**Mémoire présenté devant l'ENSAE Paris**
**pour l'obtention du diplôme de la filière Actuariat**
**et l'admission à l'Institut des Actuaires**
**le 14/03/2022**

Par : **Maude Bellugeon**

Titre: **Impact de la variable métier sur la segmentation tarifaire des contrats de prévoyance**

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*

*Entreprise :*

Nom :
*Signature :*

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

*Directeur de mémoire en entreprise :*

*Nom : Aymeric Veyron*
*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

**Mots clés :** *Absentéisme, classification, clustering, baromètre sectoriel, distance d'arbre, distance TF-IDF, segmentation*

L'objectif de ce mémoire est d'obtenir un modèle prédictif de l'absentéisme intégrant au mieux la variable métier et d'analyser les effets de cette variable sur les performances du modèle ainsi que sur la tarification des contrats de prévoyance associés.

Il s'organise en deux parties distinctes. Dans un premier temps, nous nous intéressons à l'analyse de la variable métier dans le but d'obtenir une mesure de distance pertinente entre les métiers. Pour ce faire, deux pistes principales sont explorées : l'adaptation de méthodes d'analyses textuelles et l'adaptation de modèles markoviens. Ces différentes approches permettent de quantifier plus précisément la similitude entre deux métiers et sont ensuite utilisées pour améliorer la qualité des *clusterings*. Cet objectif revêt des intérêts opérationnels pour l'entreprise : une méthodologie robuste permettant de comparer les modalités de cette variable qualitative non-ordonnée permet en effet d'améliorer la segmentation lors de la création de *benchmarks*.

Dans un second temps, nous réinjectons les résultats de cette étude dans un modèle actuariel. Nous calibrons un modèle tarifaire fondé sur une approche fréquence/sévérité sur nos données et cherchons à identifier les éventuels effets de cette nouvelle méthode de comparaison des métiers. Nous comparons ainsi les performances et les résultats de trois modèles, l'un ne prenant pas en compte la variable métier, l'autre utilisant la comparaison binaire entre métiers et le dernier en utilisant la mesure de distance entre métiers obtenue lors de la première partie. L'analyse croisée de ces différents modèles permet ainsi de confirmer l'importance de la variable métier dans la modélisation ainsi que l'efficacité de la méthodologie mise en oeuvre lors de l'élaboration de la nouvelle mesure de distance entre métiers.

A l'issue de ce mémoire, nous disposons finalement d'une nouvelle mesure de distance entre métier et d'une nouvelle variable catégorielle à cinq modalités permettant de rendre compte du métier, au même niveau de granularité que le niveau 1 du PCS. Cette variable permet d'améliorer la qualité du modèle de tarification et de segmentation calibré sur nos données, tout en conservant la cohérence avec les deux autres modèles, démontrant ainsi sa pertinence dans la modélisation et validant l'esprit de sa construction.

# Abstract

**Keywords :** *Absenteeism, classification, clustering, sector barometer, tree distance, TF-IDF distance, segmentation*

The aim of this paper is to obtain a predictive model of absenteeism that best integrates the business variable and to analyze the effects of this variable on the model's performance, as well as on the pricing of the associated insurance contracts. It is organized into two distinct parts.

First, we focus on the analysis of the business variable to obtain a relevant measure of distance between businesses. To achieve this, two main avenues are explored : the adaptation of textual analysis methods and the adaptation of Markov models. These different approaches enable us to quantify the similarity between two trades more precisely and are then used to improve clustering quality. This objective has operational benefits for the company : a robust methodology for comparing the modalities of this non-ordered qualitative variable, which can be used to improve segmentation when creating benchmarks.

Secondly, we use the results of this study in an actuarial model. We calibrate a rate model based on a frequency/severity approach on our data and seek to identify any possible effect of this new method of job comparison. We compare the performances and results of three models, one of which takes no account of the occupation variable, another using the binary comparison between trades, and the last using the distance measure between trades obtained in the first part. Cross-analysis of these different models thus enables us to confirm the importance of the business variable in modeling, as well as the effectiveness of the methodology implemented in the development of the new inter-business distance measure.

We finally have a new measure of distance between trades and a new five-modality categorical variable to account for the trade, at the same level of granularity as PCS level 1. This variable improves the quality of the calibrated pricing and segmentation model calibrated on our data, while maintaining consistency with the other two other models, thus demonstrating its relevance to the model and validating the idea of its construction.

# Table des matières

# 1 Introduction

L'absentéisme est passé d'un sujet de préoccupation à un enjeu global pour nos entreprises : son coût direct est estimé à 25 milliards d'euros annuels. A ce coût direct, qui ne représente que les prestations versées, s'ajoutent les coûts indirects, liés au remplacement des personnes en arrêt, à la perte de performance générée et la baisse de qualité de service. Selon les estimations, ces coûts indirects de l'absentéisme pourraient être jusqu'à trois fois plus élevés que les coûts directs.

En outre, l'absentéisme déséquilibre les comptes de résultats des régimes prévoyance. En effet, l'aggravation éventuelle de l'absentéisme nécessite la mise en place de provisions mathématiques conséquentes par les assureurs, qui peuvent durablement peser sur le compte de résultat. La compréhension, voire l'anticipation de l'évolution de l'absentéisme est ainsi un enjeu majeur pour les entreprises comme pour les assureurs.

Dans ce contexte, la question de la segmentation sectorielle de l'absentéisme offre un double intérêt. Du point de vue des entreprises clientes, il permet la constitution d'un baromètre de l'absentéisme beaucoup plus fin que celui obtenu en comparant uniquement les entreprises d'un même secteur via par exemple un calcul de taux d'absentéisme moyen. Cette approche est en effet discutable : elle compare des entreprises qui peuvent avoir des profils type d'employés très différents ou encore des pyramides des âges inversées. Elle est en outre inapplicable pour des secteurs monopolisés par un petit nombre d'entreprises. Du point de vue des assureurs, la mise en place d'une méthodologie de segmentation du risque en fonction des caractéristiques des entreprises s'inscrit dans une logique d'homogénéisation des classes de risques et donc d'amélioration de la tarification pour les contrats prévoyance.

L'objectif final est donc d'obtenir une modélisation prédictive la plus réaliste possible de l'absentéisme en fonction des caractéristiques individuelles. Un tel modèle est toutefois quelque peu ambitieux et se heurte à de nombreux écueils techniques, parmi lesquels la question de l'intégration du métier comme variable explicative. Cette donnée est en effet régie par une nomenclature rigide, la PCS-ESE, qui n'offre que peu de latitude quant à la comparaison de deux métiers. En outre, les approches classiques de *machine learning* traitent les données catégorielles de manière manichéenne : deux personnes ont ou n'ont pas le même métier.

Dans ce mémoire, nous tenterons d'affiner cette conclusion en développant une méthodologie

aboutissant à une mesure de distance entre les différents métiers. Il s'organise en deux parties distinctes.

Dans un premier temps, nous cherchons à construire une distance entre métiers cohérente. Cette étape préalable permet en outre de traiter les problèmes d'hétérogénéité du portefeuille. Pour cela, trois angles d'approche ont été utilisés. Le premier postule que deux emplois qui peuvent être exercés en même temps sont proches. Le second considère les transitions entre emplois, en supposant que la probabilité de passer d'un emploi à un autre fournit une mesure pertinente de leur proximité. Le troisième, enfin, compare les individus d'un point de vue plus socioéconomique en utilisant une table de données de l'INSEE. En combinant les trois, on obtient ainsi une mesure de distance entre PCS. Cette mesure de distance est ensuite utilisée pour effectuer un *clustering* entre PCS et créer une nouvelle variable catégorielle à cinq modalités, plus pertinente que la nomenclature PCS. Ces différentes approches permettent de quantifier plus précisément la similitude entre deux métiers et peuvent ensuite être utilisées pour améliorer la qualité des *clusterings*. Elles s'inscrivent également, à long terme, dans une démarche de segmentation du marché : des groupes de métiers se dégagent, invitant ainsi à considérer de nouveaux groupes de tarification dans les contrats prévoyance.

Dans un second temps, nous injectons cette mesure dans le modèle de tarification et observons son influence sur les tarifs et la segmentation. Le modèle tarifaire retenu est un modèle fréquence/sévérité, en reprenant la méthodologie des mémoires *Impact de la consommation Santé sur la probabilité d'entrée en arrêt de travail* [1] et *Risque incapacité en prévoyance collective : analyse et optimisation de la segmentation tarifaire* [2] le but n'étant pas d'obtenir une tarfication la plus précise possible mais bien d'étudier l'influence de la variable métier.

Pour appuyer cette étude, nous disposerons de données issues de la Déclaration Sociale Nominative (DSN) d'entreprises. Ces données ne sont pas assurantielles et un certain nombre de variables usuellemet utilisées lors de la tarification de contrats prévoyance (éventuels jours de carence, contrat individuel ou collectif, convention collective...) ne sont donc pas disponibles, ce qui nous conduira à simplifier le modèle de tarification utilisé.

# 2 Un peu de contexte

## 2.1 Quid de l'absentéisme ?

### 2.1.1 Définitions

La notion d'absentéisme connaît des définitions variées chez les différents auteurs. Le réseau Anact-Aract le définit ainsi comme « *toute absence qui aurait pu être évitée par une prévention suffisamment précoce des facteurs de dégradation des conditions de travail entendus au sens large : les ambiances physiques mais aussi l'organisation du travail, la qualité de la relation d'emploi, la conciliation des temps professionnels et privés, etc* » [3]. Le sociologue Wilbert E. Moore définit quant à lui l'absentéisme comme « *the practice of workers failing to report for work on some slight excuse or other, or none at all* » [4]. Le Larousse en propose une définition beaucoup plus neutre et le définit comme le « *fait d'être absent du lieu de travail, de l'école, d'une réunion, d'une assemblée, de tout lieu où, pour des raisons de travail, de participation à une action etc., la présence est obligatoire* » [5]. Dans tout ce qui suit, on se restreindra à cette définition : le terme absentéisme n'impliquera donc jamais a priori que les absences de l'individu soient répétées, de complaisance ou encore évitables. Cependant, toutes les absences ne relèvent pas de l'absentéisme. En particulier, les absences liées à des droits sociaux telles que les RTT, les congés autorisés (payés ou non), les grèves, les formations etc… ne sont pas comptabilisées. En outre, certains types d'absences ne seront pas considérés dans toute la suite du mémoire. En effet, l'absentéisme est en pratique mesuré via les arrêts de travail et ne permet donc pas de prendre en compte les absences non justifiées. Un arrêt de travail est quant à lui défini comme une « *période de suspension du contrat de travail en raison d'une maladie ou d'un accident du travail d'origine professionnelle ou non professionnelle. L'arrêt de travail est justifié par une prescription médicale*. » [6]. Les arrêts de travail se scindent en plusieurs catégories, qui seront désignées dans toute la suite comme le motif de l'absence :

— les maladies professionnelles

— les accidents de travail

— les maladies liées à un problème de santé

— les congés maternité ou paternité [1]

### 2.1.2 Quelques chiffres

D'après une étude de la DARES [7], « *sur la période 2003-2011, au cours d'une semaine de référence, 3.6% des salariés ont connu une absence au travail d'au moins une heure pour des raisons de santé ou pour la garde d'un enfant malade* ». Ces chiffres, présentés en Figure 2.1, sont cependant relativement stables en volume en ce qui concerne les accidents du travail, les accidents du trajet et les maladies professionnelles. On observe cependant une légère hausse entre 2017 et 2019 [8].

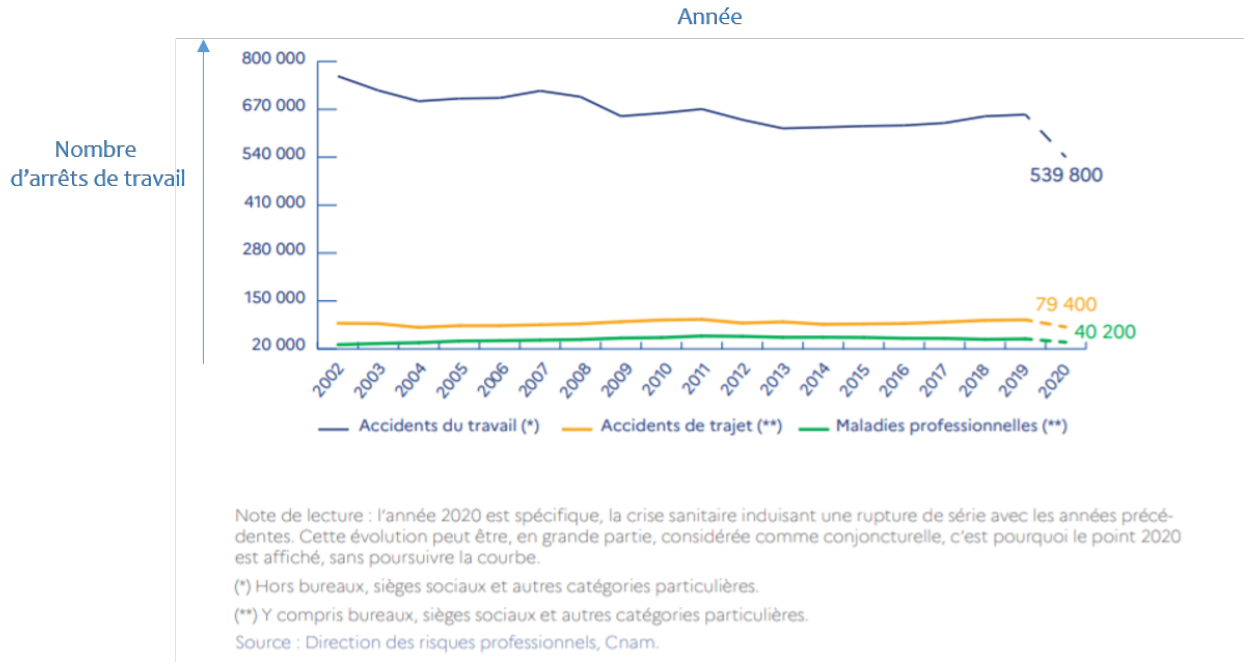


FIGURE 2.1 – Evolution du nombre d'accidents du travail, d'accidents de trajet et de maladies professionnelles avec arrêts de travail

En ce qui concerne la base de travail, qui ne contient qu'une partie du portefeuille global de SIACI SAINT HONORE, le taux d'absentéisme moyen pour l'année 2020 est 4,1% (ratio des jours d'absence calendaires par les jours calendaires théoriquement travaillés sans pondération liée au temps de travail effectif). Les chiffres 2020 sont cependant à considérer avec la méfiance la plus extrême : ils sont bien entendu biaisés par la crise sanitaire, notamment par la mise en place d'arrêts de travail dérogatoires pour les salariés à haut risque, les cas contact, les salariés cohabitant avec une personne vulnérable, etc...

1. Les congés maternité ou paternité sont exclus du champ d'étude car considérés comme non-pertinents : ils ne sont a priori par révélateurs de dysfonctionnements internes à l'entreprise, qui ne peut éthiquement se prémunir contre leur survenance.

### 2.1.3 Impact actuariel et intérêt pour l'entreprise

Dans le cadre de notre activité, la modélisation des arrêts de travail joue un rôle-clé. En sa qualité de courtier, SIACI SAINT HONORE négocie en effet, entre autre, les contrats de prévoyance pour ses clients. Or, la tarification de ces derniers est fortement impactée par l'absentéisme : une augmentation de l'absentéisme implique un recours accru aux garanties des contrats prévoyance et donc une hausse de la sinistralité pour l'assureur et une dégradation de sa rentabilité. Dans cette dynamique, les assureurs sont alors amenés à majorer les cotisations, voire à résilier les contrats les plus sinistrés au détriment des entreprises clientes. La modélisation de l'absentéisme a ainsi un intérêt double dans la défense des intérêts de nos clients. Dans une conjoncture favorable, elle fournit une marge de manœuvre lors de la négociation des contrats de prévoyance via la confrontation des chiffres obtenus avec les données des assureurs en mettant en lumière la rigueur de l'approche. A contrario, dans une conjoncture défavorable, une analyse fine de la sinistralité permet à la partie conseil RH de proposer des plans d'actions pour lutter contre l'absentéisme et de limiter l'impact d'un bilan annuel d'absentéisme lourd sur les contrats prévoyance de l'année suivante.

La question de la segmentation du risque présente quant à elle un intérêt opérationnel pour les activités de conseil RH de l'entreprise : si l'on souhaite se pencher sur l'absentéisme au sein d'une entreprise particulière, il faut un référentiel pour comparer les chiffres obtenus. Or, l'élaboration d'un outil de segmentation permet également la construction d'un *benchmark*.

# 3 Présentation des données

## 3.1 Les données internes

### 3.1.1 La base *Absences*

Les données utilisées couvrent la période du 1er janvier 2020 au 1er mars 2021. Elles contiennent des informations sur le couple individu/contrat ainsi que sur ses absences éventuelles. La granularité des données peut être choisie en fonction des besoins (annuelle, mensuelle, hebdomadaire...). Une description qualitative des données est présentée ci-dessous.

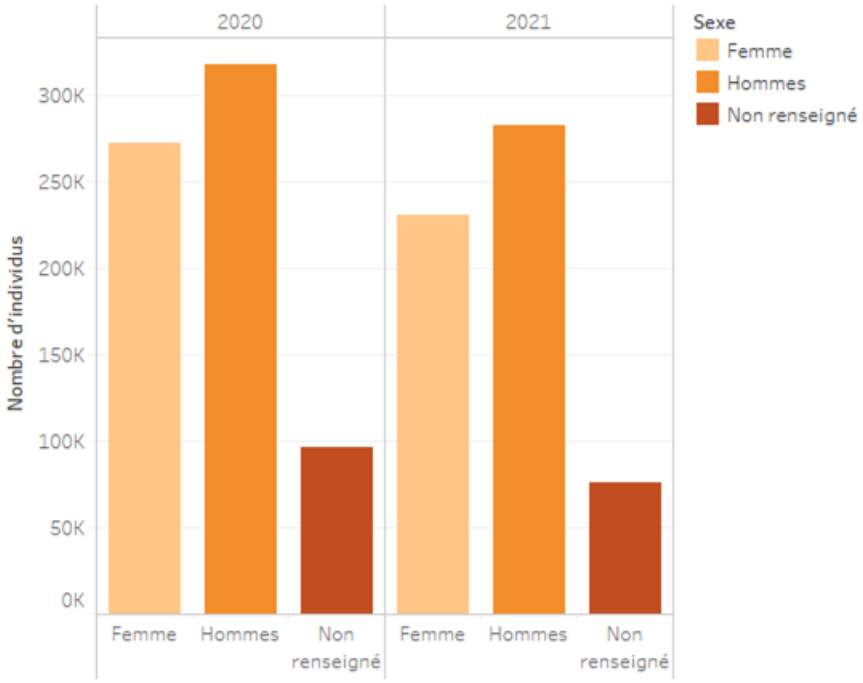


FIGURE 3.1 – Répartition des individus par genre dans la base Absences

La répartition hommes/femmes [1] au sein de la base est quasi-constante d'une année sur l'autre. Par ailleurs, le sexe n'est pas renseigné pour une partie non-négligeable des individus. La littérature

1. Figure 3.1

ayant conclu que le genre était un élément informatif important quant à l'absentéisme [9], ces individus seront radiés de l'étude.

En ce qui concerne l'âge [2], 96% du portefeuille a entre 20 ans et de 62 ans, ce qui semble cohérent puisqu'on s'intéresse à une population d'actifs. Au sein de cette tranche d'âge, les individus sont uniformément répartis. Cette homogénéité se ressent dans les statistiques descriptives : l'âge moyen est de 40,4 ans et l'âge médian de 40 ans. L'individu le plus âgé a cependant 75 ans.

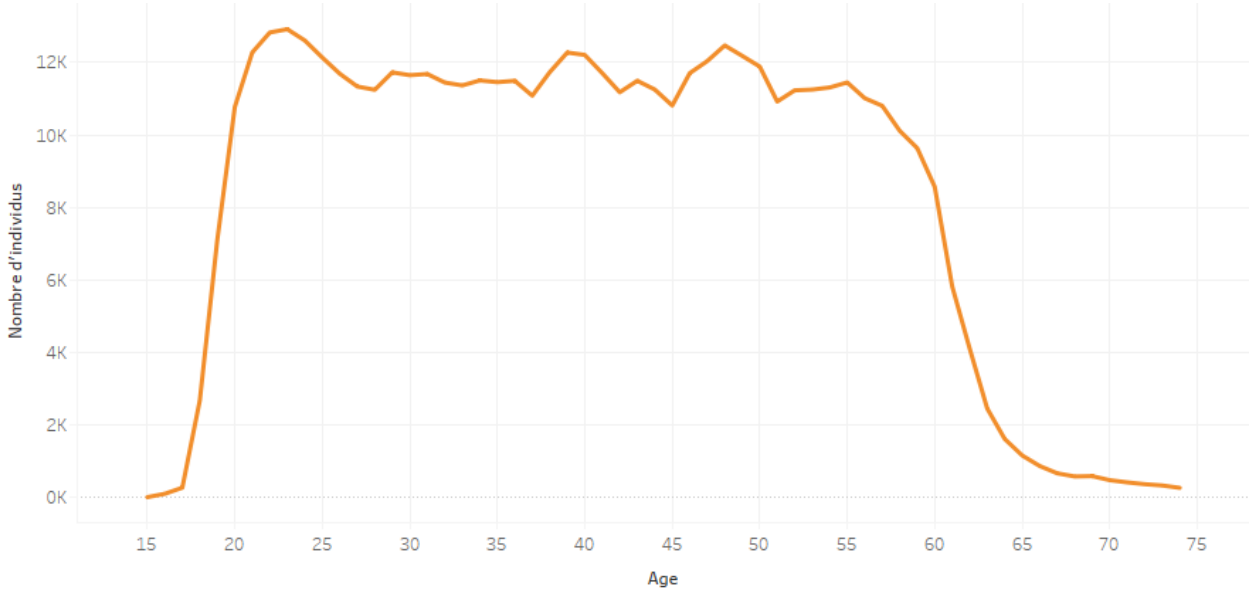


FIGURE 3.2 – Répartition des individus par âge dans la base Absences

Pour des raisons statistiques évidentes, il est clair que l'ancienneté est une variable d'intérêt lorsqu'on cherche à déterminer les facteurs qui déterminent l'absentéisme. En effet, plus un individu a d'ancienneté au sein d'une entreprise, plus il a de chances d'avoir été absent au moins une fois au sein de celle-ci. Cependant, comme le montre le graphique 3.3, la répartition de l'ancienneté est assez hétéroclite. Le mode de la série est ainsi de 1 an. L'ancienneté médiane est de 7 années et l'ancienneté moyenne de 11 années.

On s'attend également à ce que le métier exercé influe sur le comportement d'absentéisme : certaines professions sont par exemple beaucoup plus exposées aux accidents du travail. Le graphique 3.4 présente la répartition des individus en fonction de leur métier, agrégé au premier niveau de la nomenclature PCS-ESE. Cette dernière est codifiée par l'Insee et présentée plus en détail en section 5.1.

Notre portefeuille est principalement composé d'employés, d'ouvriers, de professions intermédiaires et de cadres et professions intellectuelles supérieures. Si les employés sont en majorité, les trois autres modalités sont homogènes. Cette disparité en termes de profession laisse présager des

2. Figure 3.2

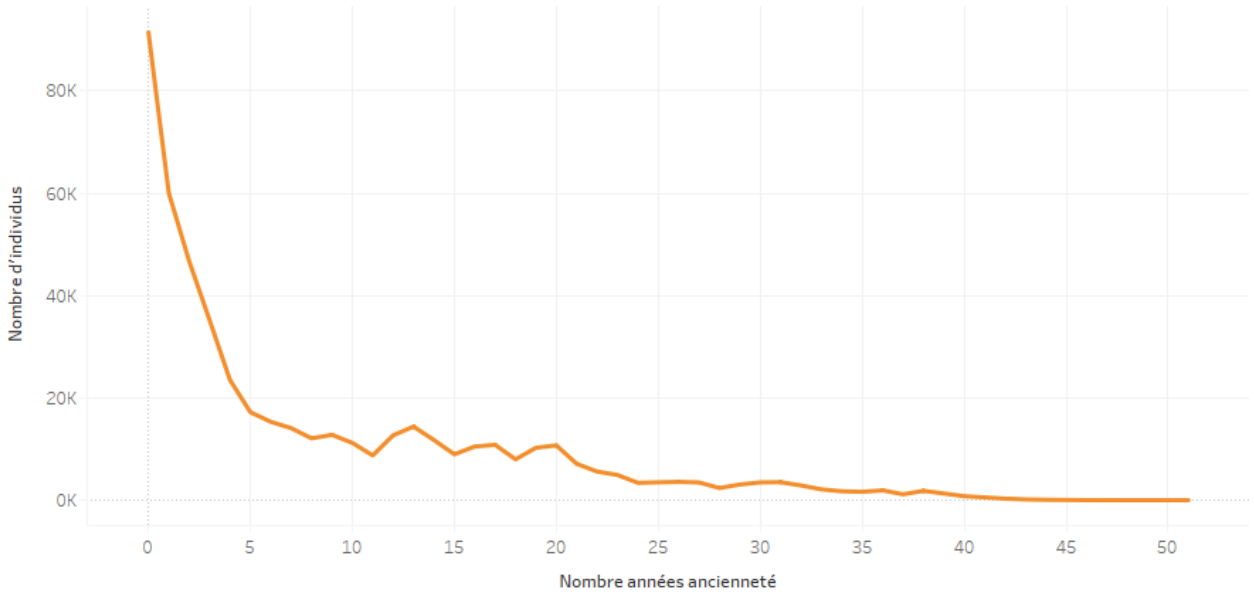


FIGURE 3.3 – Répartition des individus par ancienneté dans la base Absences

profils individuels très disparates, ce qui nécessitera probablement un traitement différencié. A une granularité inférieure, les PCS les plus présents sont les employés de commerce, les ingénieurs et cadres techniques d'entreprise, les employés administratifs d'entreprise, les cadres administratifs et commerciaux d'entreprise, les ouvriers qualifiés de type industriel, les professions intermédiaires administratives et commerciales des entreprises et les techniciens. Les disparités se retrouvent à cette échelle ce qui rend la segmentation d'autant plus pertinente. Le graphique présentant la répartition par PCS au niveau 2 est consultable en annexe A.1.

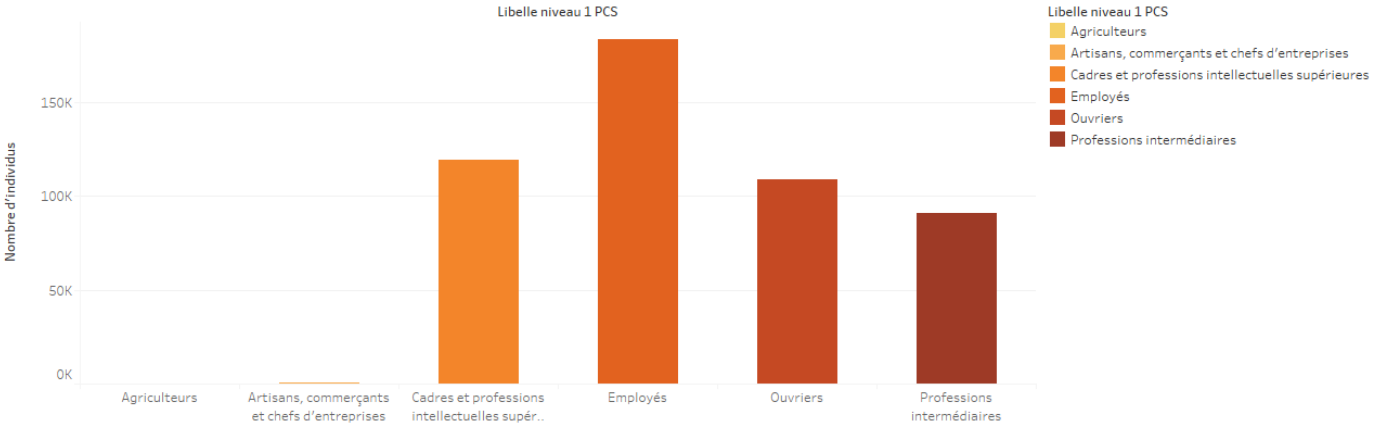


FIGURE 3.4 – Répartition des individus par PCS au niveau 1 dans la base Absences

Parmi les variables disponibles, le statut conventionnel revêt un intérêt d'un point de vue actuariel : les indemnités liées à l'absentéisme dépendent de la convention collective. Le statut le

plus représenté est celui d'ouvrier qualifié ou non qualifié, suivi des employés administratifs d'entreprises, des cadres au sens de la convention collective et des professions intermédiaires. Cette hiérarchie est différente de celle obtenue à partir des PCS : le choix de la nomenclature influe donc et sera un point de discussion important.

Enfin, on peut s'intéresser à la nature du contrat. Les résultats sont présentés en annexe A.2. L'immense majorité du portefeuille est ainsi en contrat de travail indéterminé de droit privé et la quasi-totalité des individus est regroupée dans les modalités CDI ou CDD de droit privé. Les graphiques présentant la répartition par statut conventionnel et nature du contrat sont disponibles en annexe A.

En ce qui concerne les absences, elles sont principalement dues à des arrêts maladies, comme résumé dans la figure 3.5. Dans toute la suite, les motifs d'absence conservés sont les maladies, les maladies professionnelles, les accidents de travail et les accidents de trajet. Il s'agit en effet des causes d'arrêt de travail compressibles, c'est-à-dire sur lesquels les entreprises peuvent avoir une certaine influence.


FIGURE 3.5 – Répartition des absences par motif

La répartition des absences par durée est essentielle d'un point de vue actuariel : de nombreux contrats de prévoyance prévoient une franchise d'indemnisation, souvent fixée à 90 jours. Par ailleurs, on peut suspecter que le comportement d'absentéisme différera entre les absences longues et courtes. Dans les données dont nous disposons, les absences durent majoritairement moins de 14 jours. La durée moyenne d'une absence est cependant de 58 jours du fait de certaines

valeurs extrêmes qui biaisent les données : les absences les plus longues durent plus de 5000 jours. L'étalement de la distribution se retrouve au vu des quantiles : la médiane de la distribution est de 8 jours. Si l'on se restreint aux absences de moins de 90 jours, la durée moyenne d'une absence est de 13 jours. Du fait de ces particularités de la distribution de la durée des absences et des usages de l'entreprise, la durée de 90 jours sera considérée comme pivot : les absences de plus de 90 jours sont agrégées en une seule catégorie. Dans le but de regrouper les informations de fréquence et de durée, une variable synthétique naturelle est souvent utilisée : le taux d'absentéisme. Ce taux sera ici défini par la formule suivante :

$$\text{taux d'absentéisme} = \frac{\text{Nombre de jours d'absence durant la période}}{\text{Nombre de jours travaillés durant la période}}$$

Dans cette définition du taux d'absentéisme, les week-ends et jours fériés sont comptabilisés au numérateur et au dénominateur : un arrêt qui débute le jeudi et se termine le mardi (inclus) comptera pour un arrêt de six jours et un contrat à temps plein annuel comptera pour 365 jours au dénominateur. On pourrait choisir d'autres conventions de calcul : par exemple ne travailler qu'en jours ouvrés (ou ouvrables). Cependant, le fait de travailler en jours calendaires permet de limiter les cas particuliers liés à des professions qui travaillent le dimanche (les commerçants par exemple) et simplifie grandement les calculs : c'est donc cette approche qui a été retenue. Afin de déterminer les variables d'intérêt pour modéliser l'absentéisme, on peut, en première approche, regarder comment celui-ci varie en fonction des caractéristiques individuelles. Les résultats sont présentés dans les graphiques ci-dessous.

Le taux d'absentéisme varie en fonction du genre : en moyenne, les femmes sont plus absentes que les hommes. Ce fait a été documenté par Chaupain et Guillot [9], qui l'expliquent notamment par le fait que les femmes subissent davantage de contraintes liées à la conciliation entre charges familiales et activité professionnelle. Parmi les facteurs déterminants, ils notent notamment la présence d'enfants en bas âge ainsi que le montant des revenus du ménage autres que salariés. Contrairement au cadre de l'étude de Chaupain et Guillot, nous ne disposons d'aucune donnée d'enquête sur la satisfaction au travail, le ressenti de pénibilité, la situation familiale, le niveau de revenu etc... Nous nous appuierons donc sur les conclusions de cet article pour les interprétations impliquant des données déclaratives à l'échelle individuelle.

On peut également s'intéresser à l'évolution du taux d'absentéisme en fonction de l'âge. D'après la figure 3.7 , on constate une tendance haussière du taux d'absentéisme avec le vieillissement. Les données concernant les individus de plus de 65 ans ne sont pas prises en compte : on ne dispose en effet que d'un nombre restreint d'individus, ce qui rend la moyenne peu pertinente. La tendance haussière est semblable chez les hommes et les femmes, bien que l'écart se réduise chez les plus de 55 ans. Chaupain et Guillot s'intéressent plutôt à la proportion de salariés absents au moins une fois et constatent que l'âge influe peu chez les hommes tandis que les femmes âgées de 25 à 29 ans

s'absentent nettement plus que les autres. On ne retrouve pas cette particularité dans nos données puisque les courbes des femmes et des hommes sont quasi-parallèles.

