restent toutefois très petites en 2019, ce qui permet de confirmer l'intérêt du zonier en 2019 également.

La localisation de l'entreprise apporte donc une information pertinente dans la connaissance du risque incapacité. L'ajout d'un tel critère de tarification en prévoyance collective est donc envisageable. Les avantages et limites de son utilisation pour tarifer des garanties sont détaillés dans le chapitre suivant.

Chapitre 10

Avantages et limites d'un zonier incapacité

Ce dernier chapitre a pour but de reporter les avantages et inconvénients de l'utilisation de la localisation de l'entreprise dans la tarification des contrats collectifs prévoyance.

10.1 Avantages d'utiliser la localisation de l'entreprise dans la tarification prévoyance collective

Les tarifications des garanties en prévoyance collective sont souvent basées sur peu de critères, comme l'âge, la répartition homme/femme, les CSP, la taille d'entreprise, voire le secteur d'activité. Inclure la localisation de l'entreprise dans cette tarification apparaît donc comme un critère plutôt innovant. En effet, les zoniers sont en général utilisés en assurance auto et MRH. Ajouter de nouveaux critères, s'ils sont pertinents, apporte un degré de segmentation supplémentaire et permet de proposer des tarifs plus proches du risque porté par les assurés, ce qui peut être un avantage concurrentiel. En effet, si un assureur a une tarification moins segmentée que sur le marché, il aura tendance à attirer les assurés qui portent un risque important. Par exemple, si un assureur propose le même tarif à deux groupes d'assurés ne portant pas le même risque alors le groupe le moins risqué souscrira sûrement chez la concurrence (qui propose un tarif plus segmenté et donc moins cher) et l'assureur se contentera du groupe le plus risqué, qui lui, payera une prime inférieure à son risque. C'est le phénomène d'antisélection.

Comme évoqué dans le chapitre précédent, les écarts de sinistralité sont importants entre les zones et le zonier permet d'améliorer la connaissance du risque incapacité après prise en compte des critères usuels de tarification (âge, genre, CSP, secteur d'activité). La localisation de l'entreprise apparaît donc comme un critère pertinent pour tarifer les garanties incapacité temporaire de travail dans les contrats prévoyance. D'ailleurs, l'impact du zonier sur le risque incapacité estimé par la régression multinomiale est très élevé (cf. section 9.2). De tels écarts pourraient difficilement être applicables en tarification et il faudra certainement limiter l'impact du zonier dans le tarif par rapport à ce qui est évalué par la régression multinomiale. Autrement dit, l'impact de la localisation de l'entreprise sur le tarif devrait être inférieur à celui estimé par la régression multinomiale. Le fait d'avoir cette marge de manœuvre est plutôt une bonne chose car cela prouve que la localisation de l'entreprise est un critère intéressant pour segmenter un peu plus le risque incapacité.

10.2 Limites de l'utilisation de la localisation de l'entreprise dans la tarification prévoyance collective

Ajouter un nouveau critère de tarification augmente d'une part la segmentation du risque mais réduit d'autre part sa mutualisation. Ainsi, lorsqu'un nouveau critère est ajouté dans la définition du tarif, il faut veiller à ne pas aboutir à une segmentation trop importante qui entraînerait des tarifs très élevés pour les assurés les plus à risque. Dans ce cas-là, l'assurance pourrait ne pas satisfaire un de ses objectifs fondamentaux qu'est la mutualisation.

Néanmoins, le zonier incapacité présenté dans ce mémoire ne concerne que l'assurance collective. Or, dans les contrats collectifs, les cotisations sont mutualisées parmi l'ensemble des salariés (ou l'ensemble d'un collège [1]) d'une même entreprise, ce qui limite grandement la segmentation. En effet, les salariés d'un même collège ou d'une même entreprise paient tous le même taux de cotisation pour les mêmes garanties. Un salarié de 50 ans d'une entreprise donnée ne paiera donc pas plus qu'une personne de 25 ans pour un même niveau de garantie, bien que son risque de subir un (long) arrêt de travail soit plus élevé.

En résumé, le risque d'aboutir à une segmentation trop importante suite à l'introduction de ce nouveau critère de tarification est réduit voire inexistant. De plus, comme évoqué dans la section précédente, il est possible de contrôler l'impact du zonier sur la tarification, ce qui réduit encore plus le risque de déboucher sur une segmentation trop conséquente.

Avant d'incorporer une dimension géographique en tarification, il faut aussi se poser la question de l'éthique d'un tel critère. En effet, il ne faudrait pas que le zonier discrimine certaines populations sur des critères controversés ou illégaux comme l'origine ou la religion des personnes. La zone géographique peut effectivement révéler beaucoup d'informations sur ces sujets, même si aucune variable de ce type n'ait été incorporée dans nos travaux. D'ailleurs, il existe très peu de données (voire aucune) sur la répartition des religions sur le territoire français [2]. Il est cependant connu que certaines populations (ayant la même origine et/ou religion) sont plus présentes dans certains quartiers, voire certaines villes que dans d'autres.

Néanmoins, le zonier est à la maille code postal et une sorte de lissage est effectuée pour que les codes postaux les plus proches aient plus de chance d'appartenir à la même classe de risque. Aussi, la méthode d'affectation des classes de risques aux codes postaux ayant peu d'affiliés accentue ce lissage. Un quartier ne devrait donc pas être isolé puisqu'il s'agit d'une maille plus petite que le code postal.

Par ailleurs, le périmètre étudié est la prévoyance collective, ce qui limite aussi ce risque de discrimination par rapport à tout autre type d'assurance individuelle. En effet, le critère retenu est la localisation de l'entreprise et non le lieu de résidence des salariés, qui est lui beaucoup plus corrélée avec la présence ou non de groupes éthniques. Il est en effet plutôt rare que l'ensemble ou la majorité des salariés d'une même entreprise vienne de la même ville, voire du même quartier.

Le zonier incapacité décrit dans ce mémoire a pour unité géographique le code postal. Comme évoqué à la section 4.1, ce choix présente de nombreux avantages comme le fait d'être un bon compromis entre l'IRIS [3] et le département. Néanmoins, cette maille est assez précise

1. Un contrat prévoyance collectif peut concerner l'ensemble des salariés d'une même entreprise mais il est possible qu'il ne soit proposé qu'à une certaine catégorie d'entre eux (seulement les cadres par exemple). Dans ce cas, on parle de collège. Cependant, cette distinction ne peut se faire par des critères de revenu, d'âge ou d'état de santé.

2. Des statistiques existent quant à elles concernant la répartition des origines sur le territoire français.

3. Les îlots regroupés pour l'information statistique (IRIS) sont des découpages infra-communaux, créés par l'INSEE dans le cadre du recensement de la population de 1999.

et cela complexifie la communication autour de ce zonier auprès de publics moins proches de la fonction actuarielle comme les commerciaux. Il va donc falloir faire preuve de pédagogie pour bien leur expliquer les intérêts de réaliser un zonier à une telle maille.

Enfin, l'utilisation du critère "localisation de l'entreprise" dans la tarification sera moins évident pour les entreprises ayant plusieurs établissements en France. En effet, dans une entreprise donnée, le tarif devrait être le même pour l'ensemble des salariés (ou du collège) quel que soit le code postal de l'établissement dans lequel ils travaillent. Cependant, il est tout de même possible de prendre en compte le zonier dans la tarification des garanties pour ces entreprises. Une idée serait par exemple de considérer seulement les codes postaux des établissements ayant les plus grands nombres de salariés et de calculer une prime uniquement sur ces codes postaux mais applicable à l'ensemble du personnel de l'entreprise. Une autre possibilité serait de calculer un tarif par établissement et d'obtenir la prime finale par pondération de ces tarifs. La tarification pour ces entreprises serait donc moins évidente mais tout de même réalisable.

Conclusion

Dans un objectif d'amélioration du positionnement tarifaire en prévoyance collective, il est possible d'envisager de nouveaux critères de tarification pour ce type d'assurance, comme la localisation de l'entreprise. En effet, bien que le critère géographique soit très utilisé dans les tarifications des assurances auto et MRH, il est en général peu présent en prévoyance collective. Le présent mémoire visait donc à étudier la pertinence de ce critère pour tarifer des garanties incapacité temporaire de travail. Pour cela, un zonier est construit en utilisant notamment les données de la DSN [1]. Pour rappel, la DSN est un système qui permet à chaque employeur, de déclarer de façon unique, dématérialisée et mensuelle, un ensemble d'informations liées à la protection sociale de ses salariés. Elle regroupe donc beaucoup d'informations sur les salariés, notamment ce qui concerne leur contrat de travail et arrêts maladie, autorisant ainsi la construction d'un zonier à partir de ses données [2].

Afin de rendre possible la construction de ce zonier, plusieurs étapes ont dû être mises en œuvre en amont. Tout d'abord, il a fallu tenir compte des critères déjà utilisés dans la tarification prévoyance collective avant d'élaborer ce zonier. Pour cela, une régression logistique multinomiale avec les variables âge, genre, CSP et secteur d'activité (sous forme de classes de risques) a été implémentée. Bien que ce n'était pas son objectif principal, ce modèle a confirmé l'intérêt des critères classiques de tarification puisque les niveaux de risques varient énormément en fonction des profils d'individus (basés sur ces variables). Cette modélisation avait pour but principal d'isoler les effets des variables citées ci-dessus afin de récupérer les résidus qui eux représentaient la part non expliquée de la sinistralité après prise en compte de ces variables. Étant donné que la régression multinomiale a été réalisée à la maille individuelle, une étape d'agrégation des résidus au code postal, unité géographique retenue pour l'élaboration de ce zonier, s'est avérée nécessaire. Les résidus individuels de chaque code postal ont donc été agrégés en créant des indicateurs à cette maille. Par suite, une sélection de ces indicateurs a été effectuée en utilisant le lasso pour ne retenir que les indicateurs les plus pertinents. Cette sélection a aussi permis d'accorder plus de poids aux arrêts de travail les plus longs (et à l'absence d'arrêt de travail), modalités qui intéressent le plus les assureurs.

