Mémoire présenté le :

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

Par : Corentin BATUT

Titre : Aide à la construction d'un plan de prévention santé optimisé grâce à l'exploitation des données et apports du NLP

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

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

Membre présents du jury de l'Institut des Actuaires — signature

Membres présents du jury de l'ISFA

Entreprise :

Nom : ADDACTIS

Signature :

Directeur de mémoire en entreprise :

Nom : Jean-Pascal HERMET

Signature :

Invité :

Nom :

Signature :

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

Signature du responsable entreprise

Signature du candidat

Résumé

Mots clés : assurance santé, science des données, prévention ciblée, classification non supervisée, parcours de consommation santé, traitement du langage naturel, prédiction.

L'assureur santé, par définition, est une personne physique ou morale qui s'engage, moyennant le paiement d'une prime, à rembourser à l'assuré une partie de ses frais médicaux.

Cette définition est aujourd'hui caduque et non exhaustive. L'assureur est en voie de devenir un acteur innovant et pluridisciplinaire ayant comme mission de protéger ses assurés contre les aléas de la vie. Pour ce faire, il peut, en parallèle des remboursements des frais médicaux, proposer d'autres services tels que des actions de prévention. Cela lui permet de réduire ses dépenses de soins tout en améliorant la santé de ses assurés, mais également d'anticiper un risque stratégique tel qu'une augmentation du rôle du Régime Obligatoire de Sécurité sociale. Cette situation pourrait entraîner une réduction de près de 70% du chiffre d'affaires des assureurs santé, comme cela est envisagé dans le troisième scénario du rapport publié en janvier 2022 par le Haut Conseil pour l'Avenir de l'Assurance Maladie (HCAAM), le scénario "Grande Sécu".

Cependant, pour développer des actions de prévention efficientes (avec un « bon » retour sur investissement), ces actions doivent être ciblées et adaptées aux besoins des assurés.

Après avoir montré que l'assureur santé dispose, grâce à ses bases des prestations versées, de données suffisantes pour comprendre la consommation santé des assurés, ce mémoire développera plusieurs méthodes pour comprendre et prédire la consommation santé d'un portefeuille d'assurance santé collective au cours du temps. Ces méthodes, issues de la science des données, auront pour objectif de guider l'assureur dans le ciblage des actions de prévention afin de favoriser la santé des assurés.

Abstract

Keywords: health insurance, data science, targeted prevention, unsupervised classification, health consumption trajectory, natural language processing, prediction.

The health insurer, by definition, is a physical or legal person who undertakes, in exchange for the payment of a premium, to reimburse the policyholder a part of his/her medical expenses.

This definition is now obsolete and not exhaustive. The insurer is becoming an innovative and multidisciplinary player whose mission is to protect its policyholders against life's risks. To do so, he can, in addition to reimbursing medical expenses, offer other services such as prevention actions. This allows it to reduce its medical expenses while improving the health of its policyholders, but also to anticipate a strategic risk such as an increase in the role of the compulsory social security system. This situation could lead to a reduction of nearly 70% in the turnover of health insurers, as envisaged in the third scenario of the report published in January 2022 by the "Haut Conseil pour l'Avenir de l'Assurance Maladie" (HCAAM), the "Grande Sécu" scenario.

However, to develop efficient prevention actions (with a "good" return on investment), these actions must be targeted and adapted to the needs of the policyholders.

After showing that the health insurer has, thanks to its databases of paid claims, data allowing it to understand the health consumption of the policyholders, this study will develop several methods to understand and predict the health consumption of a collective health insurance portfolio over time. These methods, derived from data science, will aim to guide the insurer in targeting prevention actions to promote the health of policyholders.

Remerciements

Je tiens à remercier toutes les personnes qui de près ou de loin ont contribué à l'élaboration de ce mémoire.

Tout d'abord, j'exprime toute ma reconnaissance envers Jean-Pascal HERMET, mon tuteur d'apprentissage, qui m'a formé, encadré et soutenu durant mon parcours chez ADDACTIS France. Je le remercie également pour son accompagnement sans faille dans ce mémoire. Malgré les doutes et les difficultés, ce fut un énorme plaisir de travailler ensemble.

Je remercie naturellement Cécile PARADIS et Alexandra BARRAL pour m'avoir fait confiance et m'avoir permis de travailler sur ce passionnant sujet de mémoire.

Je tiens aussi à remercier l'équipe pédagogique de l'ISFA, et plus particulièrement mon tuteur académique Stéphane LOISEL pour son accompagnement.

Également, je remercie l'ensemble de la practice *Pricing & Analytics - Life & Health* ainsi que tous les collaborateurs d'ADDACTIS France pour leurs conseils mais aussi pour l'ambiance incroyable dans laquelle j'ai pu évoluer.

Et enfin, un grand merci à mes parents pour leur soutien inconditionnel depuis tant d'années.

Introduction

Le système de santé français est historiquement axé sur une approche curative et non sur une approche préventive. De même, l'assureur santé agit traditionnellement en aval (à travers le remboursement des frais de santé des assurés). Or, depuis quelques années, sous le poids de dépenses de santé toujours plus importantes, cette approche est de plus en plus contestée. En effet, les personnels de santé tirent régulièrement la sonnette d'alarme. Le système de santé français est même décrit comme « au bord du gouffre » [Le Temps, 2023].

Face à ces problèmes, l'Assurance Maladie et les acteurs publics ne sont pas les seuls à avoir un rôle à jouer. Les assureurs santé peuvent et doivent aussi contribuer en proposant des services de prévention. Qui plus est, ils ont aussi tout intérêt à le faire. Il s'agit d'une stratégie gagnant-gagnant où il est possible pour l'assureur de diminuer le coût du risque et de fidéliser ses assurés tout en améliorant leur santé.

Cependant, la prévention peut se révéler coûteuse lorsqu'elle est proposée à tout son portefeuille d'assurés et inefficiente si elle n'est pas adaptée aux besoins des assurés. Elle doit en effet être ciblée et évaluée (retour sur investissement). Pour ce faire, l'assureur doit avoir une bonne connaissance du risque et ainsi une bonne compréhension de la consommation santé de son portefeuille.

Il s'agira dans ce mémoire de proposer une méthodologie afin de **comprendre et prédire la consommation santé d'un portefeuille d'assurance santé collective au cours du temps dans un but de prévention ciblée**.

Cette méthode se différenciera des travaux existants par la notion de **temporalité (au cours du temps)** et de **parcours de consommation**. En effet, seule une méthode de classification statique (cartographie du portefeuille à un instant *t*) a déjà été présentée par le passé afin d'aider au ciblage d'actions de prévention. Or, **s'il est nécessaire d'identifier la population cible, il est également important d'identifier le moment le plus opportun**.

De plus, ce mémoire utilisera des techniques innovantes afin de prédire la consommation santé. Un modèle de traitement automatique du langage naturel, BERT (*Bidirectional Encoder Representations from Transformers*), sera notamment utilisé afin de modéliser et comprendre les parcours de consommation santé des assurés.

Table des matières

I Contexte et données à disposition

La prévention en santé est un enjeu de taille pour l'assureur. Elle permet de minimiser les risques et ainsi réduire les dépenses engagées par l'assureur, notamment lorsqu'elle est bien ciblée. De plus, elle présente un double intérêt assureur et assuré en améliorant la santé des individus ciblés et en personnalisant la relation entre l'assureur et l'assuré.

Aujourd'hui, de nombreux acteurs proposent des programmes de prévention et parmi eux, beaucoup s'intéressent à l'apport du digital et des objets connectés (notamment à travers la collecte de nouvelles données). Or, l'assureur, malgré un cadre réglementaire très strict, dispose de données (exploitées ou non) lui permettant d'anticiper les risques (et ainsi cibler l'action de prévention).

La première partie de cette étude consiste à définir les termes « prévention » et « ciblage ». Les motivations et objectifs de cette étude seront ensuite explicités. Et enfin, les données utilisées tout au long du mémoire seront présentées.

I.1 La prévention santé

Afin de mieux comprendre le contenu de cette étude, il est important de retenir une définition de la prévention et d'expliquer pourquoi l'assureur doit jouer un rôle important dans la prévention santé.

I.1.1 Définition possible

Le terme « **prévention** » est un terme générique nécessitant une définition claire et précise. D'après le dictionnaire Larousse, il se rapporte au verbe « prévenir » qui signifie ici « prendre les mesures nécessaires pour éviter un mal, un danger ». Or, par cette définition, certains types de prévention peuvent être omis. Pour cette étude, il sera alors choisi de retenir la définition très complète (trois niveaux de prévention) donnée par l'OMS (Organisation mondiale de la santé). Celle-ci peut être résumée de la manière suivante :

> « ***L'ensemble des mesures visant à éviter ou réduire le nombre et la gravité des maladies, des accidents et des handicaps*** » [OMS, 1948]

Les trois niveaux de prévention définis par l'OMS sont les suivants :

- **La prévention primaire** intervient en amont des problèmes de santé et regroupe les actions visant à empêcher l'apparition d'une maladie ou à diminuer son incidence, dont par exemple :
    - Les campagnes de vaccination
    - Les mentions sanitaires dans la publicité alimentaire (« Pour votre santé, mangez au moins cinq fruits et légumes par jour »)

- **La prévention secondaire** intervient à l'arrivée des problèmes de santé et les actions visent à déceler et soigner précocement une maladie afin de prévenir de futures complications ou séquelles, dont par exemple :
    - Le dépistage du cancer du sein
    - Le dépistage du VIH et des IST
- **La prévention tertiaire** intervient au cours ou à l'issue d'un problème de santé et comprend l'ensemble des actions visant à réduire les complications, invalidités ou rechutes consécutives à une pathologie avérée, dont par exemple :
    - Les plans d'actions pour éviter les rechutes en cas de dépression
    - L'accompagnement nutritionnel des diabétiques (afin d'éviter des complications telles que les maladies cardiaques ou la cécité)

Cette définition permet de bien comprendre la diversité des actions de prévention proposables et intéressantes dans le cadre de ce mémoire.

Cependant, l'efficacité de l'action de prévention ne dépend pas seulement de son contenu. Elle dépend aussi de la temporalité. Il faut être capable de la mettre en place au bon moment. En effet, si l'action arrive trop tôt, elle peut sembler inutile donc ne sera pas bénéfique et de même si l'action arrive trop tard, alors les risques sont déjà dégradés et elle ne sera pas efficace.

Et enfin, elle dépend de la population à qui elle est destinée. Il est donc nécessaire de compléter la définition par la classification suivante : [Gordon,1982]. Gordon distingue en effet trois classes de prévention :

- **La prévention universelle** est destinée à l'ensemble de la population.
- **La prévention sélective** est destinée à un sous-groupe spécifique qui peut être caractérisé par un critère d'âge, de sexe, de profession, etc.
- **La prévention ciblée** est destinée à un sous-groupe identifié en fonction de ses facteurs de risque (conditions, pathologies ou comportements qui rendent plus probable la survenue d'une maladie).

Dans la suite de ce mémoire, le terme « prévention ciblée » sera à de nombreuses reprises utilisé pour décrire des actions de prévention dont il a été identifié les bons destinataires mais aussi le meilleur moment pour la mise en place.

I.1.2 Intérêt pour l'assureur

Historiquement, les acteurs principaux de la prévention santé sont des institutions publiques (Sécurité sociale, ministère de la Santé, collectivités locales, etc.). Dans cette partie, il s'agira de montrer que l'assureur a lui aussi un intérêt à investir dans la prévention en santé.

a. Réduction du coût du risque

La tarification technique (calcul de la prime pure) d'un contrat d'assurance santé s'appuie dans la plupart des cas sur un modèle fréquence-coût. A chaque risque est associé une fréquence de survenance d'un sinistre et le coût moyen du sinistre selon les garanties mais également les caractéristiques de l'assuré. Ensuite, un tarif commercial prenant en compte des chargements et des taxes est proposé à l'assuré.

Le premier objectif (et aussi l'objectif principal) de l'assureur à travers la prévention est la diminution du coût du risque. La mise en place d'actions de prévention de la part de l'assureur peut en effet permettre de réduire le montant total des sinistres en diminuant la fréquence ou le coût moyen des sinistres. Par exemple, un plan de vaccination peut permettre de diminuer la fréquence et un plan de dépistage le coût moyen (du fait d'un traitement précoce de la maladie). **La prévention est ainsi un investissement pour l'assureur**. Dans ce mémoire, cet investissement sera mesuré par l'indicateur suivant :

Le ROI (*Return On Investment*) qui correspond au retour sur investissement. Il est donné par la formule suivante :

$$ROI = \frac{GI - CI}{CI}$$

Avec : GI : Gain ou perte de l'investissement

CI : Coût de l'investissement

De plus, la diminution du risque engendrée par les actions de prévention n'est pas seulement favorable à l'assureur. En effet, elle apporte à l'assuré un service pouvant contribuer à l'amélioration de son état de santé.

Ainsi, une action de prévention dont le coût de mise en place est maîtrisé et raisonnable par rapport au gain obtenu (la diminution du montant total des sinistres) permet à l'assureur d'obtenir un bon retour sur investissement (supérieur à zéro) tout en améliorant la santé de ses assurés.

b. Fidélisation des assurés et image de marque

Au-delà de l'aspect financier présenté ci-dessus, la prévention peut aussi permettre aux sociétés d'assurance de se démarquer en présentant une stratégie commerciale et marketing humaine et innovante.

En effet, une action de prévention correctement ciblée permet de personnaliser la relation entre l'assureur et l'assuré. Celle-ci va donc favoriser la fidélisation. A l'heure de la résiliation infra-annuelle des contrats santé et d'une réglementation de plus en plus souple pour

les assurés, il est aussi important d'investir dans la fidélisation que dans la publicité pour attirer de nouveaux contrats.

Une politique de mise en place d'actions de prévention va permettre de jouer sur les deux axes : d'une part, fidéliser les assurés actuels et d'autre part en attirer de nouveaux. Cela peut donc permettre à l'assureur de se démarquer de ses concurrents par une proposition de services améliorant son image et ainsi se développer.

c. Mise en place complexe

Il parait évident d'après les deux points I.1.2a. et I.1.2b. ci-dessus que l'assureur a un intérêt à proposer des actions de prévention. Cependant, un plan d'actions mal ciblé ou précipité sans une bonne estimation des gains et des coûts potentiels peut ne pas être un bon investissement pour l'assureur (mauvais retour sur investissement). C'est une des principales raisons pour lesquelles les assureurs ont longtemps négligé la dimension prévention.