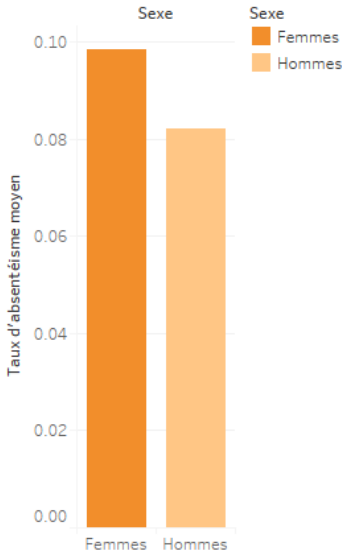


FIGURE 3.6 – Taux d'absentéisme moyen par genre

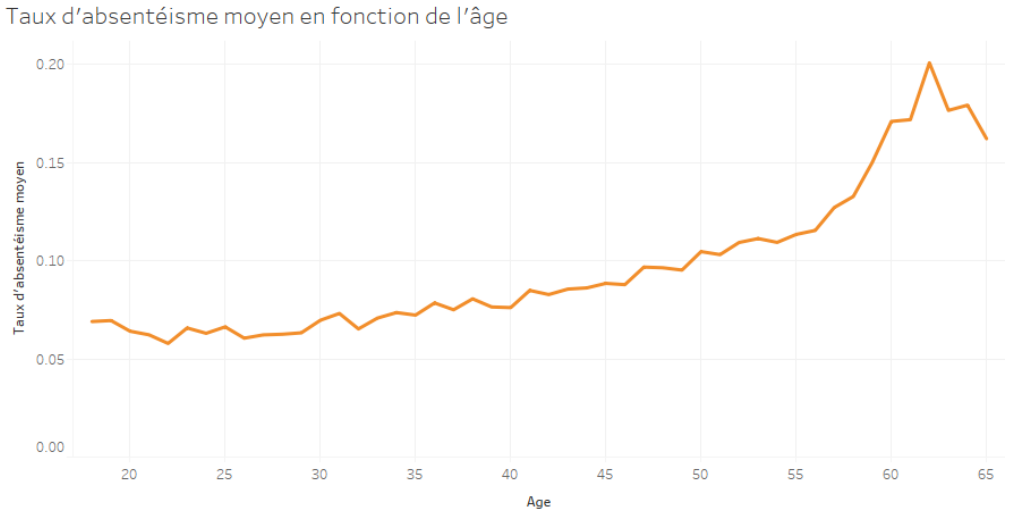


FIGURE 3.7 – Taux d'absentéisme moyen en fonction de l'âge

Le taux moyen d'absentéisme tend à croître avec l'ancienneté, bien que l'effet de l'ancienneté semble nettement moindre que celui de l'âge [3]. Cette remarque pose la question de la corrélation entre l'âge et l'ancienneté : intuitivement, on pourrait penser que ces deux variables sont très corrélées et ont un effet similaire sur l'absentéisme. Or, comme évoqué précédemment, le taux d'absentéisme croît nettement moins en fonction de l'ancienneté que de l'âge. On peut calculer le

3. Figure 3.8

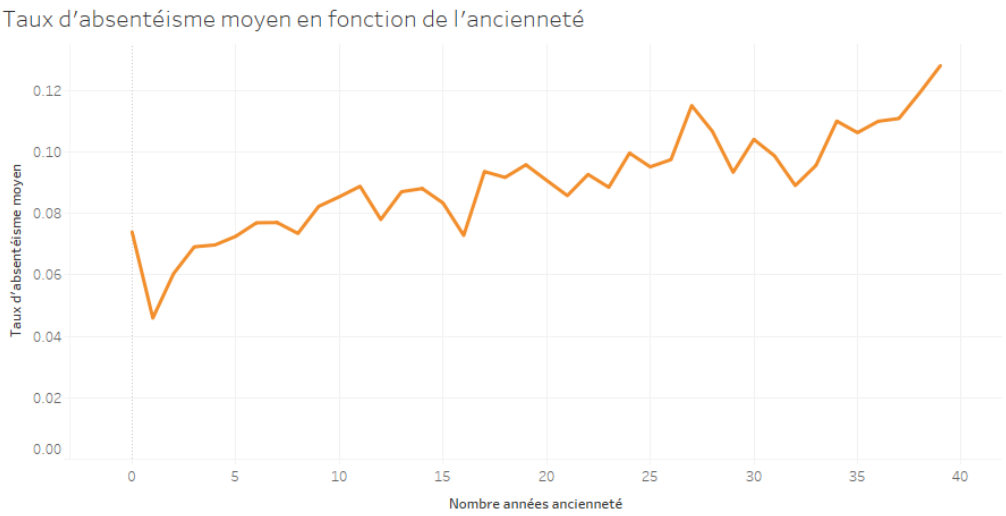


FIGURE 3.8 – Taux moyen d'absentéisme en fonction de l'ancienneté

coefficient de corrélation de Pearson entre l'âge et l'ancienneté, qui est de 0.51. Les deux variables sont donc positivement corrélées bien que le modèle linéaire ne semble pas décrire au mieux la corrélation. Le coefficient de corrélation de Spearman est de 0.42 : là encore, la corrélation est certes positive mais ne justifie pas de traitement particulier.

Le taux d'absentéisme est nettement moins élevé pour les cadres et professions intellectuelles supérieures que pour les autres PCS présents dans notre portefeuille. Les employés présentent le taux d'absentéisme le plus élevé. A une granularité plus fine, les chauffeurs, employés civils et agents de service de la fonction publique, les ouvriers agricoles et assimilés, les employés de commerce et les ouvriers non qualifiés de type artisanal sont les métiers qui présentent le taux d'absentéisme le plus élevé. Les graphiques sont disponibles en annexe. En ce qui concerne les chauffeurs, cela peut s'expliquer par le fait que cette profession est particulièrement exposée aux pathologies liées au fait d'être assis toute la journée. De même, les ouvriers agricoles et les ouvriers non qualifiés de type artisanal sont très exposés aux accidents du travail. Le fait que le taux d'absentéisme soit élevé pour ces professions pourrait donc être structurel. A contrario, les employés civils et agents de service de la fonction publique englobent les personnels des établissements d'enseignement et de santé. On peut donc raisonnablement supposer quele niveau du taux d'absentéisme est, au moins partiellement, conjoncturel et lié au contexte épidémiologique de la crise sanitaire de 2020-2021.

Enfin, le taux d'absentéisme moyen est nettement plus élevé pour les CDIs de droit public que pour les autres natures de contrats. Chaupain et Guillot relèvent eux aussi une propension à l'absentéisme plus importante chez les individus en CDI, notamment chez les femmes, et l'expliquent par la crainte d'un effet de sanction chez les individus en CDD. Le statut conventionnel influe assez

peu sur le taux d'absentéisme moyen. Les ouvriers présentent le taux d'absentéisme moyen le plus élevé, ce qui s'explique une fois encore structurellement par leur exposition plus importante aux accidents du travail et maladies professionnelles.

### 3.1.2 La base PCS

On dispose également d'une base de données mensuelles qui donne, mois par mois, le ou les emplois occupés par chaque individu. L'historique remonte au 1er avril 2016 et les dernières données concernent le mois d'août 2021. Il n'y a cependant aucune donnée entre le mois d'avril et de décembre 2016 ; pour cette raison le mois d'avril 2016 ne sera pas pris en compte dans la suite de l'étude. Finalement, les données utilisées dans cette base couvrent donc la période de décembre 2016 à août 2021. C'est à partir de cette base que le travail de comparaison des PCS a été effectué. Pour la suite de l'étude, on souhaite savoir si les volumes d'individus par PCS sont à peu près constants au cours du temps.

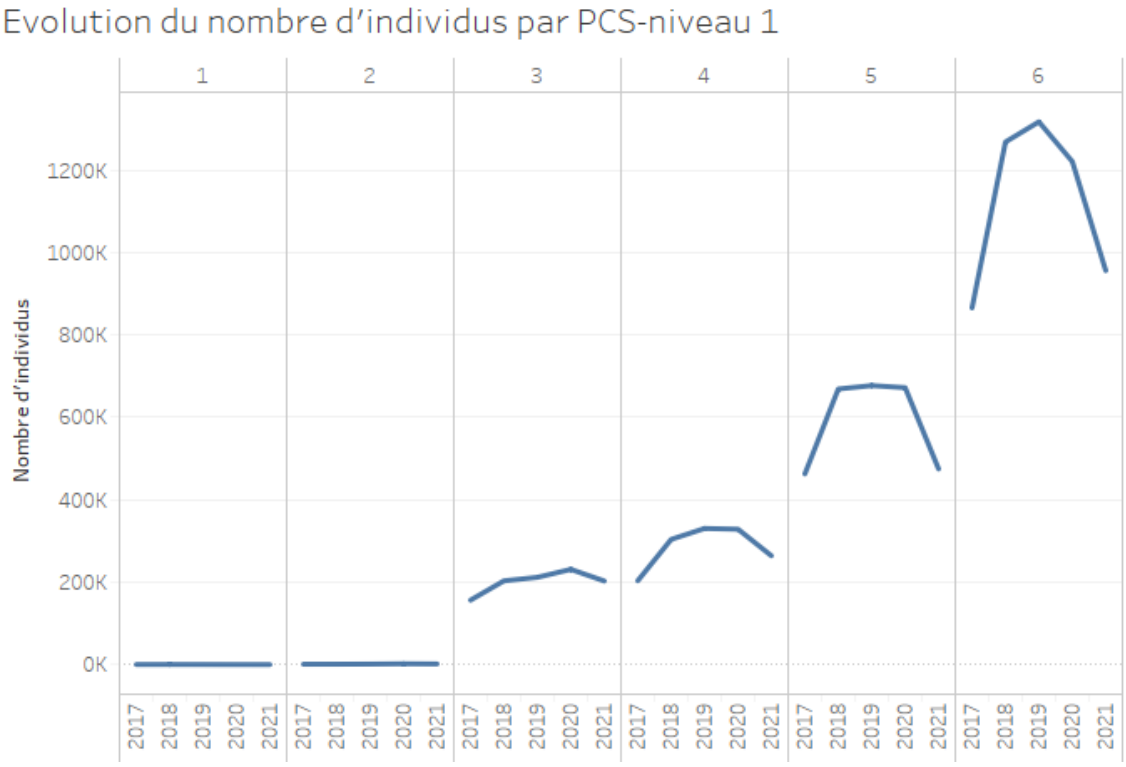


FIGURE 3.9 – Evolution du volume d'individus par PCS au niveau 1

Les volumes d'individus sont constants pour les agriculteurs (PCS 1) et les artisans, commerçants et chefs d'entreprise (PCS 2) mais ces métiers sont quasiment absents de notre portefeuille et cette information n'est donc pas très intéressante. En revanche les volumes sont croissants puis décroissants au fil du temps pour tous les autres PCS, qui sont ceux d'intérêt dans le cadre de

l'étude. Le fait que les volumes soient toujours moins importants en 2021 est cohérent puisque nous n'avons les données que jusqu'au mois d'août pour cette année. Pour tous les PCS à l'exception des ouvriers (PCS 6), les volumes sont croissants entre 2017 et 2018, et quasi constants de 2018 à 2020. L'hypothèse des volumes constants semble donc acceptable pour ces PCS. En revanche, le volume d'ouvriers diminue fortement entre 2019 et 2020 : l'hypothèse de volumes constants est ici compromise [4].

A la granularité inférieure, on retrouve le même genre de constats : les volumes sont quasi-constants pour les PCS peu représentés et connaissent de forte variations pour les PCS les plus représentatifs, avec une tendance à la baisse entre 2019 et 2020 marquée pour les PCS 56 (personnels des services directs aux particuliers), 67 (ouvriers non qualifiés de type industriel) et 68 (ouvriers agricoles et assimilés). On peut suspecter qu'il s'agisse d'une conséquence de la crise sanitaire : ces professions s'exercent souvent sous forme de contrats à durée limitée ou de missions, qui n'ont pas été renouvelés faute de besoin de main d'œuvre durant les différents confinements.

## 3.2 Les données externes : la base *Filosofi*

Les données Filosofi sont issues des déclarations de revenu pour les ménages fiscaux hors collectivités (hôpitaux, maisons de retraites, foyers…) et hors sans-domicile. Il s'agit de données annuelles, concaténant les déclarations de revenus fiscaux, le fichier de la taxe d'habitation, le fichier des personnes physiques et les fichiers sociaux contenant les données sur les prestations sociales versées au niveau de la famille ou de l'individu. Ces données sont par ailleurs complétées par celles de l'enquête Patrimoine de l'Insee, qui permettent de prendre en compte les revenus financiers non soumis à déclaration via un modèle probabiliste sur le montant de détention. Le revenu disponible est calculé à partir de l'impôt sur le revenu, de la taxe d'habitation, de la contribution sociale généralisée (CSG), de la cotisation au remboursement de la dette sociale (CRDS) ainsi que du prélèvement social sur les revenus du patrimoine. Elles permettent ainsi de comparer le revenu déclaré avant redistribution et sur le revenu disponible après redistribution à l'échelle de la commune. Elles contiennent également des indicateurs sur la distribution statistique de ces revenus (déciles, moyenne…) ainsi que sur le taux de pauvreté monétaire (proportion d'individus ayant un niveau de vie inférieur au seuil de pauvreté, fixé à 60% du niveau de vie médian) et sur la structure des revenus, notamment en ce qui concerne leur provenance : revenus salariaux, pensions ou rentes, prestations sociales, part des impôts… Le fichier contient enfin des informations sur la taille moyenne des ménages fiscaux et leur composition.

La base *Filosofi* permet ainsi d'étoffer les données dont nous disposons, en fournissant notamment des informations sur les revenus à un niveau agrégé. Or, le niveau de revenus permet

4. Figure 3.9

de différencier des métiers qui partagent un même PCS mais correspondent à une réalité opérationnelle bien différente. Un conseiller en assurances, un responsable commercial en assurance, un cadre technique et un actuaire relèvent par exemple du même PCS (376e, intitulé « cadres des services techniques et commerciaux des assurances »). Si le cadre de travail est assez similaire, le quotidien professionnel est cependant très différent et le niveau de salaire fournirait un premier élément différenciant. Le but final étant d'effectuer un *clustering*, des variables descriptives liées au lieu d'habitation et issues de données Filosofi de l'INSEE [10] ont donc été ajoutées dans l'espoir de saisir ces caractéristiques.

Cette table a été jointe avec la table issue de la DSN sur la base du code postal de résidence de l'individu. Elle permet ainsi de capter de manière indirecte de l'information sur le niveau de vie. En outre, contrairement à la rémunération individuelle, elle donne de la visibilité sur l'environnement socio-culturel, ce qui pourrait permettre de saisir, en creux, des composantes informelles quant aux comportements individuels.

## 3.3 Préparation des données

Dans les données DSN, on trouve plusieurs tables : *Individus, Contrats, Etablissements, Entreprises et Absences*. La table *Individus* contient les informations concernant les employés de la base : sexe, date de naissance, ancienneté dans l'entreprise… La table *Contrats* est la plus riche et contient les variables liées aux employés par contrat de travail. Chaque contrat de travail est identifié de manière unique par la variable *id_contrat* et chaque ligne correspond à un contrat. Un même individu peut avoir plusieurs contrats. Parmi les informations, on retrouve le type de contrat, le nombre d'heures, la rémunération, la durée du contrat, l'entreprise et l'établissement… La table *Absences* recense les absences par contrat : chaque ligne correspond à une absence. Les variables disponibles sont, entre autres : la date de début de contrat, la date de fin de contrat, le motif d'absence…

Ces deux tables *Contrats* et *Absences* sont agrégées dans une table unique à l'échelle temporelle choisie (année ou mois) : une variable « année » est par exemple créée et, pour chaque absence et chaque contrat, une ligne est créée pour chaque année durant laquelle l'absence ou le contrat existaient. Enfin, cette table est jointe avec la table *Individus* via un identifiant propre à chaque individu.

Quelques subtilités demeurent dans la table finale : une même ligne peut faire référence à un contrat ou à une absence ; on les différencie via l'identifiant rempli : si la variable *Id_absence* est remplie, alors la ligne correspond à une absence et si la variable *Id_contrat* l'est, la ligne correspond à un contrat. En outre, une ligne correspond à une unité temporelle : si les données couvrent les années 2020 et 2021 et qu'un contrat (respectivement une absence) dure durant ces deux années,

ce contrat (ou cette absence) apparaîtra dans deux lignes de la base.

Pour l'usage qui nous intéresse, nous ne conservons que les absences pour cause de maladie. Les variables *âge, ancienneté, nombre d'absences et taux d'absentéisme* sont calculées (on ne dispose initialement que de la date de naissance, de la date de début de contrat). Une variable *ind_absence* est également créée, renseignant si l'individu a ou non été absent. La granularité temporelle choisie est, dans un premier temps, l'année.

# 4 Eléments théoriques

## 4.1 Réduction de dimensionnalité

Lorsque l'on travaille avec des jeux de données contenant un grand nombre de variables, il peut être intéressant d'appliquer des techniques de réduction de dimensionnalité [1]. Cela permet d'une part de réduire drastiquement le temps de calcul, d'autre part de s'affranchir de problèmes structurels liés, par exemple, à la corrélation entre les variables. Enfin, ces techniques permettent de dégager et d'analyser les liens entre les différentes variables, apportant ainsi une plue-value notable pour l'analyse des résultats. La réduction de dimensionnalité est ainsi souvent utilisée comme traitement préliminaire avant l'utilisation de techniques de *machine learning*, notamment dans le cas de *clusterings*.
Une des techniques usuelles de réduction de dimensionnalité est *l'analyse en composantes principales* (ACP).

### 4.1.1 Principe de l'ACP

Sous sa forme initiale, l'ACP permet de traiter des jeux de données dans lesquels chaque ligne correspond à un individu et toutes les variables associées sont quantitatives. On est donc dans le cas d'une matrice $X = (x_{i,j})_{1\leq i\leq n, 1\leq j\leq p} \in \mathscr{M}_{n,p}(\mathbb{R})$ où $n$ désigne le nombre d'individus et $p$ le nombre de variables. L'analyse en composantes principales vise à réduire la dimensionnalité en créant des variables synthétiques, combinaisons linéaires des variables originelles, en nombre réduit $q \ll p$, tout en préservant au mieux l'information.

Si les différentes variables ne sont pas exprimées sur les mêmes échelles, les calculs se trouvent faussés : pour pouvoir comparer les individus, il est donc le plus souvent nécessaire de centrer et de réduire les variables.

1. Toutes les définitions de cette section proviennent du cours *Pratique des méthodes factorielles avec Python [11]*

**Inertie d'un jeu de données**

La quantité d'information contenue dans un jeu de données est généralement mesurée via l'inertie de celui-ci. L'inertie est définie comme la moyenne des carrés des distances entre paires d'observations. Formellement :

$$I = \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} d^2(i,j)$$

où $d$ désigne la distance euclidienne dans $\mathbb{R}^p$.

On peut également voir l'inertie comme la dispersion autour du barycentre du nuage de points. En notant $\mathbf{G}$ ce barycentre, qui est défini comme le vecteur des moyennes sur chacune des composantes ( $\mathbf{G} = (\overline{x_1}, \ldots, \overline{x_p})$ ), l'inertie du jeu de données vaut alors :

$$I = \frac{1}{n} \sum_{i=1}^{n} d^2(i,\mathbf{G})$$

Sous cette forme, il est facile de voir que l'inertie correspond à la somme de la variance des variables :

$$\begin{aligned} I &= \frac{1}{n} \sum_{i=1}^{n} d^2(i,G) \\ &= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{p} (x_{i,j} - \overline{x_j})^2 \\ &= \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} (x_{i,j} - \overline{x_j})^2 \\ &= \sum_{j=1}^{p} \sigma_j \end{aligned}$$

Dans le cas où les variables ont été réduites, la variance de chacune des variables vaut 1 et l'inertie est donc égale au nombre de variables : $I = p$.

**L'ACP en pratique**

Si on note $q$ la dimension du nouvel espace, on cherche $q$ variables synthétiques, appelées les *composantes principales*. Ces variables sont définies par des combinaisons linéaires des variables :

$$\text{pour } 1 \leq k \leq q,\ F_k = \alpha_{k1} x_1 + \alpha_{k2} x_2 + \cdots + \alpha_{kp} x_p$$

Le but est alors d'estimer les $(\alpha_{kj})_{k,j}$. Pour cela, on procède de manière récursive : la première

composante est construite de manière à maximiser la dispersion des projetés des points. Elle est ensuite centrée. La variance associée correspond au pouvoir explicatif de cet axe et il s'agit d'une fraction de l'inertie totale. On construit ensuite la seconde composante de la même manière, mais sur l'information résiduelle non expliquée par la première composante. Cela se traduit par le fait que les deux composantes sont orthogonales. On procède de même pour les $q$ composantes et on obtient un repère orthogonal de $\mathbb{R}^q$.

Un autre point de vue sur cette construction consiste à construire le premier axe factoriel de sorte qu'il maximise la somme des carrés des corrélations avec les variables. On construit ensuite la deuxième composante de la même manière, en travaillant sur la partie résiduelle non-expliquée par la première composante (c'est-à-dire orthogonale à la première), et ainsi de suite.

Toute l'information disponible est restituée lorsque $q = p$. Plus un nombre faible $q$ de facteurs permet de restituer une grande part de l'information, meilleure est la compression.

### 4.1.2 Choix du nombre $q$ de composantes à retenir

Le choix du nombre $q$ de facteurs utilisés lors de la représentation influe sur la qualité de celle-ci et conditionne donc la pertinence des analyses effectuées ensuite.

Pour choisir le nombre de facteurs pertinents, plusieurs règles existent [12] :

— la règle de Kaiser-Guttman, qui consiste à ne conserver que les facteurs associés à des valeurs propres strictement supérieures à 1 (dans le cas d'une ACP normée).

— la règle de Cattell, consistant à identifier les cassures (les *coudes*) dans la décroissance des valeurs propres via le diagramme des *scree plot*.

— l'étude du diagramme de la part variance expliquée en fonction du nombre de facteurs, qui permet d'identifier le seuil à partir duquel l'apport des facteurs restant devient négligeable devant celui des facteurs considérés.

— la règle de Karlis-Saporta-Spinaki : il s'agit d'une variante à la règle de Kaiser-Guttman, souvent jugée trop permissive. La règle d'acceptation d'une valeur propre $\lambda$ devient : $\lambda > 1+2\sqrt{\frac{p-1}{n-1}}$ où $p$ désigne le nombre de variables et $n$ le nombre d'individus. Le seuil est ainsi défini comme la moyenne des valeurs propres, auquel on ajoute deux fois leur écart-type.

— le test des *bâtons brisés* : ce test, dû à Frontier (1976) et Legendre-Legendre (1983) repose sur l'idée que, si l'inertie totale était répartie aléatoirement sur les axes, sa distribution suivrait une *loi des bâtons brisés*. La $k$-ème valeur propre est alors acceptable au seuil de 5% si elle est supérieure au seuil $b_k$ défini par : $b_k = \sum_{i=k}^{p} \frac{1}{i}$.

### 4.1.3 Formalisation du problème

On se place ici sous le prisme de la maximisation de la somme des carrés des corrélations avec les variables. On suppose que les $p$ variables $z_1, \ldots, z_p$ ont été centrées et réduites. On cherche alors à construire $F_1, \ldots, F_q$ tels que :

$$\begin{aligned} F_1 &= a_{11}z_1 + a_{21}z_2 + \ldots a_{p1}z_p \\ F_2 &= a_{12}z_1 + a_{22}z_2 + \ldots a_{p2}z_p \\ &\vdots \\ F_q &= a_{1q}z_1 + a_{2q}z_2 + \ldots a_{pq}z_p \end{aligned}$$

où le degré de représentativité du facteur $F_k$ est modélisé par la variance expliquée $\lambda_k$ et les facteurs sont orthogonaux deux à deux et classés par ordre décroissant d'importance. On cherche à déterminer les coefficients $(a_{ij})_{i,j}$.

Pour le premier facteur, notons $a = (a_1, \ldots, a_p)$ le vecteur des coefficients à déterminer et $R$ la matrice de covariance des variables. On cherche à résoudre le problème d'optimisation suivant :

$$\begin{cases} \max\limits_{a} \ a^T R a \\ \text{sc } a^T a = 1 \end{cases}$$

Le lagrangien associé est : $\mathscr{L} = a^T R a - \lambda(a^T a - 1)$. En dérivant par rapport à $a$, on obtient : $Ra = \lambda a$, c'est-à-dire que $a$ est vecteur propre de $R$ pour la valeur propre $\lambda$.

Ainsi, le vecteur $a_k = (a_{1k}, a_{2k}, \ldots, a_{pk})$ associés au facteur $F_k$ correspondent au vecteur propre de norme 1 associé à la $k$-ème plus grande valeur propre de la matrice $R$. Il est donc pertinent de diagonaliser la matrice $R$.

### 4.1.4 Coordonnées des individus dans l'espace factoriel

Pour calculer les coordonées des individus dans le plan factoriel, on applique les coefficients estimés sur les variables centrées et réduites. En conservant les notations précédentes et en notant $F_{ik}$ la coordonée de l'individu $i$ sur l'axe $k$, on a ainsi :

$$F_{ik} = \sum_{j=1}^{p} a_{jk} z_{ij}$$

où $z_{ij}$ désigne la valeur de la variable $z_j$ pour l'individu $i$.

A partir de ces coordonnées, on peut calculer la contribution de chaque individu à un facteur.

Cela permet notamment de comparer l'importance relative des individus dans la construction des facteurs et d'identifier d'éventuels problèmes (*outliers*...).

La contribution de l'individu $i$ au facteur $k$, notée $CTR_{ik}$ vaut :

$$CTR_{ik} = \frac{F_{ik}^2}{n \times \lambda_k}$$

### 4.1.5 Qualité de représentation des variables

La qualité de la représentation de la variable $j$ sur la composante $k$, notée $Q_{jk}$ correspond au carré du coefficient de corrélation $Q_{jk}^2 = r_j^2(F_k)$. Du fait de l'orthogonalité des facteurs, elle est additive. En sommant sur tous les facteurs, on retrouve alors l'information totale portée par la variable.

## 4.2 Clustering

Le *clustering* est une technique de classification relevant de l'apprentissage non-supervisé, c'est-à-dire que les données d'entrée ne sont pas labellisées. Le *clustering* regroupe alors les données similaires en *clusters*. Les principales applications du clustering sont :

— la connaissance client : on peut segmenter la base client selon les caractéristiques pertinentes et ainsi mieux cibler les besoins et adapter les produits ou les campagnes marketing par exemple (c'est le principe des systèmes de recommandation).

— l'analyse de données : quand on débute l'analyse d'un nouveau jeu de données, il peut être intéressant d'identifier les éventuels *clusters* et, éventuellement, de les étudier séparément.

— la mesure de l'affinité de chaque point avec chaque *cluster*. S'il y a $k$ *clusters*, on peut alors remplacer une ligne par la donnée de ses $k$ affinités avec chacun des *clusters* : un vecteur de dimension n est alors remplacé par un vecteur de dimension $k < n$.

— la détection d'anomalie : un point qui n'a que peu d'affinité avec tous les *clusters* est plausiblement un *outlier*.

— l'augmentation de données dans un cadre d'apprentissage semi-supervisé : si on ne dispose que d'un petit nombre de points labellisés, on peut étendre les labels aux points d'un même *cluster*, ce qui augmente le nombre de labels disponibles pour un algorithme d'apprentissage supervisé et améliore ses performances.

### 4.2.1 Principaux algorithmes de *clustering*

Le principe d'un algorithme de *clustering* est ainsi de segmenter les données en *clusters* : il faut donc définir ce qu'est un *cluster*. Cette définition dépend en réalité du contexte et de l'usage visé. Parmi les algorithmes les plus populaires, on trouve le K-Means [13] et le DBSCAN [14].

#### K-Means

Soit un ensemble d'observations $(x_1, \ldots, x_n)$ où chaque observation est de dimension $d$. Le *clustering* K-Means vise à segmenter les $n$ observations en ensembles $S = \{S_1, \ldots, S_k\}$ de sorte à minimiser la variance intra-cluster. Formellement, on cherche :

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x \in S_i} ||\mathbf{x} - \mu_\mathbf{i}||^2 = \underset{S}{\operatorname{argmin}} \sum_{i=1}^{k} |S_i| \mathrm{Var}(S_i)$$

où $\mu_\mathbf{i}$ est la moyenne des points de $S_i$.

Le problème peut également être formulé sous la forme du programme d'optimisation suivant[15][16] : minimiser

$$P(W, Q) = \sum_{l=1}^{k} \sum_{i=1}^{n} w_{i,l} d(X_i, Q_i)$$

sous contrainte

$$\begin{cases} \sum_{l=1}^{k} w_{i,l} = 1 \ \text{ pour tout } 1 \leq i \leq n \\ 0 \leq w_{i,l} \leq 1 \ \text{ pour } 1 \leq i \leq n \ \text{ et } 1 \leq l \leq k \end{cases}$$

où $W$ est une $n \times k$ matrice de partition et $Q = \{Q_1, \ldots, Q_k\}$ un ensemble d'objets du même domaine et $d(\bullet, \bullet)$ la distance euclidienne entre deux objets.

Ici, le nombre de *clusters* est fixé a priori. La qualité du résultat dépend donc de ce paramètre. Pour le choisir, une méthode courante est la méthode dite du coude. Elle consiste à représenter graphiquement l'inertie [2] en fonction du nombre de *clusters*. Cette fonction est strictement décroissante (plus il y a de *clusters*, plus la distance entre un point et son centroïde le plus proche sera faible et donc plus faible sera l'inertie) et présente généralement un coude, que l'on considère comme étant le nombre de *clusters* optimal.

Une autre méthode pour obtenir le nombre de clusters optimal est d'utiliser le coefficient de silhouette moyen, qui est défini par :

$$S = \frac{1}{n} \sum_{i=1}^{n} s(x_i)$$

2. L'inertie est définie comme la somme des distances euclidiennes entre chaque point et son centroïde associé. Un centroïde correspond quant à lui à un $\mu_\mathbf{i}$ dans la définition du programme d'optimisation.

avec

$$\begin{cases} s(x_i) = \frac{b_i - a_i}{\max(b_i, a_i)} \\ b_i \quad \text{la distance moyenne aux points du cluster le plus proche} \\ a_i \quad \text{la distance moyenne aux autres points du même cluster} \end{cases}$$

Le coefficient de silhouette est compris entre -1 et 1 : un coefficient proche de 1 signifie que le point est très centré dans son cluster, un coefficient proche de 0 signifie que le point est à la frontière de son cluster tandis qu'un coefficient proche de -1 signifie que le point est associé au mauvais cluster. Au vu de ces interprétations, le nombre de clusters optimal est donc celui qui maximise le coefficient de silhouette moyen.

#### DBSCAN

Cet algorithme définit un *cluster* comme une région à haute densité de points. Formellement, on fixe une petite distance $\varepsilon$ et un nombre de points minimum **min_points**. On compte pour chaque point le nombre de points présents dans un $\varepsilon$-voisinage (c'est-à-dire dans une boule de centre l'instance et de rayon $\varepsilon$ pour la norme euclidienne). Si un point a plus de **min_points** points dans son $\varepsilon$-voisinage, il est alors considéré comme central. Tous les points situés dans le $\varepsilon$-voisinage d'un point central appartiennent alors au même cluster. En suivant cette méthodologie, le jeu de données est séparé en trois types de points :

— les points centraux

— les points frontières, qui sont dans le $\varepsilon$-voisinage d'un point central mais ne sont pas centraux

— les points aberrants, qui ne sont dans le $\varepsilon$-voisinage d'aucun point central

## 4.3 Réduction de dimensionalité dans le cas de données catégorielles ou mixtes

Comme on peut le voir dans la section précédente, l'ACP et le *clustering* sont des problèmes essentiellement métriques. L'ACP repose en effet sur une représentation dans un espace euclidien de dimension réduite sur fond de maximisation de variance tandis que le *clustering* identifie des *clusters* au sein de données via un programme d'optimisation de la dispersion des *clusters*. Se pose alors la question des variables catégorielles : comment prendre en compte dans une ACP ou un *clustering* des variables qui ne sont ni numériques, ni même ordonnées ?

### 4.3.1 Analogue de l'ACP : l'AFC

L'analyse factorielle des correspondances (AFC) adapte les idées de l'ACP au cas d'un tableau croisé de variables qualitatives. Les données se présentent donc sous la forme d'un tableau, dans lequel chaque ligne correspond à une modalité de la première variable catégorielle et chaque colonne représente une modalité de la seconde variable. Une case $(i, j)$ contient alors le nombre d'individus pour lesquels la première variable prend la modalité $i$ et la deuxième la modalité $j$, noté $n_{ij}$. Une modalité $i$ de la première variable est appelée un *profil*.

Considérons que la première variable (celle représentée en ligne donc) ait $N$ modalités et la seconde $L$ modalités. La distance entre deux profils $k$ et $k'$ est donnée par une distance du $\chi_2$ :

$$d^2(k,k') = \sum_{l=1}^{L} \frac{1}{\frac{n_{.l}}{n}} \left( \frac{n_{kl}}{n_{k.}} - \frac{n_{k'l}}{n_{k'.}} \right)^2$$

où $n_{k.}$ désigne le nombre d'individus pour lesquels la première variable prend la modalité $k$.

#### Distance à l'origine

Dans ce cadre, la distance à l'origine d'un profil se calcule comme la distance au profil moyen, c'est-à-dire, pour le profil $k$ :

$$d^2(k) = \sum_{l=1}^{L} \frac{1}{\frac{n_{.l}}{n}} \left( \frac{n_{kl}}{n_{k.}} - \frac{n_{.l}}{n} \right)^2$$

#### Inertie

L'inertie traduit la quantité d'information portée par une modalité. Pour une modalité $k$, on a alors :

$$I(k) = \mathbb{P}(k) \times d^2(k)$$

où $\mathbb{P}(k)$ est le poids de la modalité, *ie* $\mathbb{P}(k) = \frac{n_{k.}}{n}$.
L'inertie totale correspond à la somme des inerties associées à chaque modalité.

#### Calcul des facteurs

Comme l'ACP, l'AFC cherche à déterminer des facteurs, ou plus exactement des axes factoriels obtenus comme combinaisons linéaires des profils, qui permettent de maximiser la dispersion des profils. Ils sont, de même que dans l'ACP, orthogonaux deux à deux.

Le premier facteur cherche ainsi à maximiser la quantité $\lambda_1$ suivante :

$$\lambda_1 = \sum_{k=1}^{N} \frac{n_{k.}}{n} \times F_{k1}^2$$

où :

— $F_{k1}$ est la coordonnée de la $k$ème modalité sur le premier facteur

— $\sum_{k=1}^{N} F_{k1} = 0$

— $\lambda_1$ représente la variance des points modalités

Pour restituer la distance entre profils, on substitue la distance euclidienne dans l'espace factoriel à la distance du $\chi^2$ entre les modalités lignes dans l'espace initial.

**Remarque :** la même étude peut être effectuée sur les profils considérés cette fois en colonne et non plus en ligne.

### 4.3.2 Analyse des correspondances multiples (ACM)

#### Généralités

L'analyse des correspondances multiples est l'analogue de l'ACP pour des variables catégorielles. Le principe est le même ; la distance entre individus considérée est la distance du $\chi^2$, qui permet de mettre en exergue les différences entre modalités rares. La distance entre individu $i$ et un individu $j$ est ainsi donnée par :

$$d^2(i,j) = \sum_{k=1}^{M} \frac{1}{\frac{n_k}{n \times p}} \left( \frac{x_{ik}}{p} - \frac{x_{jk}}{p} \right)^2$$

avec $p$ le nombre de variables, $n$ le nombre d'individus et $M$ le nombre total de modalités prises par les variables et $x_{ik}$ la modalité prise par la variable $k$ pour l'individu $i$. La distance à l'origine est là aussi définie comme la distance à l'individu moyen et l'inertie vaut encore :

$$I = \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{i \neq j} d^2(i,j)$$

Le nombre maximal de facteurs vaut $H_{\max} = M - p$. L'heuristique pour abaisser la dimensionnalité est la même que dans l'AFC.

Il est toutefois nécessaire d'accorder un point d'attention aux modalités des variables : en effet, des variables ayant un très grand nombre de modalités ou bien des modalités très rares peuvent fausser les calculs.

#### Distances entre modalités

La distance entre modalités est là encore définie par la distance du $\chi^2$ :

$$d^2(k,k') = \sum_{i=1}^{n} \frac{1}{\frac{1}{n}} \left( \frac{x_{ik}}{n_k} - \frac{x_{ik'}}{n'_k} \right)^2$$

avec $x_{ik}$ le nombre d'individus du profil $i$ prenant la modalité $k$ et $n_k$ le nombre total d'individus prenant la modalité $k$.