Nous nous sommes appuyés sur des notions liées aux données spatiales comme la structure de voisinage et l'autocorrélation spatiale pour justifier de l'intérêt de réaliser un zonier incapacité qui tient compte du voisinage de chaque code postal. Ce dernier point est rendu possible par l'utilisation de la classification ascendante hiérarchique avec contraintes de proximité géographique présentée dans (Chavent et al., 2018)[2]. Cette méthode de classification permet d'associer à chaque code postal une classe de risque. Néanmoins, ce clustering a été implémenté seulement sur les codes postaux ayant suffisamment d'affiliés pour considérer que la sinistralité du territoire était robuste. Il a donc fallu dans un second temps attribuer aux codes postaux restants une classe de risque. Pour cela, un modèle de prédiction des classes de risques a été

1. DSN : Déclaration Sociale Nominative.
2. Un zonier incapacité pouvait tout à fait être construit avant l'arrivée de la DSN mais cette "nouvelle" base de données apporte une vision plus globale des arrêts de travail.

construit à partir de données démographiques et socio-économiques de l'INSEE. Ainsi, l'affectation d'une classe de risque aux codes postaux ayant peu ou pas d'affiliés se fait notamment en utilisant leurs caractéristiques démographiques et socio-économiques. Autrement dit, le risque porté par ces codes postaux est estimé en ayant recours à de l'open data. Par ailleurs, ce modèle a également permis d'interpréter le zonier en appréhendant mieux les caractéristiques des différentes classes de risques, et ce, à l'aide des SHAP values. Par exemple, la zone 1, zone la moins risquée en termes d'arrêts de travail, est plutôt rurale alors que la zone 2 est la plus urbaine des quatre zones. La zone 2 se caractérise aussi par des revenus plus élevés que les autres alors que la zone 3 semble être la plus défavorisée.

Enfin, la dernière partie du présent mémoire justifie, grâce à différents outils statistiques (V de Cramer, valeur et significativité des coefficients de régression, tests statistiques, AIC/BIC, application sur une base de données "test"), que le zonier apporte de l'information pour améliorer la connaissance du risque incapacité. Autrement dit, le risque incapacité n'est pas équivalent partout en France. L'ajout du critère "localisation de l'entreprise" dans la tarification prévoyance collective est donc tout à fait envisageable et pertinent. Il apporterait effectivement un degré de segmentation supplémentaire dans la tarification, ce qui pourrait être un avantage concurrentiel pour l'assureur.

Outre le fait de montrer que la localisation de l'entreprise peut devenir un nouveau critère de tarification pour les garanties incapacité en prévoyance collective, ce mémoire a permis d'explorer des méthodes et notions statistiques peu utilisées dans le domaine de l'actuariat comme la régression logistique multinomiale, la classification ascendante hiérarchique avec contraintes de proximité géographique et les SHAP values.

La méthodologie implémentée pour construire ce zonier présente néanmoins quelques limites et axes d'amélioration qui ont été détaillés au fur et à mesure de ce mémoire. Une synthèse est donnée ci-dessous.

Un premier axe d'amélioration concerne l'utilisation d'une régression logistique ordinale (ou régression polytomique ordonnée) à la place de la régression logistique multinomiale. En effet, cette dernière suppose que les modalités de la variable à expliquer ne sont pas ordonnées alors que dans notre étude, un certain ordre se dégage puisque la variable cible correspond à la survenance d'arrêts de travail de différentes durées. Ainsi, implémenter un modèle faisant cette hypothèse d'ordre comme la régression logistique ordinale pourrait être approprié dans notre cas.

La classification ascendante hiérarchique avec contraintes de proximité géographique est inadaptée pour obtenir un zonier dans le cas des territoires ultramarins [3] car ils sont souvent isolés (île). Le risque dans ces territoires doit donc être modélisé à part.

Ensuite, seulement les codes postaux ayant suffisamment d'affiliés ont été utilisés dans la classification. Le fait de rajouter des données en élargissant l'étude à l'ensemble du portefeuille ou en ajoutant les données les plus récentes, pourrait consolider la sinistralité calculée sur les codes postaux ayant initialement peu d'affiliés, les permettant ainsi d'être inclus dans la classification. Augmenter le nombre de codes postaux incorporés dans la classification devrait améliorer la qualité du zonier puisque affecter la classe de risque avec la classification hiérarchique spatiale est certainement plus précis que l'attribution via le modèle de prédiction basé sur les données INSEE.

3. Les territoires ultramarins sont Guadeloupe, Guyane, Martinique, La Réunion, Mayotte, Nouvelle- Calédonie, Polynésie française, Saint-Barthélemy, Saint-Martin, Saint-Pierre-et-Miquelon, les Terres Australes et Antarctiques Françaises et les îles de Wallis-et-Futuna.

Des limites générales existent lorsqu'un nouveau critère de tarification est mis en place, et ce, quel que soit le type d'assurance étudiée. Par exemple, il faut s'assurer que la segmentation finale ne soit pas trop importante et qu'une certaine mutualisation soit conservée. Dans notre cas, le fait qu'il soit possible de contrôler l'impact du zonier sur la tarification et que seulement l'assurance collective soit concernée par ces travaux limitent grandement ce risque de segmentation disproportionnée.

Par ailleurs, ce mémoire s'est concentré sur l'application directe du zonier en tarification mais il est possible d'envisager d'autres utilisations. Par exemple, la politique commerciale peut être adaptée en privilégiant les affaires nouvelles des zones les moins risquées. Il serait aussi intéressant de vérifier si les zones les plus risquées sont aussi celles avec le plus de fraudes aux arrêts de travail. Si c'est le cas, un renforcement des contrôles sur ces zones pourrait être envisagé.

Au vu de la quantité d'informations disponibles dans les données de la DSN, de nouveaux travaux vont être certainement envisagés grâce à elles. Des critères de tarification supplémentaires pourraient notamment être implémentés à partir de la DSN, mais cette dernière pourrait aussi offrir de nouvelles opportunités, sans lien direct avec la tarification, comme par exemple la prédiction de la sinistralité incapacité puisque les arrêts de travail sont présents dans la DSN avant d'être répertoriés dans les bases sinistres.

Le présent mémoire s'est focalisé sur le risque incapacité temporaire de travail. Il pourrait donc être intéressant d'étendre ces travaux aux autres risques de la prévoyance comme l'invalidité et le décès, bien que ces risques soient liés à des évènements plus rares (et donc plus volatils) que l'arrêt de travail, ce qui nécessiterait une base de données robuste avec notamment un historique important.

Bibliographie

Ouvrages

[1] McCullagh P., Nelder J.A. (1989) Generalized linear models, 2nd Edition. Chapman and Hall/CRC, London.

Articles et contributions dans un ouvrage

[2] Chavent M., Kuentz-Simonet V., Labenne A., Saracco J. (2018). ClustGeo : an R package for hierarchical clustering with spatial constraints. Computational Statistics, 33, 1799–1822.

[3] Lundberg S., Lee S. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems (NeurIPS 2017), 4765-4774.

[4] Romeo G., Sciandra M., Chiodi M. (2015). How to define deviance residuals in multinomial regression. Conference : CLADAG 2015, 10th Scientific Meeting of the Classification and Data Analysis Group of the Italian Statistical Society.

[5] Seber G. A. F., Nyangoma S. O. (2000). Residuals for Multinomial Models. Biometrika 87, no. 1, 183-91.

[6] Shapley, L. (1953). “A value for n-person games”. Contributions to the Theory of Games, 2.28, 307–317.

[7] Wartenberg, D. (1985). Multivariate spatial correlation : a method for exploratory geographical analysis. Geographical Analysis, 17, 4, 263-283.

Mémoires

[8] Beraud-Sudreau, G. (2017) Construction d’un zonier en MRH à l’aide d’outils de Data-Science. Mémoire d’actuariat. CNAM.

[9] Pariente, J. (2017) Modélisation du risque géographique en assurance habitation. Mémoire d’actuariat. Université Paris Dauphine.

Cours

[10] Chesneau, C. (2017) Modèles de régression. Cours Master 2, Université de Caen.

[11] Koessler, F. (2007) Théorie des jeux/Valeur de Shapley

[12] Lenoir, J. (2013) Les tests statistiques dits ”non paramétriques”. Université de Picardie Jules Verne (UPJV).

[13] Ruiz-Gazen, A. (2018) Introduction to Big Data. Cours Master 1, Toulouse School of Economics.

[14] Thomas-Agnan, C. (2019) Spatial Econometrics. Cours Master 2, Toulouse School of Economics.

Textes juridiques

[15] Loi n° 89-1009, 31 décembre 1989, renforçant les garanties offertes aux personnes assurées contre certains risques, NOR : SPSX8900080L, art. 1er, https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000709057/#:~:text=L'engagement%20doit%20%C3%AAtre%20couvert,loi%20qui%20r%C3%A9git%20le%20contrat., site consulté le 24 Octobre 2020.

[16] Convention collective nationale de retraite et de prévoyance des cadres, 14 mars 1947 http://www.agirc.fr/fileadmin/agircarrco/documents/conventions_accords/CCN_14mars1947.pdf, site consulté le 24 Octobre 2020.

[17] Accord National Interprofessionnel du 17 novembre 2017 relatif à la prévoyance des cadres https://uimm.lafabriquedelavenir.fr/wp-content/uploads/2017/11/2017-11-17_ANI-relatif-a-la-prevoyance-des-cadres.pdf, site consulté le 24 Octobre 2020.

[18] Arrêté, 27 juillet 2018, portant extension et élargissement de l'accord national interprofessionnel relatif à la prévoyance des cadres, conclu le 17 novembre 2017, NOR : SSAS1821500A, https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000037311605/, site consulté le 24 Octobre 2020.