D'une part, il faut être capable de bien identifier la cible et les gains potentiels. En effet, sans cela, l'assureur se retrouverait avec un coût trop important (ou potentiellement supérieur) par rapport aux gains obtenus.

D'autre part, la mise en place d'actions de prévention est sujet aux phénomènes d'aléa moral et d'antisélection qui seront définis dans ce mémoire de la manière suivante :

- **Aléa moral** : un assuré ne consomme pas de la même manière selon ses garanties. Lorsqu'il est bien couvert, il est susceptible de consommer plus fréquemment ou pour des montants plus élevés.

- **Antisélection** : l'assureur ne dispose pas des mêmes informations que l'assuré. L'assuré détient de l'information privée et l'utilise afin de choisir le contrat le plus avantageux pour lui. Il y a ainsi asymétrie d'information.

Prenons l'exemple d'une compagnie qui propose le remboursement de licence sportive. Il s'agit de prévention primaire. Ce remboursement pourrait inciter des assurés déjà présents à s'offrir une licence sportive sans réelle motivation et ainsi ne pas aller pratiquer du sport par la suite. En effet, par la gratuité (le reste à charge nul), les assurés ne sont pas engagés personnellement. L'action de prévention ne sera donc pas bénéfique pour l'assureur. Il s'agit d'aléa moral. Parallèlement, un individu déjà sportif pourrait souscrire à un contrat car l'assureur rembourse la licence. L'individu étant déjà sportif, cela ne sera pas bénéfique pour l'assureur. Il s'agit d'antisélection.

Pour résumer, la prévention en assurance santé peut être un bon investissement pour l'assureur. Elle permet de **réduire la fréquence et les montants des sinistres tout en améliorant la santé des assurés** et en **personnalisant la relation entre l'assureur et l'assuré**. C'est une stratégie « gagnant-gagnant ». Cependant, afin d'obtenir un bon retour sur investissement, l'assureur doit s'assurer de **bien cibler** ses actions de prévention et de réaliser une **bonne évaluation** des coûts et des gains potentiels.

I.2 Etat de l'art et objectifs

Précédemment, il a été montré que l'assureur a tout intérêt à proposer des actions de prévention ciblée à ses assurés. Dans un premier temps de cette partie, il s'agira de présenter quelques programmes de prévention déjà en place sur le marché français. Ensuite, dans un second temps, une description des objectifs du mémoire sera proposée. **La mise en parallèle de l'état de l'art et des objectifs permettra de comprendre les différences et les apports entre la méthode développée dans le mémoire et les méthodes déjà existantes : cela permettra donc de mettre en évidence les avantages de cette nouvelle méthode par rapport à celles qui existent déjà.**

I.2.1 La prévention personnalisée par le digital

Un état de l'art des programmes de prévention menés par les assureurs santé en France permet de dresser le constat suivant : le ciblage et la personnalisation de la prévention est encore à ses débuts. La plupart des acteurs proposent seulement des actions de prévention universelles ou sélectives. En effet, très souvent, les services proposés sont destinés à tous les assurés. Ci-dessous quelques exemples :

- AXA Prévention : lancé en 2010, ce programme vise à sensibiliser les clients d'AXA France aux risques de la vie quotidienne : ce programme propose des conseils pour prévenir le stress, des astuces pour protéger son capital santé, des formations aux gestes de 1er secours, des bilans de santé, etc.

- Harmonie Prévention : lancé en 2013, ce programme propose des actions de sensibilisation et de prévention afin d'améliorer la qualité de vie au travail et dans la vie quotidienne. Plus concrètement, il propose aux assurés des programmes de coaching, des ateliers de sensibilisation, des bilans de santé, des conseils santé, etc.

- « Relax - bien-être » et « Boost - activité physique » par SwissLife : lancés en 2020, il s'agit de deux programmes de coaching adaptés aux activités des entreprises pour la prévention des risques physiques et de la santé mentale. Le premier est destiné aux métiers sédentaires et le second aux métiers physiques. Ces programmes sont proposés aux entreprises ayant souscrit auprès de SwissLife un contrat santé ou prévoyance collectif.

- AÉSIO mutuelle : en 2021, à l'occasion de la semaine nationale de prévention du diabète, un programme a été lancé invitant les assurés à se faire dépister dans les agences.

- La plupart des organismes s'appuient en général vers des réseaux de soins partenaires qui proposent des services de prévention : coaching sommeil, coaching nutrition, etc.

De plus, on constate que lorsque les programmes sont personnalisés afin de cibler les assurés, **la personnalisation implique systématiquement l'utilisation d'outils digitaux tels**

qu'une application ou des objets connectés. À titre illustratif, voici deux exemples de programmes de prévention ciblée qui utilisent des technologies digitales :

- Generali Vitality : déployé en France le 1er janvier 2017, il s'agit d'un programme visant à récompenser les assurés fournissant des efforts pour améliorer leur hygiène de vie. Tout d'abord, l'utilisateur doit se munir d'une montre connectée et remplir un questionnaire. Pour donner suite à cela, un programme personnalisé lui sera proposé. Il pourra par exemple lui être proposé des objectifs sportifs ou personnels tel que la réduction de consommation de tabac par exemple. Et enfin, l'assuré sera récompensé par des chèques-cadeaux ou des réductions s'il remplit ses objectifs. La prévention sera ainsi ciblée car orientée par les informations données par l'utilisateur du programme et les données collectées par la montre.

- Bandeau Dreem par Mutuelle Ociane Matmut : depuis 2020, la mutuelle et l'entreprise Dreem se sont associées afin de proposer aux assurés un bandeau connecté permettant de résoudre les problèmes de sommeil. Ce bandeau, porté pendant la nuit, collecte et analyse le sommeil de l'utilisateur. Pour donner suite à cela, un programme personnalisé sera proposé à l'utilisateur afin d'améliorer son sommeil. Ces programmes peuvent même inclure des sessions téléphoniques avec des spécialistes. Comme précédemment, le ciblage est réalisé par collecte de données.

A travers ces deux exemples, il est aisé de comprendre quelle est la méthodologie utilisée. Il s'agit **dans les deux cas de collecter de nouvelles données (par le biais du digital) et de les utiliser à des fins de ciblage. L'assureur n'utilise donc pas les données qu'il a en sa possession.**

Au-delà du fait qu'il est dommage que l'assureur n'exploite pas sa connaissance du portefeuille, ces programmes de prévention, bien qu'innovants ne font pas l'unanimité. En effet, 36% des Français se disent inquiets du traitement de leurs données par les applications mobiles de santé [Capterra, 2021]. Ceci est une limite qui pourrait encourager les assureurs à utiliser uniquement des données disponibles telles que celles issues de leurs bases de prestations.

A noter malgré tout que des données disponibles ne sont pas toujours des données utilisables. En effet, l'assureur doit respecter le RGPD (Règlement Général de la Protection des Données). Ce point sera précisé en partie I.3.3.

I.2.2 Objectifs : une autre approche que le digital

L'objectif de ce mémoire est de comprendre et prédire la consommation santé des assurés afin d'identifier **quand** et **à qui** proposer des actions de prévention à partir **de données déjà disponibles** pour l'assureur via les bases des **prestations**.

Il ne s'agira pas de collecter de nouvelles données par le biais de services digitaux mais plutôt d'exploiter des données déjà disponibles.

Dans un premier temps, des travaux de traitement de la donnée seront présentés. Il s'agira de rendre les données exploitables et compréhensibles. En effet, la base des prestations est destinée aux remboursements des actes consommés et non à la bonne compréhension de la consommation santé des assurés. Elle n'est donc pas directement interprétable.

Dans un second temps, les données traitées, obtenues par ces travaux, seront utilisées afin de prédire la consommation santé future des assurés. Plusieurs méthodes de prédiction seront testées et comparées. Il sera notamment testé BERT (*Bidirectional Encoder Representations from Transformers*), un modèle innovant qui est utilisé ici afin de modéliser et comprendre les parcours de consommation santé des assurés.

Et enfin, les prédictions et statistiques obtenues seront utilisées afin d'identifier le meilleur moment (et la bonne cible) pour proposer une action de prévention ciblée.

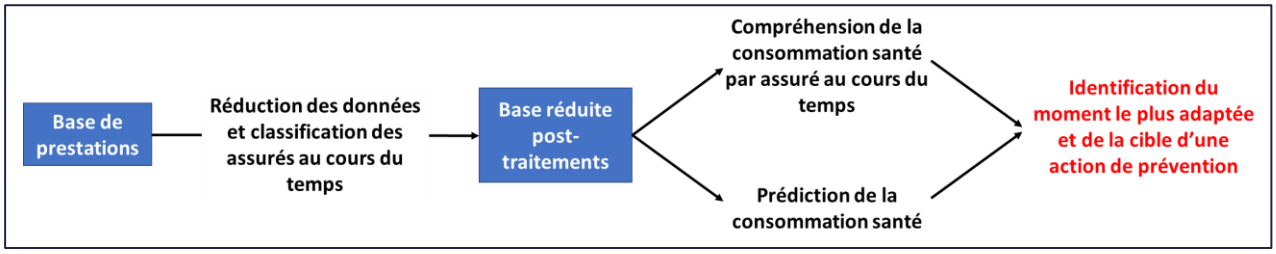


Figure 1.1 – Schéma illustrant la démarche et les objectifs du mémoire

I.3 Données utilisées pour l'étude

Le portefeuille étudié est un portefeuille de santé collective intégrant des entreprises ayant entre 45 et 320 personnes couvertes dans différents secteurs d'activité. Comme mentionné précédemment, l'assureur santé dispose de données intéressantes et non exploitées (les libellés des actes santé consommés). Ces données seront donc le point de départ des travaux de ce mémoire.

I.3.1 Présentation générale des données

Les données utilisées sont issues de deux bases de données d'un même portefeuille d'assurance santé, observé entre le 01/01/2018 au 31/12/2019 : une « base des bénéficiaires » qui recueille l'ensemble des informations sur les bénéficiaires couverts, et une « base des prestations » qui regroupe les données ligne à ligne relatives aux prestations santé indemnisées par l'organisme d'assurance.

Les variables disponibles dans chacune des deux bases de données sont les suivantes :

- **Base des bénéficiaires** :
  - L'identifiant du contrat auquel le bénéficiaire est associé
  - L'identifiant de l'adhérent/assuré principal auquel le bénéficiaire est rattaché
  - L'identifiant du bénéficiaire
  - L'identifiant de l'entreprise à laquelle l'assuré principal est rattaché
  - Le type du bénéficiaire : adhérent ou ayant-droit
  - Le sexe du bénéficiaire
  - L'âge du bénéficiaire
  - Le département de résidence du bénéficiaire
  - Les dates de début et de fin de couverture du bénéficiaire
  - La durée d'exposition du bénéficiaire par année

- **Base des prestations** :
  - L'identifiant bénéficiaire : identifiant de l'individu ayant consommé l'acte de soins
  - La date de survenance : date de survenance de l'acte de soins
  - Le libellé de l'acte de soins

- La quantité : nombre d'actes de soins (boîtes de médicaments par exemple)
- Les dépenses engagées (DE) : coût global de l'acte
- Le remboursement complémentaire (RC) : coût de l'acte indemnisé par l'organisme complémentaire
- Le remboursement Sécurité Sociale (RSS) : coût de l'acte indemnisé par la Sécurité Sociale

Des retraitements ont dû être effectués afin de sélectionner certaines de ces variables, de créer certains de ces indicateurs et de filtrer les bénéficiaires.

Au niveau de la base des bénéficiaires, les retraitements effectués sont les suivants :

- Les dates de début et de fin de couverture ont été exploitées afin de calculer une durée d'exposition comme suit :

$$\text{exposition}_{\text{année N}} = \frac{\text{Nombre de jours dans l'année N où l'individu est couvert par le contrat}}{\text{Nombre de jours dans l'année N}} \in [0,1].$$

- Seuls les bénéficiaires exposés entièrement sur les 8 trimestres ont été conservés.

  Dans ce mémoire, il sera question d'étudier l'évolution de la consommation santé des assurés. Un assuré non exposé sur toute la période d'observation est plus difficile à étudier (données manquantes sur certaines périodes). Pour simplifier l'analyse, les individus qui ne sont pas exposés pendant toute la période d'observation ne seront pas inclus dans cette étude.

- Seuls les bénéficiaires adhérents ont été conservés.

Concernant la base des prestations, ci-dessous les retraitements effectués :

- Suppression des libellés d'actes peu fréquents (présents moins de 20 fois) ou ne concernant pas plus que 10 assurés.

  Ces retraitements représentent environ 0,1% des prestations en nombre d'observations et 0,82% en montants (DE).

- Suppression des lignes anormales présentant des dépenses engagées inférieures à la somme des remboursements effectués par la Sécurité Sociale et par la complémentaire (DE < RSS + RC)

  Ce retraitement représente environ 0,13% des prestations en termes de fréquence.

La base des prestations est complétée des caractéristiques du bénéficiaires, grâce à l'information commune « Identifiant du bénéficiaire » et ce afin d'obtenir une base de données contenant 1,58 millions d'observations et 17 variables.

I.3.2 Statistiques descriptives

Les bases de données correspondent à la période d'observation du 01/01/2018 au 31/12/2019 et concernent environ **18 700 personnes protégées répartis dans près de 160 entreprises**. Au total, les données prestations regroupent 1,58 millions de lignes soit 36,46 millions d'euros de dépenses engagées et 17,53 millions d'euros de remboursements complémentaires.

Les parties suivantes visent à donner quelques statistiques descriptives sur les assurés, les entreprises et les prestations.

a. Statistiques sur les assurés

La base des bénéficiaires utilisée est une base retraitée qui ne contient que des bénéficiaires adhérents. La distinction bénéficiaire/adhérent ne sera pas réalisé. De plus, dans cette base, chaque assuré consomme au moins un acte de soin entre le 01/01/2018 et le 31/12/2019. En effet, les assurés ne consommant aucune prestation ont été exclus du périmètre de l'étude.