La distance à l'origine est définie comme la distance du $\chi^2$ à l'individu moyen :

$$d^2(k) = \sum_{i=1}^{n} \frac{1}{\frac{1}{n}} \left( \frac{x_{ik}}{n_k} - \frac{1}{n} \right)^2$$

**Inertie**

L'inertie d'une modalité alors exprimée comme le produit entre son poids relatif ($\omega_k = \frac{n_k}{n \times p}$ ) et sa distance à l'origine :

$$I(k) = \omega_k \times d^2(k)$$

L'inertie totale correspond quant à elle à la somme des inerties de chacune des modalités :

$$I = \sum_{k=1}^{M} I(k)$$

**Remarque :** Avec cette formule, on retrouve la valeur de l'inertie calculée à partir des distances entre individus.

**Construction des facteurs**

Pour construire les facteurs, on cherche là encore à maximiser l'écart des carrés à l'origine. Le premier facteur est ainsi construit de sorte à maximiser :

$$\lambda_1 = \sum_{k=1}^{M} \omega_k \times G_{k1}^2$$

où $G_{kh}$ est la coordonnée de la modalité $k$ sur le facteur $h$, $\lambda_h$ la variance associée au facteur $h$ et $\omega_k$ le poids relatif de la modalité $k$.

De même que précédemment, une fois le premier facteur construit, on construit le second à partir de la partie résiduelle par le premier facteur et ainsi de suite. La décomposition est là encore orthogonale et les valeurs propres s'additionnent donc pour obtenir la variance totale expliquée.

#### Qualité de la représentation

De même qu'en ACP, on peut calculer la qualité de représentation d'un point modalité sur un facteur. En notant $Q_h(k)$ la qualité de représentation du point modalité $k$ dans le facteur $h$, on a :

$$Q_h^2(k) = \frac{F_{kh}^2}{d^2(k)}$$

### 4.3.3 Analyse factorielle des données mixtes (AFDM)

#### Généralités

Lorsque les données sont mixtes, c'est-à-dire qu'elles contiennent à la fois des données qualitatives et quantitatives, on ne peut appliquer ni l'ACP ni l'ACM tel quel. On peut s'y ramener par des artifices (segmenter les valeurs continues puis effectuer une ACM par exemple), mais ce n'est pas toujours adapté. Une autre solution consiste à mettre en oeuvre une AFDM.

Cette méthode a l'avantage de redonner les résultats de l'ACP lorsque toutes les variables sont quantitatives et ceux de l'ACM lorsqu'elles sont toutes qualitatitves. Elle consiste à appliquer un encodage *One-Hot* amélioré sur les variables qualitatives, puis à utiliser une ACP. L'information contenue dans les variables n'est ainsi pas détériorée.

L'encodage des variables qualitatives s'effectue en deux étapes. Dans une premier temps, un encodage *One-Hot* classique est appliqué. Les fréquences relatives de chaque modalités sont ensuite calculées et les indicatrices obtenues à l'issue de l'encodage *One-hot* sont normalisées par ces fréquences relatives. Les données quantitatives sont ensuite standardisées. A l'issue de ces trois étapes, on obtient la base sur laquelle l'ACP est appliquée.

#### Formalisation

On se place dans le cas où l'on dispose de $n$ observations mêlant $C$ variables quantitatives et $D$ variables qualitatives. Les variables sont notées $\left(X_j\right)_{1\leq j\leq P}$, avec $P = C + D$. On dispose ainsi d'une matrice $(x_{ik})_{ik}$ avec $1 \leq i \leq n$ et $1 \leq k \leq P$.

Lors de la construction du premier facteur, on cherche alors à maximiser :

$$\lambda_1 = \sum_{j=1}^{C} r^2\left(F_1, X_j\right) + \sum_{j=C+1}^{C+D} \eta^2\left(F_1, X_j\right)$$

où $r^2(\bullet)$ désigne le carré du coefficient de corrélation linéaire et $\eta^2(\bullet)$ le carré du rapport de corrélation.
Le nombre maximal de facteur est $H_{max} = P - D$ et est égal à l'inertie totale.

**Analyse des résultats**

L'interprétation des résultats doit être menée avec prudence : les calculs ne sont pas les mêmes suivant que l'on s'intérese à une variable catégorielle ou continue.

Dans le cas d'une variable quantitative, la contribution de la variable $j$ au facteur $h$ vaut :

$$CTR_j(F_h) = \frac{r^2(F_h, X_j)}{\lambda_h}$$

Dans le cas d'une variable qualitative, elle vaut :

$$CTR_j(F_h) = \frac{\eta^2(F_h, X_j)}{\lambda_h}$$

De même, le calcul de la qualité de représentation d'une variable dépend de sa nature : elle vaut $r^2(F_h, X_j)$ dans le cas d'une variable quantitative et $\frac{\eta^2(F_h, X_j)}{m_j - 1}$ dans le cas d'une variable qualitative, où $m_j$ désigne le nombre de modalités prises par la variable $j$.

## 4.4 Clustering dans le cas de variables qualitatives ou mixtes

Une première solution consiste à utiliser un encodage *One-Hot* : pour une variable catégorielle $A$ ayant $k$ modalités, on ajoute artificiellement $k-1$ variables $A_1, \ldots, A_{k-1}$. La variable $A_i$ vaudra alors 1 si la variable $A$ prenait la modalité $i$ et 0 sinon. On transforme ainsi une variable catégorielle en variable numérique, qui est donc comparable avec les autres variables numériques. Cette approche présente néanmoins plusieurs problèmes. D'une part, elle ajoute un nombre possiblement élevé de variables, ce qui complexifie le traitement informatique. D'autre part, ce traitement suppose que toutes les modalités sont équidistantes les unes des autres, ce qui ne reflète pas nécessairement la réalité. Une autre solution a été apportée par Huang [17] , qui consiste à adapter l'algorithme des $K$-Means au cas d'un jeu de donnée mixtes via les modifications suivantes :

— utiliser une mesure de dissimilarité simple pour les variables catégorielles

— remplacer la moyenne des *clusters* par le mode

— utiliser une méthode fondée sur la fréquence pour trouver les modes qui résolvent le problème d'optimisation

### 4.4.1 Mesure de dissimilarité

La mesure de dissimilarité utilisée est la suivante : soient $X, Y$ deux objets catégoriels décrits par $m$ attributs : $X = (x_1, \ldots, x_m)$ et $Y = (y_1, \ldots, y_m)$. On définit la mesure de dissimilarité entre $X$ et $Y$ par :

$$d_1(X,Y) = \sum_{j=1}^{m} \delta(x_j, y_j)$$

où

$$\delta(x_j, y_j) = \begin{cases} 0 \text{ si } x_j = y_j \\ 1 \text{ si } x_j \neq y_j \end{cases}$$

### 4.4.2 Mode d'un ensemble

Soit $X$ un ensemble d'objets catégoriels décrits par leurs attributs catégoriels $A_1, \ldots A_m$. Ainsi, $X = \{X_1, \ldots, X_n\}$ et chacun des $X_i$ est de la forme $(x_1, \ldots, x_m)$ où $x_i$ est la valeur prise par l'attribut $A_i$. Un mode de $X$ est un vecteur $Q = [q_1, \ldots, q_m]$ qui minimise :

$$D(X,Q) = \sum_{i=1}^{n} d_1(X_i, Q)$$

**Remarque** : $Q$ n'est pas nécessairement un élément de $X$ !

Soit alors $\eta_{c_{k,j}}$ le nombre d'objets ayant la $k$ème catégorie dans l'attribut $A_j$ et $f_r(A_j = c_{k,j}|X) = \frac{\eta_{c_{k,j}}}{n}$ la fréquence relative de $c_{k,j}$ dans $X$.

**Théorème :** *La fonction $D(X,Q)$ est minimisée si, et seulement si, $f_r(A_j = q_j|X) \geq f_r(A_j = c_{k,j}|X)$ pour $q_j \neq c_{k,j}$ et pour tout $1 \leq j \leq m$.*

Ce théorème fournit un moyen de déterminer $Q$ pour un ensemble $X$ donné et permet ainsi d'utiliser la méthode des $K$-means pour des données catégorielles. Il implique également que le mode d'un jeu de données n'est pas unique.

### 4.4.3 Algorithme des $K$-modes

En utilisant la mesure de dissimilarité définie précédemment, la fonction de coût devient :

$$P(W,Q) = \sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m} w_{i,l}\delta(x_{i,j},q_{l,j})$$

où $w_{i,l} \in W$ et $Q_l = [q_{l,1}, q_{l,2}, \ldots, q_{l,m}] \in Q$.
On peut alors adapter l'algorithme des $K$-means pour minimiser la fonction de coût en utilisant la mesure de dissimilarité.

### 4.4.4 Algorithme des $K$-prototypes

Cet algorithme rassemble les résultats des $K$-means et des $K$-modes et permet de traiter des jeux de données mêlant données numériques et catégorielles.

Soit un jeu de données mixtes décrit par les attributs $A_1^r, A_2^r, \ldots, A_p^r, A_{p+1}^c, \ldots, A_m^c$ où les indices $r$ désignent des données numériques et $c$ des données catégorielles. La dissimilarité entre deux objets $X$ et $Y$ de ce jeu de données s'exprime alors comme :

$$d_2(X,Y) = \sum_{j=1}^{p}(x_j - y_j)^2 + \gamma\sum_{j=p+1}^{m}\delta(x_j,y_j)$$

Le terme $\gamma$ permet de rétablir l'équilibre entre les deux types d'attributs.
Le programme d'optimisation devient alors :

$$P(W,Q) = \sum_{l=1}^{k}\Big(\sum_{i=1}^{n} w_{i,l}\sum_{j=1}^{p}(x_{i,j} - q_{l,j})^2 + \gamma\sum_{i=1}^{n} w_{i,l}\sum_{j=p+1}^{m}\delta(x_{i,j},q_{l,j})\Big)$$

Soit alors $P_l^c = \gamma\sum_{i=1}^{n} w_{i,l}\sum_{j=p+1}^{m}\delta(x_{i,j},q_{l,j})$ et $P_l^r = \sum_{l=1}^{k}\Big(\sum_{i=1}^{n} w_{i,l}\sum_{j=1}^{p}(x_{i,j} - q_{l,j})^2$. Le problème se réécrit comme :

$$P(W,Q) = \sum_{l=1}^{k}(P_l^r + P_l^c)$$

Comme $P_l^r$ et $P_l^c$ sont sont tous les deux positifs ou nuls, minimiser $P(W,Q)$ revient à minimiser $P_l^r$ et $P_l^c$ séparément pour tout $1 \leq l \leq k$, ce qui peut être fait en utilisant ce qui a été évoqué dans les sections précédentes.

### 4.4.5 Mesure de similarité

Une mesure de similarité est initialement une mesure entre chaînes de caractères permettant de mesurer par un nombre la différence entre deux mots. Plus généralement, une mesure de similarité permet d'obtenir une quantification de la distance entre deux ensembles sans aucune notion d'espace métrique sous-jacente. Parmi les mesures de similarités classiques, on retrouve :

— La mesure de Hamming [18], qui décompte le nombre de lettres différentes entre les deux mots

— La mesure de Levenshtein [19], qui décompte le nombre de modifications (ajout, retrait ou changement) nécessaires pour passer d'une chaîne à une autre

Ainsi que plusieurs mesures, qui s'expriment pour des ensembles X et Y :

— Le coefficient de Dice [20] [21] : $Dice(X,Y) = \frac{2|X \cap Y|}{|X|+|Y|}$

— L'indice de Jaccard [22] : $Jacc(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$

Etc....

Une méthode de comparaison de ces distances a été proposée dans l'article *Selection of proximity measures for a Topological Correspondence Analysis* de Rafik Abdesselam [23].

## 4.5 Eléments théoriques sur les distances

### 4.5.1 Distance sur un graphe

Formellement [3], un graphe se définit comme la donnée d'un couple $G = (E,V)$ où $E \subset [V]^2$. Les éléments de $E$ sont ainsi des sous-ensembles à deux éléments de $V$. $V$ est l'ensemble des sommets et $E$ l'ensemble des arêtes.

Le nombre de sommets d'un graphe est appelé *l'ordre* et on le note $|G|$. Il s'agit du cardinal de $V$. Un graphe est dit fini si son ordre est fini et infini sinon.

Pour deux sommets $x, y \in, V$, on note $xy$ l'arête correspondante. Deux sommets $x, y \in V$ sont dits *adjacents* ou *voisins* si $xy \in E$. . Deux arêtes sont dites adjacentes si elles sont une extrémité en commun. Un graphe est dit *complet* si tous ses sommets sont deux à deux adjacents. Le *degré* d'un sommet $x \in V$ est le nombre de voisins de $x$ dans $G$.

Un graphe peut avoir plusieurs arêtes joignant les mêmes sommets : on parlera dans ce cas d'arêtes *multiples*. Il peut également avoir des *boucles*, c'est-à-dire des arêtes dont les deux extrémités correspondent au même sommet. Un graphe est dit *simple* s'il ne possède ni boucles ni arêtes

3. Toutes les définitions et figures de cette section proviennent du livre Graph Theory de Diestel R, Springer, vol 173, 2000 [24]

multiples. On se restreindra à ce cas dans toute la suite.
Un graphe est dit *connexe* si :

$$\forall x \neq y \in V,\ \exists n \in \mathbb{N},\ \exists x_1, \ldots, x_n \in V \quad \text{tels que} \quad x_1x_2,\ x_2x_3, \ldots\ x_ny \in E$$

Concrètement, cela signifie que l'on peut toujours aller d'un sommet à un autre en suivant des arrêtes.

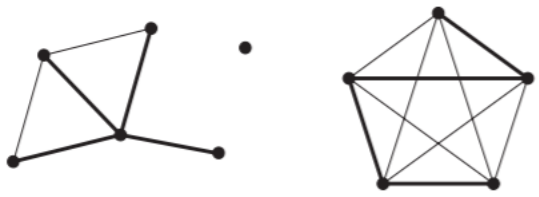

FIGURE 4.1 – Exemples de graphes : celui de gauche est non connexe tandis que celui de droite est connexe (et même complet)

Un *cycle* est un ensemble de sommets reliés par des arêtes qui bouclent, c'est-à-dire un ensemble $x_1, \ldots, x_n \in V$ tels que $x_1x_2,\ x_2x_3,\ \ldots x_nx_1\ \in E$. La *longueur* d'une chemin entre $x$ et $y$ est le nombre d'arêtes qui le composent. Pour un graphe donné, on peut définir la distance entre deux sommets $x$ et $y$ : il s'agit de la longueur d'un plus court chemin les joignant. Cette distance n'est pas toujours finie : par convention, s'il n'existe pas de chemin entre $x$ et $y$ , la distance entre eux est infinie. En particulier, dans le cas fini, un graphe est connexe si, et seulement si, tous les sommets sont à distance finie les uns des autres.
Un *arbre* est un graphe *acyclique* et *connexe*. Une *forêt* est un graphe composé de plusieurs arbres. Pour un graphe $T$, se valent alors :

(i) $T$ est un arbre

(ii) Deux sommets de $T$ sont toujours liés par un unique chemin dans $T$

(iii) $T$ est minimal connexe, c'est-à-dire que $T$ est connexe mais que $T \setminus e$ est non-connexe pour toute arête $e$.

(iv) $T$ est maximal acyclique, c'est-à-dire que $T$ ne contient pas de cycle mais que $T + xy$ en contient pour toute paire de sommets $x, y \in T$ non adjacents dans $T$

### 4.5.2 Modèle vectoriel et distance TF-IDF

#### Heuristique

On considère un corpus, constitué de $m$ documents : $\mathscr{C} = \{d_1, \ldots, d_m\}$ Chaque texte est lui-même rédigé à partir d'un dictionnaire, c'est-à-dire d'un ensemble de $N$ mots fixé : $\mathscr{D} = \{t_1, \ldots, t_N\}$.

L'idée est de comparer les textes à partir des mots qui les composent, en considérant que l'importance d'un terme au sein d'un document dépend de deux facteurs : sa spécificité et son exhaustivité [25]. L'exhaustivité de la description d'un document se définit alors comme le nombre de termes qu'il contient tandis que la spécificité d'un terme est représentée par le nombre de textes dans lequel il apparaît

On suppose alors que plus un terme est spécifique, moins il a de chance d'apparaître et plus un document est exhaustif, moins la présence d'un terme dans ce texte est significative. Pour cela, on définit le poids du terme $j$ dans le document $i$ par :

$$\text{tf-idf}_{i,j} = \frac{\rho_{j,i}}{\sum_i \rho_{j,i}} \times \log\Big(\frac{|C|}{|\{l \in [\![1,m]\!], t_j \in d_l|}\Big)$$

où $\rho_{i,j}$ désigne le nombre de fois où le terme $j$ apparaît dans le texte $i$.

On peut ensuite utiliser le modèle vectoriel[26]. Cette approche consiste à représenter chaque texte par le vecteur des poids de chaque terme au sein de ce dernier. L'exemple le plus simple de pondération consiste à associer à un terme i son nombre d'occurrence au sein du texte. Dans notre cas, on identifie le texte $i$ au vecteur $(\text{tf-idf}_{i,1}, \text{tf-idf}_{i,2}, \ldots, \text{tf-idf}_{i,N}) \in \mathbb{R}^N$

On peut alors mesurer la similarité entre deux textes via n'importe quelle mesure de similarité sur $\mathbb{R}^N$. La plus usitée est la similarité cosinus, pour laquelle la proximité entre deux documents $d_1$ et $d_2$ est donnée par : $\frac{<d_1,d_2>}{||d_1||||d_2||}$

**Formellement : la loi de Zipf**

La forme de la pondération évoquée ci-dessus a été justifiée par Zipf en 1949 [27] : en étudiant des textes de Joyce, Plaute ou encore Homère, il montre qu'empiriquement, la fréquence d'apparition du nième terme le plus fréquent au sein d'un document est de la forme $f(n) = \frac{K}{n}$, où $K$ est une constante dépendant du texte.

Mathématiquement, la loi de Zipf est une loi de probabilité sur $\mathbb{N}^*$ de paramètres $(N,s) \in \mathbb{N}^* \times \mathbb{R}_+^*$. Le paramètre $N \in \mathbb{N}^*$ représente le nombre d'éléments du dictionnaire (de mots, donc) et $s \in \mathbb{R}_+^*$ est un paramètre d'adéquation. . Sa fonction de masse est donnée par :

$$f_{N,s}(k) = \frac{1}{H_{N,s}} \frac{1}{k^s} \quad \text{avec } H_{N,s} = \sum_{n=1}^{N} \frac{1}{n^s}$$

Il est clair qu'il s'agit bien d'un mesure de probabilité sur $\mathbb{N}^*$. Elle s'interprète comme suit : soit un texte écrit à partir d'un dictionnaire de $N$ mots. On suppose que la fréquence d'apparition des mots suit la loi de Zipf de paramètres $(N,s)$. Soit alors $k \in [\![1,N]\!]$ un rang d'apparition et $t_k$ le terme correspondant. Alors, la fréquence d'apparition du terme $t_k$ au sein du document est donnée

par :

$$f_k = \frac{1}{H_{N,s}} \frac{1}{k^s}$$

En particulier, le terme $t_k$ devrait apparaître $\frac{N}{H_{N,s} \times k^s}$ fois dans le texte.

**Cas où N est infini**

Dans le cas où $N$ est infini, la loi de Zipf n'est définie que pour $s > 1$ puisque la série diverge pour $s \leq 1$. La masse totale vaut alors $\zeta(s)$ où $\zeta$ désigne la fonction zêta de Riemmann :

$$\begin{array}{rccc} \zeta & : & \{s \in \mathbb{C} \mid \Re(s) > 1 & \rightarrow & \mathbb{C} \\ & & x & \mapsto & \sum_{n=1}^{\infty} \frac{1}{n^s} \end{array}$$

**Lien avec la pondération TF-IDF**

Supposons que, pour tous les documents du corpus, la loi de fréquence des termes suive une loi de Zipf de paramètres $(N_i, 1)$ avec $N_i$ le nombre de mots du document $i$ (fini). Le terme $\frac{\rho_{j,i}}{\sum_i \rho_{j,i}}$ de la pondération TF-IDF correspond à la fréquence d'apparition du terme $j$ dans le document $i$ et on a donc :

$$\frac{\rho_{j,i}}{\sum_i \rho_{j,i}} \sim \frac{1}{H_{N,s} \times k_j^i}$$

avec $k_j^i$ le rang d'apparition du terme $j$ dans le texte $i$.

# 5 Développement de la nouvelle mesure de distance entre métiers et création de la variable agrégée associée

Dans cette section, nous nous intéressons à l'intégration de la variable métier dans les modèles. Cette variable est en effet catégorielle et non-ordonnée. Dans les modèles classiques, elle est donc traitée de manière binaire dans les algorithmes de *clustering* : deux observations ont ou n'ont pas le même métier. Nous cherchons donc à améliorer cette dichotomie.

Dans un premier temps, nous créons plusieurs mesures de distance continues entre les métiers. Pour cela, nous avons exploré trois pistes permettant de rapprocher des métiers a priori différents :

— le fait d'exercer simultanément deux professions différentes

— le fait de passer d'une profession à une autre

— les caractéristiques socioéconomiques individuelles

Pour comparer les PCS exercés simultanément, nous avons adapté la méthode de comparaison de textes fondée sur le modèle vectoriel avec la distance TF-IDF. En ce qui concerne les transitions entre PCS, nous avons choisi de les modéliser par un modèle markovien.

Dans un second temps, ces distances sont agrégées afin d'obtenir une unique distance.

Enfin, un *clustering* est effectuée sur les métiers en utilisant cette distance. Une nouvelle variable catégorielle représentant le métier est ainsi créée.

A l'issue de cette section, nous disposerons donc d'une distance permettent de quantifier plus précisément la similitude entre deux métiers, que nous réinjecterons dans les modèles de *clusterings* afin d'en améliorer la qualité.

La finalité de ce travail est de mettre en lumière des *clusters* d'individus partageant des profils de risque d'absentéisme similaires. Pour cela, deux *clusterings* sont effectués. L'un ne prend pas en compte le taux d'absentéisme et permettra de proposer un baromètre sectoriel plus pertinent que la simple moyenne au sein du secteur d'activité concerné, en se fondant sur l'hypothèse que des individus partageant des caractéristiques individuelles proches devraient avoir des comportement d'absentéisme proches hors maladies longues. L'autre, prenant cette fois en compte le risque

d'absentéisme, permet de distinguer différents profils d'absentéisme au niveau individuel.

Comme nous disposions de données mêlant variables catégorielles et numériques, notre choix algorithmique s'est porté sur le *KPrototypes*, implémenté en Python sur *scikit-learn* [28] ainsi que sur l'AFDM, implémentée manuellement.

Nous avons cependant apporté un raffinement en ce qui concerne la distance entre variables catégorielles. En effet, le but est de prendre au mieux en compte la variable métier dans le *clustering*. Pour ce faire, au lieu de conserver la distance usuellement usitée pour les variables catégorielles, une distance spécifique est implémentée. Cette distance repose sur la création d'une nouvelle mesure de la similarité entre métiers. Sa construction est l'objet de cette section.

## 5.1 Présentation de la nomenclature PCS-ESE

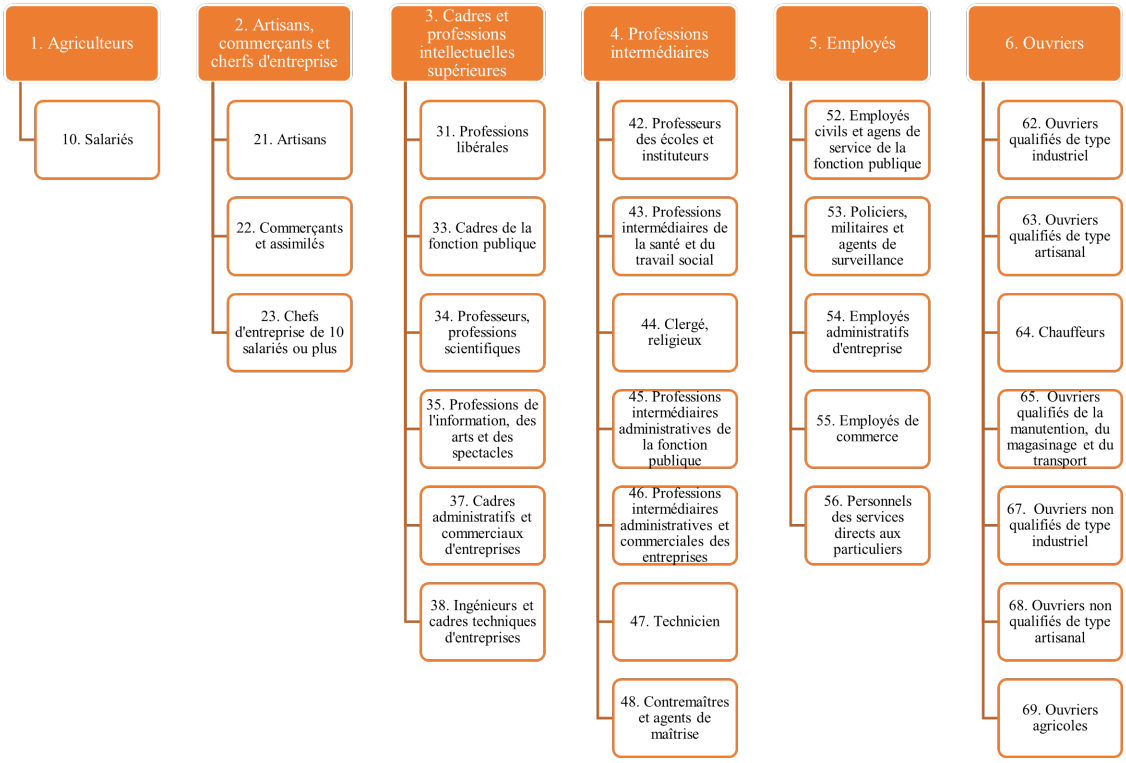


FIGURE 5.1 – Deux premiers niveaux de la nomenclature PCS-ESE

La nomenclature des professions et catégories socioprofessionnelles des emplois salariés des employeurs privés et publics (PCS-ESE) permet aux employeurs de codifier la profession de leurs salariés dans les différentes enquêtes statistiques, déclarations et formulaires administratifs, parmi lesquels la déclaration sociale nominative (DSN). Par défaut, le terme *PCS* se réfèrera à cette nomenclature dans toute la suite. La présentation générale, l'historique des modifications ainsi que le détail de la nomenclature en vigueur sont disponibles sur le site de l'Insee [29].

Cette nomenclature est hiérarchique, composée de trois niveaux. Le premier niveau contient six postes, codifiés par un chiffre entre 1 et 6, le deuxième vingt-neuf postes codifiés en accolant un deuxième chiffre à celui du premier niveau. Le troisième niveau est constitué de 429 postes et le codage est obtenu en ajoutant un chiffre et une lettre à celui du niveau deux. On dispose ainsi d'une structure d'arbre sur les PCS. La figure explicative est présentée en annexe.

## 5.2 Première approche : distance d'arbre

La nomenclature PCS-ESE est naturellement munie d'une structure de forêt au sens défini en section 4.9.10. On peut artificiellement ajouter une racine commune aux premiers niveaux pour la rendre connexe et la transformer en arbre, ce qui évite que des PCS soient à distance infinie. On peut ensuite définir la distance entre deux PCS comme la distance de graphe entre les sommets correspondants.

Cette approche ne répond cependant que partiellement au problème. En effet, la distance entre deux sommets sur l'arbre associé ne peut prendre que les valeurs 0, 2, 4 et 6 ce qui n'est pas très discriminant. On pourrait raffiner l'approche en pondérant le graphe et en donnant plus de poids aux arêtes proches de la racine, mais cela ne changerait fondamentalement pas le problème : la distance ne prendrait toujours qu'un petit nombre de valeurs.

En outre, la hiérarchie PCS-ESE est construite sur des distinctions administratives qui ne reflètent pas nécessairement la réalité opérationnelle. En considérant la distance la plus simple sans pondération, les employés de La Poste (521a) et les professions intermédiaires de La Poste (451a) sont par exemple à distance 6 tandis que les agents techniques forestiers ou gardes des espaces naturels (533b) et les convoyeurs de fonds, enquêteurs privés et gardes du corps ne seront qu'à distance 2 : cette distance rapproche donc des individus dont le quotidien professionnel est très différent, ce qui semble peu adapté pour une étude sur l'absentéisme.

En outre, cette distance rapproche des métiers dont le taux d'absentéisme est éloigné : les PCS 643a (Conducteurs livreurs, coursiers) et 641a (Conducteurs routiers et grands routiers) sont à distance 2 (donc minimale) alors que le taux d'absentéisme moyen est 2 fois plus élevé pour les Conducteurs livreurs et coursiers que pour les conducteurs routiers et grands routiers…
Il semble donc que cette mesure de distance ne soit pas la plus pertinente pour comparer des métiers, que ce soit sous le prisme de l'absentéisme ou de la réalité métier.

## 5.3 Adaptation des méthodes d'analyse textuelle et de la distance TF-IDF

### 5.3.1 Rappel du cadre

On considère un corpus, constitué de $m$ documents : $\mathscr{C} = \{d_1, \ldots, d_m\}$ Chaque texte est lui-même rédigé à partir d'un dictionnaire, c'est-à-dire d'un ensemble de $N$ mots fixé : $\mathscr{D} = \{t_1, \ldots, t_N\}$.

L'idée est de comparer les textes à partir des mots qui les composent, en considérant que l'importance d'un terme au sein d'un document dépend de deux facteurs : sa spécificité et son exhaustivité [25]. L'exhaustivité de la description d'un document se définit alors comme le nombre de termes qu'il contient tandis que la spécificité d'un terme est représentée par le nombre de textes dans lequel il apparaît.

On suppose alors que plus un terme est spécifique, moins il a de chance d'apparaître et plus un document est exhaustif, moins la présence d'un terme dans ce texte est significative.
Les détails théoriques sont explicités en section 4.5.2.

### 5.3.2 Application au problème de distance entre PCS

#### Analogie

Dans l'analogie développée, chaque PCS joue le rôle d'un document et le dictionnaire est, lui-aussi, constitué de l'ensemble de PCS. On dira qu'un « terme » apparaît dans un « document » si les deux PCS concernés ont été exercés simultanément par un même individu. On peut alors représenter chaque PCS par le vecteur des pondérations et comparer deux PCS via leur mesure de similarité cosinus, comme on le ferait pour deux documents.

#### En pratique

Cette distance entre PCS est ensuite réinjectée dans la distance vectorielle entre deux vecteurs de variables catégorielles, utilisée en entrée du *KPrototypes* lors du *clustering*. On définit la distance entre deux vecteurs de variables catégorielles contenant le PCS par :

$$d(X,Y) = \sum_{j=1}^{m-1} \delta(x_j, y_j) - \eta \frac{< pcs_1, pcs_2 >}{||pcs_1||\ ||pcs_2||}$$

où

$$\delta(x_j, y_j) = \begin{cases} 0 \text{ si } x_j = y_j \\ 1 \text{ si } x_j \neq y_j \end{cases}$$

La constante $\eta$ permet de moduler l'importance que l'on veut donner aux PCS et de l'adapter en fonction du nombre de variables catégorielles autres que le PCS. Le signe « - » permet de transformer la mesure de similarité en mesure de distance : plus la similarité entre deux PCS est grande, plus le terme $\frac{<pcs_1,pcs_2>}{||pcs_1||\ ||pcs_2||}$ est grand et donc plus $d(X,Y)$ diminue.

Cette formule permet par ailleurs de transmettre à la distance la notion d'indépendance présente dans la mesure de similarité : si deux PCS sont orthogonaux dans le modèle vectoriel, le terme $\frac{<pcs_1,pcs_2>}{||pcs_1||\ ||pcs_2||}$ s'annule et le PCS n'influe donc pas sur la distance entre les vecteurs. La distance devient donc elle aussi *indépendante* du PCS.

Dans le cadre de la comparaison et de l'agrégation avec les autres distances, la valeur de $\eta$ retenue est $\eta = 1$. La distance finale entre deux PCS est donc :

$$d(pcs_1, pcs_2) = -\frac{< pcs_1, pcs_2 >}{||pcs_1||\ ||pcs_2||}$$

On retrouve ainsi bien une relation décroissante entre la similarité de deux métiers et leur distance.

## 5.4 Distance en utilisant les transitions entre PCS

Afin de compléter l'approche statique obtenue en regardant le cumul des métiers, une approche dynamique a été mise œuvre. L'heuristique consiste ici à dire que, si on peut passer d'un métier à un autre, ces métiers concentrent des populations proches. La probabilité de passer d'un PCS à un autre fournirait ainsi une autre mesure de leur proximité.

Pour estimer la probabilité de transition entre $i$ et $j$, on compte, pour chaque date $t$, le nombre d'individus qui exercent le PCS $i$ en $t$ et le PCS $j$ en $t+1$. Cette quantité est ensuite rapportée au nombre d'individus qui exercent le PCS $i$ en $t$. On obtient ainsi une fréquence empirique de passage de $i$ à $j$ à l'instant $t$. En faisant la moyenne des fréquences empiriques sur toutes les dates disponibles, on obtient alors une estimation de la probabilité de transition.

Formellement, en notant $p_{i,j}$ la probabilité de passer du PCS $i$ au PCS $j$, on a donc, pour un modèle à $T$ périodes :

$$p_{i,j} = \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{|PCS_{t+1}^i \cap PCS_t^j|}{PCS_t^i}$$

Où la notation $PCS_t^i$ désigne l'ensemble des personnes qui occupent le PCS $i$ à l'instant $t$.

**Remarque :** La mesure ainsi obtenue n'est pas une distance ! D'une part, il s'agit plutôt d'une mesure de similarité : plus elle est proche de 0, plus la probabilité de passer de l'un à l'autre est petite, *ie* plus les métiers sont éloignés. Ce défaut est cependant facilement corrigé : en prenant $d(i,j) = 1 - p_{i,j}$, on obtient bien une fonction décroissante en la similarité.

D'autre part, cette mesure n'est pas symétrique. Cet aspect est plus ennuyeux et ses répercussions seront explicitées dans la partie suivante. Une manière de symétriser la distance serait, par exemple de considérer :

$$p_{i,j} = \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{|PCS_{t+1}^{i} \cap PCS_{t}^{j}|}{|PCS_{t}^{i}| + |PCS_{t}^{j}|}$$

Cela présenterait toutefois des inconvénients majeurs dans le cas où $|PCS_t^i| << |PCS_t^j|$ par exemple. Une autre solution serait de considérer $d(i,j) = \max\left(p_{i,j}; p_{j,i}\right)$.

## 5.5 Distance en utilisant les données *Filosofi*

Les deux approches explorées ci-dessus sont extrêmement factuelles : elles permettent de rapprocher des métiers via le nombre d'individus qui les ont exercés simultanément, indépendemment de leurs caractéristiques propres. L'idée de cette troisième approche est de renverser le raisonnement : au lieu de considérer les métiers et d'observer les individus en fonction du métier, on s'intéresse aux caractéristiques socioéconomiques des individus, que l'on cherche à relier au métier.

Nous ne disposons toutefois d'aucune donnée concernant les caractéristiques socioéconomiques des individus. Pour pallier cela, nous nous appuyons sur l'hypothèse que le lieu de résidence est un marqueur fort de ces caractérisitiques. En croisant le code postal de résidence de l'individu avec les données *Filosofi*, nous obtenons ainsi une esquisse du profil socioéconomique de l'individu. Les variables utilisées pour cette partie sont présentées en annexe, la variable géographique étant totalement gommée de l'étude une fois la jointure effectuée.