Sites internet

[19] https://freakonometrics.hypotheses.org/60635 (Régression logistique multinomiale), site consulté le 8 Juin 2020.

[20] https://www.data.gouv.fr/fr/datasets/fond-de-carte-des-codes-postaux/ (Fond de carte des codes postaux), site consulté le 18 Juin 2020.

[21] https://www.previssima.fr/question-pratique/comment-sont-calculees-les-cotisations-dans-les-contrats-sante-dentreprise.html, (Calcul des cotisations dans les contrats santé des entreprises), site consulté le 31 août 2020.

[22] http://www.dsn-info.fr/pourquoi.htm (Présentation de la DSN), site consulté le 11 Octobre 2020.

[23] http://www.dsn-info.fr/documentation/guide-demarrage-p3.pdf (Guide d'utilisation de la DSN), site consulté le 25 Octobre 2020.

[24] https://www.service-public.fr/particuliers/vosdroits/F3053 (Indemnités Journalières Sécurité Sociale), site consulté le 25 Octobre 2020.

[25] https://newsroom.malakoffhumanis.com/actualites/malakoff-humanis-presente-les-resultats-2020-de-son-barometre-annuel-absenteisme-maladie-545f-63a59.html, (Baromètre annuel Absentéisme Maladie 2020 - Malakoff Humanis), site consulté le 22 novembre 2020.

[26] https://drees.solidarites-sante.gouv.fr/IMG/pdf/8-21.pdf, (Les indemnités journalières - DREES), site consulté le 22 novembre 2020.

[27] https://www.insee.fr/fr/statistiques/1405599?geo=FRANCE-1, (Densité de population - INSEE), site consulté le 23 décembre 2020.

[28] https://www.insee.fr/fr/statistiques/3714237, (Sept salariés sur dix vont travailler en voiture - INSEE), site consulté le 23 décembre 2020.

Annexes

A Choix de la classification pour l'âge

	Dependent variable :			
	]0,15]	]15,30]	]30,90]	+90
Femme	0.217***	0.284***	0.272***	0.400***
	(0.004)	(0.007)	(0.007)	(0.009)
Âge]20,25]	0.509***	0.327***	0.343***	0.607***
	(0.012)	(0.020)	(0.022)	(0.037)
Âge]25,30]	0.738***	0.582***	0.653***	1.133***
	(0.012)	(0.020)	(0.021)	(0.035)
Âge]30,35]	0.868***	0.722***	0.806***	1.351***
	(0.012)	(0.020)	(0.021)	(0.035)
Âge]35,40]	0.849***	0.743***	0.817***	1.432***
	(0.012)	(0.020)	(0.021)	(0.035)
Âge]40,45]	0.781***	0.721***	0.841***	1.487***
	(0.012)	(0.020)	(0.021)	(0.034)
Âge]45,50]	0.709***	0.707***	0.849***	1.579***
	(0.012)	(0.020)	(0.021)	(0.034)
Âge]50,55]	0.707***	0.748***	0.939***	1.739***
	(0.012)	(0.020)	(0.021)	(0.034)
Âge]55,60]	0.507***	0.637***	0.859***	1.788***
	(0.013)	(0.021)	(0.022)	(0.034)
Âge +60	0.015	0.220***	0.551***	1.381***
	(0.023)	(0.036)	(0.035)	(0.047)
CSP Agent maîtrise	0.484***	0.699***	0.725***	0.701***
	(0.006)	(0.010)	(0.011)	(0.015)
CSP Employé	0.232***	0.636***	0.787***	0.881***
	(0.005)	(0.009)	(0.010)	(0.012)
CSP Ouvrier	0.698***	1.228***	1.349***	1.399***
	(0.006)	(0.010)	(0.011)	(0.014)
Secteur d'activité Classe 2	0.019***	0.050***	0.082***	0.153***
	(0.006)	(0.012)	(0.013)	(0.018)
Secteur d'activité Classe 3	−0.154***	0.040***	0.205***	0.456***
	(0.007)	(0.012)	(0.013)	(0.018)
Secteur d'activité Classe 4	0.110***	0.245***	0.372***	0.549***
	(0.007)	(0.013)	(0.014)	(0.020)
Constant	−2.620***	−4.191***	−4.563***	−5.984***
	(0.013)	(0.021)	(0.023)	(0.037)
Akaike Inf. Crit.	4,044,706.000	4,044,706.000	4,044,706.000	4,044,706.000
Note :				*p<0.1 ; **p<0.05 ; ***p<0.01

TABLE A.1 – Coefficients obtenus avec la régression multinomiale (10 classes pour l'âge)

B Preuve sur la régression multinomiale : équivalence des équations 3.1 et 3.2

<u>Rappel équation 3.1 :</u>

$$\forall k \in \{2, ..., m\}, \ln\left(\frac{p_k(x)}{p_1(x)}\right) = \beta_0^{(k)} + \beta_1^{(k)} x_1 + ... + \beta_p^{(k)} x_p$$

$\forall k \in \{2, ..., m\}$, on a donc :

$$p_k(x) = p_1(x) \exp\left(\beta_0^{(k)} + \beta_1^{(k)} x_1 + ... + \beta_p^{(k)} x_p\right)$$

De plus, $\sum_{k=1}^{m} p_k(x) = 1$.

$$\begin{aligned}
1 - p_1(x) &= \sum_{k=2}^{m} p_k(x) \\
&= \sum_{k=2}^{m} p_1(x) \exp\left(\beta_0^{(k)} + \beta_1^{(k)} x_1 + ... + \beta_p^{(k)} x_p\right) \\
&= p_1(x) \sum_{k=2}^{m} \exp\left(\beta_0^{(k)} + \beta_1^{(k)} x_1 + ... + \beta_p^{(k)} x_p\right) \\
\Leftrightarrow 1 &= p_1(x)\Big[1 + \sum_{k=2}^{m} \exp\left(\beta_0^{(k)} + \beta_1^{(k)} x_1 + ... + \beta_p^{(k)} x_p\right)\Big] \\
\Leftrightarrow p_1(x) &= \frac{1}{1 + \sum_{k=2}^{m} \exp\left(\beta_0^{(k)} + \beta_1^{(k)} x_1 + ... + \beta_p^{(k)} x_p\right)}
\end{aligned}$$

Or, $p_k(x) = p_1(x) \exp\left(\beta_0^{(k)} + \beta_1^{(k)} x_1 + ... + \beta_p^{(k)} x_p\right)$

On obtient donc :

$$p_k(x) = \frac{\exp\left(\beta_0^{(k)} + \beta_1^{(k)} x_1 + ... + \beta_p^{(k)} x_p\right)}{1 + \sum_{k=2}^{m} \exp\left(\beta_0^{(k)} + \beta_1^{(k)} x_1 + ... + \beta_p^{(k)} x_p\right)}$$

C Choix du poids de la contrainte spatiale dans la classification hiérarchique spatiale

α	$Q_0^*(P_4^\alpha)$	$Q_1^*(P_4^\alpha)$
0	1	0.2000421
0.05	0.944773198	0.2750321
0.1	0.815799860	0.2043353
0.15	0.700049976	0.2264806
0.2	0.543058964	0.3425976
0.25	0.439355869	0.3851224
0.3	0.345155260	0.5035956
0.35	0.270426913	0.6182537
0.4	0.219199134	0.7426139
0.45	0.198459851	0.7624696
0.5	0.164035585	0.7321006
0.55	0.142642490	0.8109902
0.6	0.146155992	0.7912296
0.65	0.107734152	0.8809706
0.7	0.100867878	0.8874484
0.75	0.051459677	0.9551765
0.8	0.048824142	0.9558345
0.85	0.044177849	0.9564400
0.9	0.004740915	0.9965125
0.95	0.004740915	0.9965125
1	0.002058059	1

Table A.2 – Part (renormalisée) d'inertie expliquée en fonction de α

D Contributions des variables (SHAP values) dans le cas du 9e arrondissement de Paris

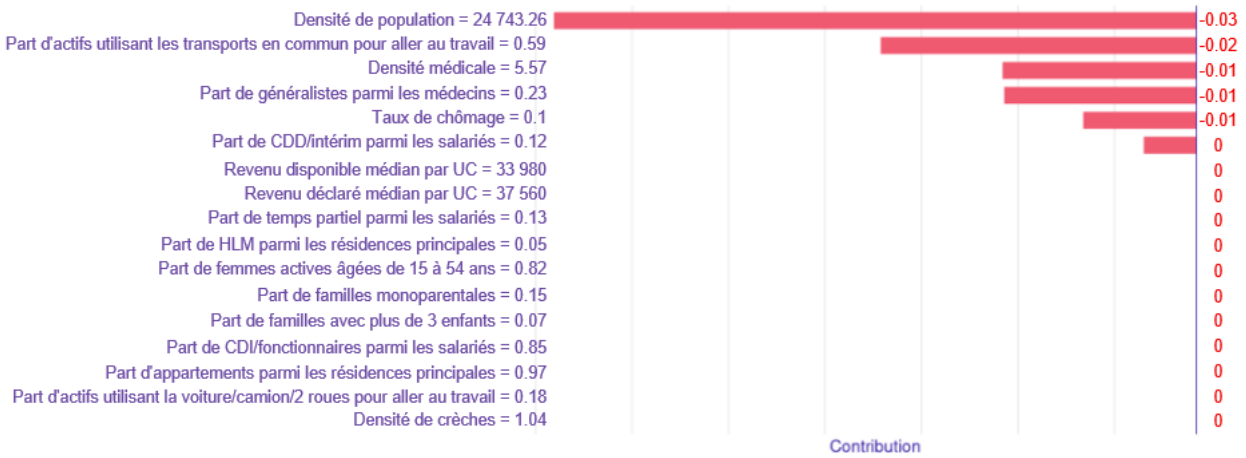

FIGURE A.1 – Contributions des variables (SHAP values) à la prédiction "Zone 1" du 9e arrondissement de Paris