Le portefeuille étudié contient 61% d'assurés homme et **l'âge moyen est de 43 ans**. La répartition par tranche d'âges et par sexe est la suivante :

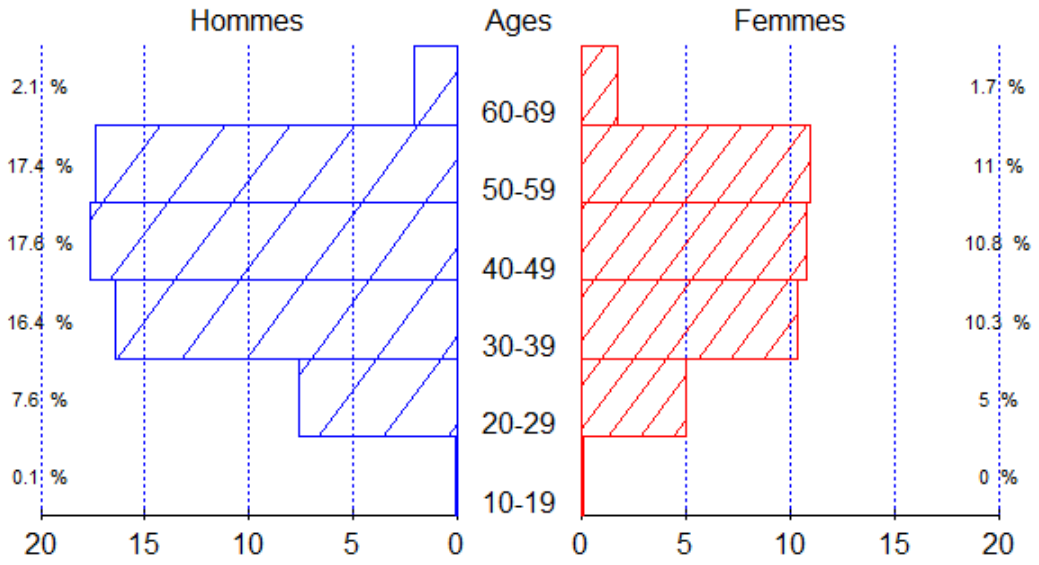


Figure 1.2 – Répartition des assurés par tranche d'âges et par sexe

b. Statistiques sur les entreprises

Il s'agit d'un portefeuille de **156 entreprises** ayant en moyenne 120 assurés. Voici ci-dessous un histogramme des effectifs par entreprises permettant d'avoir une vision plus fine du portefeuille d'entreprises :

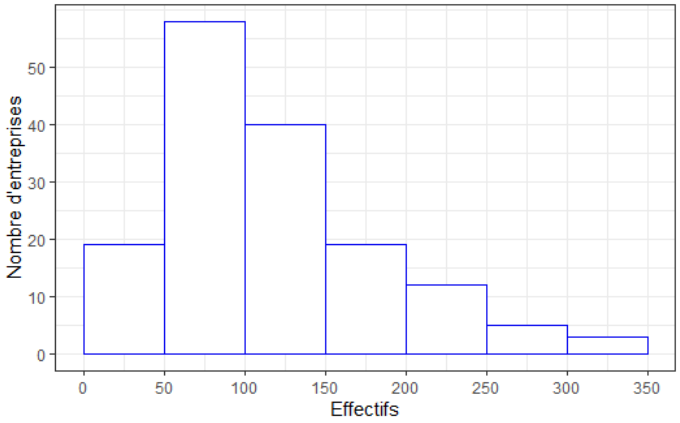


Figure 1.3 – Histogramme des effectifs par entreprise

c. Statistiques sur les libellés d'actes

Les prestations sont identifiées à l'aide des libellés des actes de soins les plus détaillés. Il existe 245 libellés différents qu'il est possible de classer en trois catégories :

- **Les libellés prépondérants**

  Certains actes sont prépondérants, autrement dit, individuellement, chacun d'entre eux représente au moins 2% du nombre d'actes consommés. C'est le cas par exemple des postes suivants :

  - Pharmacie vignette blanche 65%
  - Pharmacie vignette bleue 30%
  - Acte de kinésithérapie
  - Consultation de médecine générale

- **Les libellés plus rares**

  Certains actes, au contraire, représentent moins de 2% des observations. C'est le cas par exemple des postes suivants :

  - Acte d'anesthésie
  - Vaccin contre la grippe
  - Soins à l'étranger
  - Soins infirmiers sage-femme

- **Les libellés « honoraires » redondants avec les prestations correspondantes**

  Dans la base de prestations, une prestation ne correspond pas toujours qu'à une seule ligne. Par exemple, une seule et unique prestation de pharmacie peut amener aux trois lignes suivantes :

- Pharmacie vignette blanche 65%
- Honoraires médicament remboursable
- Honoraires de dispensation

Ces lignes « Honoraires » sont redondantes avec la première ligne en termes de fréquence et peuvent biaiser le nombre d'actes de pharmacie consommés alors qu'elles représentent la même information que la prestation liée au médicament délivré. Cependant, elles doivent être comptabilisées pour le calcul des coûts (RC et DE). Quelques précautions sont donc à prendre en compte pour travailler sur la base de prestations.

d. Statistiques sur la consommation

Dans le cadre de l'étude, il est également intéressant d'avoir une vision trimestrielle de la consommation santé par assuré. Ci-dessous deux tableaux présentant la consommation trimestrielle par assuré et par année :

	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
RC moyen annualisé*	470 €	444 €	403 €	488 €
DE moyenne annualisée*	982 €	933 €	852 €	1 011 €
RSS moyen annualisé	438 €	412 €	380 €	436 €
Nombre d'actes moyen**	10,2	9,3	8,7	10,1

* Montant moyen par assuré et par trimestre multiplié par 4

** Nombre de lignes dans la base de prestations

Figure 1.4 – Consommation santé trimestrielle par assuré en 2018

	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
RC moyen annualisé	514 €	464 €	449 €	516 €
DE moyenne annualisée	1 061 €	983 €	933 €	1 043 €
RSS moyen annualisé	463 €	436 €	407 €	435 €
Nombre d'actes moyen	12,2	11,4	10,6	12,2

Figure 1.5 – Consommation santé trimestrielle par assuré en 2019

La consommation en 2019 est plus importante que celle de 2018 : elle a augmenté en moyenne de 21% en fréquence, 6% en DE et 8% en RC. De plus, il est visible que la consommation est plus importante pendant les trimestres correspondant plutôt à des périodes hivernales (Trimestres 1 et 4) que pendant les trimestres correspondant plutôt à des périodes estivales (Trimestres 2 et 3).

Illustration graphique avec le RC moyen annualisé :

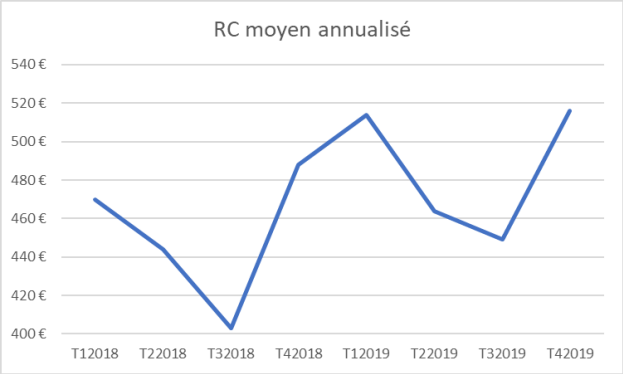


Figure 1.6 – RC moyen annualisé par trimestre

I.3.3 Cadre réglementaire et éthique

Enfin, pour terminer la présentation des données, il est important de préciser le cadre réglementaire dans lequel s'inscrit le traitement et l'exploitation de ces données.

D'après le RGPD (Règlement Général sur la Protection des Données) entré en vigueur en mai 2018, les données présentées ci-dessus sont des « **données sensibles** » car il s'agit de données de santé.

Les « données sensibles » sont une catégorie de données à caractère personnel bénéficiant d'une protection renforcée. Autrement dit, l'assureur ne peut pas les collecter ou les utiliser comme bon lui semble. **Ces données peuvent être recueillies ou utilisées seulement de manière exceptionnelle lorsque la finalité du traitement l'exige. De plus, le consentement de la personne concernée est nécessaire et, la finalité et les modalités associées doivent être déterminées en amont.**

Ainsi, afin d'exploiter les données présentées ci-dessus à des fins de ciblage d'actions de prévention, il est nécessaire d'obtenir l'accord de l'assuré. Sans cela, la base de prestations pourrait seulement être utilisée dans le but de rembourser les actes de soins ou alors les données devraient être anonymisées.

Une fois l'accord obtenu, les données peuvent être utilisées à des fins de ciblage d'actions de prévention mais restent tout de même protégées. L'assuré a notamment les droits suivants :

- Le droit de demander au responsable du traitement l'accès aux données, la rectification ou l'effacement
- Le droit de s'opposer au traitement des données (ou demander une limitation)

En pratique, cela signifie qu'il est nécessaire de mettre en place un registre donnant une vision globale mais aussi une vision exhaustive des données utilisées et des traitements réalisés. De plus, cela signifie aussi que l'accord de l'assuré peut être rompu à tout moment.

Dans ce mémoire, **les travaux présentés reposeront donc sur l'hypothèse d'un accord de l'assuré aux traitements des données**. De plus, par sécurité et par éthique, le ciblage ne sera pas réalisé par individu mais par groupe d'assurés (groupe suffisamment grand) et une pseudonymisation sera réalisée.

II Etude de la déformation de la consommation santé au cours du temps

Après une introduction consacrée à la prévention et au ciblage ainsi qu'une présentation des données utilisées, **cette nouvelle partie consiste à étudier l'évolution de la consommation santé au cours du temps des assurés**.

Un premier objectif est de justifier que l'assureur, via les libellés de prestations, dispose de données permettant de comprendre la consommation santé des assurés et ainsi de cibler des actions de prévention. Cela se traduira notamment par une recherche de lien entre les différents actes et la notion de « parcours de soins ».

Le deuxième objectif est de trouver une méthodologie afin de rendre ces données exploitables. Il s'agira d'une méthode de classification non supervisée, développée auparavant par ADDACTIS pour de la classification à un instant *t*. Dans ce mémoire, la dimension temporelle (au cours du temps) sera ajoutée et cela débouchera sur la notion de « parcours de classes de risque ».

II.1 La compréhension de la consommation santé par les libellés d'actes

En France, de nombreuses prestations sont délivrées sous ordonnance. Ainsi, les examens sont généralement réalisés dans le respect du parcours de soins et sont médicalement justifiés. De plus, indépendamment de cela, la consommation santé est structurée. Il en découle assez intuitivement l'existence de liens logiques entre les différents actes consommés.

Ainsi, **bien que l'assureur ne connaisse pas le prescripteur ou les raisons qui ont amené à la consommation d'un acte de soins, il devrait malgré tout retrouver une certaine logique dans la consommation santé de ces assurés.**

Cette hypothèse bien qu'intuitive doit être vérifiable sur notre portefeuille. Sans cela, l'hypothèse à la base de ce mémoire (compréhension de la consommation santé par les libellés de prestations) serait réfutable.

Dans les parties II.1 et II.2, les statistiques seront réalisées sur les postes de soins (regroupement de libellés d'actes) : Pharmacie, Généraliste, Spécialiste, Hospitalisation, Kinésithérapie, Dentaire, Imagerie, Analyses-Infirmiers, Médecines douces, etc.

Afin de montrer qu'il existe des liens entre les différents actes, de nombreuses études pourraient être réalisées. Trois études vont être menées :

- Etude de la consommation en pharmacie suivant une consultation chez le généraliste

- Etude de la consommation de soins de kinésithérapie suivant une consultation chez le généraliste

- Etude du nombre de séances de kinésithérapie suivant une hospitalisation ou suivant une consultation chez le généraliste

II.1.1 Etude de la consommation en pharmacie suivant une consultation chez le généraliste

L'étude du lien entre les prestations de Généralistes et les prestations de Pharmacie est ici réalisée en regardant la consommation en pharmacie à la suite d'une consultation chez le généraliste.

Tout d'abord, il est à noter que **79% des assurés du portefeuille étant allés chez un généraliste ont également consommé de la pharmacie dans les trente jours suivants**. Ces consultations seront notées « Généraliste* », et les prestations de pharmacie correspondantes « Pharmacie* ».

En parallèle de cela, **seulement 14% des assurés du portefeuille consomment de la pharmacie au moins une fois par mois**, soit bien moins que les 79% précédents. Cela démontre un lien de causalité entre la consultation chez le généraliste et la consommation de pharmacie.

Afin de préciser ce lien de causalité entre les prestations de Généralistes et les prestations de Pharmacie, la distribution des actes Pharmacie* est tracée ci-dessous :

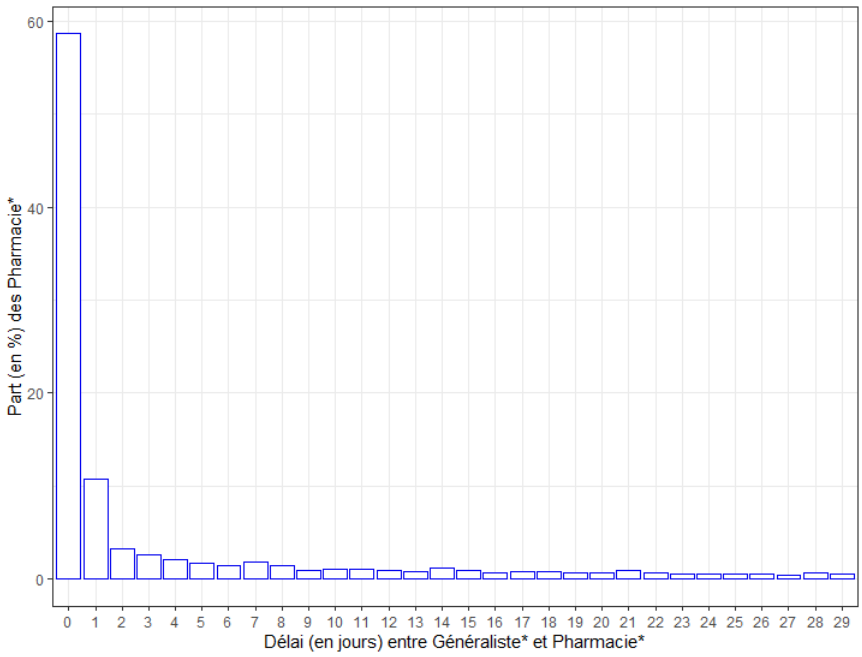


Figure 2.1 – Distribution des délais entre Généraliste* et Pharmacie*

La distribution se lit ainsi :

Parmi les 79% des assurés consommant de la pharmacie dans les trente jours suivants une consultation généraliste, 59% d'entre eux consomment le jour même de la consultation.

Ainsi, il est notable que plus le délai entre Généraliste* et Pharmacie* est élevé, moins la part des 79% des assurés est importante. De plus, une tendance hebdomadaire est présente bien que très légèrement visible sur le graphe car écrasée par la part en 0. Cela valide l'hypothèse de causalité entre les prestations Généraliste et les prestations Pharmacie.