Les individus sont ainsi représentés par les caractéristiques de leur lieu d'habitation, ainsi que par quelques caractéristiques individuelles (âge et genre principalement) indépendemment de leur métier. Une ACP est effectuée pour réduire la dimensionnalité, puis les points obtenus sont coloriés par PCS. Un PCS est alors représenté par l'individu moyen parmi ceux partageant ce métier. La distance entre deux PCS est finalement calculée comme la distance entre les individus moyens correspondants.

## 5.6 Concaténation des distances

Intuitivement, il paraît clair que les approches évoquées précédemment se complètent. Cependant, comment les concilier ? Une manière simple de le faire est de considérer comme distance entre PCS la moyenne simple ou pondérée des distances obtenues. Cela pourrait cependant gommer des différences : si l'une est nulle et l'autre proche de 1, comment l'interpréter ? En outre, il

n'est a priori absolument pas évident que les distances soient du même ordre de grandeur. Elles sont certes toutes les deux théoriquement comprises entre 0 et 1 (en valeur absolue pour la distance statique), mais on pourrait imaginer que toutes les valeurs de l'une soient proches de 0 tandis que toutes les valeurs des autres soient proches de 1. Dans ce cas, considérer la moyenne gommerait toute l'information contenue dans la distance proche de 0 alors qu'il ne s'agit que d'une question d'échelle. On pourrait remédier à cela en pondérant la moyenne, mais on ne fait que décaler le problème : comment déterminer la pondération ? Existe-t-il un moyen d'estimer statistiquement cette pondération ? Existe-t-il un moyen déterministe de fixer cette pondération ou bien n'y a-t-il que des moyens empiriques ?

En ce qui concerne le problème de différence d'échelle, la solution retenue a été de normaliser les distances : elles ont été retraitées via un *Min-max* et standardisées. Ce choix se justifie également par l'usage final : on souhaite en effet obtenir une nouvelle variable catégorielle pour les métiers et passer pour cela par un *clustering*, pour lequel il faut que les données soient standardisées. Enfin, la normalisation permet de justifier le choix arbitraire de la constante $\eta$ dans le cas de la distance obtenue via le cumul de métiers.

Pour concaténer les distances, il a été choisi d'utiliser une approche encore une fois visuelle. Si l'on dispose de $n$ distances $d_1, \ldots, d_n$ que l'on souhaite agréger, on se place dans $\mathbb{R}^n$. Pour obtenir la distance entre un PCS $i$ donné et les autres PCS, on considère alors le PCS $i$ comme l'origine d'un repère dans $\mathbb{R}^n$ : chaque PCS $j$ devient alors un point de cet espace, dont les coordonnées sont données par les différentes distances. Formellement, la coordonnée du point $j$ sur l'axe $k$ sera donnée par $d_k(i, j)$. On obtient alors la distance entre le PCS $i$ et le PCS $j$ comme la norme euclidienne du vecteur représentant le point $j$ dans le repère de $\mathbb{R}^n$ d'origine $i$.

En d'autres termes, la distance finale entre deux PCS est obtenue comme la moyenne quadratique des distances, à constante près.

**Remarque :** La distance obtenue via les transitions n'étant pas symétrique, la distance concaténée ne l'est pas non plus.

# 6 Résultats

## 6.1 Traitement des données

Pour les distances fondées sur les transitions ou le cumul entre métiers, les données ne sont pas retraitées. Le seul traitement effectué concerne les individus conservés pour le calcul : lorsqu'un individu exerce plusieurs fois un même métier sur la même période, une seule occurrence est conservée et, pour le calcul de distance fondé sur les transitions, seuls les individus ayant changé de métier durant la période d'observation sont considérés.

Pour la distance utilisant les données *Filosofi*, les lignes contenant des valeurs manquantes pour les variables qualitatives sont retirées et les valeurs manquantes restantes sont complétées par la moyenne de la variable sur la base.

## 6.2 Obtention de la distance finale

### 6.2.1 Distance fondée sur le cumul des métiers

Pour cette distance, on applique la formule présentée en 5.3 à l'intégralité de la base de données. La distance obtenue est ensuite normalisée et standardisée pour des besoins de lisibilité et de comparabilité. Les résultats obtenus sont présentés en Figure 6.1.

Si les métiers sont globalement éloignés les uns des autres ; quelques clusters se dégagent toutefois à première vue. La granularité à laquelle l'étude est effectuée (PCS 3) est toutefois à double tranchant. Si elle permet de saisir des cumuls qui n'apparaîtraient pas à un niveau plus agrégé, notamment entre deux PCS pour lesquels les différences sont plus administratives qu'opérationnelles, elle nuit toutefois à la lisibilité et à l'interprétabilité des résultats.

Par comparaison, les résultats obtenus sur le PCS au niveau 2 sont présentés en Figure 6.2. Cette matrice est à interpréter comme une matrice de corrélation : plus la similarité est proche de 1, plus les PCS sont proches. Si la similarité est nulle, ils sont indépendants, c'est-à-dire qu'ils ne sont jamais cumulés : le PCS n'apporte donc aucune information supplémentaire sur les individus. Enfin, si la similarité est proche de -1, les PCS sont opposés c'est-à-dire que des individus exerçant

respectivement ces PCS sont distants.

Dans cette matrice, les similarités sont toujours positives ou nulles et la similarité est souvent nulle. Il faut cependant lire cette matrice en gardant à l'esprit la répartition des PCS au sein de notre portefeuille. Le fait que le PCS 67 (Ouvriers non qualifiés de type industriel) soit indépendant de tous les autres est intéressant : il s'agit du PCS le plus représenté dans notre base et ce résultat incite donc à traiter cette catégorie socio-professionnelle indépendamment des autres. A contrario, les PCS 68 (Ouvrier qualifiés de type industriel) et 47 (Techniciens) sont proches pour cette mesure de similarité : il pourrait donc être intéressant de les regrouper dans une même classe dans une perspective de segmentation. Cet exemple permet par ailleurs de souligner l'avantage de cette approche par rapport à une simple distance d'arbre : ces deux professions seraient à distance maximale pour cette mesure alors qu'elles sont parfois cumulées par des individus.

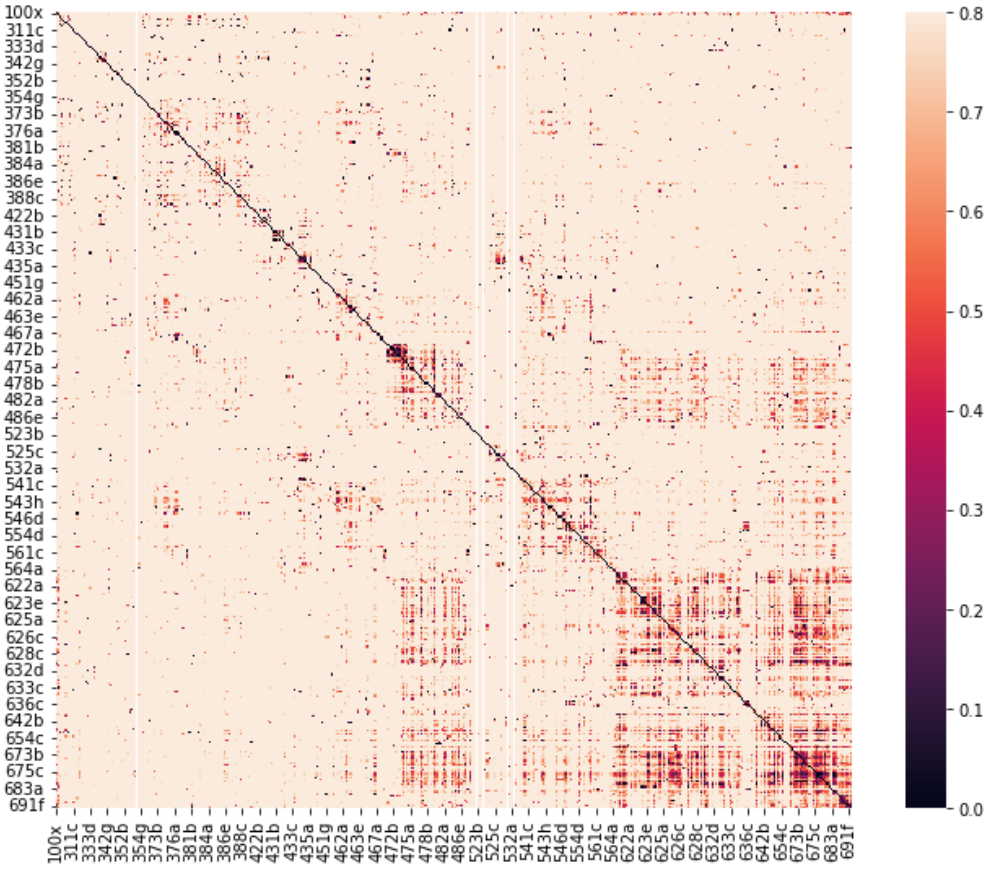


FIGURE 6.1 – Matrice de distance normalisée et standardisée en utilisant le cumul de métiers

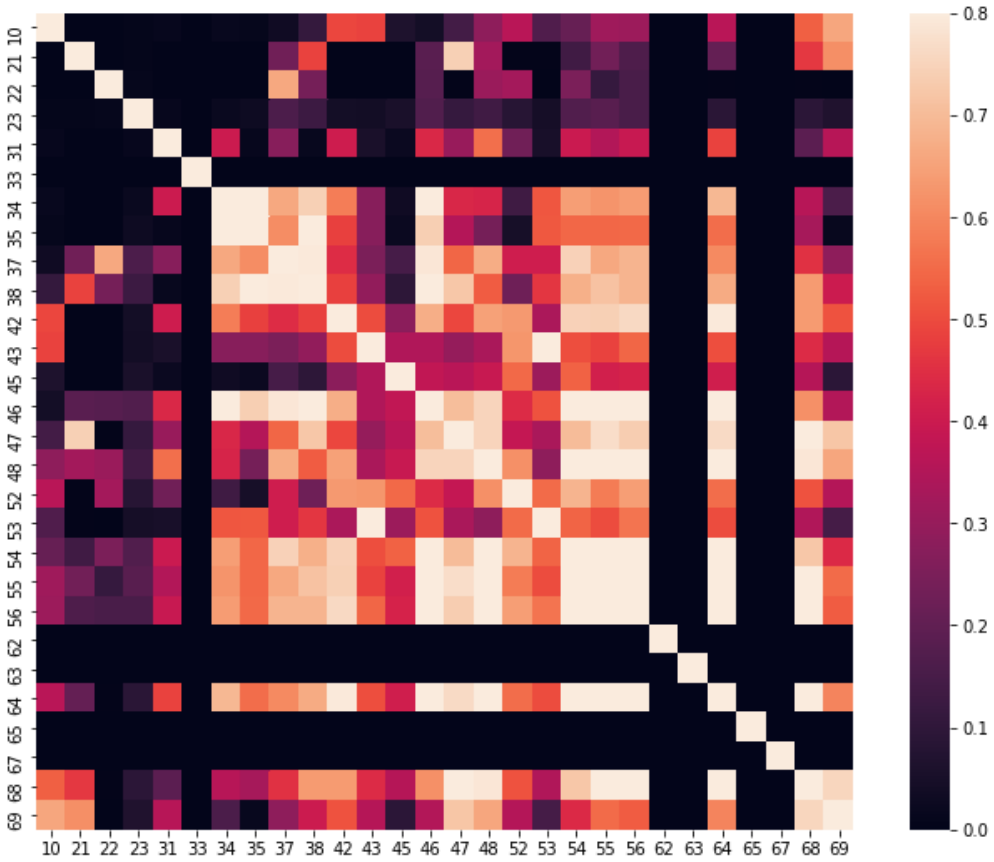


FIGURE 6.2 – Distances entre PCS au niveau 2 avec la nouvelle mesure de similarité

### 6.2.2 Distance fondée sur les transitions entre métiers

Pour cette distance, on applique la formule présentée en 5.4 à une base restreinte : on ne considère en effet que les individus ayant changé de métier au moins une fois au cours de la période d'observation. Ce choix permet de limiter le biais lié aux personnes ne changeant jamais de métier : au vu de la formule utilisée, la prise en compte de ces individus ne ferait que diminuer toutes les distances sans modifier fondamentalement les liens qui existent entre métiers.
La distance obtenue est également normalisée et standardisée et les résultats sont présentés en Figure 6.3.

Là encore, les métiers sont globalement tous éloignés les uns des autres. Cela s'explique en partie par la granularité à laquelle la matrice est calculée (niveau 3 du PCS). Quelques petits groupes plus proches se dégagent cependant, ce qui fournit une base pour des analyses plus approfondies. Elle s'explique d'autre part par le fait que les transitions entre métiers sont peu fréquentes, et ne sont captées que via les changements de PCS dans les bases de données, ce qui ne fournit pas une vision exhaustive.

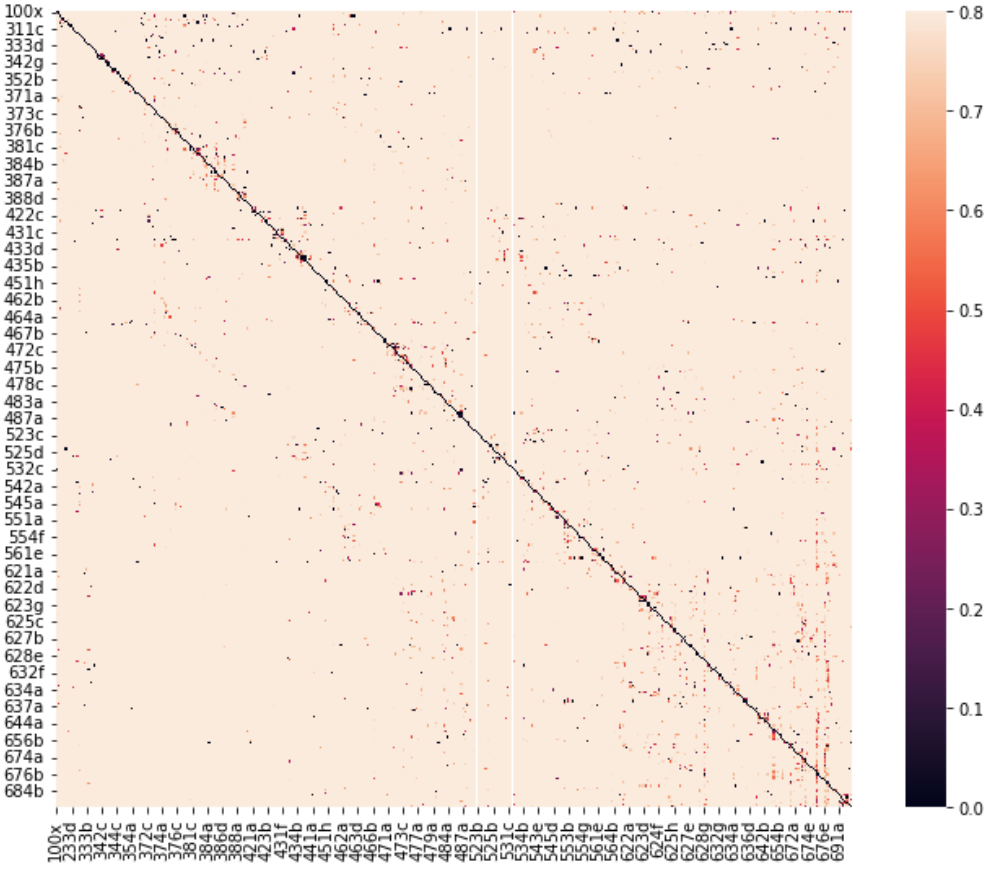


FIGURE 6.3 – Matrice de distance normalisée et standardisée en utilisant les transitions entre métiers

### 6.2.3 Distance en utilisant les données *Filosofi*

Le calcul de cette distance s'effectue en deux temps : une ACP permet d'abord de réduire la dimensionnalité, limitant ainsi l'impact de la corrélation entre les variables. Les individus moyens par PCS sont ensuite placés dans l'espace factoriel obtenu et les distances calculées à partir de ces individus moyens.

#### Etude de la corrélation entre les variables

La matrice de corrélation est présentée en Figure 6.4. La variable représentant le premier décile du niveau de vie semble ainsi particulièrement polarisante puisqu'elle est fortement corrélé au taux de pauvreté, au taux de chômage ainsi qu'aux prestations sociales versées. Les taux de pauvreté par classe d'âge sont également fortement corrélés entre eux.
Avant d'effectuer l'ACP, les valeurs sont normées et standardisées.

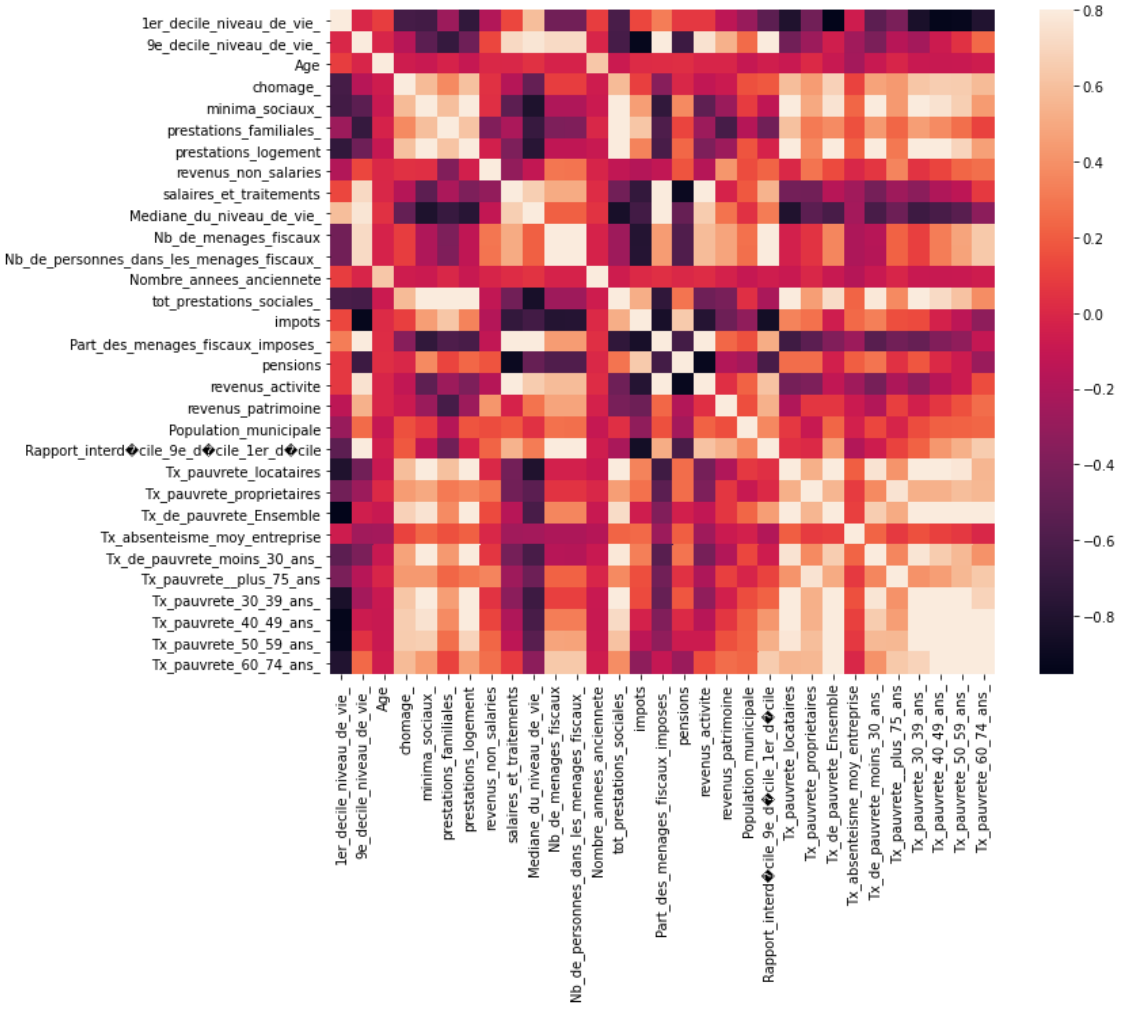


FIGURE 6.4 – Corrélation entre les variables

**AFDM**

Les figures 6.5 à 6.7 présentent le taux de variance expliqué et l'inertie totale en fonction du nombre de facteurs retenus. Les résultats sont résumés dans la Table 6.1.
Le nombre de facteurs à retenir varie énormément en fonction des tests. Nous choisissons $k = 2$ facteurs pour la visualisation pour des raisons pratiques et, au vu du faible pourcentage de variance expliquée pour ce choix de $k$, nous retenons $k = 14$ facteurs pour le calcul de distance.
La Figure 6.8 représente les individus moyens par PCS sur les deux premiers axes factoriels. Sur cette visualisation, un outlier se distingue très nettement. Pour les autres individus, deux groupes se distinguent clairement ; une segmentation plus fine en quatre *clusters* semble également admissible.

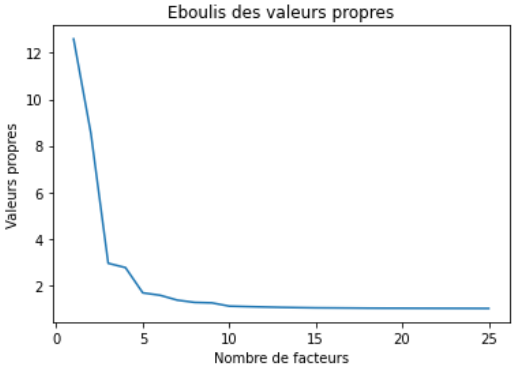


FIGURE 6.5 – Eboulis des valeurs propres

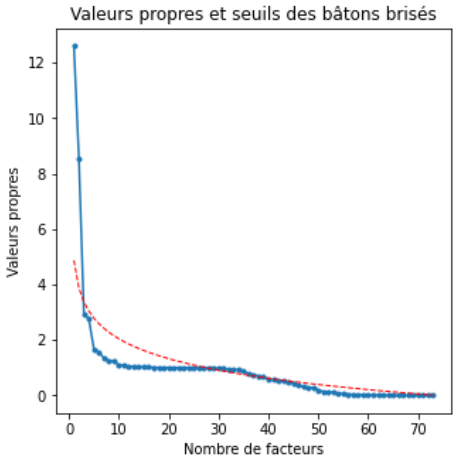


FIGURE 6.6 – Test des bâtons brisés

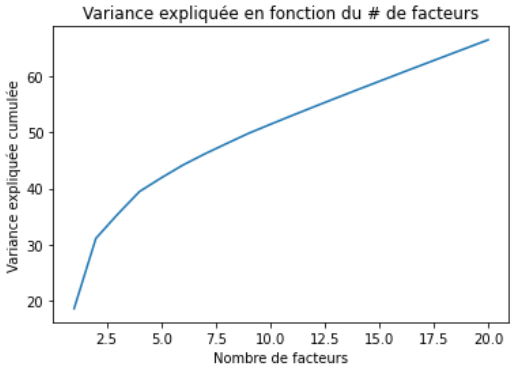


FIGURE 6.7 – Variance expliquée en fonction du nombre de facteurs

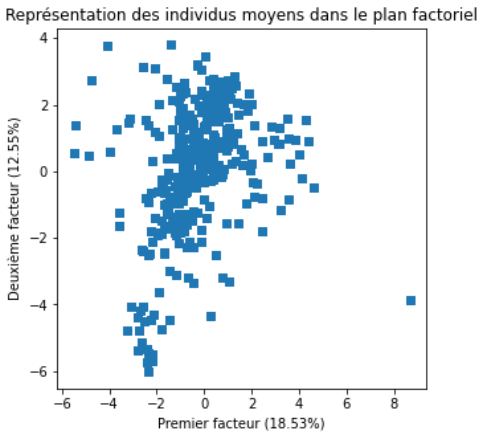


FIGURE 6.8 – Représentation des individus moyens par PCS sur les deux premiers axes factoriels

| Méthode | Seuil | # facteurs à retenir | % variance expliquée |
|---|---|---|---|
| Kaiser-Guttman | $\lambda > 1$ | 19 | 64% |
| *Elbow curve* | Visuel | 5 | 42% |
| Seuil bâtons brisés | $\lambda_k > \sum_{i=k}^{p} \frac{1}{i}$ | 2 | 31% |
| Karlis-Saporta-Spinaki | $\lambda > 1 + 2 * \sqrt{\frac{p-1}{n-1}}$ | 14 | 60% |

TABLE 6.1 – Résultats de l'AFDM en fonction de la méthode utilisée

La matrice de distance obtenue avec les 14 axes factoriels est représentée en Figure 6.9. Par comparaison, celle obtenue avec 2 axes factoriels est présentée en Figure 6.10.

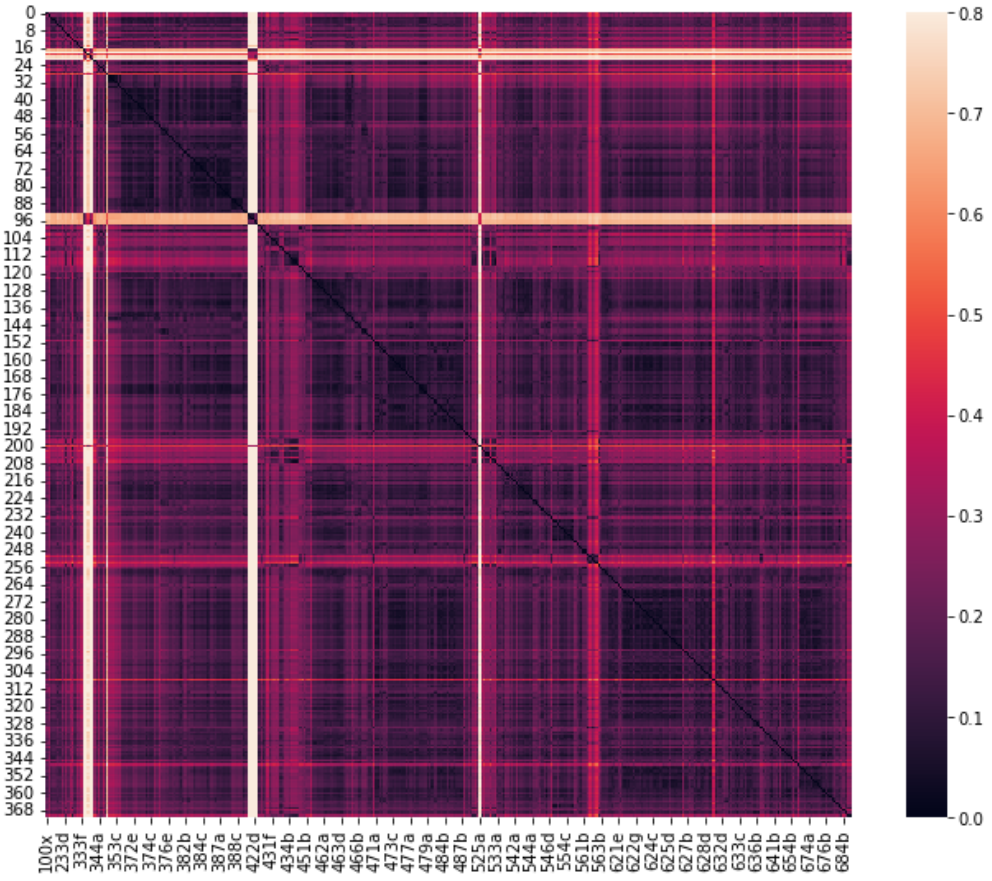


FIGURE 6.9 – Matrice de distance avec 14 axes factoriels

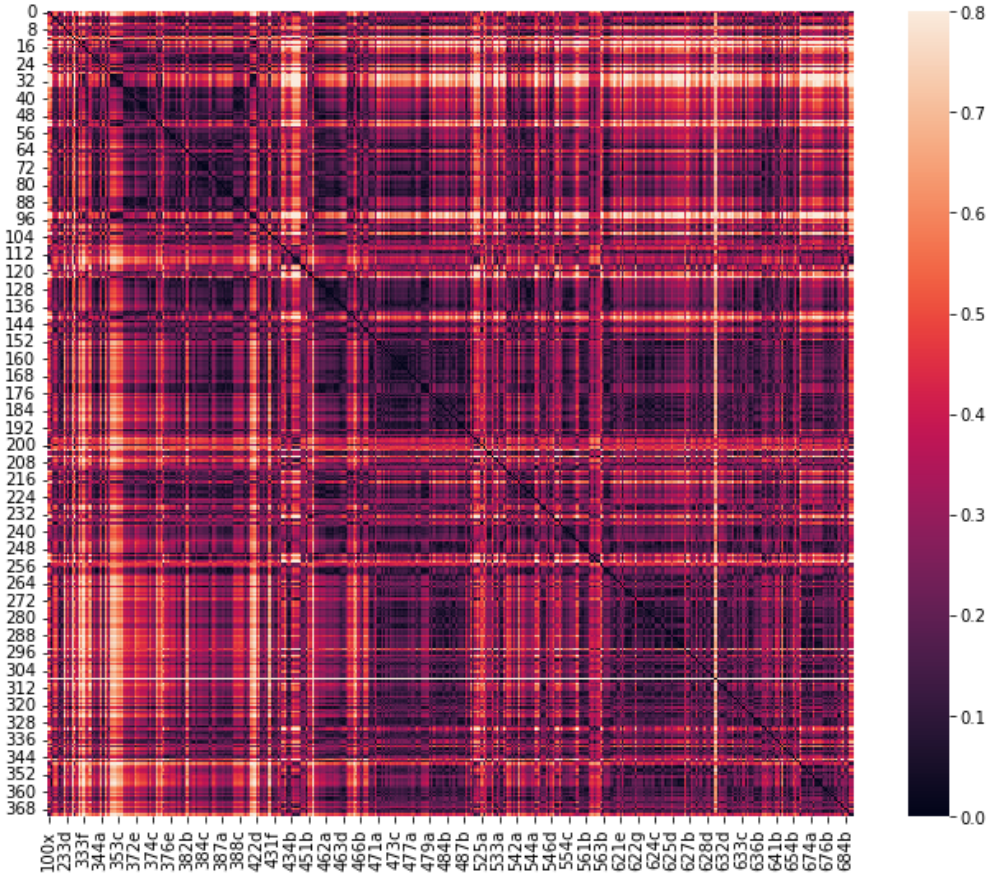


FIGURE 6.10 – Matrice de distance avec deux axes factoriels

### 6.2.4 Distance finale

La matrice de distance finale est présentée en Figure 6.11. Les métiers sont globalement tous très éloignés les uns des autres. Quelques groupes se dégagent cependant autour de la diagonale ; ils sont assez compacts, laissant présager d'éventuels clusters regroupant quelques métiers. On peut cependant anticiper que les résultats du clustering ne seront pas forcément très concluants : si les individus sont très dispersés, ils seront affectés à un cluster par défaut sans que cela ne soit réellement représentatif d'une réelle interaction entre eux.

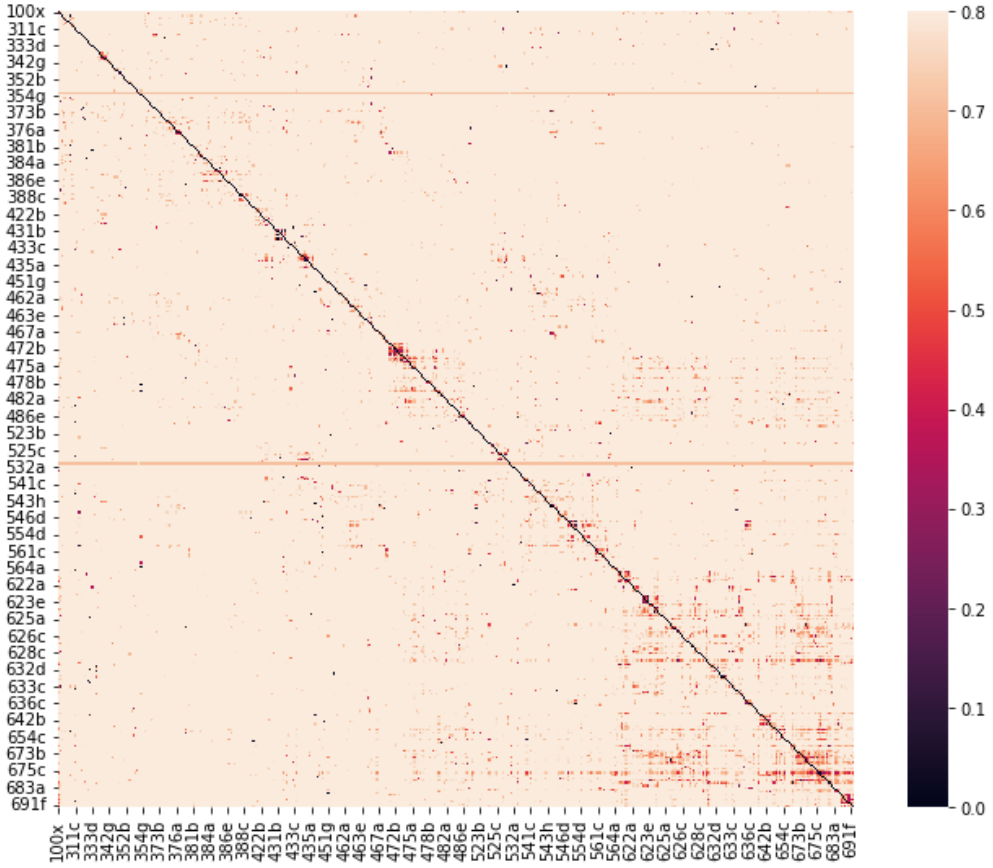


FIGURE 6.11 – Matrice de distance finale

## 6.3 Segmentation des métiers

#### Visualisation avec réduction de dimensionalité par l'algorithme TSNE

Ici, la forme des groupes est particulière : les PCS semblent cependant regroupés en cercles plus ou moins concentriques, fournissant une première intuition de regroupement.

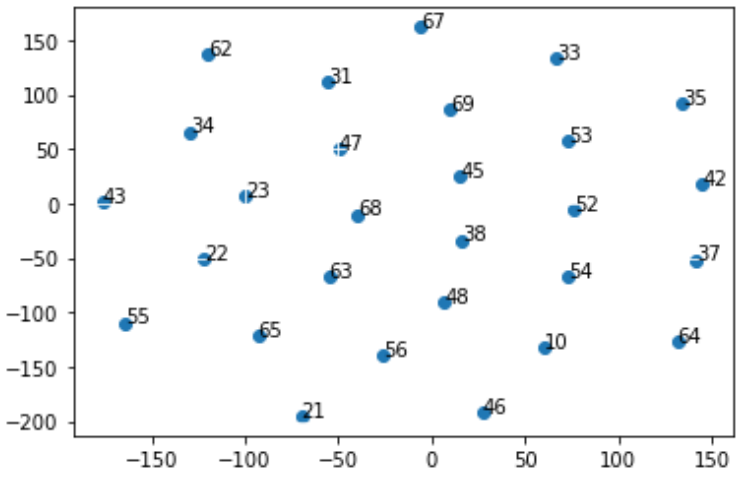


FIGURE 6.12 – Visualisation des données en deux dimensions par TSNE

L'algorithme TSNE est cependant sensible à la dimension ; une visualisation en trois dimensions est présentée en Figures 6.13 et 6.14 afin de vérifier si d'autres groupes se dégagent.
En regardant uniquement les PCS au niveau 2 [1], il semble difficile de dégager des groupes à partir de la visualisation en trois dimensions. On pourrait tenter de faire des projections en deux dimensions afin de voir si des groupes se dégagent plus clairement.