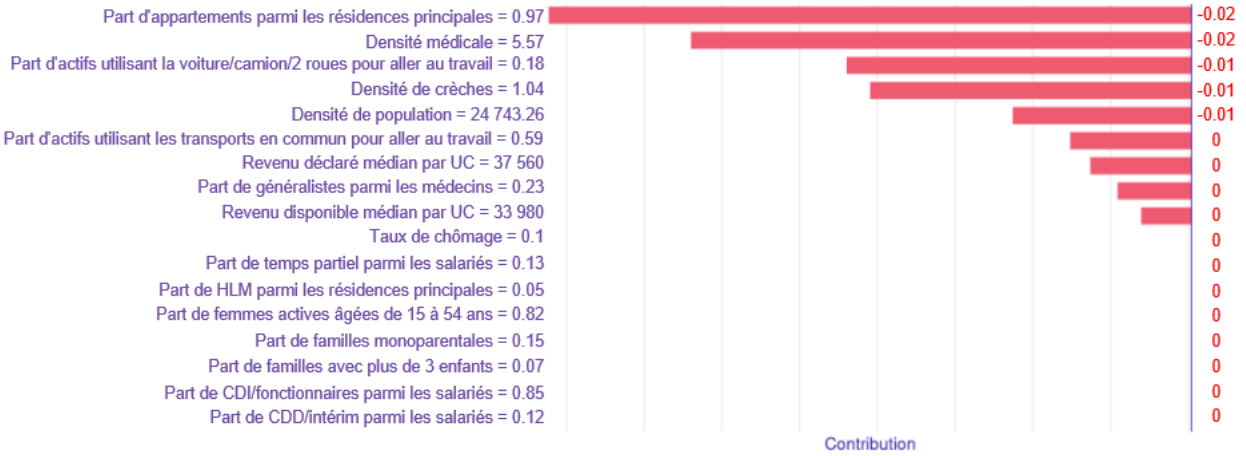

FIGURE A.2 – Contributions des variables (SHAP values) à la prédiction "Zone 4" du 9e arrondissement de Paris

E SHAP values pour les zones 3 et 4 ("summary plot")

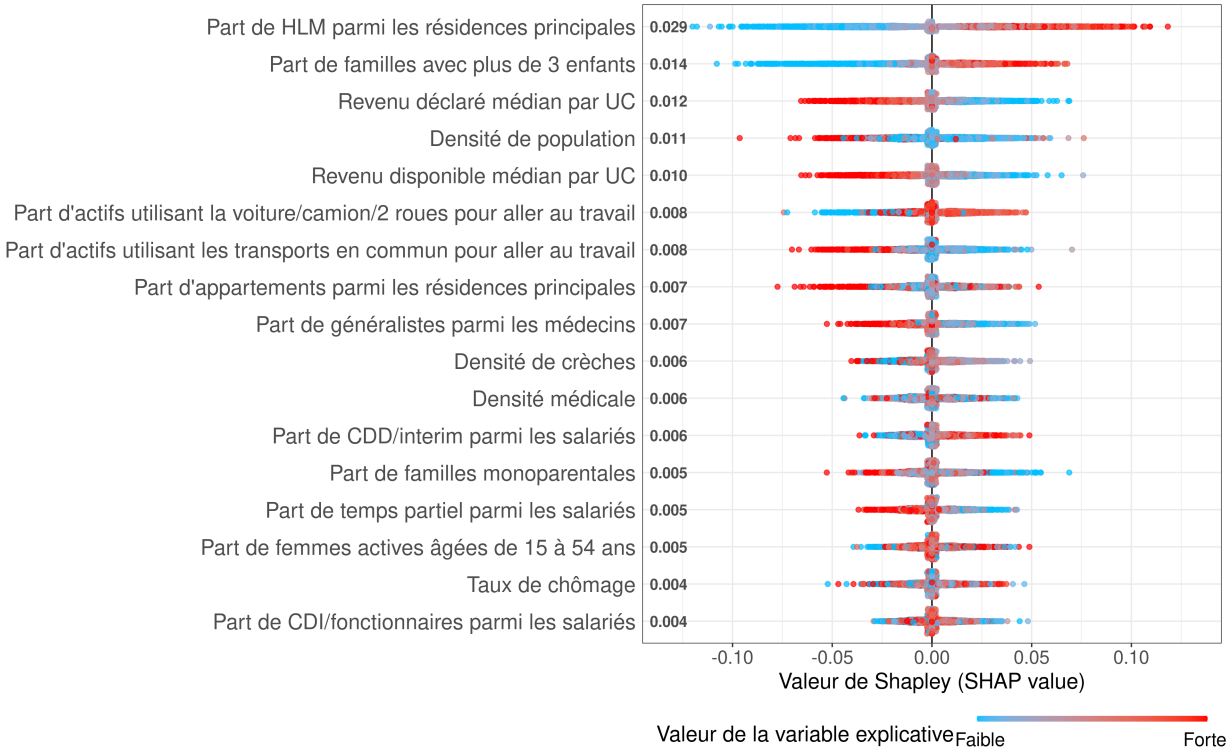

FIGURE A.3 – SHAP values pour la zone 3 ("summary plot")

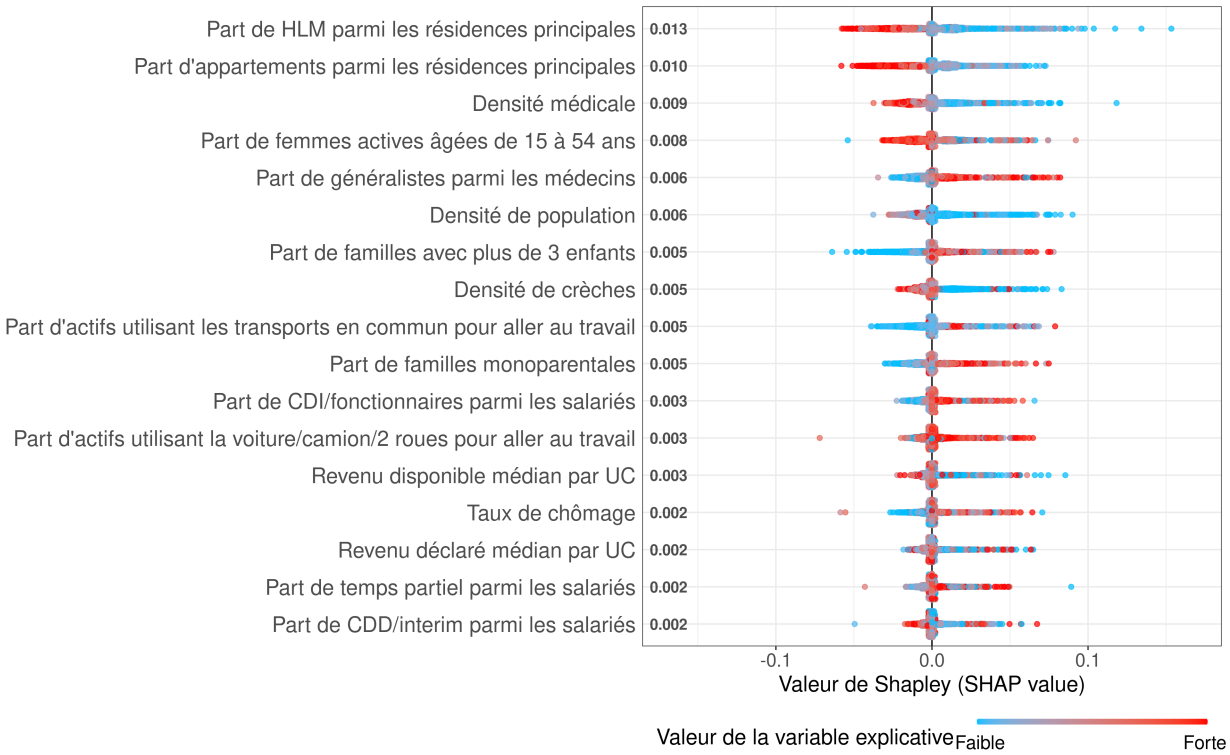

FIGURE A.4 – SHAP values pour la zone 4 ("summary plot")

Note de synthèse

La Déclaration Sociale Nominative (DSN) est un système permettant à tout employeur, de déclarer de façon unique, dématérialisée et mensuelle, un ensemble d'informations liées à la protection sociale de ses salariés. Bien que la simplification administrative soit l'objectif principal de ce dispositif, la DSN est aussi une réelle opportunité pour les assureurs de personnes puisqu'elle regroupe beaucoup d'informations sur les salariés comme des renseignements sur leur contrat de travail, arrêts maladie, rémunérations, etc. Certaines de ces données étaient déjà accessibles par les assureurs avant l'arrivée de la DSN mais cette dernière en facilite l'accès puisqu'elle centralise à elle seule beaucoup d'informations. La DSN permet donc aux assureurs prévoyance d'avoir une connaissance plus fine de leurs affiliés sous contrat collectif et du risque qu'ils portent, ce qui pourrait se concrétiser par l'ajout d'un (ou plusieurs) degré(s) de segmentation dans leur tarification prévoyance collective. Pour rappel, avoir une tarification plus segmentée que le marché peut être un avantage concurrentiel, en attirant notamment les profils les moins à risque.

L'objectif de ce mémoire est de déterminer si la localisation de l'entreprise peut devenir un critère de tarification pertinent pour les garanties incapacité temporaire de travail en assurance collective. Pour cela, un zonier est construit en utilisant les données de la DSN. Les zoniers sont peu présents dans les tarifications des assurances prévoyance alors qu'ils sont incontournables en assurance MRH et auto. Plus précisément, les critères de tarification utilisés chez Malakoff Humanis pour les garanties incapacité de travail en assurance collective, sont l'âge moyen, la répartition homme/femme, le secteur d'activité et la répartition des différentes catégories socio-professionnelles dans l'entreprise. Ainsi, pour élaborer le zonier incapacité, il faut tenir compte de ces précédents critères puisque la localisation de l'entreprise viendrait en complément de ces derniers pour tarifer. C'est la raison pour laquelle le zonier est construit à partir des résidus d'une régression ayant pour variables explicatives les critères précédemment cités [1].