II.1.2 Etude de la consommation de soins de kinésithérapie suivant une consultation chez le généraliste

De même, l'étude du lien entre les prestations de Généralistes et les prestations de Kinésithérapie est réalisée en regardant la consommation en Kinésithérapie à la suite d'une consultation chez le généraliste.

Tout d'abord, il est notable que **63% des assurés du portefeuille étant allés chez un généraliste ont également consommé au moins une séance de kinésithérapie dans les trente jours suivants**. Ces consultations seront notées « Généraliste** », et les prestations de pharmacie correspondantes « Kinésithérapie * ».

En parallèle de cela, **seulement 7% des assurés du portefeuille consomment au moins une séance de kinésithérapie par mois**. Cela démontre un lien de causalité entre la consultation chez le généraliste et la consommation de kinésithérapie.

Ci-dessous la distribution des Kinésithérapie* :

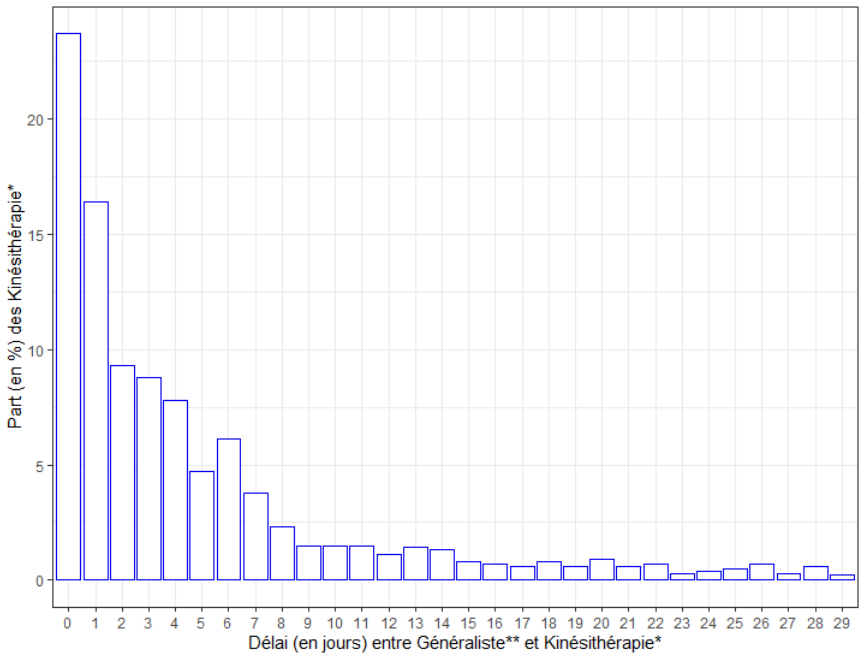


Figure 2.2 – Distribution des délais entre Généraliste** et Kinésithérapie*

Il est notable sur le graphe que plus le délai entre Généraliste** et Kinésithérapie* est élevé, moins la part est importante. Cela valide ainsi le lien de causalité entre la consultation chez le généraliste et la consommation de kinésithérapie.

II.1.3 Etude du nombre de séances de kinésithérapie suivant une hospitalisation ou une consultation chez le généraliste

Au-delà des liens entre les différentes prestations, la compréhension de la consommation santé inclut la compréhension des libellés d'actes selon le contexte. En effet, un seul et unique libellé d'acte ne signifie pas la même chose selon la prestation consommée juste avant.

Exemple :

> Une personne allant chez l'orthodontiste pourrait consommer une prestation Radiologie en effectuant une radiographie des dents. En parallèle, une personne victime d'un accident de la route pourrait consommer elle aussi une prestation Radiologie à la suite de douleurs cervicales. Il s'agit dans les deux cas de radiologie. Cependant, elles peuvent être différenciées dans la base de prestations car la première personne aura consommé une prestation Dentaire auparavant.

Dans cette partie, le contexte sera observé à l'aide de la fréquence (nombre de séances) pour les prestations Kinésithérapie. Précédemment, il a été montré qu'il existe un lien de causalité entre les prestations Généraliste et les prestations Kinésithérapie. Un lien de la même nature existe aussi entre les prestations Hospitalisation et les prestations Kinésithérapie. Il s'agira ici de comparer le nombre de séances de kinésithérapie suivant une consultation chez le généraliste avec le nombre de séances suivant une hospitalisation.

Ci-dessous, la distribution du nombre de séances de kinésithérapie suivant une prestation Généraliste ou suivant une prestation Hospitalisation par tranche de 5 séances.

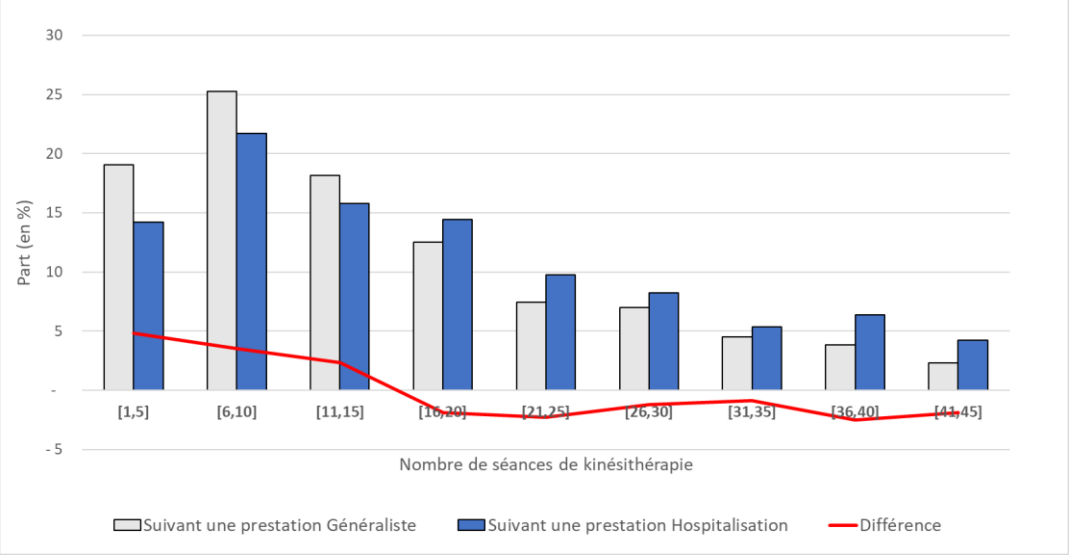


Figure 2.3 – Distribution du nombre de séances de kinésithérapie suivant une prestation Généraliste vs suivant une prestation Hospitalisation

Il est manifeste sur la figure 2.3 que l'hospitalisation amène à des parcours de kinésithérapie généralement plus long que ceux suivant une consultation chez le généraliste. Par exemple, **55% des parcours suivant une hospitalisation nécessitent plus de quinze séances de kinésithérapie contre seulement 43% à la suite d'une consultation chez le généraliste**.

En conclusion, de très simples statistiques permettent rapidement d'identifier le **lien entre les différents actes consommés** et ainsi le potentiel de compréhension de la consommation santé des assurés par les libellés d'actes. Il s'agira dans la suite de ce mémoire d'approfondir l'étude de ces données afin de **comprendre et anticiper la déformation de la consommation santé des assurés.**

II.2 Les parcours de soins

Dans la partie précédente, il a été démontré qu'il est possible de visualiser à l'aide de nos données des liens logiques entre les différents actes consommés et ainsi introduire la notion de parcours. Dans un premier temps, il s'agira dans cette partie de définir la notion de « parcours de soins ». Ensuite, il sera présenté très rapidement les tentatives d'utilisation des parcours de soins. Et enfin, il sera expliqué pourquoi il a été choisi de ne pas poursuivre ces travaux afin de s'orienter vers les classes de risque.

II.2.1 Définition

Dans ce mémoire, **le terme « parcours de soins » est utilisé pour définir une succession logique de prestations consommées par l'assuré.**

Exemple :

Date de survenance	Famille d'acte
15/07/2018	Généraliste
15/07/2018	Pharmacie
16/08/2018	Spécialiste
18/08/2018	Optique

Date de survenance	Famille d'acte
01/02/2019	Généraliste
05/02/2019	Kinésithérapie
12/02/2019	Kinésithérapie
17/02/2019	Kinésithérapie

Figure 2.4 – Exemple illustratif de consommation santé d'un assuré à des fins d'identification de parcours de soins

Dans cet exemple, il est aisé d'identifier (par logique) trois parcours de soins :

- Généraliste → Pharmacie
- Spécialiste → Optique
- Généraliste → Kinésithérapie → Kinésithérapie → Kinésithérapie

En effet, les délais entre les parcours sont suffisamment importants pour ne pas hésiter sur les choix de fin de parcours et il est compréhensible que :

- Dans le premier parcours, le généraliste a prescrit des médicaments à son patient.
- Dans le deuxième parcours, la prestation Spécialiste est en fait une consultation ophtalmologique.

- Dans le troisième parcours, le généraliste a prescrit plusieurs séances de kinésithérapie.

II.2.2 Comment identifier et séparer les parcours ?

La consommation santé d'un assuré, présentée en II.2.1 est un exemple illustratif où l'identification des parcours de soins était plutôt simple. Cependant, en pratique, l'identification des parcours est beaucoup plus complexe. Ils peuvent en effet s'entremêler ou tout simplement être incompréhensibles. De plus, il est nécessaire d'automatiser le processus d'identification de parcours de soins. Pour ce faire, une définition claire et programmable est exigée. Après de nombreux tests, **la définition suivante a été retenue** :

- **Un parcours débute par une prestation dite « pivot »**

  Il est possible de définir plusieurs prestations « pivot » mais il sera choisi la prestation Généraliste car le médecin généraliste est souvent prescripteur de la suite du parcours de soins.

- **Un parcours a une durée maximale**

  Il sera choisi de retenir arbitrairement trente jours (durée permettant d'avoir suffisamment d'actes tout en restant interprétable).

En pratique, un programme va donc parcourir la base de prestations par assuré et par ordre chronologique. Dès lors qu'il va rencontrer la prestation Généraliste, il débutera un nouveau parcours de soins. En parallèle de cela, il terminera un parcours de soins si la durée de celui-ci dépasse trente jours où s'il débute un nouveau parcours de soins.

Cette définition imparfaite présente notamment les limites suivantes :

- Les parcours entremêlés ne sont pas pris en compte
- Des prestations ne sont incluses dans aucun parcours si elles ne sont pas précédées d'une prestation Généraliste
- Les parcours ne débutent pas toujours par l'acte pivot et peuvent durer plus de trente jours

La définition retenue est ainsi grandement critiquable mais il s'agit du moins mauvais choix et d'une définition implémentable. En effet, la complexité temporelle a été prise en compte afin de choisir cette définition. Une définition trop complexe serait difficilement implémentable sur une base de prestations volumineuse, telle la base étudiée et ses 1,58 millions de lignes.

II.2.3 Proposition de simplification et de classification

Les parcours de soins obtenus par la définition précédente ne sont pas directement exploitables car trop nombreux. En effet, il existe des milliers de combinaisons possibles. Des premières simplifications ont donc été réalisées afin de simplifier les parcours.

Exemple : transformer une succession de Kinésithérapie (au-delà de trois consécutifs) en « Kinésithérapie régulière ».

Par la suite, **une méthode de partitionnement de données (*clustering*) a été utilisée afin de déterminer des groupes de parcours de soins présentant des risques similaires**. La méthode est la suivante :

- On représente chaque parcours par des variables de comptage d'actes et de liaisons d'actes

Exemple :

« Généraliste → Kinésithérapie → Kinésithérapie → Kinésithérapie → Pharmacie → Kinésithérapie → Généraliste »

La représentation du parcours ci-dessus est la suivante :

Lien	Généraliste → Kinésithérapie	Kinésithérapie → Kinésithérapie	Kinésithérapie → Pharmacie	Pharmacie → Kinésithérapie	Kinésithérapie → Généraliste
Nombre d'apparitions	1	2	1	1	1

Acte	Généraliste	Kinésithérapie	Pharmacie
Nombre d'apparitions	2	4	1

Figure 2.5 – Représentation d'un parcours de soins par des variables de comptage d'actes et de liaisons d'actes

- On effectue ensuite une CAH[1] sur les variables de comptage

L'idée à travers la démarche de classification est de pouvoir interpréter les parcours de soins et ainsi associer un risque à chaque parcours.

[1] La classification ascendante hiérarchique (CAH) est une méthode de partitionnement de données (*clustering*).

II.2.4 Résultats

Les résultats issus de la méthode de classification des parcours présentée ci-dessus ne sont pas satisfaisants. Quelques classes regroupent la majorité des parcours et les autres seulement quelques parcours facilement interprétables. Il n'est pas ressorti d'information par la classification.

De plus, comment interpréter une classe ayant un nombre important de parcours ? Il est possible de regarder la moyenne par classe des variables de comptage.

Cependant, en réalisant cela, plusieurs problèmes peuvent apparaitre :

- Il est possible d'obtenir des moyennes semblables pour différentes classes
- La discrimination par variable de comptage n'est pas toujours adaptée

  Il suffit en effet d'une interversion entre deux actes ou l'insertion d'un acte indépendant pour que deux parcours ne soient pas dans la même classe.

  Exemple :

  Les deux parcours ci-dessous, bien que similaires, ne sont pas dans la même classe à cause du caractère discriminant de la liaison « Généraliste → Analyses-Infirmiers ».

  - « Généraliste → Pharmacie → Analyses-Infirmiers → Généraliste »
  - « Généraliste → Analyses-Infirmiers → Pharmacie → Généraliste »

A la suite de ces résultats, un constat a été dressé : aucune définition satisfaisante et implémentable permettant de travailler sur les parcours de soins n'a pu être retenue.

Ainsi, il sera donc retenu pour la suite des travaux une autre approche que celle des parcours de soins. Cette **nouvelle approche (« les classes de risque »)** sera présentée dans la partie suivante et sera ensuite complétée afin de **retrouver la dimension temporelle qui caractérise l'approche parcours de soins.**

II.3 Méthode de classification : d'une approche statique à une approche dynamique

Dans cette partie, la méthode qui sera présentée est une méthode développée par ADDACTIS [Gauchon, 2020]. Elle sera utilisée dans les travaux présentés dans la suite du mémoire. Il s'agit d'une **méthode de classification non supervisée permettant de cartographier à un instant *t* un portefeuille d'assurance santé**.