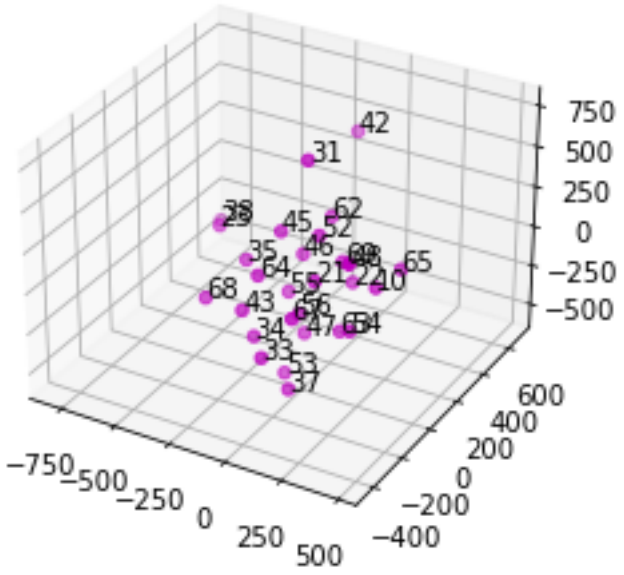


FIGURE 6.13 – Visualisation en trois dimensions par TSNE au niveau 2

Les données avec les PCS au niveau 3 sont difficiles à représenter du fait du nombre impor-

1. Figure 6.13

tant de PCS. Sur la Figure 6.14, cinq PCS se dégagent cependant nettement des autres. Il s'agit des PCS 626a (Pilotes d'installation lourde des industries de transformation : métallurgie, production verrière, matériaux de construction), 543g (Employés administratifs qualifiés des autres services des entreprises), 461f (Maîtrise et techniciens administratifs des autres services administratifs), 385c (Ingénieurs et cadres technico-commerciaux des industries de transformations (biens intermédiaires)) et 376f (Cadres des services techniques des organismes de sécurité sociale et assimilés). Les PCS 376f et 385c sont quasiment absents de la base de données et leur isolement n'est donc pas très significatif. Les trois autres regroupent en revanche respectivement 51 000 individus (626a), 116 000 individus (543g) et 31 000 individus (461f) sur une base totale de : le fait qu'ils se détachent complètement des autres PCS est donc intéressant et invite à les traiter séparément des autres par la suite.

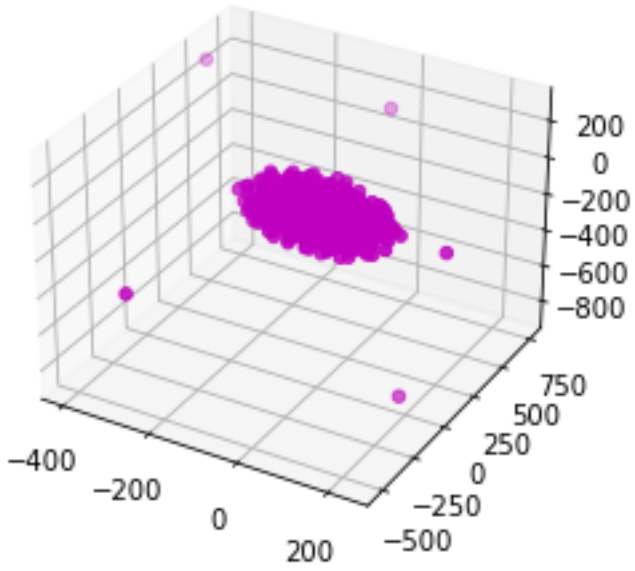


FIGURE 6.14 – Visualisation en trois dimensions des PCS au niveau 3

**Création de la nouvelle variable métier**

La variable catégorielle finale regroupant les métiers est obtenue via un *clustering* à partir de la matrice de distance obtenue précédemment.
Les figures 6.15 et 6.16 présentent respectivement l'inertie et le score en fonction du nombre de *clusters*. Ces graphiques sont utilisés pour appliquer une méthode du coupde et déterminer le nombre de *clusters* à conserver. On choisit ici d'en conserver 5.
La Table 6.2 présente les caractéristiques des clusters et la répartition par PCS au niveau 1.

Nous obtenons donc un nouvelle variable catégorielle permettant de coder le métier à un niveau

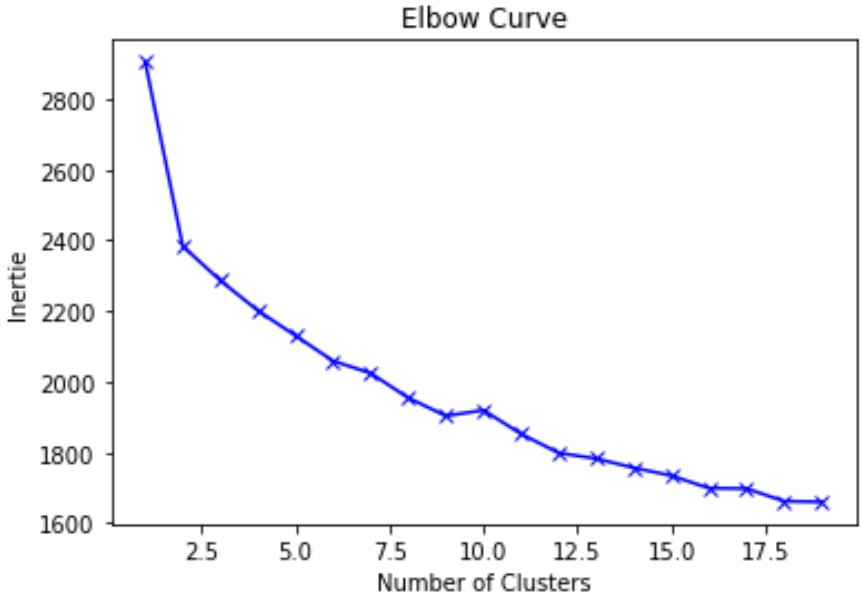


FIGURE 6.15 – Méthode du coude avec l'inertie

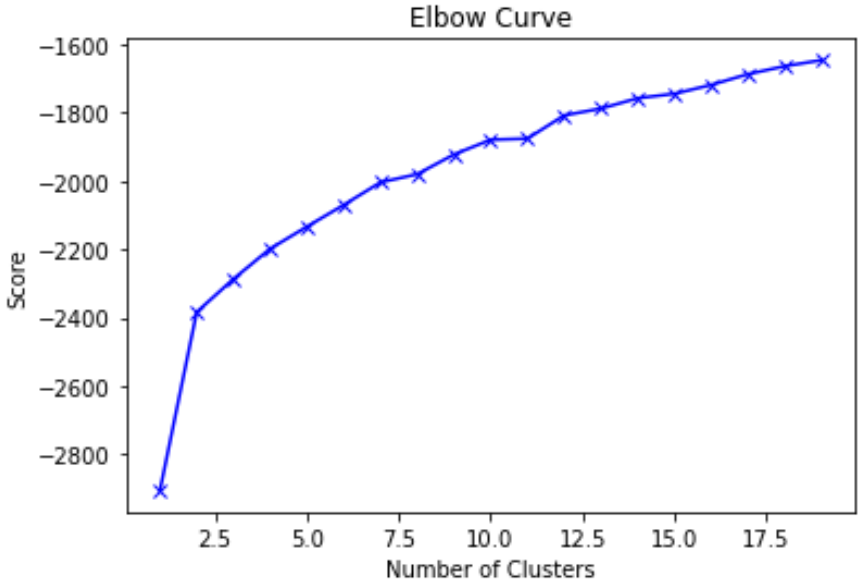


FIGURE 6.16 – Méthode du coude avec le score

| Cluster | Taille | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| Cluster 1 | 21 | 0% | 0 % | 4.8% | 57.1% | 38.1% | 0% |
| Cluster 2 | 212 | 0% | 3.3% | 35.8% | 28.8% | 18.4% | 13.7% |
| Cluster 3 | 63 | 0% | 0% | 28.6% | 27% | 44.4% | 0% |
| Cluster 4 | 50 | 0% | 0% | 2% | 14% | 2% | 82% |
| Cluster 5 | 79 | 1.3% | 1.3% | 0% | 24% | 2.5% | 70.9 % |

TABLE 6.2 – Résultats du *clustering* en utilisant la nouvelle distance sur le PCS au niveau 1

de granularité équivalent au niveau 1 du PCS.
La structure de ses modalités reprend quelques éléments du PCS : le PCS 1 (Agriculteurs) est par exemple intégralement regroupé dans le cinquième cluster et le PCS 2 (Artisans, commerçants et chefs d'entreprises) partagé entre les clusters 2 et 5.

## 6.4 *Back-testing* sur des PCS particuliers

Un premier test quant à la pertinence de cette distance a été réalisé en analysant les distances entre PCS calculées avec la nouvelle méthodologie sur un échantillon bien choisi :

- Manutentionnaires non-qualifiés (676a)
- Ouvriers non qualifiés divers de type industriel (676 e)
- Ouvriers non qualifiés du gros œuvre du bâtiment (681a)
- Ouvriers non qualifiés des travaux publics et du travail du béton (671c)
- Maçons qualifiés (632a)
- Magasiniers qualifiés (653a)
- Agents d'accueil qualifiés, hôtesses d'accueil et d'information (541b)
- Caissiers de magasin (552a)
- Conducteurs routiers et grands routiers (641a)
- Conducteurs livreurs, coursiers (643a)
- Nettoyeurs (684a)

Nous disposons en effet d'un nombre suffisant de données concernant ces métiers pour que les résultats ne soient pas purement fortuits. En outre, ces professions peuvent se séparer en groupes éloignés d'un point de vue tant opérationnel que profil d'absentéisme : on souhaiterait donc qu'une distance appropriée les discrimine.
Intuitivement, on souhaiterait que les PCS soient groupés comme suit :

- 676a et 676 e
- 681a, 671c et 632a
- 653a, 541b et 552a
- 641a et 643a
- 684a devrait être éloigné de tous les autres

En analysant les résultats présentés dans la Figure 6.17., on constate que le PCS 684a n'est pas isolé des autres ; il est cependant quasi-indépendant du 641a. Les 676a et 676e sont relativement

| PCS | 684a | 681a | 676e | 676a | 671c | 653a | 643a | 641a | 632a | 552a | 541b |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 684a | 1.0 | 0.340 | 0.230 | 0.488 | 0.310 | 0.260 | 0.153 | 0.09 | 0.165 | 0.123 | 0.197 |
| 681a | 0.340 | 1.0 | 0.251 | 0.533 | 0.630 | 0.236 | 0.132 | 0.142 | 0.544 | 0.071 | 0.108 |
| 676e | 0.230 | 0.251 | 1.0 | 0.427 | 0.273 | 0.284 | 0.093 | 0.098 | 0.161 | 0.092 | 0.113 |
| 676a | 0.488 | 0.533 | 0.427 | 1,0 | 0.573 | 0.527 | 0.234 | 0.185 | 0.321 | 0.268 | 0.284 |
| 671c | 0.310 | 0.630 | 0.273 | 0.573 | 1.0 | 0.255 | 0.140 | 0.234 | 0.472 | 0.083 | 0.109 |
| 653a | 0.260 | 0.236 | 0.284 | 0.527 | 0.255 | 1.0 | 0.134 | 0.094 | 0.144 | 0.221 | 0.178 |
| 643a | 0.153 | 0.132 | 0.093 | 0.234 | 0.140 | 0.134 | 1.0 | 0.132 | 0.061 | 0.035 | 0.079 |
| 641a | 0.09 | 0.142 | 0.098 | 0.185 | 0.234 | 0.094 | 0.132 | 1.0 | 0.115 | 0.025 | 0.322 |
| 632a | 0.165 | 0.544 | 0.161 | 0.321 | 0.472 | 0.144 | 0.061 | 0.115 | 1.0 | 0.047 | 0.051 |
| 552a | 0.123 | 0.071 | 0.092 | 0.268 | 0.083 | 0.221 | 0.035 | 0.025 | 0.047 | 1.0 | 0.236 |
| 541b | 0.197 | 0.108 | 0.113 | 0.284 | 0.109 | 0.178 | 0.079 | 0.322 | 0.051 | 0.236 | 1.0 |

FIGURE 6.17 – Distance statique pour l'échantillon de PCS

proches au regard de l'ordre de grandeur des autres valeurs dans le tableau, avec mesure de similarité de 0.427. Les PCS 641a et 643a semblent quant à eux assez éloignés puisque leur mesure de similarité est de 0.132. En ce qui concerne le groupe 681a, 671c et 632a, le tableau indique une forte proximité en adéquation avec ce qui était espéré. Pour le groupe 653a, 541b et 552a la similarité est assez faible de l'ordre de 0.2. Cependant, ces mesures brutes nous renseignent finalement assez peu : la valeur est difficilement interprétable si elle n'est pas très proche de 0 ou de 1.

Un moyen plus pertinent d'interpréter les résultats est regarder PCS par PCS quel sont les métiers les plus proches et comment se classent les distances. Les résultats sont présentés en Figure 6.18. En analysant ce tableau, on constate que les différents blocs de couleurs sont relativement compacts, ce qui est de bon augure : la mesure de distance conserve donc la proximité qu'on souhaiterait empiriquement voir apparaître.

| PCS de référence | 684a | 681a | 676E | 676A | 671c | 653a | 643a | 641a | 632a | 552a | 541b |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Distance croissante | 676a | 671c | 676a | 671c | 681a | 676a | 676a | 541b | 681a | 676a | 641a |
| | 681a | 632a | 653a | 681a | 676a | 676e | 684a | 671c | 671c | 541b | 676a |
| | 671c | 676a | 671c | 653a | 632a | 684a | 671c | 676a | 676a | 653a | 552a |
| | 653a | 684a | 681a | 684a | 684a | 671c | 653a | 681a | 684a | 684a | 684a |
| | 676e | 676e | 684a | 676e | 676e | 681a | 681a | 643a | 676e | 676e | 653a |
| | 541b | 653a | 632a | 632a | 653a | 552a | 641a | 632a | 653a | 671c | 676e |
| | 632a | 641a | 541b | 541b | 641a | 541b | 676e | 676e | 641a | 681a | 671c |
| | 643a | 643a | 641a | 552a | 643a | 632a | 541b | 684a | 643a | 632a | 681a |
| | 552a | 541b | 643a | 643a | 541b | 643a | 632a | 653a | 541b | 643a | 643a |
| | 641a | 552a | 552a | 641a | 552a | 641a | 552a | 552a | 552a | 641a | 632a |

FIGURE 6.18 – Classification par distance décroissante

## 6.5 En résumé

A l'issue de cette partie, nous disposons donc de deux éléments principaux :

— Une mesure de distance synthétique entre métiers qui permet de s'affranchir de la nomenclature PCS-ESE

— Une nouvelle variable agrégée représentant le métier

Lors de la construction de la mesure de distance synthétique, plusieurs points ont été mis en exergue. D'une part, la difficulté de trouver un niveau de granularité pertinent pour calculer cette mesure de distance, notamment pour la distance utilisant les transitions. A une granularité trop fine, les transitions ne sont en effet pas visibles car elles ne concernent qu'un nombre très réduit d'individus. A un niveau trop agrégé, les transitions n'existeront plus et ne seront donc pas observées.

Par ailleurs, la qualité de la variable construite est difficile à estimer, puisque nous ne disposons pas de méthode quantitative permettant de la mesurer. Les résultats sont ainsi difficiles à interpréter et il est très complexe d'estimer la pertinence de la méthode en tant que tel.

Les analyses menées sur l'échantillon de PCS permettent cependant de valider *a priori* la méthodologie : les résultats obtenus sont cohérents avec ce qui est empiriquement attendu et apportent une partie des améliorations désirées.

Afin d'approfondir l'étude des résultats, nous choisissons de mettre en oeuvre une approche opérationnelle et empirique. La variable représentant le métier, obtenue à l'issue du *clustering* est ainsi réinjectée dans un modèle de tarification afin de quantifier son impact sur la qualité du modèle, par comparaison au PCS. Si cette nouvelle variable améliore le modèle, nous pourrons ainsi conclure qu'elle capte de nouvelles informations par rapport au PCS et que sa construction est valide.

# 7 Segmentation et tarification : théorie

## 7.1 Prime pure et tarification

La tarification consiste à estimer le tarif permettant de couvrir le risque assuré par le produit d'assurance en fonction de l'individu ou du groupe d'individus. Pour ce faire, le concept le plus couremment utilisé est celui de prime pure : la prime pure correspond à l'espérance de coût annuel de la police pour l'assureur. Dans le cas d'un contrat individuel, cette espérance est calculée individu par individu ; dans le cas d'un contrat collectif elle est calculée sur l'ensemble du périmètre couvert par la police.
Le calcul de cette prime s'effectue en général sur la base des données dont dispose l'assureur, c'est-à-dire sur l'historique de son portefeuille.
On a ainsi, en notant $\Pi_i$ la prime pure pour l'individu $i$ et $\mathscr{C}_i$ le coût annuel associé :

$$\Pi_i = \mathbb{E}(\mathscr{C}_i)$$

## 7.2 Modèles de tarification

### 7.2.1 Approche fréquence-sévérité

Ce modèle consiste à supposer que la fréquence des sinistres est indépendante de la sévérité. Le coût total annuel d'une police pour l'assureur se calcule comme suit :

$$\mathscr{C} = \sum_{k=1}^{N} c_k$$

où $N$ représente le nombre de sinistres dans l'année et $c_k$ le coût du sinistre $k$.

Sous hypothèse d'indépendance entre le coût et la fréquence des sinistres (*ie* entre $N$ et $c_k$ pour tout $k$) et en supposant de plus que l'espérance du coût d'un sinistre ne dépend pas du sinistre, on a alors :

$$
\begin{aligned}
\mathbb{E}(\mathscr{C}) &= \mathbb{E}\Big(\sum_{n=1}^{\infty} c_n \mathbb{1}_{1\leq n\leq N}\Big) \\
&= \sum_{n=1}^{\infty} \mathbb{E}\big(c_n \mathbb{1}_{1\leq n\leq N}\big) \quad \text{par linéarité de l'espérance, la somme étant finie} \\
&= \sum_{n=1}^{\infty} \mathbb{E}\big(c_n\big)\mathbb{E}\big(\mathbb{1}_{1\leq n\leq N}\big) \quad \text{par indépendance et car } \mathbb{E}(c_k) = \mathbb{E}(c_1) \text{ pour tout } k \\
&= \mathbb{E}(c_1) \sum_{n=1}^{\infty} \mathbb{P}\big(n \leq N\big) \\
&= \mathbb{E}(c_1)\mathbb{E}(N)
\end{aligned}
$$

Ainsi, sous ces hypothèses, pour estimer la prime pure, il suffit d'estimer indépendemment la loi de la fréquence et la loi de sévérité des sinistres et de multiplier les espérances.

Nous choisissons cette approche pour des raisons pratiques : elle est en effet particulièrement simple à mettre en oeuvre et nous permetra d'obtenir un premier aperçu de l'influence de la nouvelle variable métier sur la tarification et la segmentation. Elle repose cependant sur des hypothèses fortes qui ne sont en pratique pas vérifiées : la fréquence d'entrée en incapacité n'est pas indépendante de la durée de celui-ci, pour des raisons de temporalité.

### 7.2.2 Modèles de régression

#### Régression linéaires simples et GLM

Le modèle de régression linéaire simple est donné par l'équation suivante :

$$\hat{y} = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n$$

ou encore, vectoriellement : $\hat{y} = \theta \bullet \mathbf{x}$
Ici, $\hat{y}$ est la variable prédite, $n$ le nombre de variables explicatives, $x_i$ la valeur de la $i$ème variable explicative et $\theta_j$ est le $j$ème paramètre du modèle.
Il permet ainsi de modéliser des relations de la forme :

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n + \varepsilon$$

avec $\varepsilon$ le terme d'erreur, le plus souvent supposé gaussien : $\varepsilon \sim \mathscr{N}\big(0,1\big)$.
Cependant, cette forme étant assez restrictive, on utilise plutôt des modèles de régression linéaire généralisée (GLM).
Dans ce cadre, en conservant les notations précédentes, on suppose une relation de la forme :

$$\sigma(y) = \sum_{i=1}^{n} \theta_i x_i + \varepsilon$$

où $\sigma$ est une fonction monotone, différentiable, inversible et telle que $\mathbb{E}(y) = \sigma^{-1}\Big(\sum_{i=1}^{n} \theta_i x_i\Big)$. Cette fonction est appelée *fonction de lien*.

**Régression logistique**

La régression logistique est un cas particulier de GLM, dans lequel l'inverse de la fonction de lien est donnée par la fonction Logit : $\sigma^{-1}(t) = \frac{1}{1+e^{-t}}$.

Cette fonction a la particuliarité de prendre ses valeurs dans l'intervalle $[0;1]$. Elle est donc particulièrement adaptée aux deux cas d'intérêt de ce mémoire : la modélisation de probabilité et la classification binaire.

En effet, lorsque la variable à prédire $y$ est une variable binaire prenant ses valeurs dans $\{0;1\}$, les modèles de régressions usuels, même généralisés, seront peu performants. Ils estiment en effet une quantité continue et non pas une quantité discrète. Au lieu de travailler sur $y$, on estime alors $p = \mathbb{P}(y = 1)$.

La variable $p$ est alors une variable continue, prenant ses valeurs dans l'intervalle $[0;1]$, que l'on peut estimer via la régression logit : $\hat{p} = \sigma^{-1}(\mathbf{x^T} \bullet \theta)$,

On retourne ensuite à $y$ via la règle de classification suivante :

$$\hat{y} = \begin{cases} 1 & \text{si } \hat{p} \geq 0.5 \\ 0 & \text{si } \hat{p} < 0.5 \end{cases}$$

La régression logistique a ainsi une double casquette : il s'agit à la fois d'un modèle de régression et d'une méthode de classification. Les manières de vérifier l'adéquation du modèle avec les données dépendent donc naturellement de l'aspect usité en pratique.

**Régression : pseudo-R2**

A l'instar du R2 pour les régressions linéaires, le pseudo-R2 permet d'évaluer l'adéquation de modèle de régression pour des variables cibles discrètes notamment. Les principaux pseudo-R2 usités sont présentés dans le tableau ci-dessous.

| Méthode | R2 |
|---|---|
| Efron | $1 - \frac{\sum(y_i - \pi_i)^2}{\sum(y_i - \bar{y})^2}$ |
| McFadden | $1 - \frac{\ln\left(\hat{\mathscr{L}}_{\text{modèle}}\right)}{\ln\left(\hat{\mathscr{L}}_{\text{nul}}\right)}$ |
| McFadden ajusté | $1 - \frac{\sum(y_i - \pi_i)^2}{\sum(y_i - \bar{y})^2}$ |
| McFadden | $1 - \frac{\ln\left(\hat{\mathscr{L}}_{\text{modèle}}\right) - K}{\ln\left(\hat{\mathscr{L}}_{\text{nul}}\right)}$ |
| Comptage | $\frac{C}{T}$ |

### 7.2.3 Classification

#### Courbe ROC

La *reciever operatting characteric curve* (courbe ROC) est la courbe représentative de la fonction qui associe le taux de vrai positifs (souvent appelé *sensibilité*) au taux de faux positifs ( qui vaut 1- *spécificité*, où la spécificité désigne le taux de vrais négatifs) dans une classification binaire. On la représente généralement avec la première bissectrice, qui correspond au cas d'un classificateur purement aléatoire.

Afin de mesurer les performances d'un classificateur, on calcule souvent l'aire sous la courbe ROC, appelée AUC (*area under curve*). Un classificateur parfait aura donc une AUC de 1 tandis qu'un classificateur purement aléatoire aura une AUC de 0.5. On compare ensuite deux modèles via leur courbe ROC, puis leur AUC : si la courbe ROC d'un modèle est toujours au-dessus de celle du second modèle, alors le premier est préférable. Dans le cas où les courbes se croisent, le modèle avec l'AUC la plus élevée est préférable.

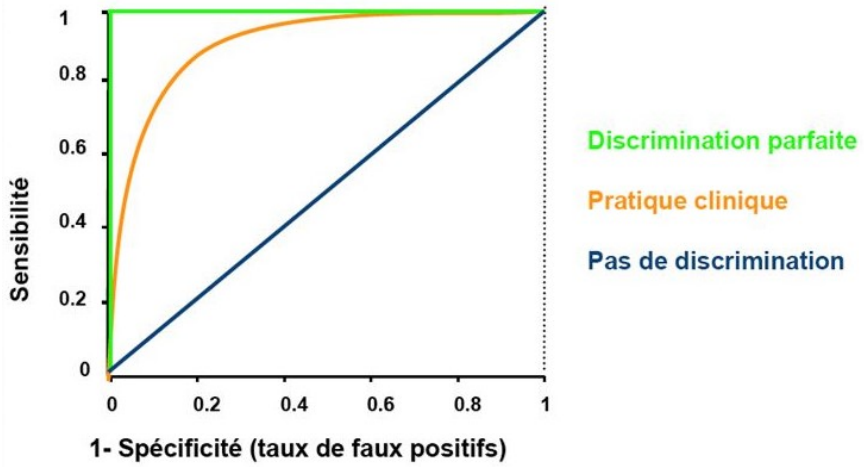


FIGURE 7.1 – Exemple de courbe ROC

Sur la Figure 7.1, la courbe verte correspond au cas d'un modèle de classification parfait. La

courbe bleue correspond quant à elle à une classification purement aléatoire.

#### Matrice de confusion

La matrice de confusion est un élément d'analyse important pour les modèles de classifications : pour une classification en $n$ classes, elle s'écrit sous la forme $(c_{i,j})_{i,j}$ avec $c_{i,j}$ le nombre d'éléments de la classe $i$ ayant été classifiés dans la classe $j$. Elle permet notamment une compréhension plus fine des erreurs du modèle et donc des améliorations ciblées.

### 7.2.4 Cadre du modèle

Dans notre cas particulier, la survenance d'un sinistre conrrespond à l'entrée en arrêt maladie et le coût d'un sinistre est fonction directe de la durée de l'absence. Dans tout ce qui suit, on ne considèrera pas de jour de carence : cette variable dépend en effet de la convention collective et du statut, ce qui rend sa prise en compte malaisée.

En outre, le but est d'estimer un coût global utilisable également par les entreprises : on considèrera donc qu'il n'y a pas de franchise et on prendra en compte les absences dès le premier jour. Ce détail technique est toutefois facilement modifiable pour des usages assurentiels via la modification des durées d'arrêt lors de la préparation des données.

### 7.2.5 Modèle tarifaire

Nous reprenons ici le modèle tarifaire envisagé dans le mémoire d'actuariat *Risque incapacité en prévoyance collective : analyse et optimisation de la segmentation tarifaire* d'Iaroslav HUBLIN. Le modèle construit est ainsi un modèle individuel, qui permettra d'obtenir le tarif collectif comme la moyenne des tarifs appliqués aux individus de la classe.

#### Hypothèses du modèle

Le taux de cotisation est défini en fonction de la probabilité d'entrée en incapacité et de la durée probable d'un arrêt. Il est fonction de l'âge, de l'ancienneté au sein de l'entrerprise, du statut conventionnel, de la région d'emploi, du genre, du code d'activité principal de l'établissement (APET), de la nature du contrat et, dans les comparatifs, le PCS au niveau 2 ou la variable catégorielle obtenue en partie 3.

La durée d'un arrêt est segmentée en fonction des durées d'intérêt pour l'entreprise : elle est donc représentée par une variable aléatoire discrète, prenant cinq modalités. On ne considère ni jours de carence, ni franchise.

Les montants ne sont pas actualisés dans cette étude, le but n'étant pas d'obtenir le tarif le plus exact possible mais de comparer les tarifs obtenus sous le prisme de la variable métier.

On suppose par ailleurs que les modalités d'indemnisations sont indépendantes des caractéristiques individuelles et correspondent au maintien du salaire complet de l'employé durant sa période d'incapacité. L'indemnité versée par la Sécurité Sociale est par ailleurs supposée dépendre uniquement du salaire, constante quelle que soit la durée d'indemnisation et valant 50 % du salaire brut. Cette dernière hypothèse ne sert qu'à pouvoir mener le calcul et n'est absolument réaliste. Elle conduit à sous-estimer la charge pour les assureurs (ou les entreprises) puisque l'indemnité versée par la Sécurité Sociale est en réalité plafonnée.
La prime pure annuelle pour un individu $i$ $\Pi_i$ vaut ainsi :

$$\Pi_i = \mathbb{P}(\mathbb{1}_A = 1) * \big(d_1(A) + P_2 d_2(A)\big) * (SAB_i - ISS_i)/365$$

avec :

- $\mathbb{1}_A$ la probabilité d'entrée en incapacité
- $d_1(A)$ la durée probable en jours d'un arrêt de moins de 90 jours
- $P_2$ la probabilité qu'un arrêt dure strictement plus de 90 jours
- $d_2(A)$ la durée probable en jours d'un arrêt de strictement plus de 90 jours
- $SAB_i$ le salaire annuel brut de l'individu
- $ISS_i$ l'indemnité versée par la Sécurité Sociale à l'individu i

L'hypothèse faite sur l'indemnité de la Sécurité Sociale permet ainsi d'exprimer la prime en pourcentage du salaire :

$$\Pi_i = \mathbb{P}(\mathbb{1}_A = 1) * \big(d_1(A) + P_2 d_2(A)\big) * SAB_i * 50\%$$

Le taux de prime devient donc indépendant du salaire, ce qui permettra de comparer les individus et donc de segmenter les tarifs.
Le calcul du taux de prime se ramène ainsi au calcul suivant :

$$¶_i = \frac{\mathbb{E}\big(\text{Durée d'arrêt en jours pour l'individu } i\big)}{365}$$

**Remarque :** Au vu de la formule utilisée, il est tout à fait possible que le taux de prime soit strictement supérieur à 1. Ce problème vient du fait que le modèle de tarification considéré est annuel : le taux de prime est donc considéré comme une fraction du salaire annuel, là où la durée d'un arrêt de travail peut être supérieure à 365 jours. Ces arrêts de très longue durée n'étant que peu représentés dans notre base, ils sont considérés comme des *outliers* et exclus.

**Estimation de la probabilité d'entrée en incapacité**

La probabilité d'entrée en incapacité $\mathbb{P}(\mathbb{1}_A = 1)$ est estimée grâce à une régression logistique, dont la formule est :

$$\mathbb{1}_A \sim \text{âge} + \text{ancienneté} + \text{genre} + \text{région} + \text{apet} + \text{statut conventionnel} + \text{nature du contat}$$

L'exposition des individus est prise en compte via une pondération mesurant la durée d'observation. La profondeur d'historique étant de deux ans, l'exposition de l'individu $i$ est représenté par le poids $w_i = \frac{\text{durée théorique de présence de l'individu en jours}}{365*2}$.

**Estimation du modèle de durée des arrêts de moins de 90 jours**

Afin de simplifier le modèle, on considère un modèle périodique et non journalier. La durée $d$ d'un arrêt, exprimée en jours, est ainsi agrégée en cinq catégories : $1 \leq d \leq 3$ ; $4 \leq d \leq 10$ ; $11 \leq d \leq 30$ ; $31 \leq d \leq 90$ et $d > 90$. On note $t_1, \ldots, t_5$ les dates $\{0, 4, 10, 30, 90\}$.
On s'intéresse dans un premier temps aux arrêts de moins de 90 jours, c'est-à-dire à l'estimation de la loi de $d_1(A)$.
On définit ensuite les variables suivantes :

$$I_{j,t_k} = \begin{cases} 1 & \text{si l'individu } j \text{ est absent à la date } t_k \\ 0 & \text{sinon} \end{cases}$$

et

$$S_{j,k} = \begin{cases} 1 & \text{si } I_{j,k} = 0 \text{ et } I_{j,t_{k-1}} = 1 \\ 0 & \text{sinon} \end{cases}$$

La variable $S_{j,k}$ représente ainsi l'indicatrice de sortie d'incapacité à l'instant $t + t_k$ , avec $t$ la date d'entrée en incapacité. Modéliser la loi de la durée d'incapacité revient ainsi à estimer la loi des $S_{j,k},\ k \in \{0, 4, 10, 30, 90\}$. On a en effet, pour l'individu $j$ :

$$d_1(A) = 3 * \mathbb{P}\big(S_{j,1} = 1 | \mathbb{1}_A = 1\big) + 10 * \mathbb{P}\big(S_{j,2} = 1\big) + 30 * \mathbb{P}\big(S_{j,3} = 1\big) + 90 * \mathbb{P}\big(S_{j,4} = 1\big)$$

La loi de chacun des $S_{j,k}$ est estimée par une régression logistique selon le modèle suivant :

$$S_{j,k} \sim \text{âge} + \text{ancienneté} + \text{genre} + \text{région} + \text{apet} + \text{statut conventionnel} + \text{nature du contat}$$

Le tableau 7.2 présente plus en détail les variables utilisées.

| Nom de la variable | Type | Caractéristiques |
|---|---|---|
| Genre | Chaîne de caractères | Deux modalités : 0 et 1 |
| Age | Numérique | Dans l'intervalle [16 ; 70] |
| Ancienneté | Numérique | |
| APET | Chaîne de caractères | 18 modalités |
| Nature Contrat | Chaîne de caractères | 3 modalités : CDD, CDI et Autres |
| Statut conventionnel | Chaîne de caractères | 6 modalités : Agents de la fonction publique, Cadres, Professions intermédiaires, Ouvriers qualifiés, Autres cadres, Employés administratifs |
| Région | Chaîne de caractères | 12 modalités |

FIGURE 7.2 – Variables utilisées dans la régression

**Estimation de la probabilité d'incapacité de plus de 90 jours**

On a :

$$P_2(A) = \mathbb{P}\big(S_{j,5} = 0 | S_{j,4} = 0 \ \& \ S_{j,3} = 0 \ \& \ S_{j,2} = 0 \ \& \ S_{j,1} = 0 \ \& \ \mathbb{1}_A = 1\big)$$

En effet, la probabilité qu'un individu soit en arrêt strictement plus de 90 jours correspond à la probabilité qu'il ne soit pas sorti d'arrêt au 90-ème jour ($S_{j,5} = 0$) sachant qu'il est absent ($\mathbb{1}_A = 1$) et qu'il n'est pas sorti d'incapacité strictement avant le 90-ème jour ($S_{j,4} = 0$ & $S_{j,3} = 0$ & $S_{j,2} = 0$ & $S_{j,1} = 0$).
Elle est également estimée via un GLM logistique.

**Estimation du modèle de durée des arrêts de plus de 90 jours**

La durée probable d'un arrêt de 90 jours est estimée via la table de provisionnement BCAC[30] et ne dépend donc que de l'âge.

## 7.3 Généralités sur la segmentation

La segmentation consiste à créer des classes d'individus afin d'obtenir des groupes de risque homogène. Cela permet un compromis entre l'approche individuelle et collective : les tarifs ne sont pas réellement individuels, mais la prime est néanmoins différenciée en fonction du risque associé à la classe. La segmentation est obtenue en fonction des variables qualitatives et quantitatives qui

caractérisent les individus. Le tarif appliqué à une classe d'assurés correspond alors à la moyenne des tarifs individuels composant cette classe.