Afin de pouvoir mener cette régression, il faut définir la variable à expliquer qui reflète la sinistralité incapacité de travail. Le sujet est abordé avec les dimensions fréquence et durée des arrêts maladie [2]. Plus précisément, la variable cible correspond à la survenance d'un arrêt de travail d'une certaine durée, catégorisée de la façon suivante :

— Absence d'arrêt de travail ("Absence d'AT"),
— Arrêt de travail de 0-15 jours ("]0,15]"),
— Arrêt de travail de 15-30 jours ("]15,30]"),
— Arrêt de travail de 30-90 jours ("]30,90]")
— Arrêt de travail de plus de 90 jours ("+90")

Ce choix de variable est lié à notre ambition d'avoir une meilleure connaissance globale du risque incapacité et pas seulement d'un point de vue fréquence ou durée.

La construction du zonier a nécessité plusieurs étapes. La méthodologie implémentée est

1. Pour rappel, les résidus de la régression contiennent la part non expliquée de la sinistralité après prise en compte des variables explicatives.

2. Les congés maternité, paternité et adoption ont été exclus du périmètre d'étude car ils sont pris en charge dans le cadre de garanties spécifiques qui ne sont pas l'objet de ce zonier.

détaillée dans la figure NDS.1.

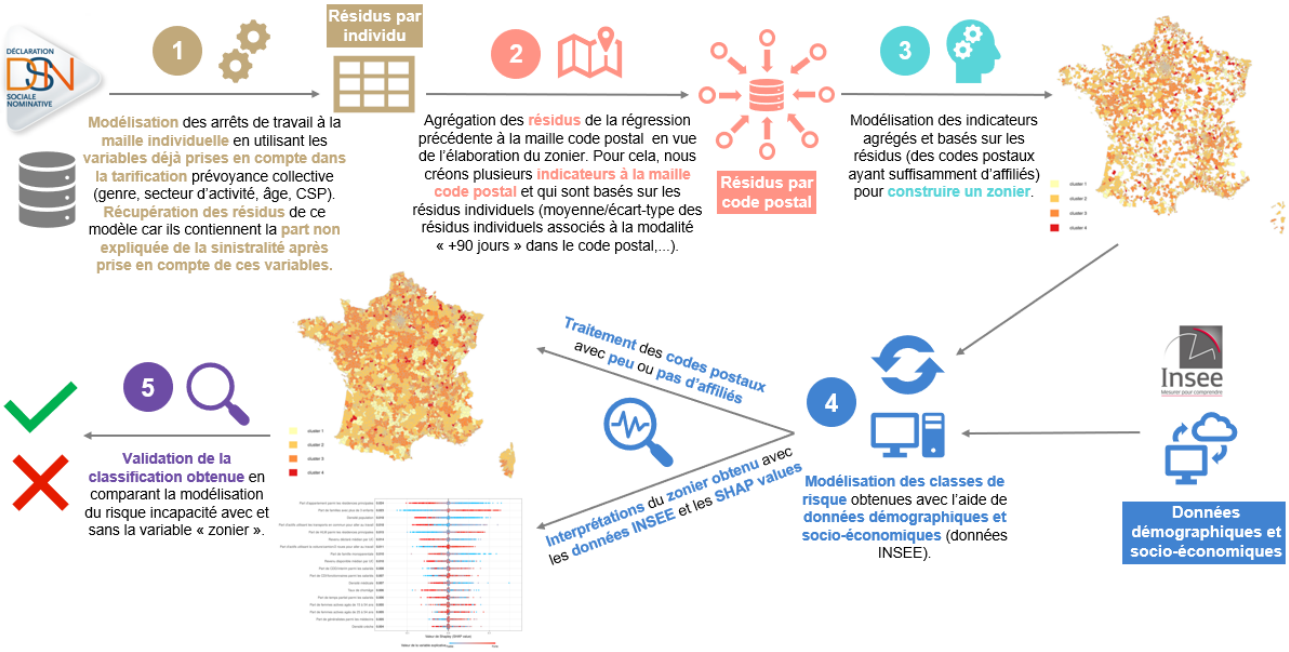

FIGURE NDS.1 – Méthodologie d'élaboration du zonier

La première étape consiste à contrôler les effets sur la sinistralité des variables déjà utilisées en tarification. Pour cela, une régression logistique multinomiale avec les variables âge, genre, CSP et secteur d'activité (sous forme de classes de risques) est implémentée. Cette modélisation s'est effectuée à la maille individuelle, c'est pourquoi la variable "genre" est utilisée à la place de la variable "répartition homme/femme". Ce modèle a, d'une part, permis de confirmer la pertinence des critères usuels de tarification. D'autre part, il a isolé la partie non expliquée de la sinistralité après prise en compte de ces critères. Cette partie correspond aux résidus de la régression.

Étant donné que la régression multinomiale est réalisée à la maille individuelle, une étape d'agrégation des résidus au code postal, unité géographique retenue pour l'élaboration de ce zonier, s'est avérée nécessaire. Les résidus individuels de chaque code postal ont donc été agrégés en créant des indicateurs à cette maille. La moyenne et l'écart-type des résidus individuels associés à la modalité "+90" jours dans le code postal sont deux exemples d'indicateurs. Au vu du nombre important d'indicateurs créés, une sélection est effectuée en utilisant le lasso pour ne retenir que les indicateurs les plus pertinents. Cette sélection a aussi permis d'accorder plus de poids aux arrêts de travail les plus longs (et à l'absence d'arrêt de travail), modalités qui intéressent le plus les assureurs puisque c'est la survenance d'arrêts longs (ceux dépassant le délai de franchise) qui représente le risque le plus important.

La troisième partie de la méthodologie traite de la construction du zonier incapacité à partir des indicateurs sélectionnés précédemment. Au lieu d'avoir recours à un classique lissage spatial d'un seul indicateur, la méthode utilisée dans cette étude est une classification ascendante hiérarchique avec contraintes de proximité géographique [3], basée sur plusieurs indicateurs. Cette classification n'est réalisée que sur les codes postaux ayant suffisamment d'affiliés [4] et l'objectif

3. Cette méthode est présentée dans (Chavent et al., 2018)[2].
4. L'idée de cette contrainte est de ne considérer dans un premier temps que les codes postaux dont les indicateurs sont robustes.

de cette méthode est de tenir compte de la proximité géographique dans la constitution des clusters, qui deviendront les futures classes de risques associées à la localisation de l'entreprise.

Le zonier obtenu avec cette méthode est donné en figure NDS.2. La classification est exprimée par ordre croissant du risque, c'est-à-dire que la classe 1 contient les codes postaux les moins risqués et la classe 4 les plus risqués. En guise de première validation de cette classification, la figure NDS.3 représente la répartition des fréquences des arrêts de travail de plus de 90 jours (par code postal) en fonction des classes de risques.

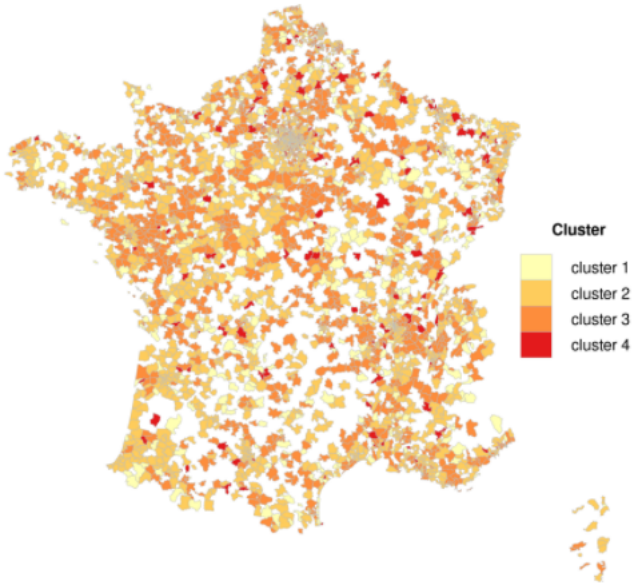

FIGURE NDS.2 – Zonier avant traitement des codes postaux avec peu ou pas d'affiliés (France métropolitaine)

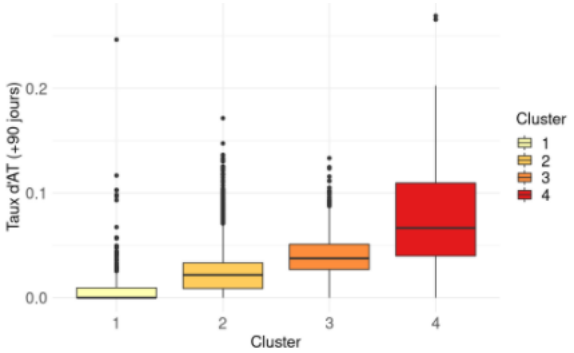

FIGURE NDS.3 – Fréquence des arrêts de travail de plus de 90 jours par cluster

Comme évoqué précédemment, seulement les codes postaux considérés comme ayant suffisamment d'affiliés sont affectés à une classe de risque. Ainsi, les codes postaux représentés en blanc sur la carte sont ceux dont le calcul de la sinistralité est jugé insuffisamment robuste. Le lissage entre les codes postaux voisins est plutôt limité. Les codes postaux voisins ne sont pas forcément affectés à la même classe de risque. Néanmoins, il y a certaines zones où cela se produit. C'est notamment le cas pour les régions Pays de la Loire et Centre-Val de Loire où les codes postaux sont dans la grande majorité classés dans le groupe 3. Les codes postaux n'étant pas encore associés à une classe de risque devront l'être dans un second temps puisque Malakoff Humanis se doit de pouvoir proposer un tarif partout en France. Cette affectation est un des objectifs de la quatrième étape de la méthodologie, détaillée plus loin. La méthode

d'affectation tiendra compte des classes de risques du voisinage et permettra ainsi d'obtenir un zonier davantage lissé.