II.3.1 Présentation générale de la méthode

La méthode, **non supervisée**, permet d'**associer à chaque assuré,** par l'analyse de sa consommation santé, **une classe de risque (groupe d'assurés homogènes au niveau de leurs risques).** Elle utilise une approche fréquence, autrement dit, la classification n'est ni réalisée à l'aide des montants des prestations, ni à l'aide des caractéristiques des assurés (âge, sexe…) mais seulement grâce à la **fréquence de consommation de chaque acte pour chacun des assurés**.

Il est présenté ci-dessous un exemple illustratif de cartographie de portefeuille obtenue par la méthode de classification non supervisée.

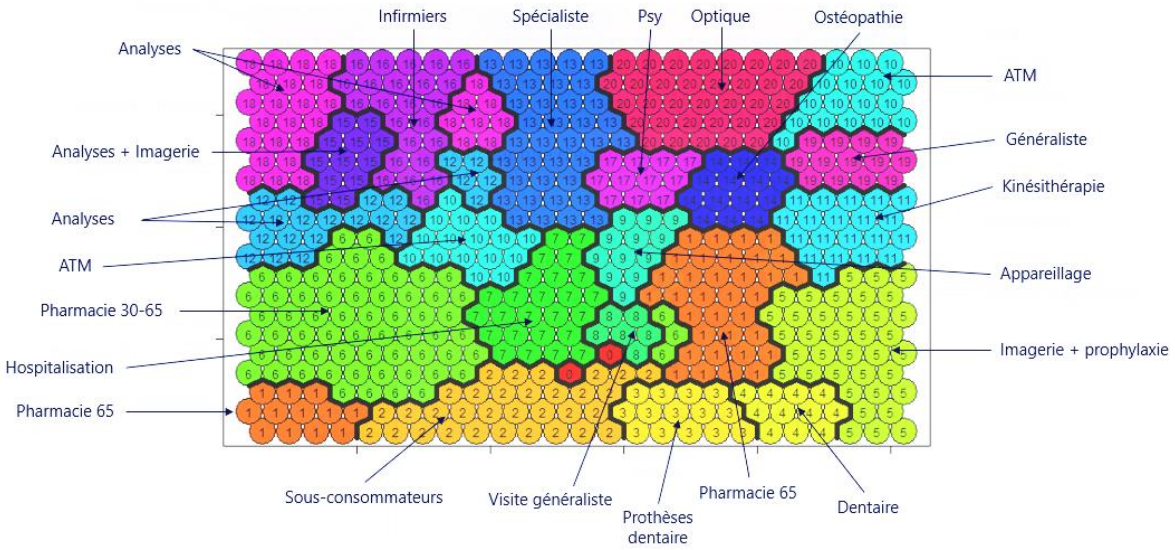


Figure 2.6 – Exemple de cartographie de portefeuille

Sur cette carte, chaque groupe d'assurés est associé à une classe de risque (groupe de risques identifiables) indiquant une homogénéité de leur consommation santé.

Cette méthode, brièvement présentée ici, est une succession de plusieurs étapes :

- Création d'une matrice de fréquence de consommation par assuré

- Réduction de dimension par mSDA[2]
- Cartographie par cartes auto-organisatrices de Kohonen et classification par CAH (classification ascendante hiérarchique)[3]

Il ne sera pas présenté en détails les méthodes et outils utilisés dans ce mémoire. En effet, tout cela est présenté et expliqué dans les mémoires cités ci-dessus et n'est pas l'objet de ces travaux. Cependant quelques points sont à décrire afin de pouvoir comprendre comment et pourquoi cette méthode sera utilisée dans le cadre de ce mémoire.

a. Données utilisées

Les données en entrée du mSDA doivent être sous la forme de matrice de fréquence. La première étape consiste à transformer la base de prestations en matrice de fréquence :

Identifiant bénéficiaire	Date de survenance	Libellé d'acté	DE	RC	RSS	Quantité
3	01/01/2018	Pharmacie	1,02€	0,71€	0,31€	1
2	01/01/2018	Pharmacie	1,02€	0,71€	0,31€	1
1	03/01/2018	Auxiliaires médicaux	3,78€	1,51€	2,27€	1
2	01/02/2018	Prothèses	350,00€	215,00€	75,25€	1
1	10/02/2018	Auxiliaires médicaux	23,43€	11,17€	12,26€	1
3	11/02/2018	Pharmacie	2,92€	1,02€	1,90€	1
…	…	…	…	…	…	…

	Auxiliaires médicaux	…	Pharmacie	Prothèses
1	2	…	0	0
2	0	…	1	1
3	0	…	2	0
…	…	…	…	…

Figure 2.7 – Transformation de la base de prestations en matrice de fréquence

[2] mSDA : *marginalized Stackled Denoising Autoencoder* ; NMF : *Non-negative Matrix Factorization*. Le lecteur pourra se rapporter à l'annexe c) pour une brève présentation de ces deux concepts.
[3] Le lecteur pourra se rapporter à l'annexe d) afin de comprendre les cartes de Kohonen.

Il est ainsi important de comprendre que la classification en classes de risque ne prend en compte que les fréquences d'apparition des libellés d'actes par assuré.

b. Données en sortie

Une fois la classification réalisée (pour un nombre de classes choisi), il est obtenu :

- **Une représentation de classes d'assurés par carte de Kohonen**

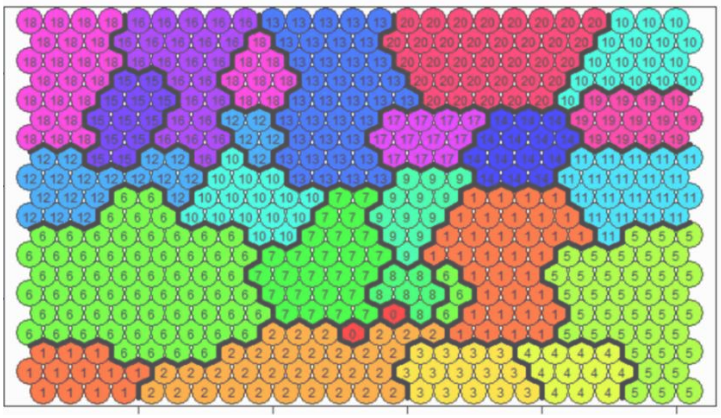


Figure 2.8 – Représentation de classes d'assurés par carte de Kohonen

- **Une matrice d'interprétation des classes d'assurés**

La figure ci-dessous représente une matrice d'interprétation simplifiée des classes d'assurés. Les colonnes correspondent aux libellés d'actes et les lignes aux classes de risque. Les couleurs, quant à elles, correspondent aux proportions de consommants par libellé d'actes et par classe. Plus la cellule est foncée, plus la proportion est élevée.

Le curseur a été volontairement positionné sur la cellule en ligne 17 et en colonne « consultation neuro psych » afin de montrer comment utiliser cette matrice. La classe 17, par exemple, correspond au risque Psy[4] car plus de 80% des assurés consomment des prestations « consultation neuro psych ». Il est à noter que les classes sont définies à partir du risque le plus discriminant mais cela n'empêche pas les assurés de la classe de consommer d'autres types de prestations dans des proportions plus faibles.

[4] Psy : classe de risque composée principalement de consultation de psychiatre et de psychologue, reflétant les troubles de santé mentale

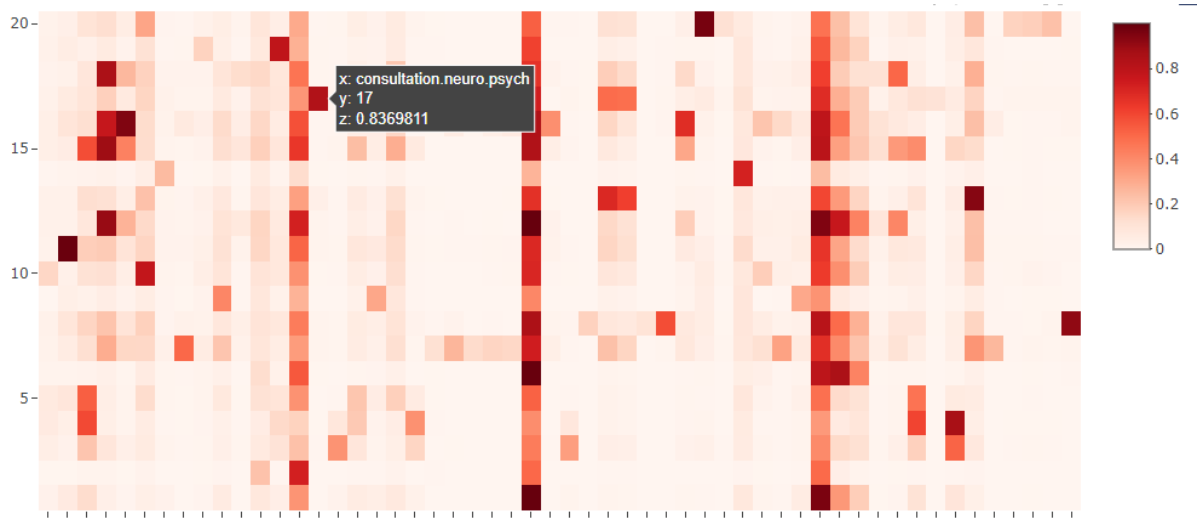


Figure 2.9 – Matrice d'interprétation simplifiée des classes d'assurés

- Des statistiques par classe : âge moyen, répartition par sexe, RC moyen, DE moyenne, effectif, etc.

Une étude de la matrice d'interprétation et des statistiques par classe devra être réalisée afin d'interpréter les classes. De plus, il est important de souligner que la méthode présente des résultats interprétables, les classes correspondent bien à des groupes de risques facilement identifiables. Une fois ce travail accompli, chaque assuré consommant appartient à une classe de risque sur l'ensemble de la période considérée :

Identifiant bénéficiaire	Classe
1	**Infirmiers**
2	**Hospitalisation**
3	**Dentaire**
…	**…**

Figure 2.10 – Données en sortie de la classification en classes de risque

Remarques :

- Appartenir à une classe de risque signifie présenter des fréquences de consommation d'actes de soins similaires aux autres assurés appartenant à la même classe.

- L'interprétation de la classe de risque n'est pas une vérité absolue pour un assuré donné mais est un résumé de la consommation santé (en fréquence) de l'ensemble des assurés de la classe.

Par exemple : un assuré appartenant à la classe Dentaire ne consomme pas obligatoirement du Dentaire. A priori, Il a seulement une consommation santé se rapprochant (en fréquence) de la consommation santé d'assurés consommant du Dentaire.

II.3.2 Application à l'étude de la déformation de la consommation santé au cours du temps

La méthode de classification présentée ci-dessus fournit une vision statique du portefeuille (à un instant *t*). Afin de conserver le même « esprit « que les parcours de soins, autrement dit, la dimension temporelle (évolution au cours du temps), il sera effectué plusieurs cartographies à différents instants via une discrétisation du temps (une cartographie par trimestre). Ces cartographies seront ensuite concaténées afin de créer des « **parcours de classes de risque** ». Il s'agit d'une analogie avec les parcours de soins à une maille plus élargie.

Exemple de parcours de classes de risque :

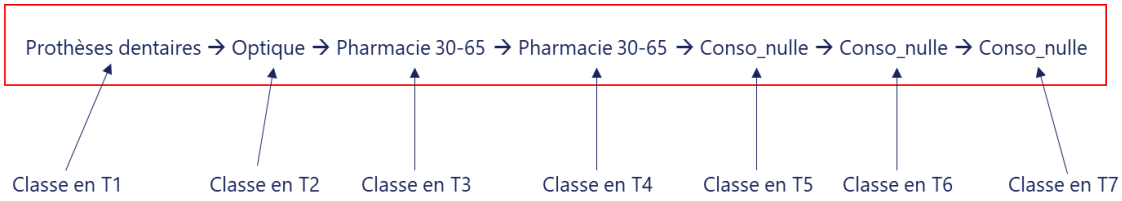


Figure 2.11 – Parcours de classes de risque d'un assuré du trimestre 1 au trimestre 7

Tout cela sera expliqué plus en détails dans la partie II.5.

II.4 Application de la méthode de classification

Dans cette partie, la méthode utilisée afin de réaliser plusieurs cartographies à différents instants sera d'abord présentée. Ensuite, les classes de risque obtenues seront étudiées afin de dresser une première analyse de portefeuille.

II.4.1 Création des classes de risque

Tout d'abord, avant d'entrer dans les détails de la méthodologie développée, il est important de définir la maille et les données utilisées :

- **Maille trimestre**

  Il sera effectué une cartographie par trimestre. Cette hypothèse a été choisie comme compromis entre la durée maximale supposée d'un parcours et la volonté de pouvoir construire des parcours de classes de risque suffisamment grands. D'un point de vue technique, elle permet de disposer de suffisamment de données afin de réaliser une série de cartographies robustes.

- **La cartographie est réalisée du trimestre 1 au trimestre 7** (01/01/2018-30/09/2019)

  Les données du trimestre 8 (01/10/2019-31/12/2019) ne sont pas incluses dans la cartographie en raison de leur date de réception, qui était début janvier. Nous avons pris cette décision car nous considérons que ces données n'étaient pas encore stabilisées à cette date et qu'elles pourraient fausser l'analyse.

Méthodologie

La méthode la plus simple serait de scinder la base de données par trimestre et de réaliser sept classifications. Cependant, les classes obtenues ne seraient pas les mêmes entre les différents trimestres (du fait de l'instabilité des cartes de Kohonen [Hermet, 2020]). Il ne serait donc pas possible de les comparer. Par exemple, deux classes ayant la même interprétation pourraient ne pas se ressembler en termes de fréquence et de statistiques (âge, sexe, montants, etc.). Il sera donc réalisé qu'une seule classification de la manière suivante :

- **Préparation des données**

  Pour chaque prestation, le trimestre correspondant à la date de survenance de l'acte et l'identifiant du bénéficiaire ayant consommé l'acte sont concaténés.

  Exemple :

Identifiant bénéficiaire	Date de survenance	Nouvel ID
2240	01/01/2018	**T1**2240

Figure 2.12 – Création d'une nouvelle variable « Nouvel ID » prenant en compte la temporalité

- **Classification unique**

  Une seule et unique classification est réalisée à partir des identifiants « Nouvel ID ». Autrement dit, dans ce cadre, un assuré est considéré comme sept assurés différents. Pour reprendre l'exemple ci-dessus, l'assuré 2240 ne sera pas classé mais les individus **T1**2240, …, **T7**2240 le seront.