En pratique, la segmentation s'obtient grâce à des méthodes de classification : on retrouve donc les habituels alogrithmes de *clustering* : *K-Means*, arbres de classification, etc. Le nombre de classes à retenir ainsi que la méthode utilisée pour les obtenir permettent de rendre compte de la volonté de différenciation des tarifs de l'assureur et de son aversion au risque.

### 7.3.1 Elements théoriques sur les arbres binaires de décision

Un arbre de décision est une méthode de classification classique. Le principe est de représenter les données sous forme d'arbre : chaque nœud correspond à un test simple permettant de segmenter les données. Les branches issues d'un nœud correspondent quant à elles aux différentes réponses possibles au test affecté à celui-ci. Les feuilles représentent quant à elles la segmentation finale des données.

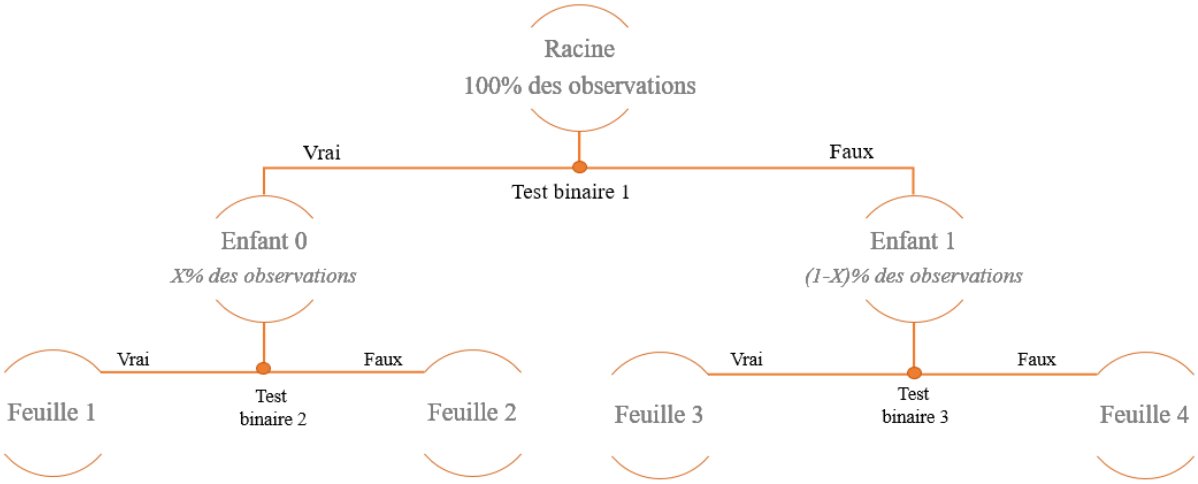


FIGURE 7.3 – Fonctionnement d'un arbre binaire de décision

Une fois l'arbre construit, on calcule le pourcentage d'individu de chaque classe dans chacune des feuilles. Pour savoir à quelle classe appartient un individu de la base de données, on parcourt l'arbre en partant de la racine jusqu'à arriver à une feuille et on lui attribue ensuite la classe majoritaire au sein de cette feuille. La méthode de construction d'arbres de *scikit-learn* ne fournit que des arbres binaires, c'est-à-dire que chaque nœud (hormis les feuilles) a exactement deux enfants : les tests sont donc nécessairement des tests binaires. En outre, chacun des tests n'utilise qu'une seule des variables disponibles. Dans le cas où toutes les variables sont numériques, l'algorithme utilise

des seuils comme tests : un test se présente donc sous la forme ($variable\ k \leq t_k$ où $t_k$ est un seuil à fixer.

Le choix de la variable $k$ et du seuil $t_k$ se fait en minimisant la fonction de coût suivante :

$$J(k,t_k) = \frac{m_0}{m}G_0 + \frac{m_1}{m}G_1$$

Pour un test binaire ayant 0 et 1 comme issues avec $m$ le nombre d'individus en entrée du test, $m_i$ le nombre d'individus dans le sous-ensemble $i$ et $G_i$ qui mesure l'impureté du sous-ensemble $i$. On cherche donc un test qui produit les sous-ensembles les plus *purs* en pondérant l'impureté d'un sous-ensemble par sa taille.

Une fois que l'ensemble a été séparé en deux, l'algorithme cherche à séparer chacun des sous-ensembles obtenus en deux, etc... L'algorithme s'arrête lorsqu'il a atteint la profondeur maximale de l'arbre (ie le nombre maximal de tests désirés) ou qu'il ne parvient plus à réduire l'impureté en effectuant de nouveaux tests.

La mesure d'impureté communément utilisée est l'indice de Gini, qui vaut :

$$G = 1 - \sum_{k=1}^{n} p_k^2$$

avec $p_k$ le ratio du nombre d'individus appartenant à la classe $k$ parmi les individus en entrée du test.

Dans le cas qui nous intéressera, on distinguera deux classes d'individus : ceux qui ont été absents durant la période d'intérêt et ceux qui n'ont jamais été absents. L'indice de Gini est défini plus généralement comme le ratio de l'aire entre la première bissectrice et courbe de Lorenz et l'aire sous la première bissectrice. La courbe de Lorenz d'une variable Y est obtenue comme la représentation graphique de la fonction qui associe à un fractile de la population associe la part de Y détenue par celui-ci. Ici, Y est le nombre total de personnes entrées en arrêt de travail. L'indice de Gini varie entre 0 et 1, 0 correspondant à une égalité parfaite et 1 à une inégalité parfaite.

Cependant, les arbres de décision ont une certaine tendance à l'*overfitting* et donnent généralement des résultats assez pauvres dès lors qu'on calcule un score de validation croisée. Les forêts aléatoires permettent de contourner ce problème.

### Forêts aléatoires

Une forêt aléatoire est un ensemble d'arbres aléatoires. Chaque arbre aléatoire est entraîné sur un échantillon de données. Le plus souvent, l'échantillonnage est effectué avec remise (*bagging sampling*). Une fois cette étape effectuée, la prédiction pour un nouvel individu est effectuée en agrégeant les prédictions obtenues pour chaque arbre : la classe prédite est le mode de toutes les prédictions obtenues.

# 8 Tarification et segmentation : pratique

Dans cette partie, nous calibrons le modèle tarifaire présenté dans la partie 7 sur nos données. Trois modèles de tarification sont en réalité construits : le premier n'intègre aucune variable concernant le métier, le second intègre le PCS et le troisième la nouvelle variable métier construite à l'issue des parties 5 et 6.

A l'issue de cette phase, une segmentation est effectuée via un arbre de décision pour chacun des modèles et leurs performances sont comparées.

La comparaison entre les trois modèles permet de mesurer l'impact de la nouvelle variable métier : si elle améliore les performances, elle permet de saisir de l'information qui n'est pas portée par le PCS. Cet impact est quantifié lors de la tarification en comparant les niveaux de risques et les tarifs obtenus et lors de la segmentation en comparant les scores *accuracy*.

## 8.1 Construction du tarif

Les résultats des différents GLM sont présentés en annexe B. Les statuts conventionnels *Ouvrier* et *Employé administratif* augmentent le risque d'occurrence d'une absence et d'absence longue, ce qui semble cohérent au moins pour les ouvriers. Cet effet est commun aux trois modèles. En ce qui concerne le PCS, les effets sont cohérents avec le statut conventionnel, notamment pour les ouvriers et employés administratifs. Les agriculteurs sont également exposés.

Les clusters de métiers semblent quant à eux regrouper des métiers avec des comportements d'absentéisme similaires.

Le premier cluster semble ainsi favoriser les métiers souvent absents mais pour des durées courtes (moins de trois jours). Ce cluster regroupe les métiers d'aide à la personne et d'éducation ; ce schéma semble donc cohérent avec la pénibilité inhérente pour des affections de courte durée.

Les métiers du deuxième cluster semblent être exposés aux absences longues et ce peu fréquemment. Il regroupe néanmoins des professions très différentes (boucher et chirurgien-dentiste…) ce qui rend les conclusions délicates.

Le troisième cluster regroupe les professions plutôt techniques ou de bureau ; le comportement d'absentéisme associé correspond à des individus jamais absents ou alors avec des absences très

courtes. Cet effet peut s'expliquer par un effet Covid : ces professions ont eu massivement accès au télétravail durant la période considérée, réduisant ainsi l'exposition et la nécessité de se déclarer en arrêt maladie.

Le quatrième cluster augmente la probabilité d'absence, pour des durées entre 4 et 10 jours et regroupe majoritairement des ouvriers. Cela est cohérent avec d'une part le facteur de pénibilité inhérent à ces métiers, d'autre part le fait que ces professions ont été plus exposées au covid que les autres, du fait de l'impossibilité du télétravail.

En ce qui concerne la nature du contrat, les différentes modalités ont des impacts très différenciés : cette variable semble donc être déterminante. Les CDI sont plus exposés que les autres, notamment pour les absences longues. Cela peut s'expliquer structurellement : les contrats courts ne peuvent connaître d'absence plus longue que la durée du contrat. Les effets sont similaires dans les trois modèles.

L'âge influe surtout sur les absences longues et augmente le risque. Cela peut se comprendre par l'augmentation du risque de maladies longues liées au vieillissement. En outre, l'âge influe négativement sur les absences courtes. Cet effet peut s'expliquer d'une part par les évolutions de carrière d'une part : en vieillissant, on accède à des positions où les arrêts courts n'ont plus nécessairement besoin d'être déclarés et d'autre part par l'allégement des contraintes familiales et notamment la disparition des absences liées à la garde des enfants [9].
Les trois modèles sont cohérents entre eux, ce qui est encourageant.

### 8.1.1 Evolution du risque par modèle

Pour le modèle sans métiers, les variables utilisées sont : le genre, l'âge, l'ancienneté, l'APET de l'entreprise, le statut conventionnel, la nature du contrat ainsi que la région.

La figure 8.1, présentée ci-dessous, représente le risque moyen en fonction des différentes variables ; l'APET est représenté sur une figure à part pour des questions de lisibilité.

Sur la figure 8.1, le risque moyen est croissant de gauche à droite. Les régions les plus à risque sont l'Ile de France et l'Occitanie et les femmes sont une population plus à risque que les hommes. En ce qui concerne le statut conventionnel, le résultat est étonnant puisque les cadres représentent la population la plus à risque tandis que les ouvriers sont les moins risqués après les agents de la fonction publique...

Dans le modèle avec le PCS, le risque est classé de manière quasiment identique au modèle sans le métier, à l'exception de la variable nature du contrat : dans le modèle sans le métier, les CDIs sont une population plus à risque que les CDDs alors que les rôles sont inversés dans la modélisation avec le PCS. Le classement du risque en fonction du PCS1 semble cohrent avec celui obtenu avec le statut conventionnel. Les résultats sont présentés en Figure 8.3.

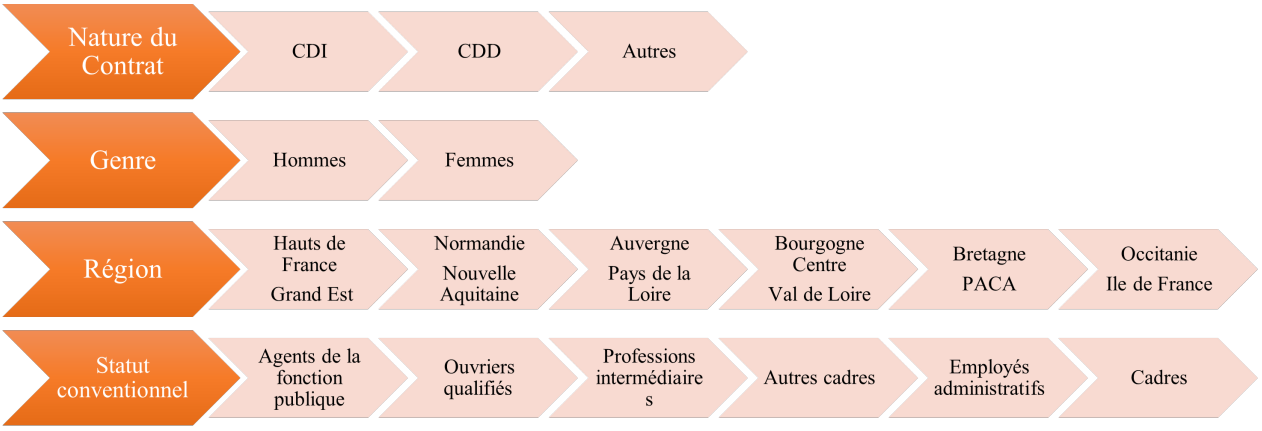


FIGURE 8.1 – Evolution du risque en fonction des variables - modèle sans le métier


FIGURE 8.2 – Evolution du risque en fonction de l'APET

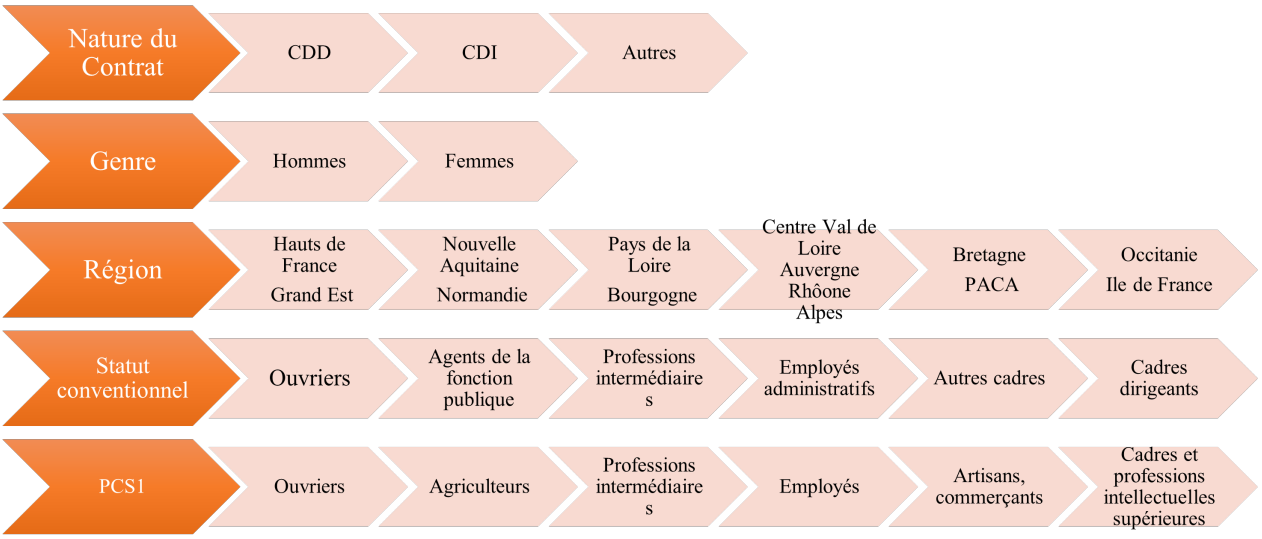


FIGURE 8.3 – Evolution du risque en fonction des variables - modèle avec pcs

Dans le modèle avec la nouvelle variable métier, l'évolution du risque en fonction des différentes variables est très semblable à celle observée dans le cas du modèle avec le PCS. Le détail est présenté en Figure 8.4.

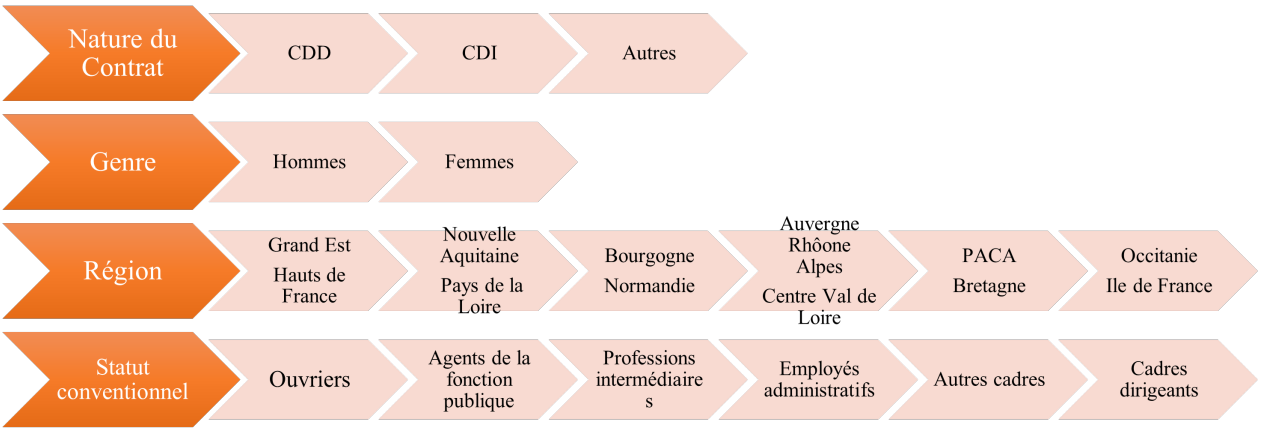


FIGURE 8.4 – Evolution du risque en fonction des variables - modèle avec la nouvelle variable métier

On observe cependant des éléments surprenants, notamment le fait que les cadres soient une population plus à risque que les ouvriers. La modalité *Cadre dirigeant* ne contient que peu d'observations et il n'est donc pas pertinent de chercher à l'analyser. En revanche, la modalité *Autres cadres* contient assez d'individus pour préseter un intérêt et demeure plus exposée que les ouvriers d'après les tarifs obtenus. Cela ne provient pas d'une erreur de modélisation mais d'un effet de portefeuille. En effet, dans les trois GLM, le coefficient associé à la modalité *Autres cadres* dans la probabilité d'être absent est négatif tandis que celui associé à la variable *Ouvrier* est positif. En ce

qui concerne la probabilité d'absence, on observe donc bien la corrélation attendue. C'est donc la distribution de la durée qui influe sur le tarif moyen. En regardant de plus près les coefficients des GLM, on remarque que le statut conventionnel *Ouvrier* influe positivement sur la probabilité que la durée d'absence soit comprise entre 31 et 90 jours ou plus de 90 jours et négativement sur le reste. La modalité *Autres cadres* a quant à elle un poids positif pour les durées de moins de 10 jours et négatif pour le reste.

La Figure 8.5 présente la répartition des absences par durée en fonction du statut conventionnel et permet ainsi de mieux comprendre les effets évoqués ci-dessus. Dans la base utilisée, les ouvriers sont en effet les plus absents, mais ils sont aussi ceux qui présentent le plus d'absences longues (de plus de 30 jours). Les cadres, même s'ils sont moins absents, ont une espérance de durée d'absence inférieure à celle des ouvriers. Or, la durée d'absence étant considérée comme indépendante de la probabilité d'absence, les longues absences sont structurellement moins probables que les absences courtes, ce qui diminue mécaniquement le tarif pour les ouvriers puisque les lois sont calibrées sur l'ensemble du portefeuille. Il s'agit donc d'un effet de mutualisation du tarif, qui justifie la nécessité d'une segmentation.

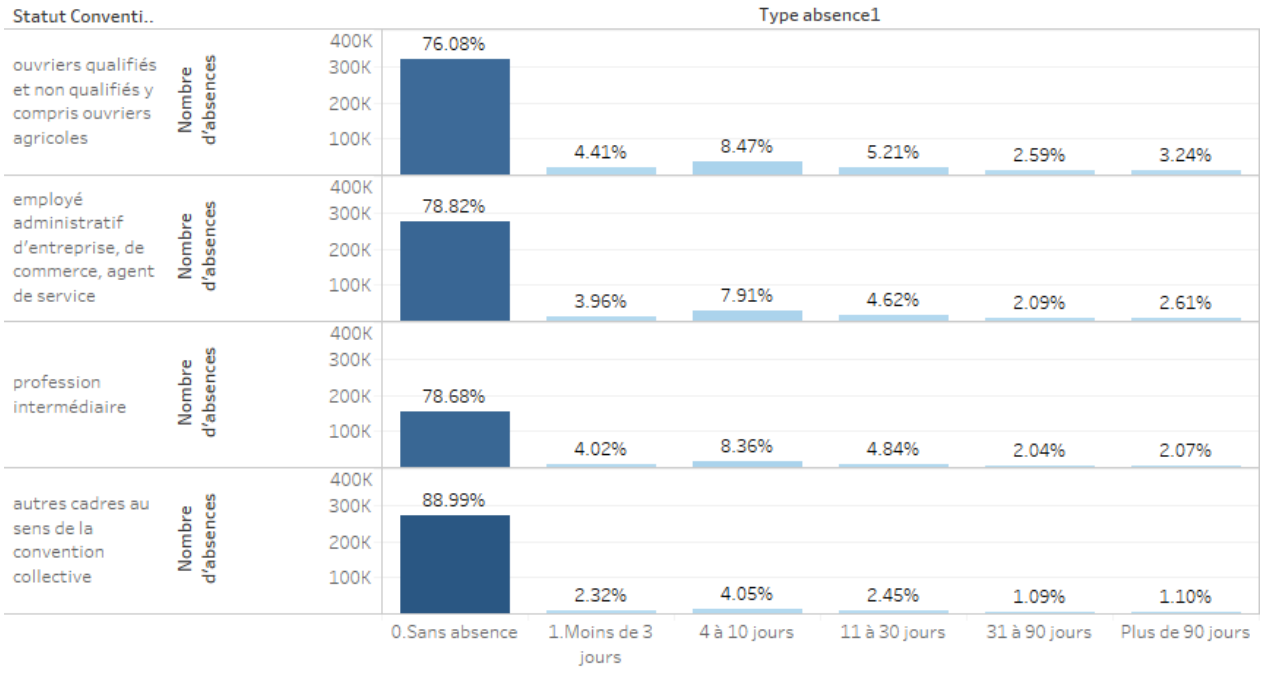


FIGURE 8.5 – Répartition des absences par durée et par statut conventionnel

Le fait que les contrats CDD soient moins risqués que les contrats CDI peut quant à lui s'expliquer par des facteurs plus sociologiques : les contrats à durée déterminée étant plus précaires que ceux à durée indéterminée, les individus en CDD s'absentent moins que les individus en CDI. L'ensemble de ces conclusions est cependant encore une fois très dépendant du portefeuille : dans

l'étude *Les absences au travail : une anayse à parttir des données françaises du Panel européen des ménages*, Chaupain-Guillot et Guillot [9] arrivent parfois à des conclusions inverses, concernant les CDI et les CDD par exemple. La Figure 8.6 présente la répartition des absences par durée en fonction de la nature du contrat. On y retrouve bien le fait que les personnes en CDD sont moins absents et pour des durées plus courtes que celles en CDI.

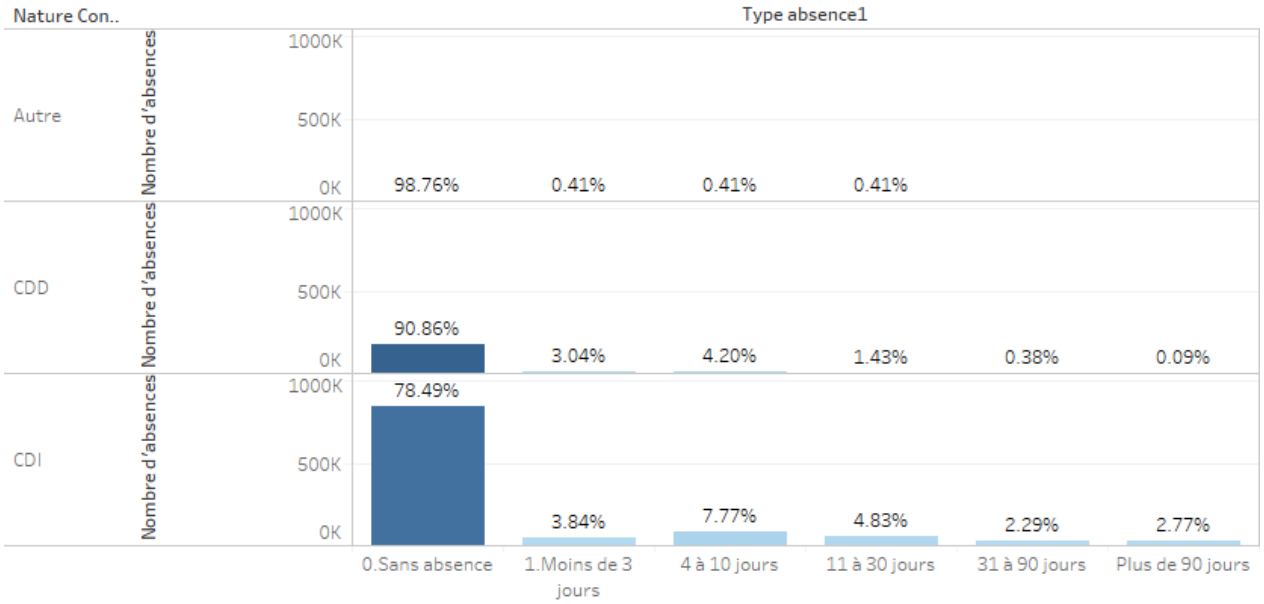


FIGURE 8.6 – Répartition des absences par durée et nature du contrat

Enfin, la nouvelle variable de métier semble séparer les individus en fonction de leur durée probable d'absence. En effet, l'un des *clusters* influe positivement sur la probabilité d'observer une durée d'arrêt et négativement sur les autres. L'influence sur la probabilité d'absence est toujours négative avec toutefois de fortes différences d'ordre de grandeur (d'un facteur 10).

### 8.1.2 Comparaison des trois modèles

Les Figures 8.7 à 8.11, présentée ci-dessous, comparent les tarifs moyens obtenus par chacune des trois modélisations.

L'évolution du tarif en fonction de l'âge (Figure 8.7) pour les modèles sans le métier et avec le PCS semble étonnante : le tarif moyen diminue fortement à partir de 25 ou 28 ans. Le pic du début de distribution peut s'expliquer par le fait que ces classes d'âge sont peu représentées au sein de la base : la modélisation est donc sensible aux éventuels outliers au sein de ces classes.

En ce qui concerne la modélisation avec la variable métier, l'évolution semble plus cohérente : la forte décroissance du tarif au-delà de 65 ans peut également s'expliquer par le fait que ces classes d'âges sont peu représentées au sein de la base.

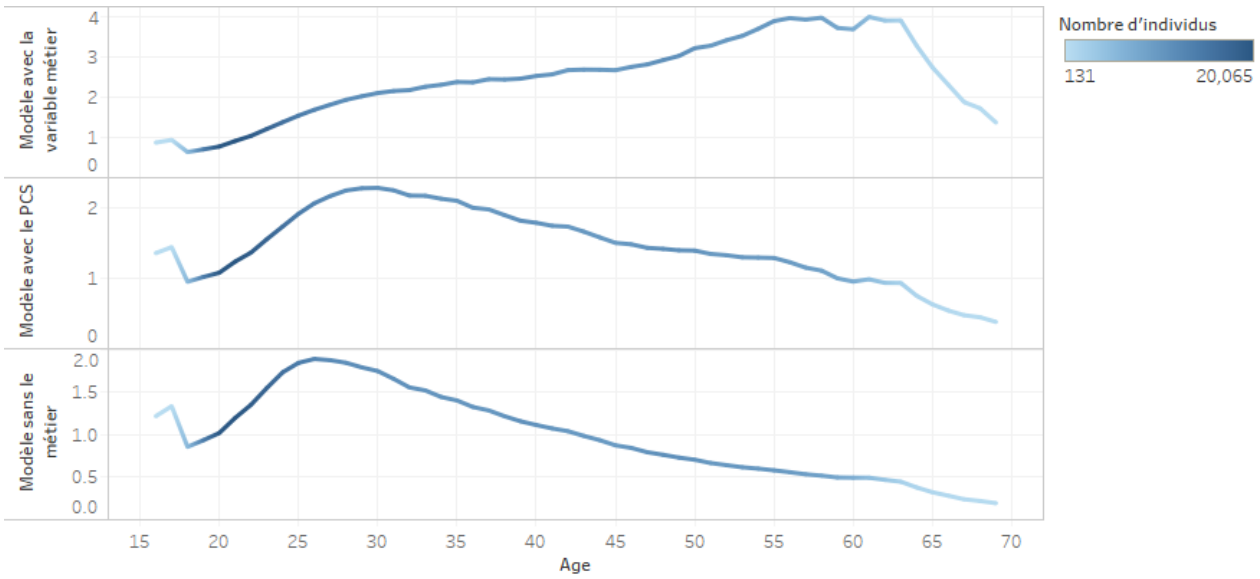


FIGURE 8.7 – Tarif moyen en fonction de l'âge

Les résultats des GLM permettent de comprendre plus en détail les effets de l'âge sur le tarif : l'âge a, dans les trois modélisations, un poids négatif dans la probabilité d'absence. Cet aspect peut se comprendre de deux manières. D'une part, en prenant en compte la proportion de cadres dans notre portefeuille : une partie importante des personnes les plus âgée sont ainsi des cadres en fin de carrière qui n'ont pas nécessairement besoin de se déclarer en incapacité lorsqu'ils sont malades. D'autre part, pour des raisons plus sociologiques, les personnes les plus âgées n'ont plus la responsabilité d'enfants en bas âge. Or, toujours d'après l'étude de Chaupain-Guillot et Guillot, ce facteur est déterminant pour expliquer les taux d'absences des 25-29 ans. En cela, nos résultats sont cohérents avec leur étude. Par ailleurs, l'âge influe positivement sur la probabilité qu'un arrêt dure plus de 10 jours et négativement sur les autres durées et il impacte fortement la probabilité d'un arrêt de plus de 90 jours. On retrouve donc une fois encore un effet de portefeuille lié à la faible probabilité des arrêts très longs. La modélisation avec la nouvelle variable métier semble limiter cet effet de portefeuille.
On peut ici noter que l'indépendance entre la probabilité d'arrêt et la durée de l'arrêt semble minimiser l'impact des arrêts longs.

Le tarif moyen est presque constant en fonction de l'ancienneté à l'exception des petites valeurs (Figure 8.8). Avant cinq années d'ancienneté, le risque est ainsi croissant en l'ancienneté (ce qui peut s'expliquer par un facteur d'exposition par exemple) ; ensuite l'ancienneté n'influe presque plus.

La nature de contrat *Autres* est la plus risquée. Elle n'est cependant pas très bien représentée au sein de la base et regroupe des modalités très différentes, ce qui induit une forte variance et explique le tarif élevé (Figure 8.9).

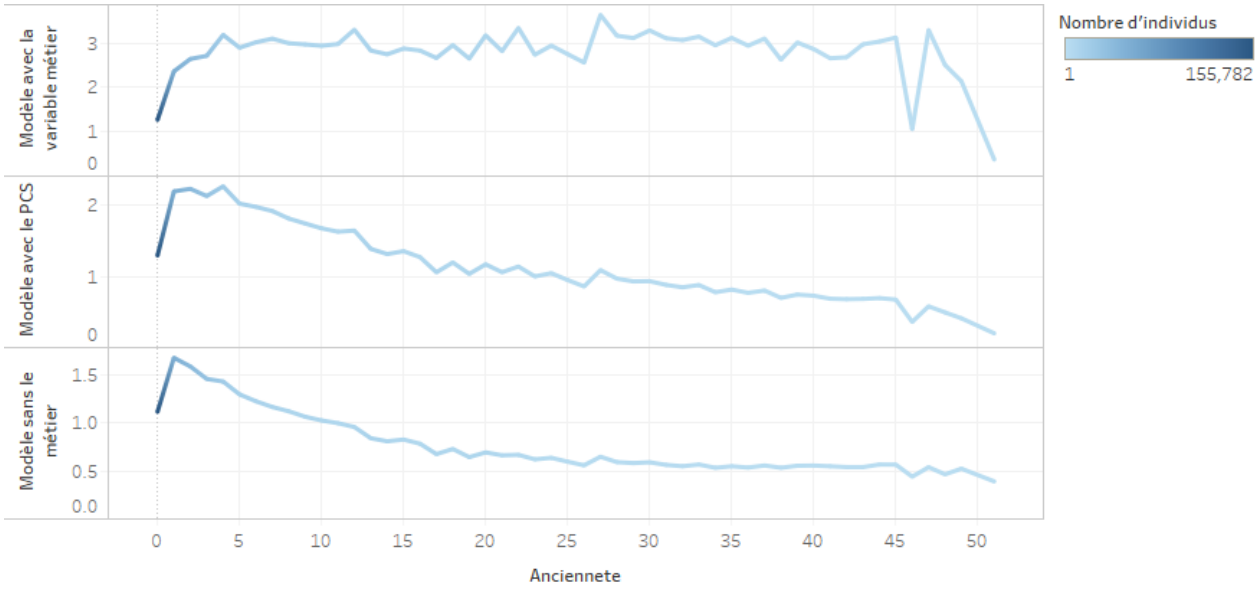


FIGURE 8.8 – Tarif moyen en fonction de l'ancienneté

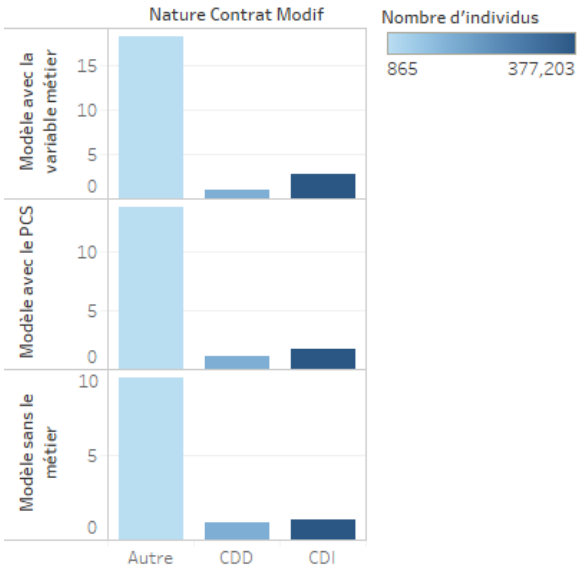


FIGURE 8.9 – Tarif moyen en fonction de la nature du contrat

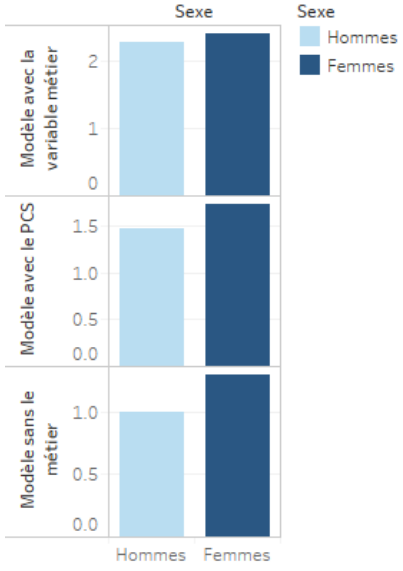


FIGURE 8.10 – Tarif moyen par genre

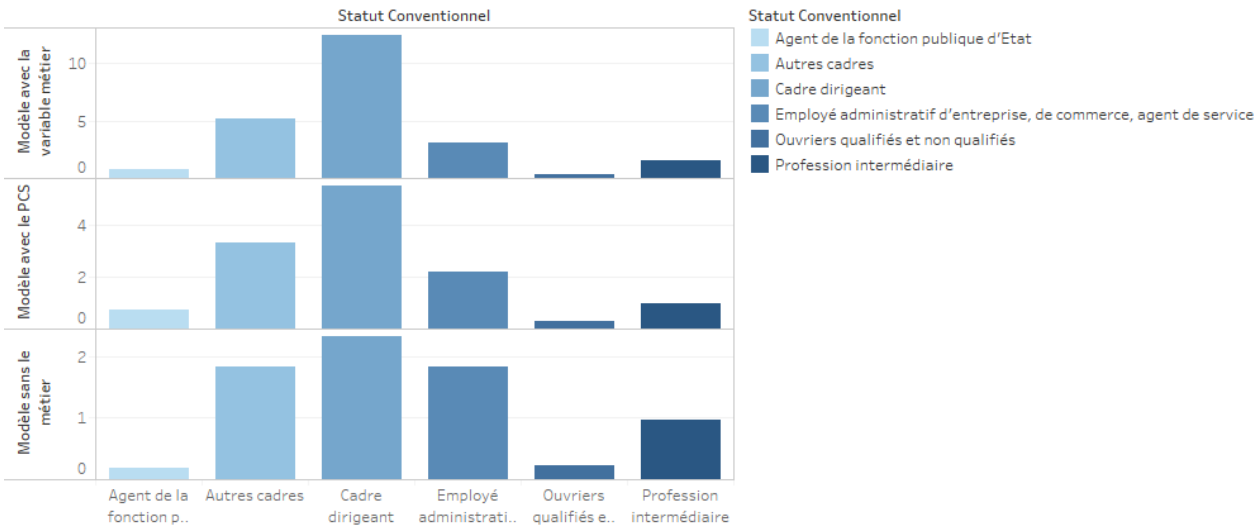


FIGURE 8.11 – Tarif moyen par statut conventionnel

Les tarifs moyens par région (Figure 8.12) suivent la même évolution dans les trois modélisations ; le tarif avec les métiers est cependant toujours plus onéreux que les deux autres.