Les résultats donnés dans la figure NDS.3 sont plutôt satisfaisants. La sinistralité croît avec les classes de risques (les boîtes à moustaches sont globalement plus élevées au fur et à mesure que les classes de risques augmentent). Les clusters 3 et 4 ont bien une sinistralité plus importante que les clusters 1 et 2. Cependant, il s'agit d'une analyse bivariée qui ne tient donc pas compte des potentielles corrélations avec d'autres variables (âge, CSP, genre et secteur d'activité). Des résultats prenant en compte ces corrélations sont évoqués dans la dernière étape de la méthodologie.

La quatrième phase de la méthodologie (cf. figure NDS.1) a pour but de conduire une analyse du zonier et d'en dégager des interprétations. Pour cela, une modélisation des classes de risques obtenues précédemment est réalisée avec l'aide de données démographiques et socio-économiques de l'INSEE. L'objectif de cette modélisation basée sur les forêts aléatoires (random forest) est double :

— Interpréter, avec l'aide des SHAP [5] values, le clustering spatial obtenu précédemment. Cela permettra de comprendre si l'appartenance d'un code postal à une certaine classe de risque est corrélée avec son niveau de richesse, ses équipements, etc.

— Prédire la classe de risque pour les codes postaux avec peu ou pas d'affiliés, en utilisant leurs caractéristiques démographiques et socio-économiques, tout en tenant compte de la proximité géographique.

Le premier objectif fait appel à la notion de valeur de Shapley, introduite dans un premier temps dans le domaine de la théorie des jeux. Dans le contexte du Machine Learning, la valeur de Shapley correspond à la contribution marginale d'une variable explicative à la prédiction d'une observation. Par exemple, une valeur de Shapley de 0.1 pour une observation et une variable explicative données signifie que cette variable fait augmenter de 0.1 la prédiction de cette observation. Afin d'obtenir une interprétation globale du modèle, les SHAP values doivent être calculées sur l'ensemble des observations. La zone 2 est prise en exemple dans le graphique NDS.4, appelé "summary plot" et qui répertorie les SHAP values associées à la prédiction de cette zone, et ce, pour chaque variable et chaque observation.

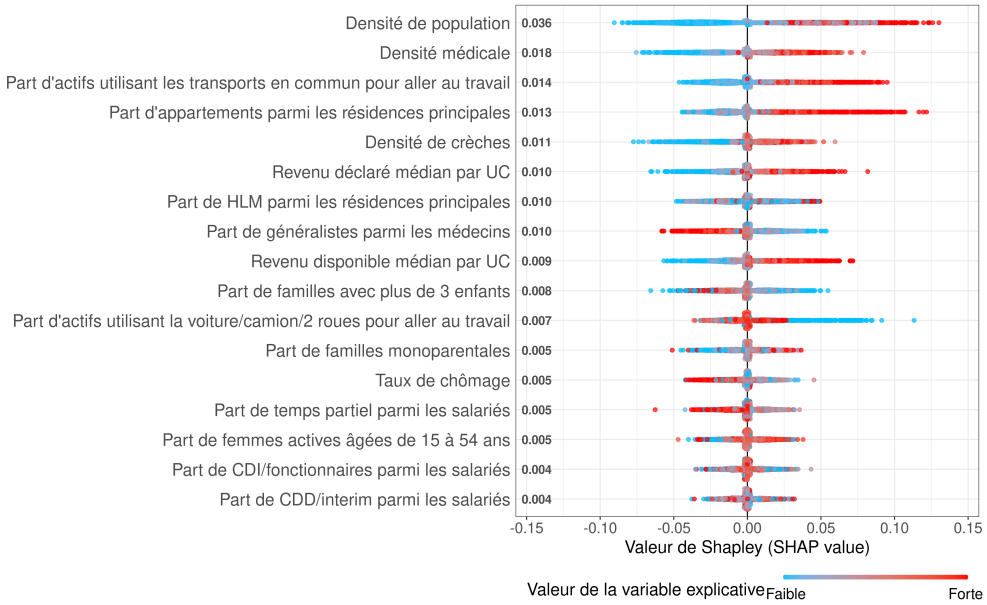

FIGURE NDS.4 – SHAP values pour la zone 2 ("summary plot")

5. SHapley Additive exPlanations.

Quelques détails sont donnés ci-dessous pour faciliter l'interprétation du graphique précédant. Tout d'abord, les variables explicatives sont triées dans l'ordre décroissant d'importance pour la zone en question. Ensuite, chaque point correspond à la valeur de Shapley pour une variable et un code postal. Enfin, ce point est coloré en fonction de la valeur de la variable explicative donnée (forte valeur en rouge et faible valeur en bleu). Ainsi, la densité de population, variable la plus importante pour discriminer la zone 2, serait corrélée positivement avec cette zone. En effet, les codes postaux avec une forte densité de population (les points rouges) ont globalement des SHAP values positives, ce qui signifie qu'une forte densité de population augmente la probabilité d'être en zone 2. De manière complémentaire, les valeurs de Shapley sont généralement négatives pour les codes postaux à faible densité de population (les points bleus), signe qu'une faible densité de population diminue la probabilité d'être en zone 2. La zone 2 est donc plutôt composée de codes postaux urbains. L'analyse de ce type de graphique pour chacune des zones permet de dresser le panorama suivant [6] :

— **Zone 1.** La zone la moins risquée en termes de probabilité de tomber en arrêt de travail est aussi la plus rurale de toutes les zones. Elle réunit une majorité des critères associés à une zone rurale : faible densité de population, peu de HLM (en proportion des résidences principales), peu d'actifs utilisant les transports en commun pour aller travailler et des territoires faiblement équipés (peu de médecins et de crèches).
— **Zone 2.** Cette zone est la plus urbaine parmi les quatre. Elle rassemble une grande partie des caractéristiques liées aux zones urbaines : forte densité de population, beaucoup d'actifs utilisant les transports en commun pour se rendre au travail, des résidences principales plutôt composées d'appartements et des territoires plutôt équipés (beaucoup de crèches, de médecins et plus particulièrement de médecins spécialistes). C'est aussi dans cette zone que les habitants sont globalement les plus riches.
— **Zone 3.** Cette zone semble être la plus défavorisée. C'est clairement celle où les habitants ont les revenus les plus faibles et une importante présence de HLM est observée dans cette zone.
— **Zone 4.** La zone la plus risquée en termes de probabilité de tomber en arrêt de travail est aussi une zone plutôt rurale. Elle se caractérise par des faibles parts de HLM et d'appartements parmi les résidences principales, des territoires faiblement équipés (peu de médecins et de crèches) mais une densité de population plus importante que la zone 1.

Maintenant que les interprétations des classes de risques ont été données, il est possible de détailler le deuxième objectif de cette modélisation des classes de risques à partir des données INSEE, à savoir l'attribution d'une classe de risque aux codes postaux avec peu ou pas d'affiliés.

Le random forest utilisé est un modèle de Machine Learning qui renvoie pour toute nouvelle observation, une probabilité d'appartenance à chacune des zones. Il est donc possible de l'utiliser pour déterminer la classe de risque des codes postaux avec peu ou pas d'affiliés, en leur affectant la zone ayant la probabilité prédite la plus élevée. Cette méthode présente le défaut de ne pas prendre en compte la proximité géographique dans l'affectation des classes de risques. En effet, le random forest pourrait attribuer pour un code postal donné, une classe de risque différente de celles de ses voisins, ce qui réduirait le lissage du zonier et créerait des territoires isolés. Pour résoudre ce problème et tenir compte de la proximité géographique, une idée est d'attribuer pour un code postal donné (et non encore affecté à une zone), la classe de risque ayant la plus grande probabilité parmi celles de ses 4 plus proches voisins. Ainsi, les classes de risques seraient attribuées aux codes postaux concernés en fonction de leurs caractéristiques démographiques et socio-économiques mais aussi de leur voisinage. La figure NDS.5 présente le résultat de cette méthode qui permet d'obtenir un zonier complet.

6. Ces résultats peuvent facilement se retrouver à partir du tableau 8.5 ou en analysant les "summary plots" de chaque zone (cf. figures 8.8, A.3 et A.4).

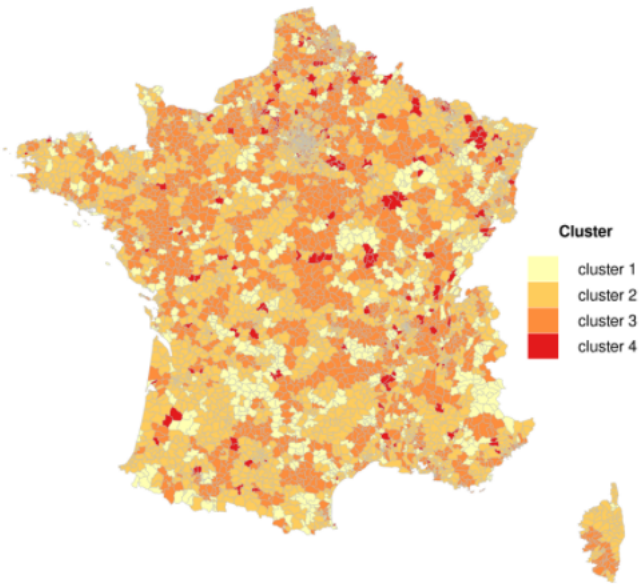

FIGURE NDS.5 – Zonier après traitement des codes postaux avec peu ou pas d'affiliés (France métropolitaine)

Comme attendu, la carte ci-dessus est bien plus lissée que celle présentée avant traitement des codes postaux avec peu ou pas d'affiliés (cf. figure NDS.2).