II.4.2 Etude des classes de risque obtenues

a. Vue d'ensemble

Après application de la méthode présentée en II.3, les classes de risque obtenues sont les suivantes (représentées sous forme de carte de Kohonen) :

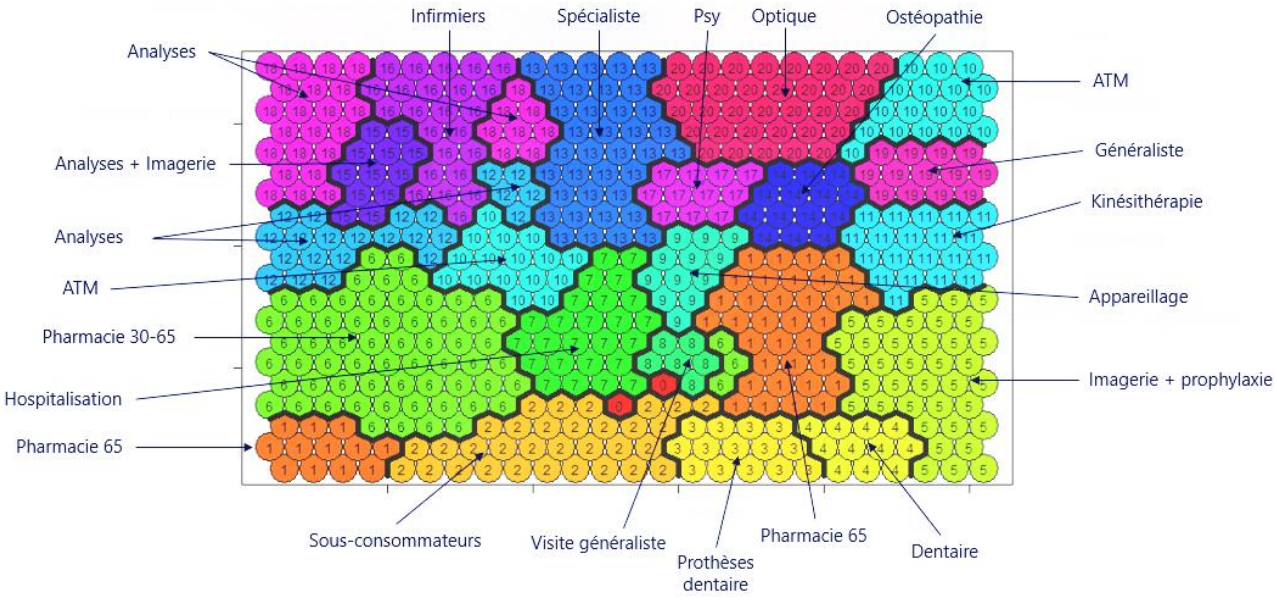


Figure 2.13 –Cartographie du portefeuille par carte de Kohonen

Cette carte de Kohonen présente les **20 classes de risque** construites par application de la méthodologie présentée plus haut, il s'agit de 20 « représentations » de comportements homogènes de consommation reflétant le risque santé. Elles seront reprises tout au long de cette étude.

A noter que parmi ces vingt classes, trois contiennent à elles seules un tiers des consommants :

- Pharmacie 30-65 (13% des consommants)
- Pharmacie 65 (12% des consommants)
- Sous-consommateurs (9% des consommants)

Remarque : la classification étant réalisée à partir de la base de prestations, seuls les assurés consommants sont classés. Afin d'obtenir une base complète, une classe sera ensuite artificiellement attribuée aux assurés non consommants : la classe nommée *Consommation nulle*.

b. Des profils spécifiques selon les classes

L'analyse des statistiques par classe permet d'associer à chacune des classes un profil type. Ici, deux classes correspondent à un profil type plus jeune que la moyenne et quatre classes à un profil type plus âgé que la moyenne :

Classe	Age moyen	▲Age*
Sous-consommateurs	41 ans	-3 ans
Analyses (classe 18)	42 ans	-2 ans
…	…	…
Kinésithérapie	46 ans	2 ans
Analyses (classe 12)	46 ans	2 ans
Analyses + Imagerie	46 ans	2 ans
Infirmiers	46 ans	2 ans

*▲Age = Age moyen – Age moyen du portefeuille

Figure 2.14 – Ages moyens atypiques par classe de risque

Cette analyse est intéressante car elle confirme des réalités bien connues. En effet, il parait logique que les classes Infirmiers ou Kinésithérapie soient associées à des assurés plutôt âgés. D'autre part, elle permet de différencier les deux classes Analyses.

De même, deux classes sont atypiques par la répartition par sexe :

Classe	% d'hommes	▲% d'hommes*
Spécialiste	39%	-22%
Ostéopathie	70%	+9%

*▲% d'hommes = % d'hommes – % d'hommes du portefeuille

Figure 2.15 – Répartitions par sexe atypiques par classe de risque

Il s'agit ici d'une sélection de résultats, le lecteur pourra se référer à l'annexe f pour la caractérisation statistique complète de chacune des 20 classes de risque.

c. Des dépenses en santé différentes selon les classes

Les statistiques des dépenses moyennes par classe permettent de mieux comprendre les spécificités de chacune des classes. Il est possible par exemple de les comparer à la moyenne du portefeuille.

En moyenne, par trimestre, sur l'ensemble du portefeuille, le RC moyen par assuré consommant est de 159 € et le reste à charge moyen par assuré consommant est de 44 €.

Certaines classes sont remarquables car plus coûteuses pour l'assureur ou pour l'assuré que la moyenne du portefeuille :

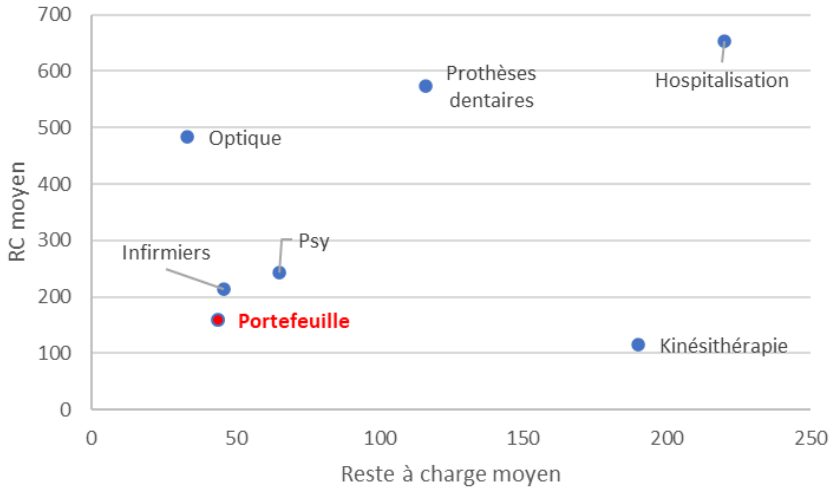


Figure 2.16 – Classes de risque coûteuses pour l'assureur ou l'assuré

A noter : les trois classes les plus représentées présentent des montants moyens bien en dessous de la moyenne du portefeuille.

d. Saisonnalité et dérive

Les classes étant les mêmes à chaque trimestre, il est possible d'observer des évolutions intra-classes entre le trimestre 1 et le trimestre 7.

Tout d'abord, il est possible d'observer de la saisonnalité en fréquence (nombre de personnes par classe) et en montants (en DE) :

- Quatre classes sont sur-représentées en hiver (Trimestres 1, 4 et 5) :

  Les graphes ci-dessous représentent les fréquences par trimestre divisées par la fréquence moyenne des quatre classes.

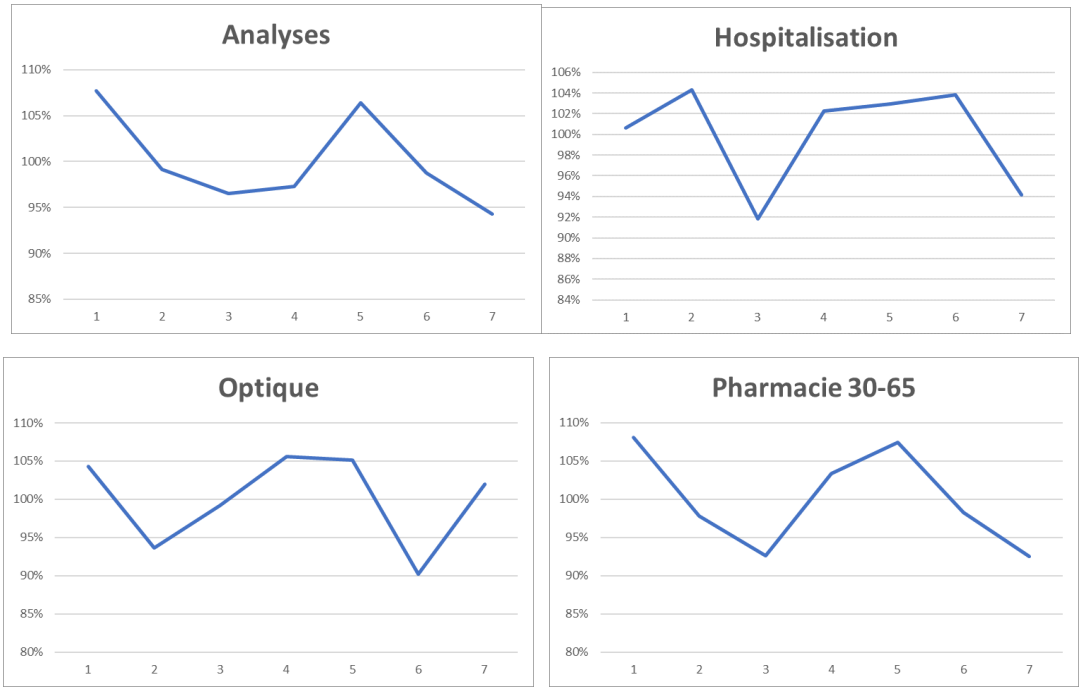


Figure 2.17 – Classes de risque sur-représentées en hiver

- Deux classes sont sur-représentées en été (Trimestres 3 et 7) :

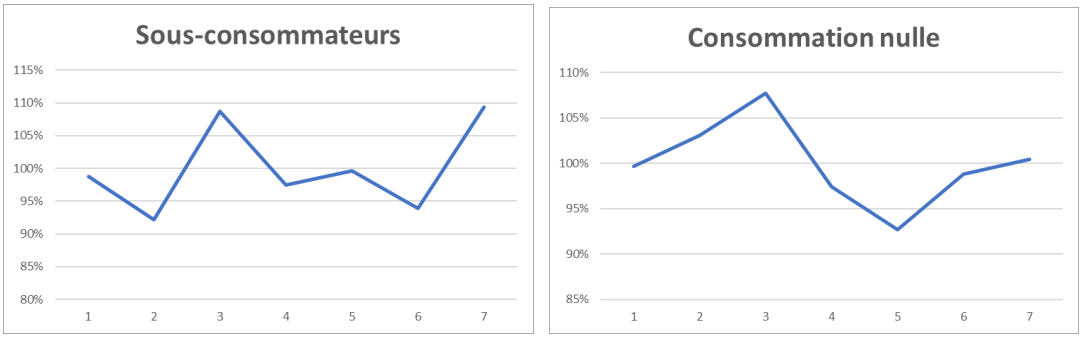


Figure 2.18 – Classes de risque sur-représentées en été

- Six classes sont plus coûteuses en hiver

  Les graphes ci-dessous représentent les dépenses engagées en euros par trimestre des six classes.

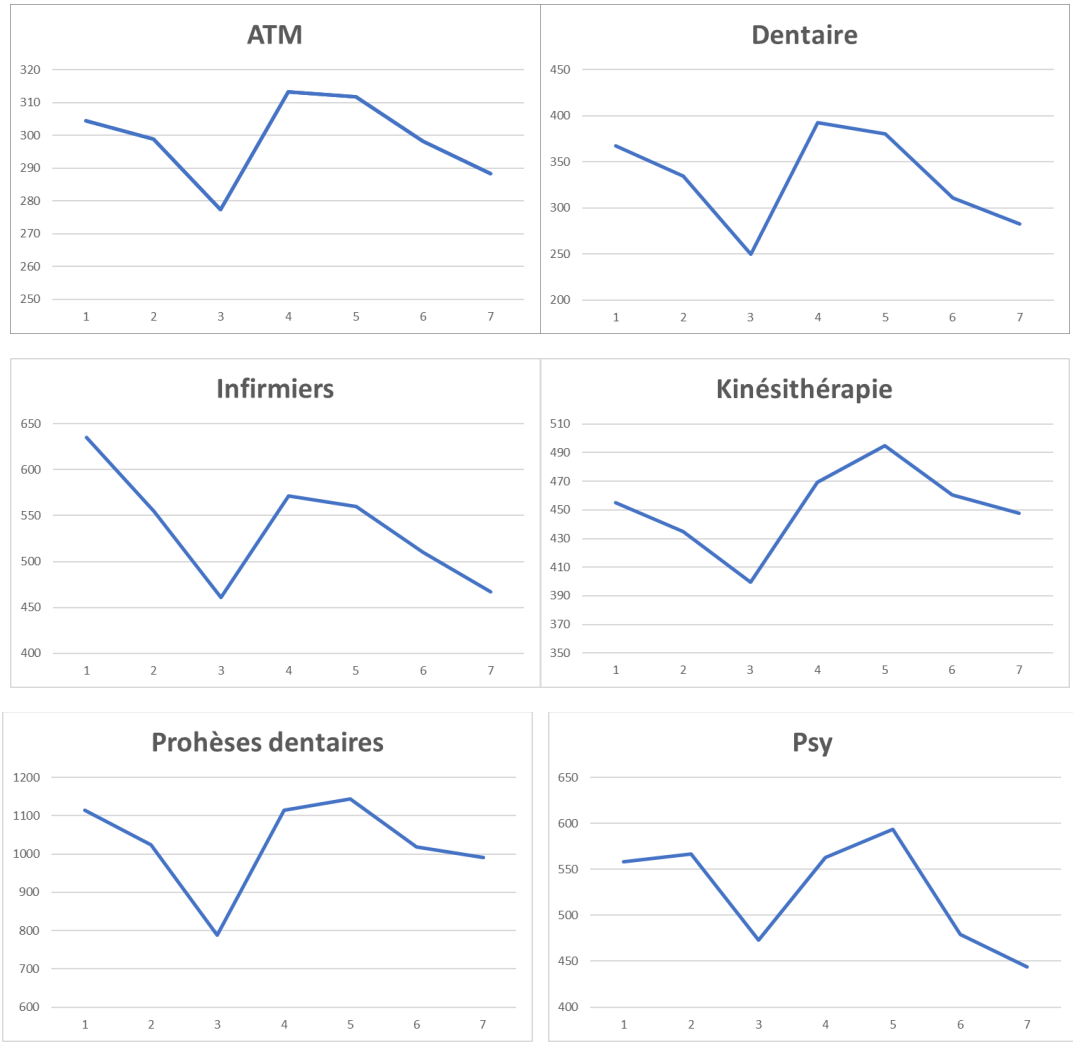


Figure 2.19 – Classes de risque plus coûteuses en hiver

De plus, il est possible d'observer les tendances par classe sur les sept trimestres afin d'observer ou non des dérives en fréquence et en DE :