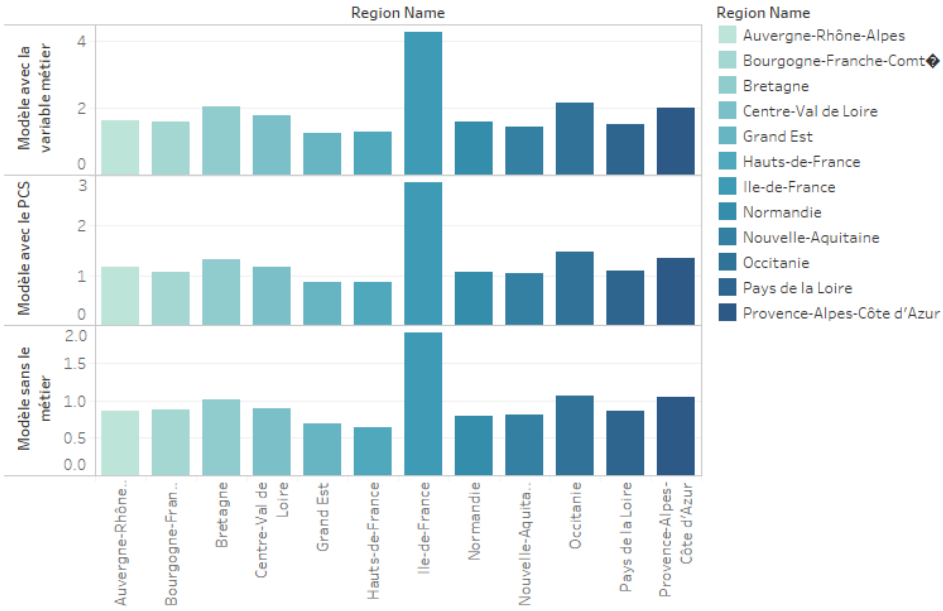


FIGURE 8.12 – Tarif moyen par région

## Conclusion

Les trois tarifs classent globalement les individus dans le même ordre de risque : l'ajout du PCS dans le modèle ne modifie donc pas fondamentalement la structure de risque du portefeuille.

Le fait que le passage du PCS à la nouvelle variable de métier conserve la hiérarchie de risque montre également que cette variable ne perturbe pas le modèle.

Par ailleurs, le tarif avec la nouvelle variable métier est toujours plus onéreux que celui avec le PCS, qui est lui-même toujours plus onéreux que celui sans le métier. Le tarif avec le métier est donc plus favorable aux bons risques que les deux autres.

## 8.2 Segmentation

La segmentation est obtenue en deux temps : un algorithme des *K*-means est tout d'abord appliqué sur les tarifs afin de déterminer les regroupements tarifaires optimaux. Une variable catégorielle correspondant à la catégorie de tarif est ainsi créée. Dans un second temps, un arbre binaire de classification est entraîné sur les données afin de prédire la classe tarifaire à laquelle un nouvel individu appartient à partir de ses caractéristiques individuelles (sans calculer son tarif individuel donc). Le tarif d'une classe est calculé comme la moyenne des tarifs associés aux individus de l'échantillon d'entraînement la composant.

### 8.2.1 Regroupement tarifaire

Ce regroupement est effectué grâce à une classification *K-Means* appliquée sur les tarifs. Les différentes *Elbow curves* permettant de sélectionner le nombre de clusters sont présentées ci-dessous. Au vu des graphiques ci-dessous, cinq classes tarifaires sont retenues pour toutes les modélisations.

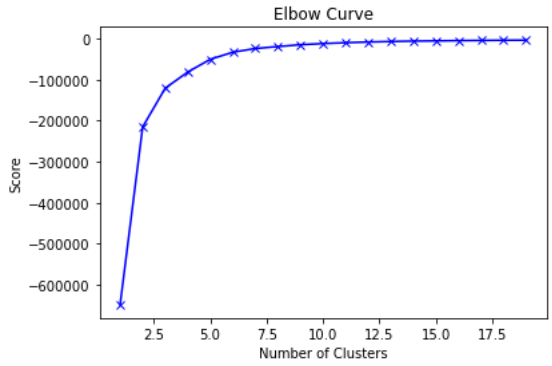


FIGURE 8.13 – Elbow curve - modèle sans variable métier

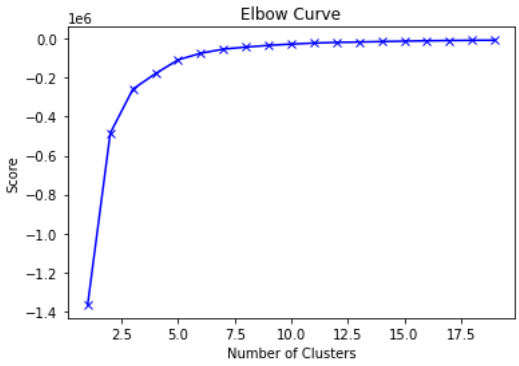


FIGURE 8.14 – Elbow curve - modèle avec PCS

Le tableau ci-dessous présente le taux de cotisation par groupe en fonction de la modélisation choisie. Les groupes ne sont cependant pas identiques en fonction de la modélisation ; il s'agit simplement de comparer les taux de cotisations obtenus.

| Modèle | Classe 1 | Classe 2 | Classe 3 | Classe 4 | Classe 5 |
|---|---|---|---|---|---|
| Sans le métier | 11.3% | 3.97% | 2.43% | 1.26% | 0.34% |
| Avec le PCS 1 | 20.1% | 5.49% | 3.16% | 1.70% | 0.411% |
| Avec la variable métier | 47.6% | 9.82% | 5.33% | 2.61% | 0.547% |

Les taux de cotisations changent ainsi drastiquement en fonction de la modélisation. L'ajout de la nouvelle variable métier et, dans une moindre mesure, l'ajout du PCS dans la modélisation

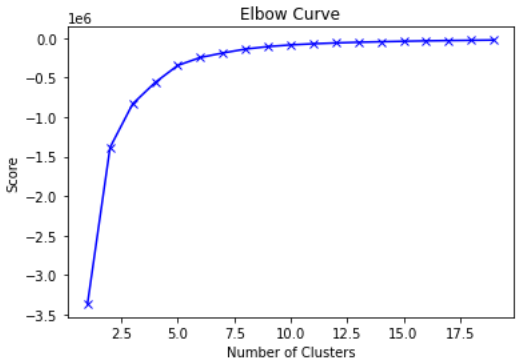


FIGURE 8.15 – Elbow curve - modèle avec la nouvelle variable métier

augmentent ainsi les différences de tarifs entre les classes : ces variables ont donc un impact fort sur la segmentation.

### 8.2.2 Arbres de décisions

La seconde étape consiste à réinjecter cette segmentation dans un arbre de décision. On peut ainsi déterminer à quelle catégorie tarifaire un individu appartient sans avoir besoin de calculer son tarif individuel. Cette approche est également intéressante car les décisions permettant d'affecter un individu à une catégorie sont explicites. Elle permet ainsi de comprendre l'importance des variables dans le processus décisionnel.

Les figures 8.16 à 8.18 présentent l'importance des différentes variables dans la construction de l'arbre. Les variables non mentionnées ont une importance nulle. L'ancienneté n'influe dans aucun des processus décisionnels : on retrouve ainsi le résultat de la partie précédente.

| Variable | Importance |
|---|---|
| SC_ouvriers qualifiés et non qualifiés | 0.566 |
| Age | 0.318 |
| Region_Ile-de-France | 0.0681 |
| SC_profession intermédiaire | 0.0481 |
| Apet_C | 4.678e-09 |

FIGURE 8.16 – Importance des variables - modèle sans variable métier

| Variable | Importance |
|---|---|
| SC_ouvriers qualifiés et non qualifiés | 0.612 |
| Age | 0.164 |
| SC_profession intermédiaire | 0.112 |
| Region_Ile-de-France | 0.0866 |
| CL_4 | 0.0255 |

FIGURE 8.17 – Importance des variables - modèle avec la nouvelle variable métier

Les variables d'intérêt sont ainsi quasimment identiques dans toutes les modélisations. Dans le cas de la modélisation avec le PCS, les modalités *Ouvriers* et *Cadres et professions intellectuelles supérieures* influent sur la modélisation. Il est cependnt intéressant de noter que, dans le cas de la modélisation avec la nouvelle variable métier, seule une des modalités influe. Le poids de cette dernière est comparable à celui des modalités *Ouvriers* et *Cadres et professions intellectuelles*

| Variable | Importance |
|---|---|
| SC_ouvriers qualifiés et non qualifiés | 0.598 |
| SC_profession intermédiaire | 0.164 |
| Region_Ile-de-France | 0.109 |
| PCS_Ouvriers | 0.0609 |
| PCS_Cadres et professions intellectuelles supérieures | 0.0346 |
| Age | 0.0328 |
| Apet_C | 2.011e-09 |

FIGURE 8.18 – Importance des variables - modèle avec PCS

*supérieures* de la modélisation avec le PCS.
Les Figures 8.19 à 8.21 présentent les arbres de décision obtenus pour chacune des modélisations.

**Modèle sans métier**

L'âge est utilisé dans le processus de décision pour les statuts conventionnels autres que les ouvriers. Les valeurs seuils utilisées sont : 32 ans, 34 ans, 43 ans et 53 ans. La matrice de confusion est donnée ci-dessous.

$$\begin{bmatrix} 48148 & 4 & 4280 & 165 & 125 \\ 16 & 4722 & 21 & 6 & 982 \\ 2930 & 762 & 15452 & 15 & 7570 \\ 4 & 81 & 16 & 1436 & 21 \\ 19 & 3712 & 2100 & 18 & 10495 \end{bmatrix}$$

Le modèle a un score *accuracy* de 0.78.

**Modèle avec PCS**

Sur ce modèle, il est intéressant de noter que les individus ayant le PCS *Ouvriers* et le statut conventionnel *Professions intermédiaires* sont classifiés dans la même classe que les individus ayant le statut conventionnel *Ouvriers qualifiés et non-qualifiés*. En revanche, les individus ayant le PCS *Ouvriers* et dont le statut conventionnel n'est pas *Professions intermédiaires* ne sont pas tous classifiés avec les individus ayant le statut conventionnel *Ouvriers*. Il semble donc que le PCS ne recoupe pas exactement réellement la variable statut conventionnel. Dans la mesure où cette dernière semble être profondément déterminante dans la classification, la création d'une autre variable métier paraît donc pertinente.

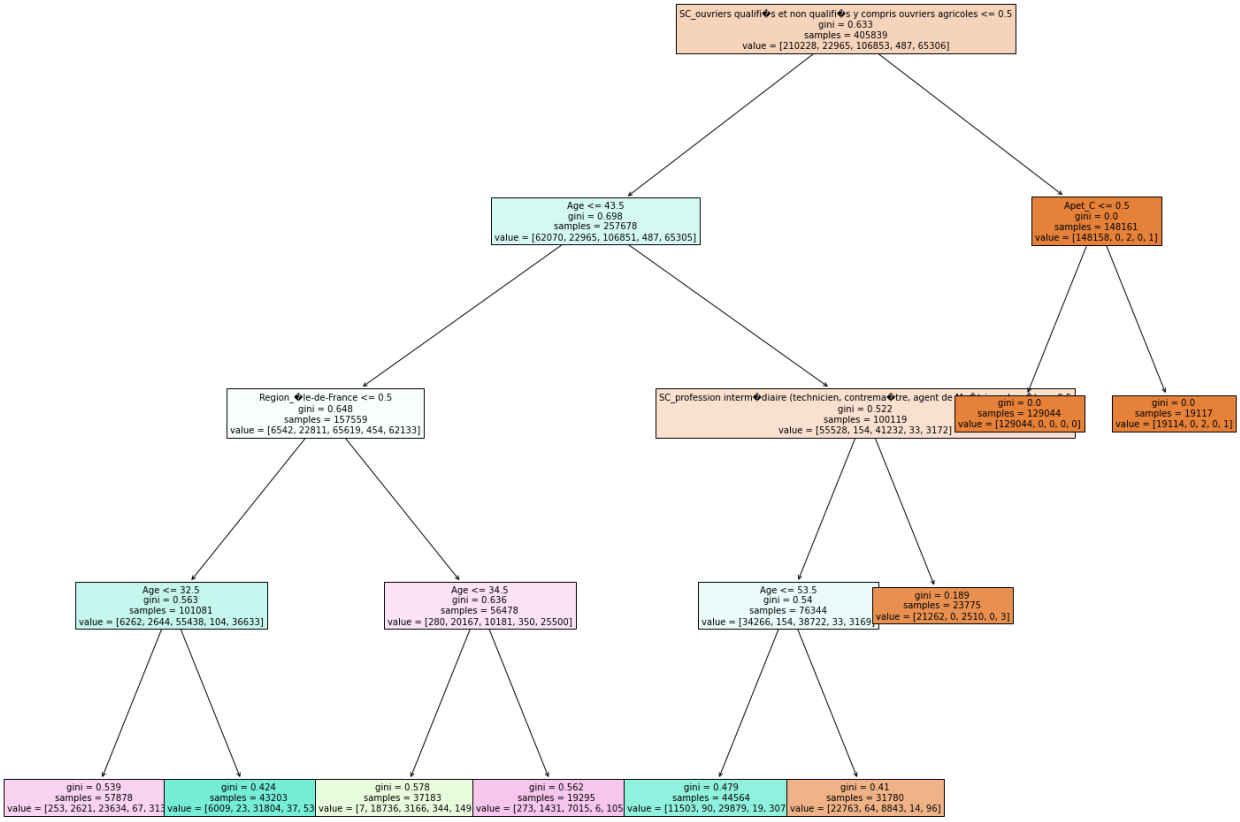


FIGURE 8.19 – Arbre de décision - modèle sans variable métier

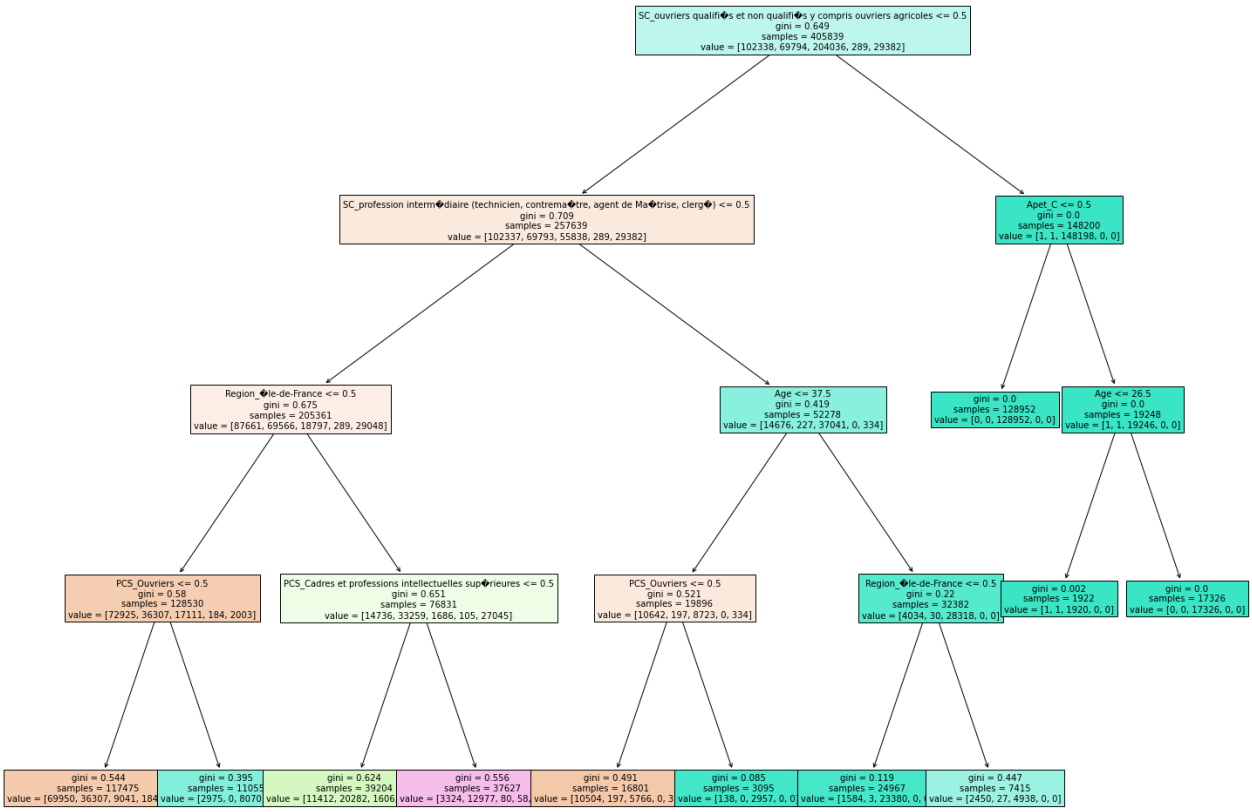


FIGURE 8.20 – Arbre de décision - modèle avec le PCS

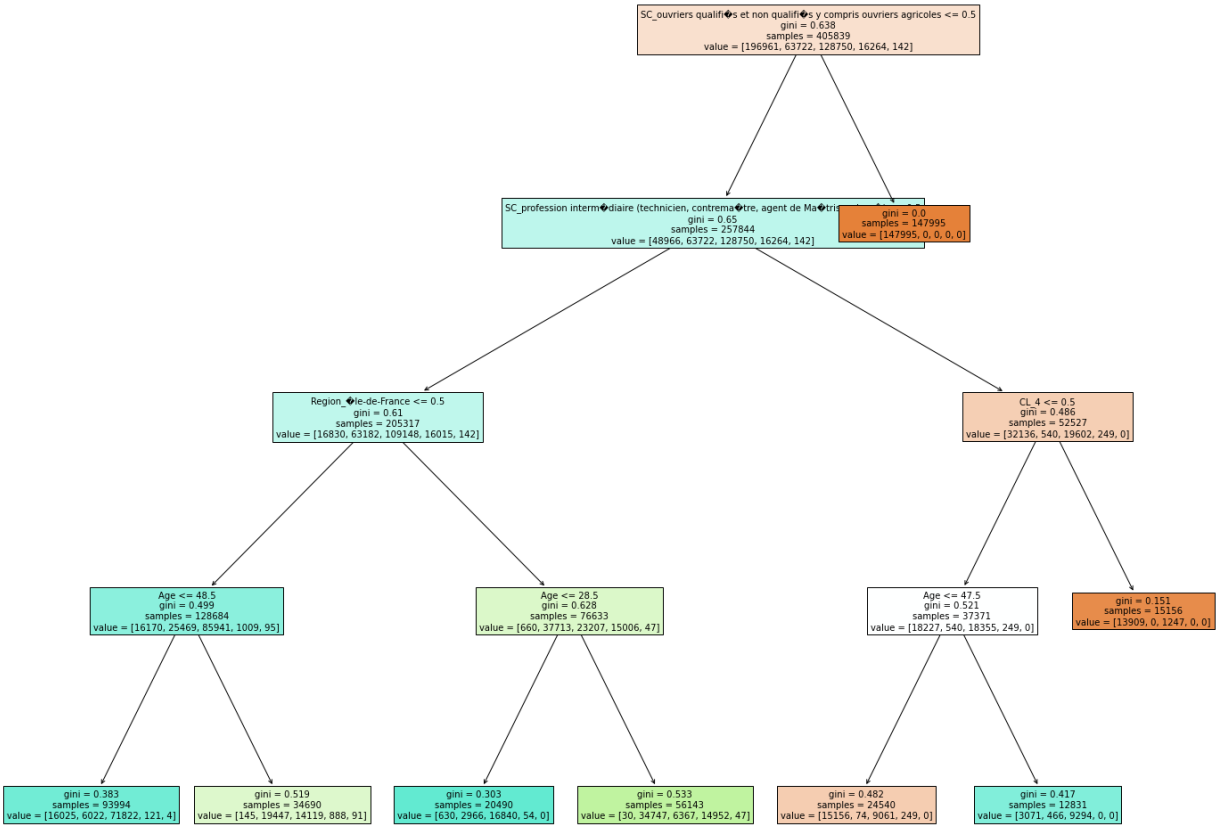


FIGURE 8.21 – Arbre de décision - modèle avec variable métier

La matrice de confusion est donnée ci-dessous.

$$\begin{bmatrix} 20125 & 2820 & 1821 & 120 & 819 \\ 9244 & 4985 & 5 & 263 & 3215 \\ 3674 & 388 & 46921 & 6 & 26 \\ 42 & 15 & 0 & 2541 & 15 \\ 559 & 1562 & 1 & 0 & 5223 \end{bmatrix}$$

Le modèle a un score *accuracy* de 0.76.

**Modèle avec variable métier**

Une seule des modalités de la nouvelle variable pèse dans le processus décisionnel. Le modèle a un score *accuracy* de 0.81.

$$\begin{bmatrix} 44429 & 46 & 4765 & 0 & 0 \\ 19 & 13501 & 2411 & 6 & 5 \\ 2497 & 5086 & 24605 & 89 & 36 \\ 60 & 3963 & 43 & 1536 & 9 \\ 0 & 34 & 1 & 63 & 12 \end{bmatrix}$$

**Conclusions**

Les trois modèles de classification donnent des résultats corrects. L'ajout de la variable PCS n'apporte cependant rien à la modélisation puisque le score produit est comparable à celui obtenu sans variable métier. L'ajout de la variable métier construite permet quant à elle d'améliorer la qualité de la modélisation : elle capte donc de l'information qui n'est pas contenue dans le PCS, ce qui était le résultat souhaité. L'approfondissement des travaux sur cette variable semble donc pertinent, afin d'en améliorer la pertinence et, indirectement, d'améliorer le modèle.

### 8.2.3 En résumé

A l'issue de cette partie, trois modèles de tarification et les segmentations associées ont été calibrés sur les données. Le premier, qui sert de *benchmark*, ne prend pas en compte la variable métier ; le second utilise le PCS pour rendre compte du métier et le troisième la variable créée dans les sections 5 et 6.

En ce qui concerne les niveaux de risques, ils sont similaires entre le deuxième et le troisième modèle, ce qui valide partiellement le processus de construction de la nouvelle variable : elle ne dénature pas l'information contenue dans le métier. En étudiant les niveaux de risques par durée

d'absence, on constate par ailleurs que la nouvelle variable discrimine bien les individus en fonction de leur durée probable d'absence.

Les trois modèles tarifaires sont également cohérents entre eux, et toujours classés dans le même ordre : le tarif avec la nouvelle variable métier est plus onéreux que le tarif avec le PCS, qui est lui même plus onéreux que le tarif sans la variable métier. Cet effet n'est, à ce jour, pas expliqué et est en cours d'investigation.

Enfin, la nouvelle variable métier influe positivement sur les performances du modèle de segmentation, qui est le niveau le plus abouti de modélisation. Elle modifie par ailleurs fortement les niveaux de cotisation par groupe, en augmentant les écarts entre les groupes : elle permet donc de mieux différencier les niveaux de risques (représentés par les taux de cotisations) au sein de la population. Les performances du modèle, si elles sont correctes, ne sont cependant pas exceptionnelles et il serait intéressant d'affiner la construction de la nouvelle variable et de vérifier si cela permet d'améliorer la qualité du modèle global.

# 9 Conclusion

## 9.1 Principaux résultats

L'objectif de ce mémoire était de questionner l'influence du métier sur l'absentéisme et donc sur la tarification des contrats de prévoyance. A l'issue de cette phase d'étude, il apparaît que la nomenclature PCS-ESE n'est effectivement pas pertinente pour décrire les liens entre les différents métiers sous le prisme de l'absentéisme, puisqu'elle dégrade (ou du moins n'améliore pas) les performances de tarification.

La construction d'une mesure de distance permettant de remplacer la distance naïve obtenue par la nomenclature était ainsi au coeur de ce projet. Les trois pistes explorées (cumul de métier, transition entre métiers et comparaison des caractéristiques individuelles) fournissent des résultats à la fois cohérents et complémentaires. Leur agrégation au sein d'une mesure de distance unique permet ainsi la prise en compte de trois points de comparaison : l'aspect statique, l'aspect dynamique et l'aspect structurel.

La modélisation de tarification associée permet de mesurer l'intérêt pratique d'une telle mesure. Elle améliore en effet les performances prédictives du modèle et semble permettre de limiter les biais de données, pregnants dans les autres modélisations. Elle revêt ainsi un intérêt plus large que le seul intérêt théorique. Des vérifications plus avancées demeurent cependant nécessaires pour confirmer le bénéfice théorique.

## 9.2 Limites et axes d'amélioration

### Pour les mesure de distance

L'approche fondée sur les transitions entre PCS présente plusieurs inconvénients. Tout d'abord, elle n'est pas symétrique : pour $i \neq j, p_{i,j} \neq p_{j,i}$.

Cet aspect est cohérent avec la réalité observable mais problématique d'un point de vue opérationnel. Dans ce mémoire, cet aspect a été corrigé en symétrisant la distance de manière artificielle : la distance considérée entre $i$ et $j$ est le maximum de $p_{i,j}$ et de $p_{j,i}$.

En outre, elle est sensible au volume de personnes occupant un PCS donné. Or, ces volumes ne sont pas fixes. En effet, des individus apparaissent ou disparaissent de la base de données au fil du temps : on ne travaille pas en système fermé. Ces variations peuvent biaiser les probabilités calculées : si à la date $t_1$ la moitié des individus exerçant un PCS $i$ donné en $t_0$ sortent de la base de données, alors la probabilité de passer de ce PCS à n'importe quel autre PCS $j$ sera inférieure à 0.5. En réalité, il ne s'agit, dans le cas général, même pas d'une mesure de probabilité : dès lors que des individus sortent du champs d'observation, pour tout PCS $i$ exercé par un des individus sortants :

$$\sum_j p_{i,j} < 1$$

Le fait qu'on n'obtienne pas une matrice stochastique n'est pas gênant tant que l'on cherche juste à mesurer une similitude entre métiers. Il empêche cependant d'aller plus loin d'un point de vue théorique en appliquant par exemple des résultats de chaînes de Markov. Il pourrait être facilement contourné en ajoutant artificiellement un PCS « OUT » et en considérant simultanément tous les individus un jour présent dans la base.

Une problématique encore plus raffinée se pose : comment traiter les individus qui sortent de la base de données pendant une période, puis entrent de nouveau ? L'ajout d'un état « Out » permet de régler ce problème mais agit comme une boîte noire qui gomme le passif de l'individu. La solution retenue est de considérer qu'un individu qui sort de la base puis y revient a conservé les mêmes métiers qu'à son départ durant la période manquante mais des méthodes de complétion des données d'enquêtes de panels sont étudiées.

De plus, si l'on souhaitait développer un modèle markovien, il faudrait trouver un moyen de modéliser le fait qu'une transition entre métier est plus significative que le fait de conserver le même métier : on ne travaille pas réellement dans un univers uniforme. Reste à savoir comment calibrer cette différence...

Par ailleurs, les approches présentées dans ce mémoire s'affranchissent totalement de la problématique de l'absentéisme, ce qui est questionnable. Le risque est ainsi de retomber dans le travers reproché à la distance de graphe, à savoir rapprocher des métiers qui présentent des comportements d'absentéisme très différents. Cependant, prendre en compte le comportement d'absentéisme dans la distance injecterait la variable à prédire dans les variables explicatives et induirait ainsi un biais qu'il serait très difficile de quantifier.

Enfin, la pertinence de cette mesure reste à évaluer. En effet, les premiers tests, à savoir le *backtesting* sur des PCS particuliers et le fait qu'elle améliore les prédictions dans le modèle de segmentation, laissent penser que cette distance est cohérente et pertinente. Cependant, avant de l'utiliser à une plus grande échelle, il faut pouvoir quantifier si l'apport de cette distance en termes de modélisation justifie le coût de production. La mise en oeuvre de ces vérifications est en cours.

## Pour la calibration des modèles

Plusieurs limites théoriques demeurent dans les modèles calibrés. La première concerne l'hypothèse faite lors de la construction du tarif sur l'indemnité versée par la Sécurité Sociale. Il est en effet très approximatif de supposer qu'elle est identique pour tout le monde en proportion de la rémunération. Pour pallier cela, il faudrait utiliser des bases de données permettant d'estimer la rémunération moyenne par PCS (par exemple) et ainsi modéliser correctement l'indemnité versée par la Sécurité Sociale. Cet aspect est d'autant plus ennuyeux qu'il conduit à une sous-estimation du risque pour l'assureur.

Une autre limite du modèle concerne la modélisation des durées. Cette modélisation par tranche permet en effet d'obtenir une première approximation mais demeure assez grossière. En outre, la question du découpage des périodes se pose : il est pour l'instant totalement arbitraire et fondé uniquement sur les besoins de l'entreprise. La dernière tranche de durée mériterait par ailleurs d'être subdivisée : entre trois mois et trois ans d'absence, les comportements ne sont probablement pas identiques... De plus, la durée d'absence est supposée indépendante du fait d'être absent et de l'historique de l'individu. Une piste d'amélioration du modèle serait ainsi d'inclure l'historique d'absence des individus, en ajoutant par exemple une variable indiquant si l'individu a ou non été absent le mois précédent.

Par ailleurs, nous avons utilisé le temps théorique de présence dans l'entreprise pour représenter l'exposition des individus. Or, cette pondération est corrélée au profil des individus. Il serait donc nécessaire d'explorer davantage les liens entre les deux variables afin de quantifier le biais induit. La question de la pertinence de cette dernière se pose également : nous aurions pu choisir un poids d'exposition mesurant l'ancienneté dans la base de données et donc l'exposition annuelle au risque d'incapacité. A défaut, nous pourrions également ne conserver dans la base que les individus présentenant une profondeur d'historique semblable et comparable.

L'ensemble de ces éléments permettrait d'approfondir les différences observées entre la tarification individuelle, plutôt attendue, et les tarifs moyens plutôt inattendus quant à eux et de comprendre si le problème provient de la construction du modèle de durée ou bien est intrinsèquement lié au portefeuille.

Enfin, ce modèle ne se préoccupe pas des coûts indirects de l'absentéisme ni des éventuels effets de masse. La prédiction donne en effet une espérance de sinistralité individuelle, que l'on peut agréger au niveau de l'entreprise. Elle ne prend cependant pas du tout en compte l'impact de l'absence simultannée de plusieurs individus alors que cela peut avoir des conséquences fortes sur l'activité de l'entreprise voire favoriser de manière indirecte l'absentéisme au sein d'autres équipes.

Malgré les nombreuses questions et problématiques qui demeurent, cette étude fournit une nouvelle étape vers une segmentation plus pertinente du portefeuille en fonction des ressemblances.

Elle se positionne ainsi comme une base de réflexion pour développer un modèle de scoring plus adéquat concernant le risque volumétrique de l'absentéisme dans les contrats prévoyance.