Pour finir, la dernière étape de la méthodologie (cf. figure NDS.1) est une étape de validation de la classification obtenue. À ce titre, une comparaison de la modélisation du risque incapacité avec et sans la variable «zonier» est effectuée. Grâce à différents outils statistiques (V de Cramer, valeur et significativité des coefficients de régression, tests statistiques, AIC/BIC), il a été démontré que le zonier apporte de l'information pour améliorer la connaissance du risque incapacité. Autrement dit, le risque incapacité n'est pas le même partout en France. Ceci a été confirmé par l'application du zonier sur une base de données "test". En effet, bien que le zonier soit moins performant sur les données de test, les écarts de sinistralité entre les zones restent largement significatifs. L'ajout du critère "localisation de l'entreprise" dans la tarification prévoyance collective est donc tout à fait envisageable et pertinent.

Executive summary

The nominative social declaration (DSN for Déclaration Sociale Nominative in French) is a system allowing each employer to transmit every month in a unique and dematerialised way, a set of information regarding the social protection of their employees. Although administrative simplification is the main objective of this system, the DSN is also a real opportunity for insurers since it gathers a large amount of information on employees such as details on their employment contract, sick leaves, wages and so on. Part of this data was already available to insurers before the implementation of the DSN, but the latter facilitates access by centralising information. The DSN therefore enables insurers to have a better knowledge of their policyholders and the level of risk borne, which could result in the addition of one (or more) degree(s) of segmentation in their pricing system of group insurance contracts. As a reminder, having a more segmented pricing than the market can be a competitive advantage, especially by attracting the least risky profiles.

The aim of this dissertation is to determine if the company location can become a relevant pricing criterion for the temporary work disability cover in group insurance. For this purpose, a zoning is developed by using data from the DSN. Geographic criterion is not very present in the pricing of life insurance policies while it is essential in car and house insurance. More specifically, the pricing criteria used at Malakoff Humanis for temporary work disability cover in group insurance are the average age, the gender distribution, the business sector and the distribution of the different socio-professional categories within the company. Thus, to develop a zoning, it is necessary to take into account these preceding criteria since the company location would complement them in the pricing system. This is the reason why the zoning is built from the residuals of a regression having as explanatory variables the criteria previously mentioned [1].

In order to carry out this regression, we need to define an explained variable which reflects the risk of sick leaves. The subject is approached with the dimensions of frequency and duration of sick leaves [2]. More precisely, the target variable corresponds to the sick leave occurrence of different durations, categorized as follows:

— No sick leave ("No SL"),
— 0-15 days of sick leave ("]0,15]"),
— 15-30 days of sick leave ("]15,30]"),
— 30-90 days of sick leave ("]30,90]")
— More than 90 days of sick leave ("+90")

This choice of variable is linked to our ambition to have an overall knowledge of the sick leave risk and not only from a frequency or a duration point of view.

The construction of the zoning required several stages. The methodology implemented is detailed in the figure ES.1.

1. As a reminder, the residuals of the regression contain the unexplained part of the occurrence of sick leaves after taking into account explanatory variables.

2. Maternity, paternity and adoption leave were excluded from the scope of study because they are covered by specific guarantees which are not concerned by this zoning.

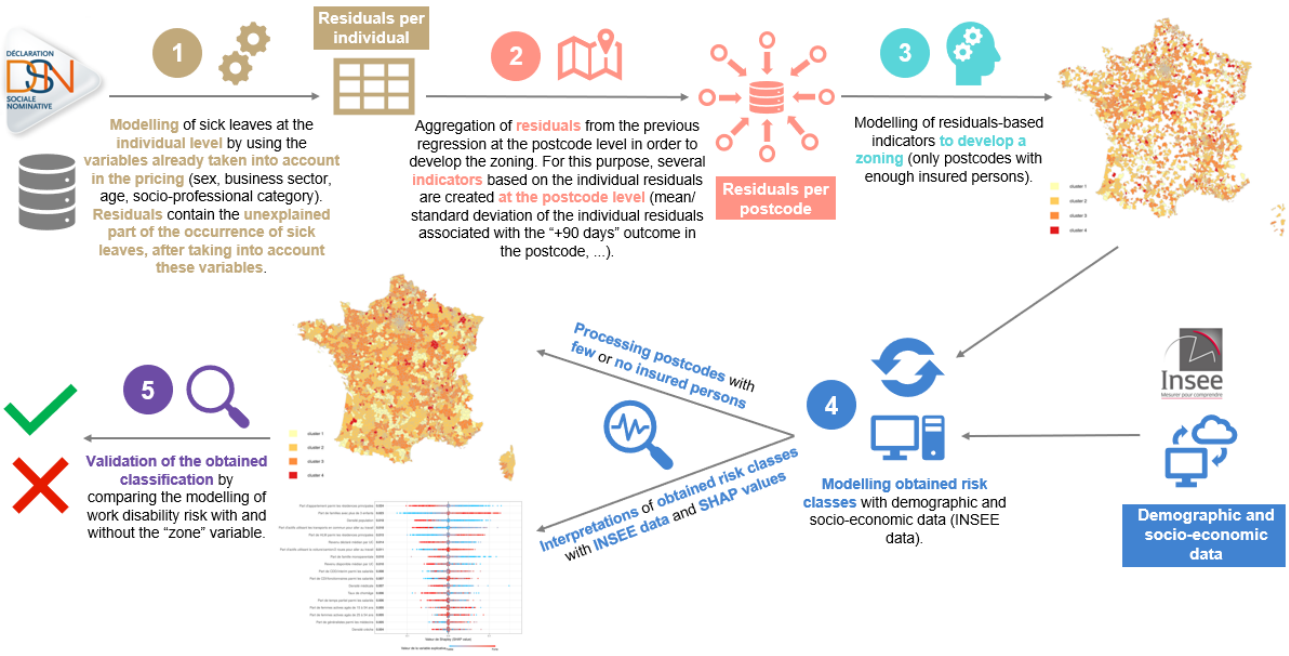

Figure ES.1 – Methodology for developing the zoning

The first step consists in controlling the effects of the variables already used in the pricing system on sick leave risk. For this purpose, a multinomial logistic regression with the variables age, sex, socio-professional category and business sector (in the form of risk classes) is implemented. This modelling was carried out at the individual level. This is the reason why "sex" variable is used instead of the "gender distribution" variable. This model has, on the one hand, confirmed the relevance of the usual pricing criteria. On the other hand, it isolated the unexplained part of the sick leaves occurrence, after taking into account these criteria. This part corresponds to the residuals of the regression.

Since the multinomial regression is carried out at the individual level, aggregating the residuals at the postcode level, geographical unit chosen for the zoning, constitutes a necessary step. Individual residuals of each postcode were therefore aggregated by creating indicators at this level. The mean and standard deviation of the individual residuals associated with the "+90 days" category in the postcode are two examples of indicators. Given the large number of indicators created, a selection is done using the lasso to retain only the most relevant indicators. This selection also enabled to give more weight to sick leaves of long duration, categories that particularly interest insurers since occurrence of long sick leaves (those exceeding the deductible period) represents the highest risk.

The third part of the methodology deals with the development of the zoning based on the indicators selected previously. Instead of using a classic spatial smoothing of one indicator, the method implemented in this study is the hierarchical clustering with spatial constraints [3], based on several indicators. This clustering is carried out only on postcodes having a sufficient number of policyholders [4]. The aim of this method is to take into account the geographical proximity in the composition of clusters, which will become the future risk classes associated with the company location.

The zoning obtained with this method is given in figure ES.2. The classification is ranked in ascending order of risk, that is, class 1 contains the least risky postcodes and class 4 the

3. This method is described in (Chavent et al., 2018) [2].
4. The idea of this constraint is to consider only the postcodes whose indicators are robust.

riskiest. As a first validation of this clustering, the figure ES.3 represents the distribution of the frequencies of sick leaves exceeding 90 days (by postcode) according to the risk classes.

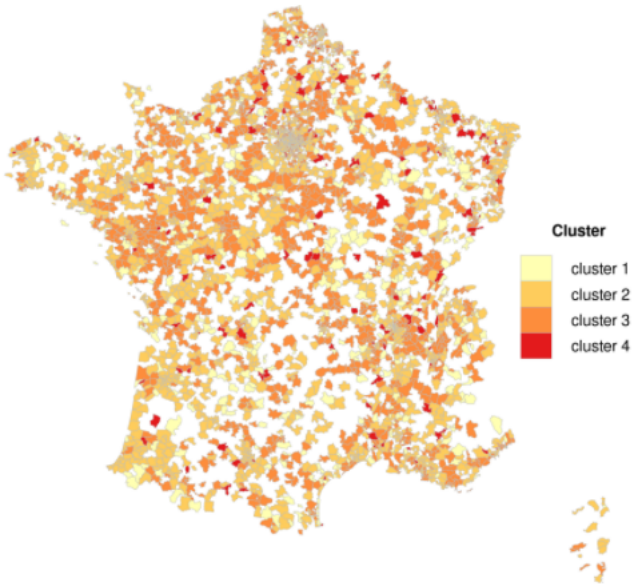

Figure ES.2 – Zoning before processing postcodes with few or no insured persons (Metropolitan France)

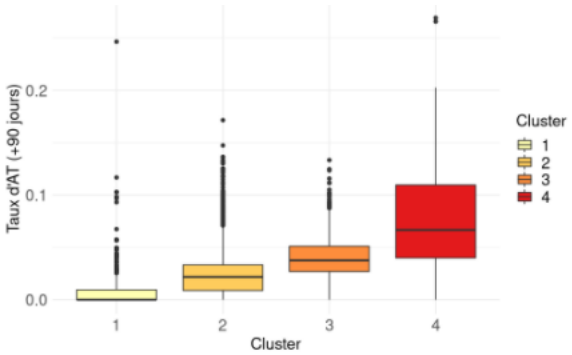

Figure ES.3 – Frequency of sick leaves exceeding 90 days per cluster

As mentioned above, only postcodes having a sufficient number of policyholders are assigned to a risk class. Thus, the postcodes shown in white on the map are those for which the indicators are not robust. The smoothing between neighbouring postcodes is rather limited. Neighbouring postcodes are not necessarily assigned to the same risk class. However, there are some areas where this scenario happens. This is particularly the case for the "Pays de la Loire" and "Centre-Val de Loire" regions where the postcodes are mostly classified in cluster 3. Postcodes which are not associated to a risk class yet should be processed in a second step because Malakoff Humanis has to be able to offer a price everywhere in France. This processing is one of the aims of the fourth step of the methodology, detailed below. It will take into account the risk classes of the neighbourhood and will therefore make it possible to obtain a smoother zoning.