- Il n'y a pas de dérive majeure en DE
- Les effectifs de 2 classes ont suivi une tendance haussière sur les sept trimestres observés :

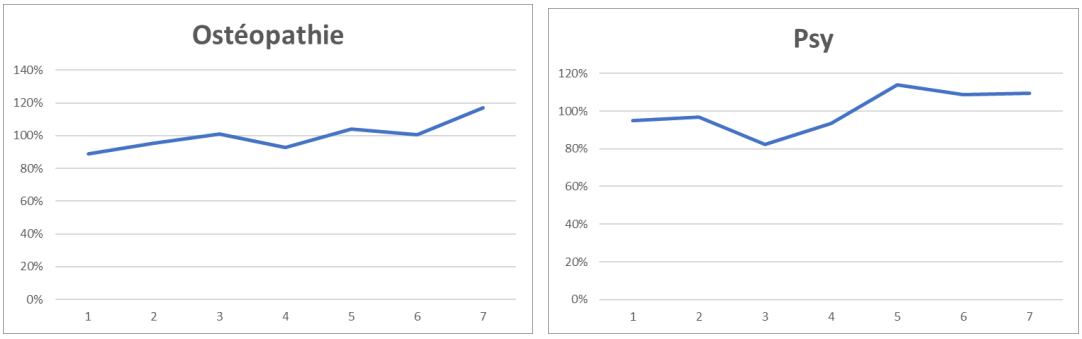


Figure 2.20 – Classes de risque avec une tendance haussière en fréquence

II.5 Etude des parcours de classes de risque

Pour rappel, la cartographie du portefeuille est réalisée du T1 (trimestre 1) au T7 (trimestre 7). Il est donc obtenu une classe par assuré et par trimestre. Un parcours de classes de risque est tout simplement la concaténation des classes par assuré.

Exemple :

Voici ci-dessous le résultat de la cartographie pour le bénéficiaire 3 :

Identifiant bénéficiaire	Classe de risque T1	Classe de risque T2	Classe de risque T3	Classe de risque T4	Classe de risque T5	Classe de risque T6	Classe de risque T7
3	Prothèses dentaires	Optique	Pharmacie 30-65	Pharmacie 30-65	Conso nulle[5]	Conso nulle	Conso nulle
...	...	...	...	...	...	...	...

Figure 2.21 – Exemple de cartographie d'un assuré au cours du temps

Son parcours de classes de risque est donc le suivant :

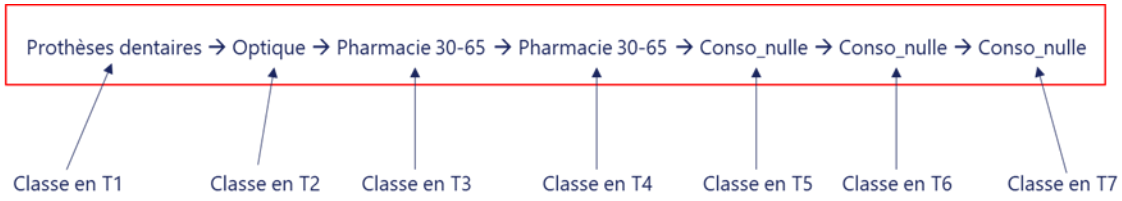


Figure 2.22 – Exemple parcours de classes de risque d'un assuré

Dans cette partie, il s'agira d'étudier les parcours de classes de risques obtenus afin de mieux comprendre la consommation santé du portefeuille.

II.5.1 Matrice de transition inter-classes

Un premier outil intéressant afin d'analyser les parcours de classes de risques est la matrice de transition, autrement dit, les probabilités de passage d'une classe de risque à une autre.

a. Création de la matrice de transition

Les probabilités de passage sont calculées au global (sur les six paires de trimestres 1-2, 2-3, ..., 6-7) par concaténation. Ci-dessous, l'explication par un exemple (portefeuille avec deux assurés) :

[5] Conso nulle : Consommation nulle (Absence de consommation)

Identifiant bénéficiaire	Classe de risque T1	Classe de risque T2	Classe de risque T3	Classe de risque T4	Classe de risque T5	Classe de risque T6	Classe de risque T7
1	Prothèses dentaires	Optique	Pharmacie 30-65	Pharmacie 30-65	Conso nulle	Conso nulle	Conso nulle
2	Pharmacie 30-65	Conso nulle	Pharmacie 30-65	Spécialiste	Conso nulle	Pharmacie 65	Pharmacie 65

Figure 2.23 – Exemple illustratif : classes de risque d'un portefeuille de deux assurés

Les données vont être transformées en paires de trimestres comme ci-dessous :

	T_i	T_{i+1}
T1 à T2	Prothèses dentaires	Optique
	Pharmacie 30-65	Conso nulle
T2 à T3	Optique	Pharmacie 30-65
	Conso nulle	Pharmacie 30-65
T3 à T4	Pharmacie 30-65	Pharmacie 30-65
	Pharmacie 30-65	Spécialiste
T4 à T5	Pharmacie 30-65	Conso nulle
	Spécialiste	Conso nulle
T5 à T6	Conso nulle	Conso nulle
	Conso nulle	Pharmacie 65
T6 à T7	Conso nulle	Conso nulle
	Pharmacie 65	Pharmacie 65

Figure 2.24 – Exemple illustratif : paires de classes de risque d'un portefeuille de deux assurés

Une fois les données sous ce format, les probabilités de passage d'une classe a à une classe b sont notées P_{a-b} et sont calculées de la manière suivante :

$$P_{a-b} = \frac{Fréquence\ de\ la\ paire\ a-b}{Fréquence\ de\ la\ classe\ a\ en\ T_i}$$

On obtient ainsi la matrice de transition suivante :

$$\begin{pmatrix} P_{1\to1} & \cdots & P_{1\to M} \\ \vdots & \ddots & \vdots \\ P_{M\to1} & \cdots & P_{M\to M} \end{pmatrix}$$

Avec M le nombre de classes de risque

$P_{i\to j}$ la probabilité de passage de la classe i à la classe j comme définie ci-dessus

b. Analyse des résultats

Par la matrice de transition[6] décrite précédemment, il est possible de dresser quelques conclusions :

- **Certaines classes sont plus absorbantes que d'autres**

  Une **classe absorbante** présente une probabilité de maintien élevée (probabilité de rester dans la même classe).

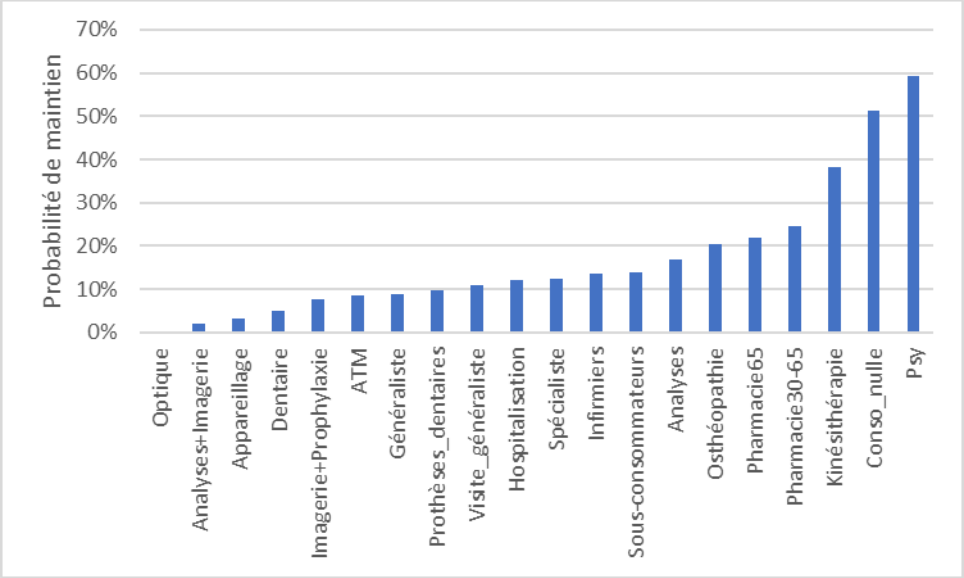


Figure 2.25 – Probabilités de maintien du portefeuille

De plus, il est possible de construire la matrice de transition sur les populations de moins de 50 ans et les populations de plus de 50 ans. On peut ainsi dresser d'autres conclusions :

- **Des classes sont plus absorbantes sur la sous-population des plus de 50 ans**

[6] Par souci de lisibilité, seuls les résultats principaux sont présentés ici. Le lecteur pourra retrouver l'intégralité de la matrice de transition en annexe g.

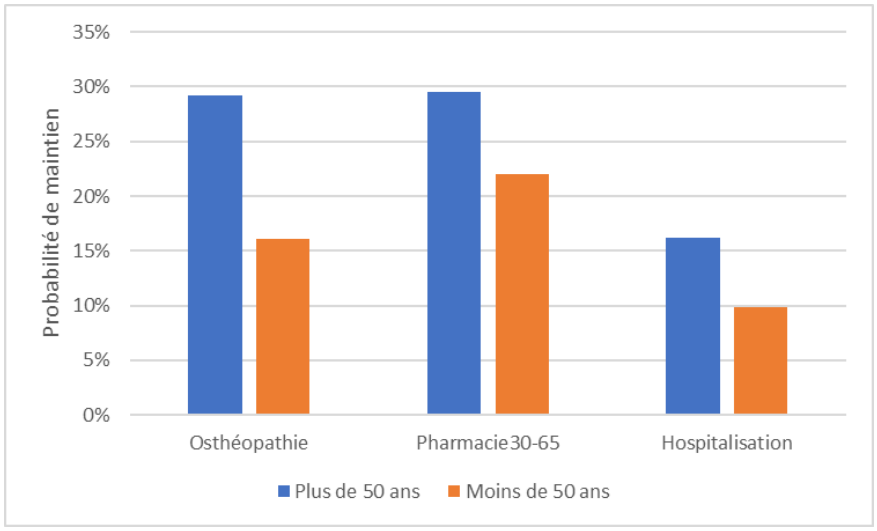


Figure 2.26 – Probabilités de maintien plus élevées chez les plus de 50 ans

- **La probabilité de passage en Consommation nulle est plus élevée chez les moins de 50 ans** (quelle que soit la classe à l'origine)

II.5.2 Déformation du coût du risque selon le parcours de consommation

Dans la partie II.1.3, il a été montré par un exemple que selon le parcours de consommation santé, un acte de soins peut être plus ou moins grave (la gravité était mesurée avec la fréquence). En effet, le nombre de séances de kinésithérapie suivant une hospitalisation différait statistiquement du nombre de séances suivant une visite chez le généraliste.

Ici, la notion de gravité selon le parcours de consommation santé sera observée à travers les dépenses engagées moyennes et les classes de risque. On peut ainsi dresser rapidement quelques constats :

- **La classe Psy est une classe systématiquement aggravante**[7].

  En effet, un assuré ayant appartenu à la classe Psy en T_{i-1} (trimestre précédent) aura en moyenne des DE de 566 € tandis que la moyenne des DE quelle que soit la classe en T_{i-1} est de 242 €. Cette différence s'observe en fonction de la classe de risque en T_i :

[7] Classe aggravante : Les dépenses engagées d'un assuré ayant été classé dans cette classe de risque en T_{i-1} sont supérieures aux dépenses engagées d'un assuré moyen du portefeuille.

Classe en T_{i-1}	Classe en T_i	Moyenne des DE en T_i	Moyenne des DE en T_i pour les assurés venant de Psy en T_{i-1}	Différence	Effectifs des assurés venant de Psy en Ti-1
Psy	Analyses	310€	670€	+360€	57
Psy	Pharmacie 65	176€	222€	+46€	44
Psy	Spécialiste	235€	470€	+235€	73
…	…	…	…	…	…

Figure 2.27 – Exemple pour 3 classes de l'aggravation du risque par la classe Psy en T_{i-1}

- **La classe Kinésithérapie est une classe aggravante dans la grande majorité des cas mais dans une moindre mesure que la classe Psy.**

  Un assuré ayant appartenu à la classe Kinésithérapie en T_{i-1} aura en moyenne des DE de 434 € tandis que la moyenne des DE quelle que soit la classe en T_{i-1} est de 242 €.

- **A contrario, les classes Sous-consommateurs et Consommation nulle sont des classes diminuant les DE moyennes quelle que soit la classe observée.**

  En effet, un assuré ayant appartenu à la classe Consommation nulle en T_{i-1} aura en moyenne des DE de 104 € (169 € pour Sous-consommateurs) tandis que la moyenne des DE quelle que soit la classe en T_{i-1} est de 242 €.

- **Le maintien en classe Psy n'est pas la situation la plus aggravante.**

  Un assuré ayant appartenu à la classe Généraliste en T_{i-1} et classé en Psy en T_i aura en moyenne des DE de 1 096 € tandis qu'un assuré ayant appartenu à la classe Psy en T_{i-1} et classé en Psy en T_i aura en moyenne des DE de 584 €. Autrement dit, une personne ayant un suivi psychiatrique/psychologique sur au moins 2 trimestres consomme moins (en DE) qu'une personne venant de se faire prescrire un suivi psychiatrique/psychologique. Il est ainsi possible de supposer qu'un suivi psychiatrique/psychologique permet d'améliorer d'état de santé général d'un assuré. Cela reste malgré tout une situation légèrement aggravante car la moyenne des dépenses engagées par un assuré classé en Psy est de 525 €.

Ces analyses permettent de mieux comprendre la consommation santé du portefeuille. Mêlées à des avis d'experts (médecins, préventeurs, etc.), elles peuvent notamment aiguiller des actions de prévention.