# A Description des données

## A.1 Base des professions et catégories socioprofessionnelles (PCS)

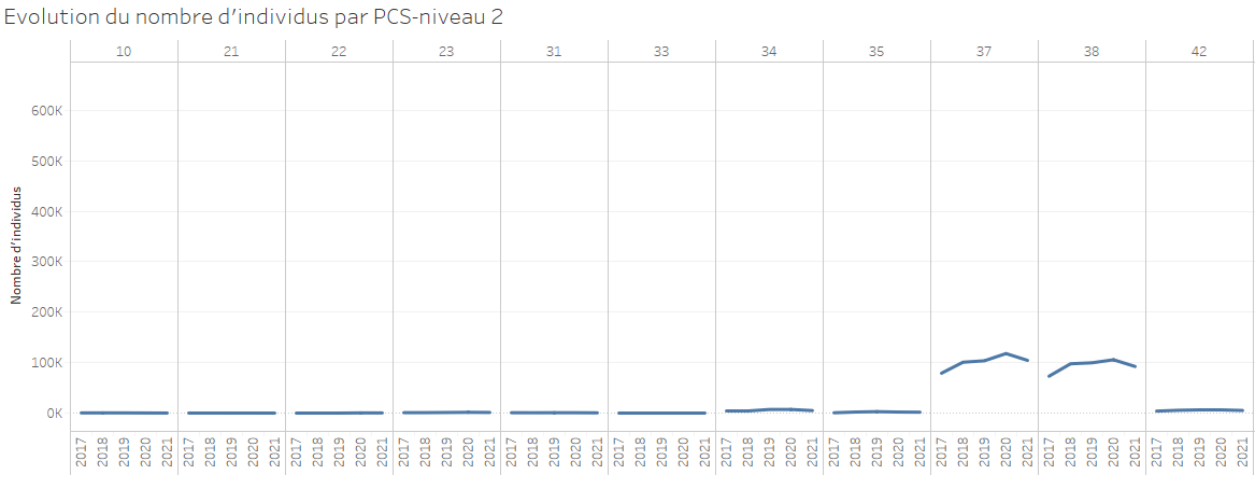


FIGURE A.1 – Evolution du volume d'individus par PCS au niveau 2 - Partie 1

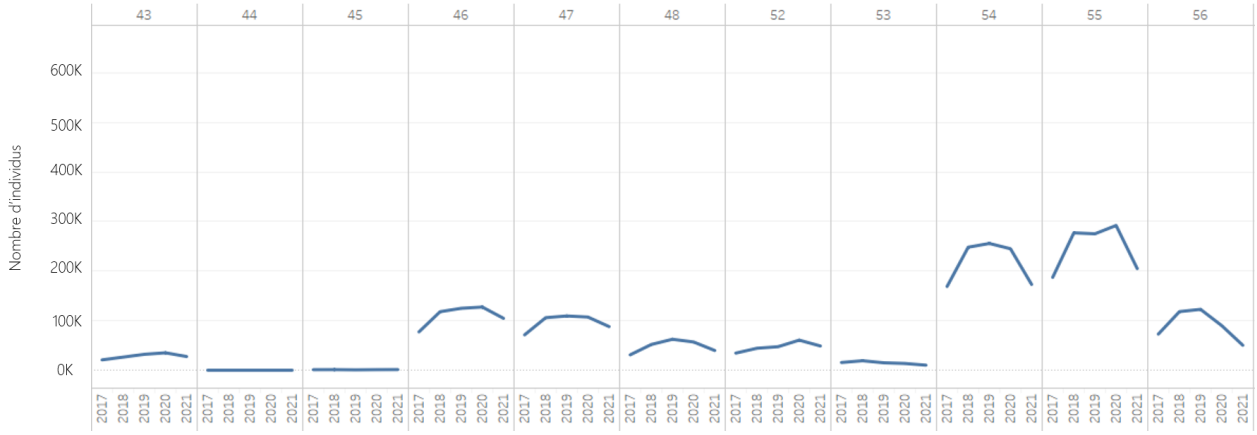


FIGURE A.2 – Evolution du volume d'individus par PCS au niveau 2 - Partie 2

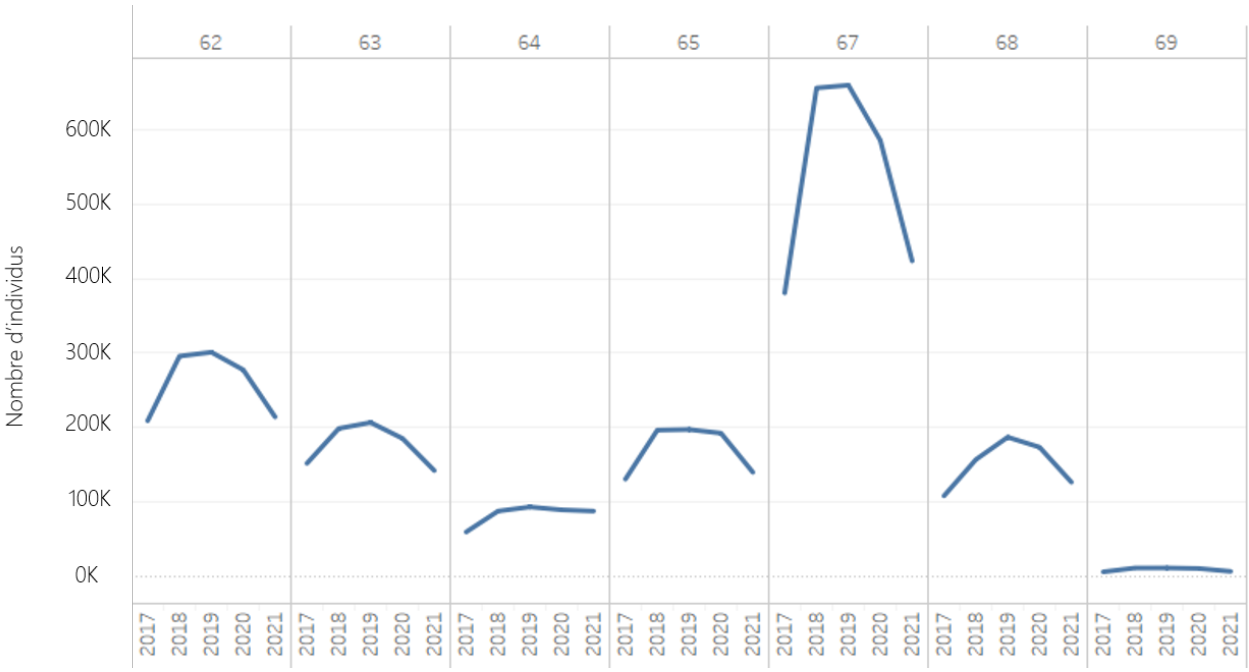


FIGURE A.3 – Evolution du volume d'individus par PCS au niveau 2 - Partie 3

## A.2 Base Absences

| Nature Contrat | |
|---|---|
| Autre nature de contrat, convention, mandat | 518 |
| Contrat de travail à durée déterminée de droit privé | 187,814 |
| Contrat de travail à durée indéterminée de Chantier ou d'opération | 56 |
| Contrat de travail à durée indéterminée de droit privé | 1,085,021 |
| Contrat de travail à durée indéterminée de droit public | 597 |
| Contrat à durée indéterminée intermittent | 2 |
| Convention de stage (hors formation professionnelle) | 781 |
| Mandat social | 286 |
| [FP] Détachement d'un agent d'une Fonction Publique donnant lieu à pension (ECP) | 2 |
| [FP] Détachement d'un agent d'une Fonction Publique ne donnant pas lieu à pension (ENCP) | 4 |

FIGURE A.4 – Répartition des individus par nature du contrat

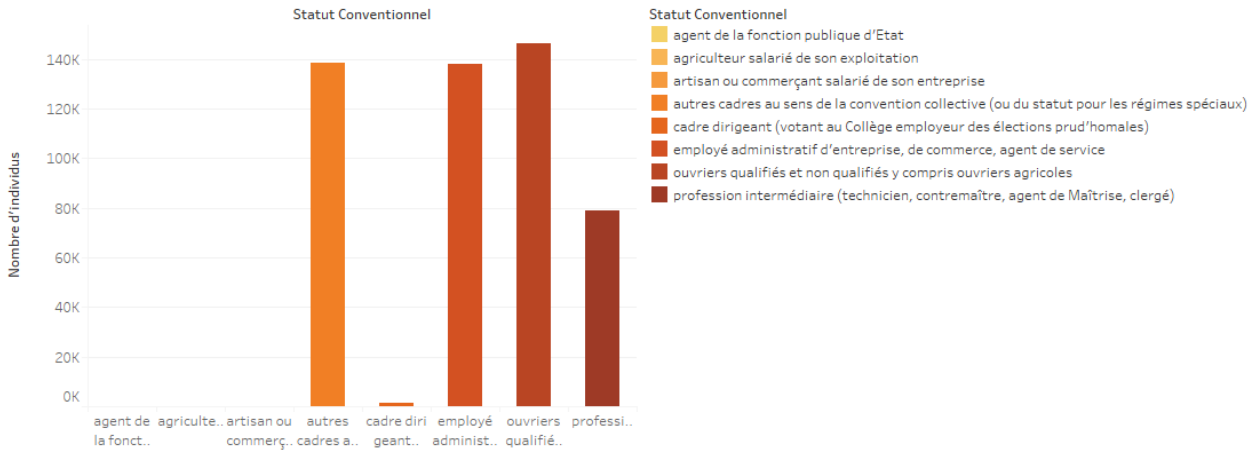


FIGURE A.5 – Répartition des individus par statut conventionnel

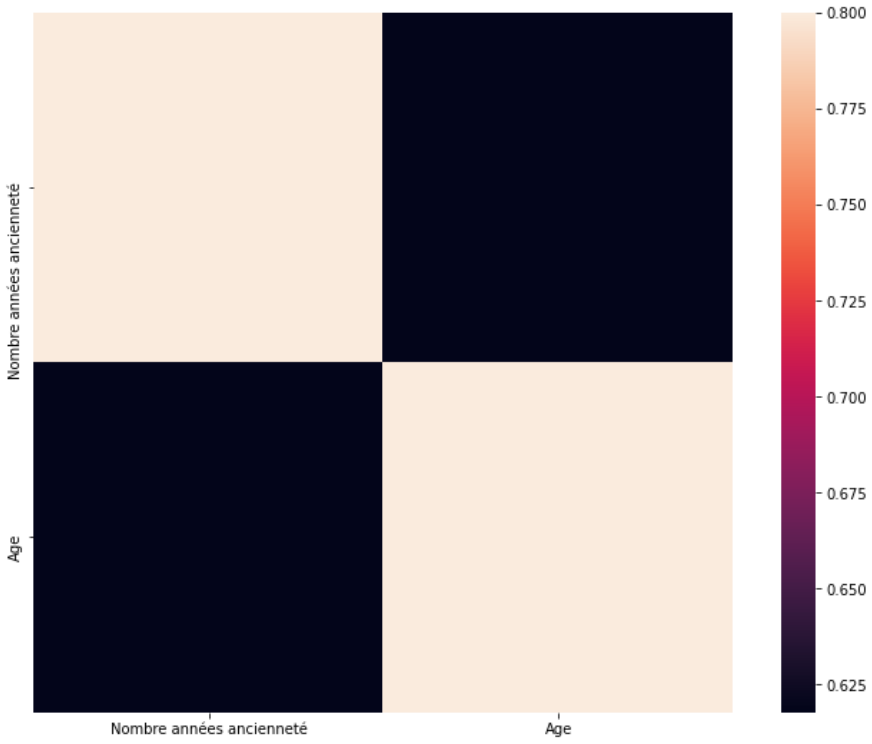


FIGURE A.6 – Corrélation entre l'âge et l'ancienneté

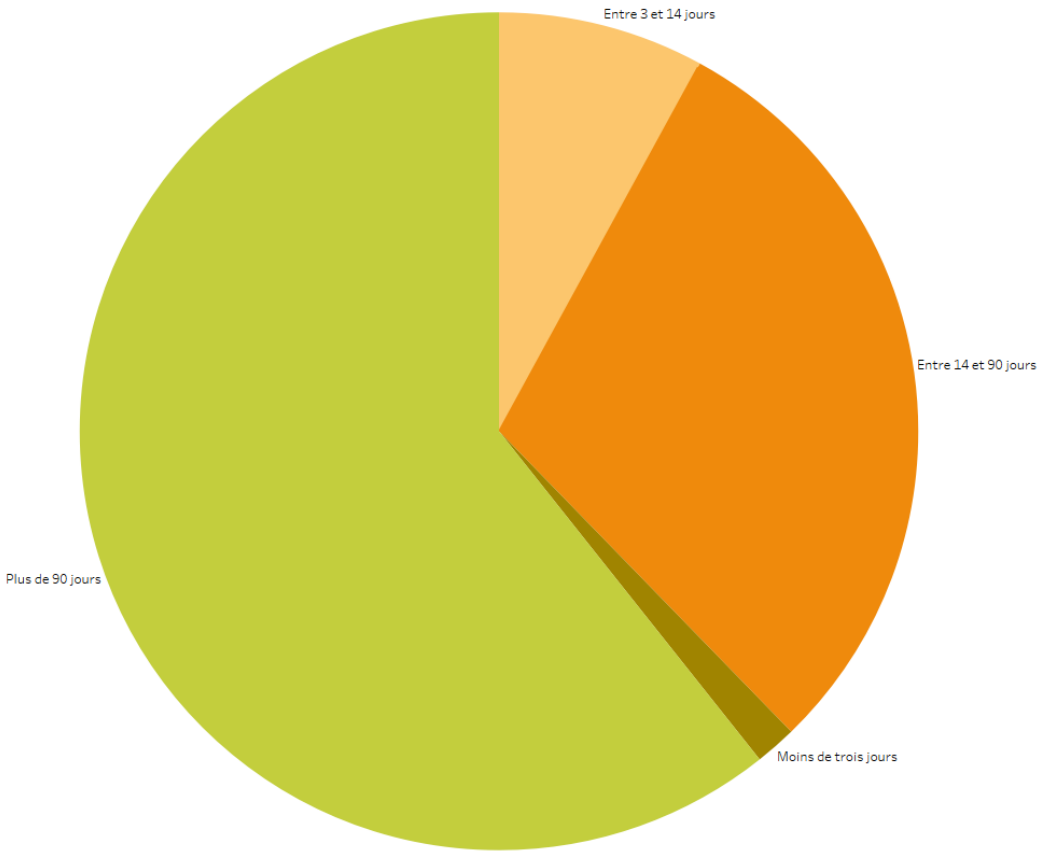


FIGURE A.7 – Répartition des absences par durée (volume en nombres de jours)

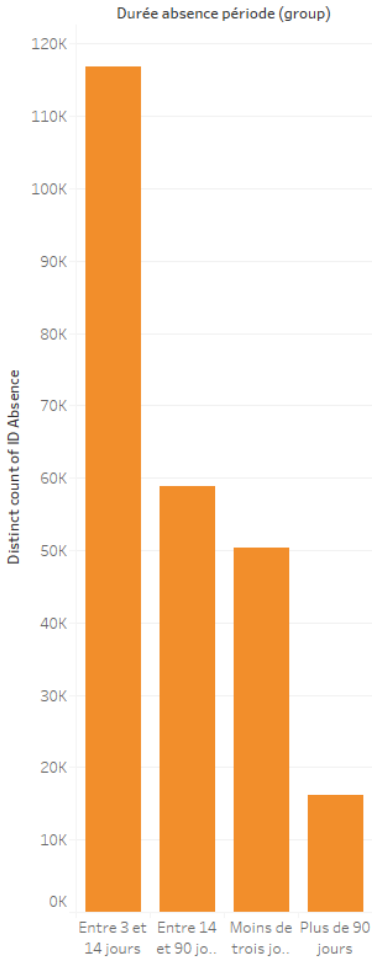


FIGURE A.8 – Répartition des absences par durée (volume en nombre d'absences)

# B Résultats des GLM

# B.1 Coefficients du modèle sans le métier

| Variable | Ind_absence | Moins_3 | 4_10 | 11_30 | 31_90 | Plus_90 |
|---|---|---|---|---|---|---|
| **Intercept** | -1.182 | -1.272 | -0.525 | -0.533 | -2.535 | -3.083 |
| **SC_agent de la fonction publique d'Etat** | -0.037 | 0.758 | -0.242 | -0.123 | -0.122 | -0.194 |
| **SC_autres cadres** | -0.676 | 0.014 | 0.012 | -0.084 | -0.088 | -0.082 |
| **SC_cadre dirigeant** | -1.323 | -0.182 | -0.021 | 0.090 | 0.247 | -0.398 |
| **SC_employé administratif** | 0.318 | -0.383 | 0.014 | -0.008 | -0.073 | 0.182 |
| **SC_ouvriers qualifiés et non qualifiés** | 0.560 | -0.324 | -0.104 | -0.073 | 0.013 | 0.320 |
| **SC_profession intermédiaire** | 0.091 | -0.157 | 0.019 | -0.034 | -0.131 | 0.014 |
| **NC_Autre** | -2.479 | -0.222 | -1.001 | 1.131 | -0.333 | -0.375 |
| **NC_CDD** | 0.325 | 0.182 | 0.426 | -0.858 | -0.265 | -0.903 |
| **NC_CDI** | 1.088 | -0.233 | 0.253 | -0.505 | 0.444 | 1.119 |
| **Sexe_1.0** | -0.664 | -0.149 | -0.083 | -0.127 | -0.129 | -0.170 |
| **Sexe_2.0** | -0.403 | -0.124 | -0.239 | -0.104 | -0.025 | 0.012 |
| **Apet_B** | -0.646 | 0.245 | -0.560 | 0.590 | -0.096 | -0.347 |
| **Apet_C** | 0.094 | 0.098 | 0.173 | 0.034 | -0.011 | -0.508 |
| **Apet_D** | -0.083 | -0.169 | 0.063 | -0.011 | 0.082 | 0.044 |
| **Apet_E** | 0.558 | -0.152 | 0.337 | -0.120 | -0.131 | -0.298 |
| **Apet_F** | 0.170 | 0.211 | -0.019 | -0.231 | 0.015 | 0.042 |
| **Apet_G** | 0.030 | 0.092 | -0.095 | -0.040 | 0.107 | 0.053 |
| **Apet_H** | 0.035 | 0.088 | 0.016 | -0.049 | 0.011 | 0.005 |
| **Apet_I** | 0.480 | -0.013 | 0.126 | -0.123 | -0.042 | 0.026 |
| **Apet_J** | -0.077 | 0.231 | -0.007 | -0.140 | -0.085 | -0.094 |
| **Apet_K** | -0.130 | 0.298 | -0.111 | -0.074 | -0.122 | -0.244 |
| **Apet_L** | -0.320 | -0.306 | 0.057 | -0.089 | 0.240 | 0.235 |
| **Apet_M** | -0.691 | -0.253 | -0.263 | -0.090 | 0.131 | 0.510 |
| **Apet_N** | -0.085 | 0.354 | -0.115 | -0.219 | -0.019 | 0.081 |
| **Apet_O** | 0.374 | -0.055 | 0.074 | 0.099 | -0.011 | -0.201 |
| **Apet_P** | -0.042 | -0.291 | -0.174 | 0.308 | -0.180 | 0.316 |
| **Apet_Q** | 0.255 | 0.001 | 0.076 | -0.157 | -0.251 | 0.285 |
| **Apet_R** | -0.755 | -0.241 | 0.144 | 0.209 | 0.134 | -0.603 |
| **Apet_S** | -0.232 | -0.411 | -0.045 | -0.127 | 0.074 | 0.540 |
| **Region_Auvergne-Rhône-Alpes** | -0.066 | 0.032 | -0.058 | 0.058 | 0.097 | -0.133 |
| **Region_Bourgogne-Franche-Comté** | -0.103 | 0.106 | -0.021 | -0.005 | -0.019 | -0.133 |
| **Region_Bretagne** | -0.341 | 0.383 | -0.221 | -0.071 | 0.103 | -0.026 |
| **Region_Centre-Val de Loire** | -0.094 | 0.482 | -0.134 | -0.016 | -0.070 | -0.332 |
| **Region_Grand Est** | 0.084 | 0.439 | -0.059 | -0.068 | -0.044 | -0.340 |
| **Region_Hauts-de-France** | 0.123 | 0.210 | -0.085 | 0.023 | 0.006 | -0.128 |
| **Region_Normandie** | -0.115 | 0.177 | -0.030 | -0.008 | 0.076 | -0.232 |
| **Region_Nouvelle-Aquitaine** | -0.141 | 0.497 | -0.210 | -0.122 | 0.043 | -0.118 |
| **Region_Occitanie** | -0.103 | 0.269 | -0.153 | -0.046 | 0.080 | -0.059 |
| **Region_Pays de la Loire** | -0.102 | 0.515 | -0.065 | -0.191 | -0.074 | -0.236 |
| **Region_Provence-Alpes-Côte d'Azur** | -0.077 | 0.020 | -0.179 | 0.099 | 0.058 | 0.049 |
| **Region_Ile-de-France** | -0.131 | 0.348 | -0.046 | 0.009 | -0.103 | -0.327 |
| **Anciennete** | 0.015 | -0.025 | 0.011 | -0.002 | -0.038 | 0.037 |
| **Age** | -0.030 | -0.404 | -0.248 | 0.072 | 0.282 | 0.638 |

# B.2 Coefficients du modèle avec le PCS

| Variable | Ind_absence | Moins_3 | 4_10 | 11_30 | 31_90 | Plus_90 |
|---|---|---|---|---|---|---|
| **Intercept** | -1,068 | -0,731 | -0,422 | -0,543 | -2,591 | -3,399 |
| **PCS_Agriculteurs** | 0,103 | 0,316 | 0,110 | -0,460 | 0,580 | -0,511 |
| **PCS_Artisans, commerçants et chefs d'entreprises** | -0,813 | 0,318 | -0,321 | 0,549 | -1,293 | -0,376 |
| **PCS_Cadres et professions intellectuelles supérieures** | -0,472 | -0,215 | 0,001 | -0,068 | 0,058 | 0,157 |
| **PCS_Employés** | 0,128 | -0,089 | -0,059 | -0,099 | 0,109 | 0,104 |
| **PCS_Ouvriers** | 0,217 | -0,425 | 0,007 | -0,065 | 0,200 | 0,298 |
| **PCS_Professions intermédiaires** | -0,123 | -0,219 | 0,001 | -0,153 | 0,153 | 0,239 |
| **SC_agent de la fonction publique d'Etat** | 0,048 | 0,580 | -0,205 | -0,120 | -0,115 | -0,128 |
| **SC_autres cadres au sens de la convention collective** | -0,345 | 0,029 | 0,022 | -0,126 | -0,065 | -0,064 |
| **SC_cadre dirigeant** | -1,138 | -0,151 | -0,104 | 0,113 | 0,226 | -0,332 |
| **SC_employé administratif d'entreprise, de commerce** | 0,115 | -0,379 | 0,055 | -0,032 | -0,083 | 0,174 |
| **SC_ouvriers qualifiés et non qualifiés** | 0,301 | -0,246 | -0,073 | -0,097 | -0,009 | 0,278 |
| **SC_profession intermédiaire** | 0,058 | -0,145 | 0,043 | -0,033 | -0,147 | -0,015 |
| **NC_Autre** | -2,205 | -0,329 | -0,473 | 0,735 | -0,301 | -0,208 |
| **NC_CDD** | 0,237 | 0,213 | 0,185 | -0,665 | -0,299 | -0,944 |
| **NC_CDI** | 1,006 | -0,197 | 0,027 | -0,366 | 0,407 | 1,064 |
| **Sexe_1,0** | -0,617 | -0,153 | -0,056 | -0,159 | -0,154 | -0,150 |
| **Sexe_2,0** | -0,344 | -0,161 | -0,205 | -0,137 | -0,039 | 0,062 |
| **Apet_B** | -0,625 | 0,116 | -0,347 | 0,466 | -0,284 | -0,057 |
| **Apet_C** | 0,060 | 0,140 | 0,162 | 0,034 | -0,025 | -0,583 |
| **Apet_D** | -0,027 | -0,215 | 0,088 | 0,061 | 0,045 | -0,054 |
| **Apet_E** | 0,532 | -0,112 | 0,276 | -0,036 | -0,207 | -0,324 |
| **Apet_F** | 0,117 | 0,221 | 0,002 | -0,269 | 0,003 | 0,045 |
| **Apet_G** | 0,050 | -0,039 | -0,074 | -0,032 | 0,144 | 0,117 |
| **Apet_H** | 0,005 | 0,117 | -0,012 | -0,035 | 0,019 | -0,043 |
| **Apet_I** | 0,478 | -0,053 | 0,084 | -0,048 | -0,008 | -0,007 |
| **Apet_J** | -0,045 | 0,167 | -0,003 | -0,131 | -0,031 | -0,075 |
| **Apet_K** | -0,162 | 0,244 | -0,078 | -0,059 | -0,132 | -0,243 |
| **Apet_L** | -0,314 | -0,367 | 0,060 | -0,071 | 0,293 | 0,194 |
| **Apet_M** | -0,687 | -0,207 | -0,252 | -0,059 | 0,170 | 0,463 |
| **Apet_N** | -0,069 | 0,262 | -0,096 | -0,218 | 0,021 | 0,101 |
| **Apet_O** | 0,467 | 0,047 | -0,001 | 0,009 | 0,104 | -0,148 |
| **Apet_P** | -0,080 | 0,004 | -0,271 | 0,262 | -0,383 | 0,433 |
| **Apet_Q** | 0,367 | 0,022 | 0,089 | -0,175 | -0,251 | 0,226 |
| **Apet_R** | -0,758 | -0,305 | 0,140 | 0,139 | 0,334 | -0,638 |
| **Apet_S** | -0,271 | -0,357 | -0,030 | -0,137 | -0,006 | 0,505 |
| **Region_Auvergne-Rhône-Alpes** | -0,061 | -0,275 | 0,030 | 0,053 | 0,062 | 0,028 |
| **Region_Bourgogne-Franche-Comté** | -0,079 | -0,185 | 0,063 | -0,015 | -0,062 | 0,043 |
| **Region_Bretagne** | -0,356 | 0,078 | -0,132 | -0,071 | 0,077 | 0,102 |
| **Region_Centre-Val de Loire** | -0,100 | 0,165 | -0,052 | -0,040 | -0,080 | -0,138 |
| **Region_Grand Est** | 0,102 | 0,131 | 0,011 | -0,060 | -0,090 | -0,162 |
| **Region_Hauts-de-France** | 0,148 | -0,123 | 0,001 | 0,037 | -0,045 | 0,048 |
| **Region_Normandie** | -0,101 | -0,146 | 0,055 | -0,015 | 0,081 | -0,085 |
| **Region_Nouvelle-Aquitaine** | -0,131 | 0,166 | -0,128 | -0,112 | 0,018 | 0,055 |
| **Region_Occitanie** | -0,113 | -0,060 | -0,074 | -0,030 | 0,043 | 0,116 |
| **Region_Pays de la Loire** | -0,086 | 0,240 | -0,002 | -0,180 | -0,095 | -0,114 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Region_Provence-Alpes-Côte d'Azur** | -0,062 | -0,316 | -0,068 | 0,106 | 0,012 | 0,194 |
| **Region_Ile-de-France** | -0,122 | 0,012 | 0,034 | 0,029 | -0,115 | -0,175 |
| **Anciennete** | 0,013 | -0,034 | 0,001 | -0,002 | -0,033 | 0,041 |
| **Age** | -0,031 | -0,387 | -0,239 | 0,079 | 0,285 | 0,628 |

## B.3 Coefficients du modèle avec le métier

| Variable | Ind_absenc | Moins_3 | 4_10 | 11_30 | 31_90 | Plus_90 |
|---|---|---|---|---|---|---|
| Intercept | -1,014 | -1,016 | -0,320 | -0,582 | -2,597 | -3,201 |
| CL_0 | -0,067 | 0,012 | -0,053 | -0,044 | -0,028 | -0,037 |
| CL_1 | -0,332 | -0,042 | -0,098 | -0,053 | 0,007 | 0,028 |
| CL_2 | -0,360 | 0,059 | -0,115 | -0,045 | -0,026 | -0,020 |
| CL_3 | -0,040 | -0,111 | 0,006 | -0,041 | -0,013 | -0,013 |
| CL_4 | -0,148 | -0,095 | 0,024 | -0,092 | -0,039 | -0,029 |
| SC_agent de la fonction publique d'Etat | 0,021 | 0,771 | -0,135 | -0,135 | -0,134 | -0,124 |
| SC_autres cadres au sens de la convention collective | -0,571 | 0,020 | 0,048 | -0,107 | -0,098 | -0,095 |
| SC_cadre dirigeant | -1,406 | -0,118 | -0,106 | 0,125 | 0,208 | -0,365 |
| SC_employé administratif d'entreprise, de commerce | 0,344 | -0,388 | 0,028 | -0,025 | -0,034 | 0,182 |
| SC_ouvriers qualifiés et non qualifiés | 0,547 | -0,300 | -0,101 | -0,080 | 0,051 | 0,306 |
| SC_profession intermédiaire | 0,118 | -0,162 | 0,030 | -0,053 | -0,092 | 0,026 |
| NC_Autre | -2,354 | -0,178 | -0,340 | 0,806 | -0,333 | -0,183 |
| NC_CDD | 0,320 | 0,197 | 0,135 | -0,689 | -0,236 | -0,945 |
| NC_CDI | 1,086 | -0,196 | -0,031 | -0,392 | 0,470 | 1,058 |
| Sexe_1,0 | -0,622 | -0,101 | -0,045 | -0,144 | -0,099 | -0,123 |
| Sexe_2,0 | -0,326 | -0,076 | -0,191 | -0,131 | 0,001 | 0,053 |
| Apet_B | -0,563 | 0,057 | -0,398 | 0,467 | -0,311 | -0,025 |
| Apet_C | 0,023 | 0,121 | 0,142 | 0,051 | 0,004 | -0,537 |
| Apet_D | -0,120 | -0,154 | 0,046 | 0,068 | 0,057 | -0,055 |
| Apet_E | 0,420 | -0,087 | 0,238 | -0,033 | -0,185 | -0,291 |
| Apet_F | 0,155 | 0,210 | 0,008 | -0,243 | 0,012 | 0,038 |
| Apet_G | 0,129 | 0,035 | -0,051 | -0,037 | 0,117 | 0,042 |
| Apet_H | 0,014 | 0,113 | -0,011 | -0,028 | 0,030 | -0,030 |
| Apet_I | 0,453 | -0,075 | 0,110 | -0,059 | 0,000 | 0,002 |
| Apet_J | -0,054 | 0,186 | 0,017 | -0,127 | -0,039 | -0,093 |
| Apet_K | -0,093 | 0,268 | -0,063 | -0,090 | -0,097 | -0,226 |
| Apet_L | -0,305 | -0,334 | 0,079 | -0,091 | 0,291 | 0,177 |
| Apet_M | -0,619 | -0,243 | -0,214 | -0,049 | 0,180 | 0,473 |
| Apet_N | -0,008 | 0,303 | -0,074 | -0,223 | 0,009 | 0,054 |
| Apet_O | 0,381 | 0,045 | -0,024 | 0,014 | 0,127 | -0,116 |
| Apet_P | -0,017 | 0,020 | -0,264 | 0,269 | -0,404 | 0,400 |
| Apet_Q | 0,211 | 0,035 | 0,090 | -0,196 | -0,243 | 0,245 |
| Apet_R | -0,733 | -0,299 | 0,168 | 0,136 | 0,337 | -0,650 |
| Apet_S | -0,219 | -0,378 | -0,035 | -0,104 | 0,016 | 0,523 |
| Region_Auvergne-Rhône-Alpes | -0,062 | -0,263 | 0,030 | 0,056 | 0,070 | 0,030 |
| Region_Bourgogne-Franche-Comté | -0,068 | -0,180 | 0,064 | -0,008 | -0,050 | 0,050 |
| Region_Bretagne | -0,363 | 0,074 | -0,128 | -0,070 | 0,089 | 0,115 |
| Region_Centre-Val de Loire | -0,098 | 0,181 | -0,050 | -0,037 | -0,074 | -0,140 |
| Region_Grand Est | 0,109 | 0,141 | 0,014 | -0,059 | -0,082 | -0,160 |
| Region_Hauts-de-France | 0,141 | -0,097 | 0,001 | 0,038 | -0,042 | 0,040 |
| Region_Normandie | -0,095 | -0,127 | 0,058 | -0,014 | 0,088 | -0,086 |
| Region_Nouvelle-Aquitaine | -0,130 | 0,177 | -0,124 | -0,111 | 0,026 | 0,057 |
| Region_Occitanie | -0,110 | -0,046 | -0,073 | -0,029 | 0,051 | 0,117 |
| Region_Pays de la Loire | -0,085 | 0,243 | 0,001 | -0,177 | -0,087 | -0,112 |
| Region_Provence-Alpes-Côte d'Azur | -0,052 | -0,310 | -0,063 | 0,107 | 0,022 | 0,198 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Region_Ile-de-France | -0,133 | 0,031 | 0,034 | 0,029 | -0,110 | -0,179 |
| Anciennete | 0,010 | -0,028 | 0,000 | -0,004 | -0,034 | 0,037 |
| Age | -0,028 | -0,393 | -0,239 | 0,082 | 0,287 | 0,633 |

# Bibliographie


[1] A. Saadaoui and E. Dumont, "Impact de la consommation santé sur la probabilité d'entrée en arrêt de travail," *Mémoire d'Actuariat*, 2019. [Online]. Available : https://www.institutdesactuaires.com/se-documenter/memoire-d-actuariat-38?id=fb11a46edded274d5dbd739a3894ea5a

[2] I. Hublin, "Risque incapacité en prévoyance collective : analyse et optimisation de la segmentation tarifaire," *Mémoire d'Actuariat*, 2019. [Online]. Available : https://www.institutdesactuaires.com/se-documenter/memoire-d-actuariat-38?id=cca307f94dc5e1a342002d5b90ce239d

[3] Anact-Aract, "Plaquette 10 questions sur l'absentéisme," https://www.anact.fr/10-questions-sur-labsenteisme.

[4] W. E. Moore, "Industrial relations and the social order," 1947.

[5] "Site du Larousse," https://www.larousse.fr/dictionnaires/francais/absent%C3%A9isme/261.

[6] "Code du travail numérique," https://code.travail.gouv.fr/glossaire/arret-de-travail.

[7] "Site officiel de la DARES," https://travail-emploi.gouv.fr/IMG/pdf/2013-009.pdf.

[8] "Chiffres clés de la Sécurité Sociale 2019," https://www.securite-sociale.fr/files/live/sites/SSFR/files/medias/DSS/2020/CHIFFRES%20CLES%202020%20ED2019.pdf.

[9] S. Chaupain-Guillot and O. Guillot, "Les absences au travail : une analyse à partir des données françaises du panel européen des ménages," *ÉCONOMIE ET STATISTIQUE N° 408-409*, 2007.

[10] "Site officiel de l'INSEE," https://www.insee.fr/fr/metadonnees/source/serie/s1172.

[11] R. Rakotomalala, "Pratique des méthodes factorielles avec Python." [Online]. Available : http://eric.univ-lyon2.fr/~ricco/cours/cours/Pratique_Methodes_Factorielles.pdf

[12] D. Ajar, "Le problème de la détermination du nombre de facteurs en analyse factorielle," *Revue des sciences de l'éducation*, p. 45–62, 1982.

[13] J. A. Hartigan and M. A. Wong, "Algorithm as 136 : A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979. [Online]. Available : http://www.jstor.org/stable/2346830

[14] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *Proceedings of Second International Conference on Knowledge Discovery and Data Mining*, pp. 226–231, 1996.

[15] S. S.Z and M. Ismail, "K-means type algorithms : A generalized convergence theorem and characterization of local optimality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 81–87, 1984.

[16] L. Bobrowski and J. Bezdek, "c-means clustering with the ll and l∞ norms," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 21, pp. 545 – 554, 06 1991.

[17] Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining and Knowledge Discovery*, 1998.

[18] M. Norouzi, D. J. Fleet, and R. R. Salakhutdinov, "Hamming distance metric learning," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012. [Online]. Available : https://proceedings.neurips.cc/paper/2012/file/59b90e1005a220e2ebc542eb9d950b1e-Paper.pdf

[19] L. Yujian and L. Bo, "A normalized levenshtein distance metric," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1091–1095, 2007.

[20] T. Sørensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons," *Biologiske SkrifterbKongelige Danske Videnskabernes Selskab*, vol. 5, no. 4, p. 1–34, 1948.

[21] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, p. 297–302, 1945.

[22] P. Jaccard, "Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines," *Bulletin de la Société vaudoise des sciences naturelles*, vol. 37, pp. 241–272, 1901.

[23] R. Abdesselam, *Selection of Proximity Measures for a Topological Correspondence Analysis*. John Wiley & Sons, Ltd, 2020, ch. 6, pp. 101–121. [Online]. Available : https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119721871.ch6

[24] R. Diestel, "Graph theory," *Springer, vol 173*, 2000.

[25] K. Spärck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation, vol. 28, no 1*, 1972.

[26] G. Salton, A. Wong, and C. Yang, "A vector space model for automatic indexing," *Communication of the ACM, v18 n 11*, 1975.

[27] W. Li, "Zipf's law everywhere," *Glottometrics*, vol. 5, 2002.

[28] "Site de l'index des packages Python (PyPi)," https://pypi.org/project/kmodes/.

[29] “Site officiel de l’INSEE,” https://www.insee.fr/fr/information/2497958.

[30] “Site de ressources actuarielles,” http://www.ressources-actuarielles.net/bcac.