The results given in figure ES.3 are quite good. The rate of long-term sick leaves increases with the risk classes (boxplots are generally higher as the risk class increases). Clusters 3 and 4 have a greater rate of long-term sick leaves than clusters 1 and 2. However, this bivariate analysis does not take into account the potential correlations with other variables (age, socio-professional category, sex and business sector). Results taking into account these correlations

are mentioned in the last step of the methodology.

The fourth step of the methodology (see figure ES.1) aims to develop an analysis of the different risk classes and draw interpretations. To do so, risk classes obtained previously are modelled with demographic and socio-economic data from INSEE. This modelling, based on random forest, has two objectives:

— To interpret, with SHAP [5] values, the spatial clustering obtained previously. This will make it possible to understand whether the belonging of a postcode to a certain risk class is correlated with its wealth, its equipment and so on.

— To predict the risk class for postcodes with few or no policyholders, by using their demographic and socio-economic characteristics, while taking into account geographic proximity.

The first objective uses the concept of Shapley value, first introduced in the field of game theory. In the context of Machine Learning, the Shapley value corresponds to the marginal contribution of an explanatory variable to the prediction of an observation. For example, a Shapley value of 0.1 for a given observation and explanatory variable means that this variable increases the prediction of that observation by 0.1. In order to obtain a global interpretation of the model, SHAP values must be calculated on all observations. Zone 2 is taken as an example in the graph ES.4, called "summary plot" and which lists the SHAP values associated with the prediction of this zone, for each variable and each observation.

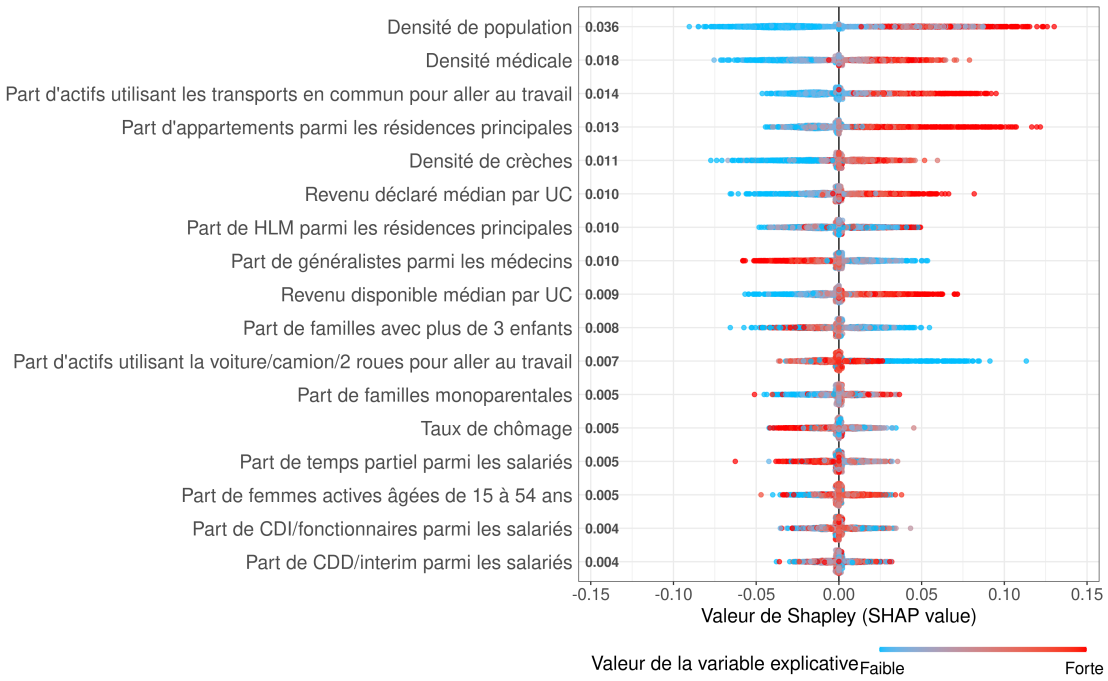

Figure ES.4 – SHAP values for the zone 2 ("summary plot")

A few details are given below to facilitate the interpretation of the previous graph. First of all, the explanatory variables are sorted in descending order of importance for the area in question. Next, each point corresponds to the Shapley value for a variable and a postcode. Finally, this point is coloured according to the value of a given explanatory variable (high value in red and low value in blue). Thus, the population density, the most important variable to discriminate zone 2, would be positively correlated with this zone. Indeed, postcodes with a high population density (the red dots) have globally positive SHAP values, which means

5. SHapley Additive exPlanations.

that a high population density increases the probability of being in zone 2. At the same time, Shapley values are generally negative for postcodes with low population density (the blue dots), indicating that a low population density decreases the probability of being in zone 2. Zone 2 is therefore rather composed by urban postcodes. The analysis of this type of graph for each of the zones makes it possible to draw up the following panorama [6]:

— **Zone 1.** The least risky zone in terms of probability of sick leaves is also the most rural of all zones. It meets most of the criteria associated with a rural area: low population density, few HLM [7] (as a proportion of main residences), few workers using public transport to go to work and poorly equipped areas (few doctors and nurseries).

— **Zone 2.** This zone is the most urban of the four. It has many of the characteristics associated with urban areas: high population density, many workers using public transport to go to work, main residences rather composed of flats and territories rather well-equipped (many nurseries, doctors and particularly specialist doctors). It is also in this area that the inhabitants are globally the richest.

— **Zone 3.** This area seems to be the most disadvantaged one. It is clearly the area where the inhabitants have the lowest incomes and a significant presence of HLM is observed in this zone.

— **Zone 4.** The riskiest area in terms of probability of sick leaves is also a rather rural area. It is characterised by a low proportion of low-rent housing (HLM) and a low proportion of flats among the main residences. It is also associated to poorly equipped areas (few doctors and nurseries) but with a higher population density than zone 1.

Now that the interpretations of the risk classes have been given, it is possible to detail the second objective of this modelling of risk classes from INSEE data, namely the assignment of a risk class to postcodes with few or no policyholders.

The random forest used is a Machine Learning model that returns a probability of belonging to each zone for each new observation. It is therefore possible to use it to determine the risk class of postcodes with few or no policyholders by assigning them to the area having the highest probability predicted by the model. The drawback of this method is that it does not take geographical proximity into account when assigning risk classes. Indeed, the random forest could assign for a given postcode a different risk class than its neighbours, which would reduce zoning smoothing and create isolated territories. To solve this problem and to take into account geographical proximity, one idea is to assign to a given postcode (not yet assigned to a zone) the risk class with the highest probability among its 4 closest neighbours. In this way, risk classes would be assigned to the postcodes concerned according to their demographic and socio-economic characteristics but also to their neighbourhood. Figure ES.5 shows the result of this method, which gives a complete zoning.

6. These results can easily be found from the table 8.5 or by analysing the "summary plots" of each zone (see figures 8.8, A.3 and A.4).

7. HLM for Habitation à Loyer Modéré in French.

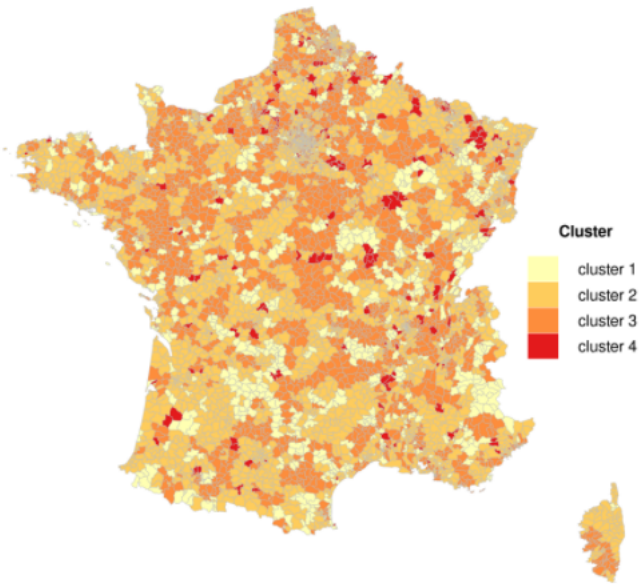

Figure ES.5 – Zoning after processing postcodes with few or no insured persons (Metropolitan France)

As expected, the map above is much smoother than the one presented before processing postcodes with few or no policyholders (see figure ES.2).

Finally, the last step of the methodology (see figure ES.1) is a validation step of our classification. In this respect, a comparison of the modelling of sick leave risk with and without the zoning variable is carried out. Thanks to several statistical tools (Cramer's V, value and significance of regression coefficients, statistical hypothesis tests, AIC/BIC), it has been shown that the zoning provides information to improve knowledge of sick leave risk. In other words, the sick leave risk is not the same everywhere in France. It was confirmed by applying the zoning to a test dataset. Although the zoning is less efficient on the test dataset, differences in claims occurrence between zones remain significant. Adding the "company location" as a criterion in the group insurance pricing system for the sick leave coverage is therefore entirely conceivable and relevant.