Ainsi, dans cette partie, il a été montré qu'il est possible d'analyser et comprendre statistiquement **l'évolution de la consommation santé des assurés** grâce aux libellés d'actes de soins présents dans la **base des prestations**. Pour ce faire, une méthode de **classification non supervisée** est proposée**.** De cette façon, il est obtenu des **classes de risque par assuré et par trimestre.** Ces classes sont ensuite concaténées afin de créer des **parcours de classes de risque.**

III Construction d'une méthodologie de prédiction du risque santé par entreprise

Dans la partie précédente, une méthode a été développée afin d'étudier la consommation santé du portefeuille par assuré, via la création de classes de risque sur les périodes passées.

Ces travaux peuvent permettre d'orienter des plans d'actions de prévention mais il est aussi nécessaire d'anticiper au mieux les risques futurs, et donc les classes de risque sur les périodes futures.

Cette nouvelle partie consiste en la présentation de méthodologies afin de réaliser des prédictions de classes de risque. Il s'agira de prédictions de l'évolution du risque santé pour chaque entreprise.

Tout d'abord, il sera expliqué pourquoi les prédictions seront réalisées à la maille entreprise. Ensuite, le modèle BERT (*Bidirectional Encoder Representations from Transformers*), au cœur de certaines méthodes, sera introduit. Et enfin pour finir, les méthodes développées et testées dans ce mémoire seront présentées.

III.1 Contexte et objectifs

La suite des travaux est réalisée à la maille entreprise et non plus par assuré. Dans un premier temps, ce choix sera justifié. Dans un second temps, un outil d'étude des entreprises atypiques sera présenté afin de montrer l'intérêt de l'étude de la consommation santé par entreprise.

III.1.1 Choix de la « maille entreprise »

Il est possible, à l'aide de statistiques ou de modèles de *Machine Learning*, de déterminer la probabilité qu'un individu soit classé au trimestre suivant dans une classe de risque donnée. Or, une probabilité faible ne signifie pas que l'individu ne sera pas classé dans cette classe en réalité (Exemples : accident ou maladie virale). De même, une probabilité élevée ne signifie pas que l'individu sera réellement classé dans cette classe. Et enfin, un individu peut être associé à des probabilités élevées pour plusieurs classes. Il est donc très difficile de conclure et de réaliser des prédictions exactes de classes de risque à la maille assuré.

Cependant, du fait de la loi des grands nombres, il est possible statistiquement de prédire la répartition par classe de risque d'un groupe d'assurés suffisamment important.

De plus, les données à disposition proviennent d'un portefeuille de santé collective. Il est donc plutôt cohérent de chercher à cibler des actions de prévention à l'échelle d'une entreprise.

Et enfin, si les données sont regroupées par entreprise et anonymisées de manière appropriée, il n'est plus nécessaire d'obtenir le consentement individuel de chaque assuré.

Pour toutes ces raisons, les prédictions seront réalisées par entreprise (prédictions de la répartition par classe de risque). **L'intérêt sera ensuite de proposer l'action de prévention à l'ensemble de l'entreprise.**

III.1.2 Détection d'entreprises atypiques

Afin de montrer un intérêt de l'étude de la consommation santé par entreprise, cette sous-partie traitera de la détection des entreprises atypiques[8]. Pour ce faire, plusieurs méthodes sont possibles. Il est possible d'analyser les données par entreprise (montants, effectifs par classe de risque, âge, sexe, etc.) et de dresser des statistiques et analyses détaillées.

Une autre méthode sera présentée ici, elle permettra de mettre en évidence les entreprises ayant des effectifs par classe de risque anormalement élevés (par rapport à l'ensemble du portefeuille) ou bien les entreprises ayant des risques qui s'aggravent au cours du temps (croissance de l'effectif de la classe de risque d'un trimestre à un autre).

Il sera ainsi tracé un nuage de points permettant de repérer les entreprises les plus atypiques selon deux indicateurs :

- En ordonnée, un indicateur « moyenne » : positionnement relatif de la part des consommants de l'entreprise dans la classe de risque choisie par rapport à la même part pour l'ensemble du portefeuille.
- En abscisse, un indicateur « évolution » : indicateur représentant l'évolution dans le temps du risque au sein de l'entreprise.

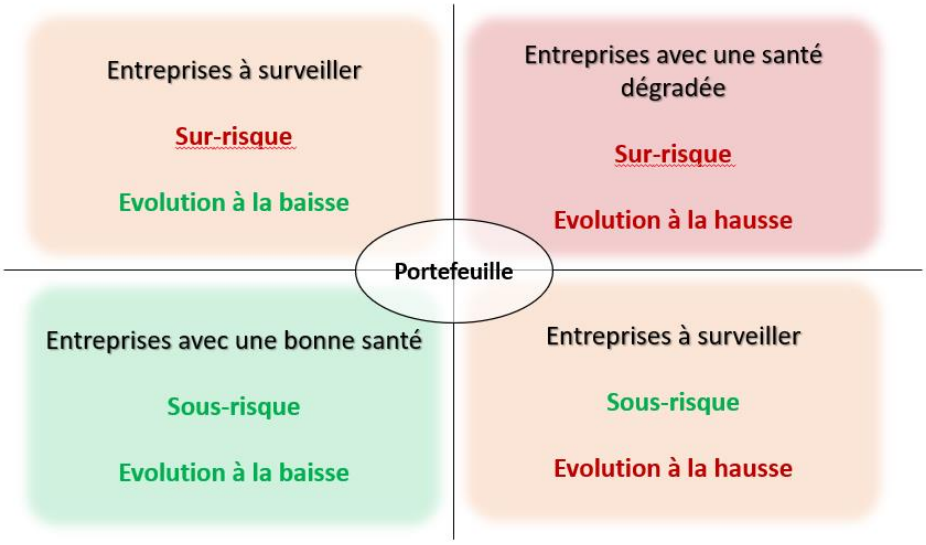


Figure 3.1 – Proposition de cartographie des entreprises

[8] Entreprise atypique : entreprise se distinguant des autres entreprises par la consommation santé de ses salariés. Ici, il s'agira d'entreprises dont le risque est anormalement élevé au sens des indicateurs définis et où une proposition d'action de prévention serait justifiée.

a. Création de l'indicateur « moyenne »

On notera $\overline{Part}_{i,j}$ la moyenne entre les trimestres 1 à 6 de la part des assurés de l'entreprise *i* classés en classe *j*. Elle est définie comme suit :

$$\overline{Part}_{i,j} = \frac{1}{6}\sum_{t=1}^{6} Part_{i,j}^{t}$$

$Part_{i,j}^{t} = \frac{|E_i \cap C_j^t|}{|E_i|}$; E_i : effectif de l'entreprise *i* ; C_j^t : effectif de la classe *j* au trimestre *t*

L'indicateur « moyenne » est calculé de la manière suivante :

$$indic_{moyenne} = \overline{Part}_{i,j} - \overline{Part}_{\text{portefeuille},j}$$

b. Création de l'indicateur « évolution »

Aucune méthode présente dans la littérature ne répondait à tous les critères que nous cherchions pour représenter l'évolution du risque par entreprise. Un indicateur reposant sur une régression linéaire a ainsi été développé et est calculé en deux étapes de la manière suivante :

- Détermination des coefficients de régression linéaire a et b (e_t = at+b avec e_t : effectif au trimestre *t*) et calcul du coefficient de détermination R^2

La modélisation a été réalisée sur les effectifs (par classe et par entreprise) de T1 à T6.

- Calcul de l'indicateur « évolution » :

$$indic_{évolution} = aR^2.\mathbb{1}_{\{R^2>1\%\}}$$

Le seuil des 1% est obtenu empiriquement et ne sera pas justifié théoriquement dans ce mémoire.

c. Identifications des entreprises atypiques

Une fois les indicateurs calculés, il est construit deux graphes par classe de risque :

- Un graphe complet avec en abscisse l'indicateur « moyenne » et en ordonnée l'indicateur « évolution »

- Un graphe se focalisant sur les quinze entreprises maximisant la somme des deux indicateurs

Illustration avec la classe de risque Psy :

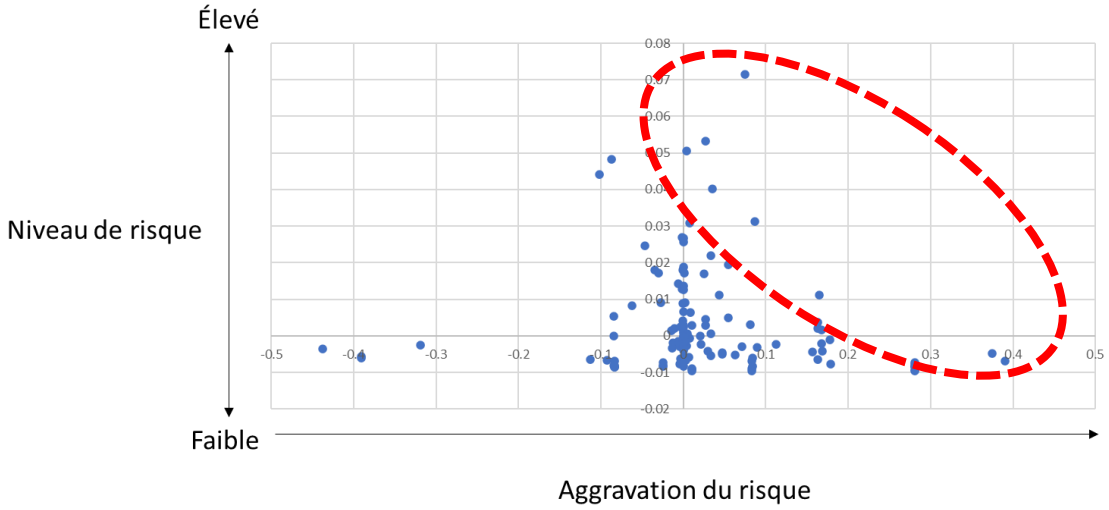


Figure 3.2 – Détection entreprises atypiques Psy

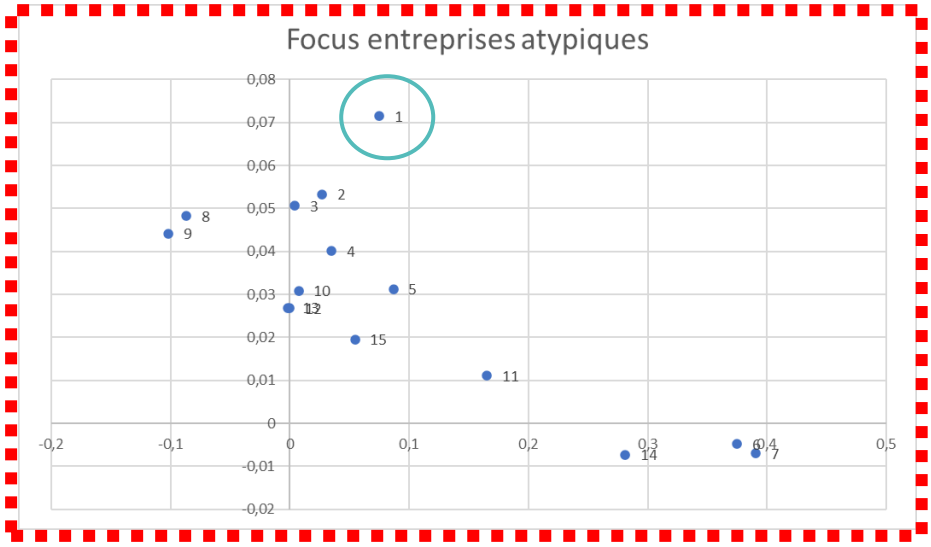


Figure 3.3 – Détection entreprises atypiques Psy (Focus)

Ainsi il est possible d'identifier les entreprises atypiques (potentiellement cibles de nos actions de prévention). Par exemple, on remarque que l'entreprise 1 est à risque (pour la Psy) +7% par rapport à la moyenne du portefeuille et en évolution à la hausse d'un trimestre à l'autre.

Cette méthode rapide permet donc d'avoir une vision d'ensemble avant d'affiner les analyses. Il sera ensuite conseillé de regarder un peu plus en détails les données et statistiques correspondant à ces entreprises atypiques.

III.2 Introduction au modèle BERT

Les travaux et les prédictions à la maille entreprise étant expliqués et motivés, la suite de cette partie concernera les méthodes développées et testées dans le cadre de ce mémoire afin de prédire la répartition par classe.

Cette partie présentera rapidement un modèle de NLP[9] : BERT[10] sous la forme d'une introduction afin de pouvoir comprendre plusieurs méthodes (utilisant ce modèle) qui seront développées en partie III.3. Il s'agit d'un modèle extrêmement complexe et ce mémoire n'aura pas vocation à présenter l'intégralité des spécificités de BERT mais plutôt son fonctionnement général et son application dans le cadre de la prévention en assurance santé.

III.2.1 Le *Natural Language Processing* (NLP)

Le ***Natural Language Processing*** (traitement automatique du langage naturel) est une discipline mêlant linguistique, informatique et intelligence artificielle. Elle vise à créer des outils permettant à une machine de comprendre et traiter le langage naturel.

Exemples d'application :

- Traduction de textes sans intervention humaine
- Résumé automatique de textes
- *Chatbots* (agent conversationnel paramétré pour accomplir une tâche précise)

III.2.2 Présentation du modèle BERT

BERT est un modèle de NLP développé en 2018 par Google [Devlin & al,2018] et considéré aujourd'hui comme une référence dans la discipline (présentant d'excellentes performances dans de nombreuses tâches de NLP).

Un exemple d'application très connu et aussi une des raisons pour laquelle Google a développé ce modèle est la suggestion de requêtes :

Figure 3.4 – Illustration BERT requêtes Google

[9] Natural Language Processing
[10] Bidirectional Encoder Representations from Transformers

Comme il est possible de le voir à travers l'illustration ci-dessus, **BERT est capable de capter les relations entre les mots et ainsi de prédire la suite d'une phrase**. Il sera présenté rapidement dans cette partie le fonctionnement du modèle BERT.

a. Présentation de la famille des *Transformers*

Tout d'abord, BERT est un modèle de type *Transformer*. Un *Transformer* est un réseau de neurones composé d'un nombre constant de petits blocs dans une architecture *Encoder-Decoder*. Chacun de ces blocs est constitué d'un réseau de neurones à propagation avant (*Feed Forward*)[11] et d'une couche de *Self-attention* permettant de conserver l'interdépendance des mots dans les phrases. Ces modèles sont très intéressants car ils n'utilisent pas de réseaux récurrents[8] et cela les rend parallélisables.

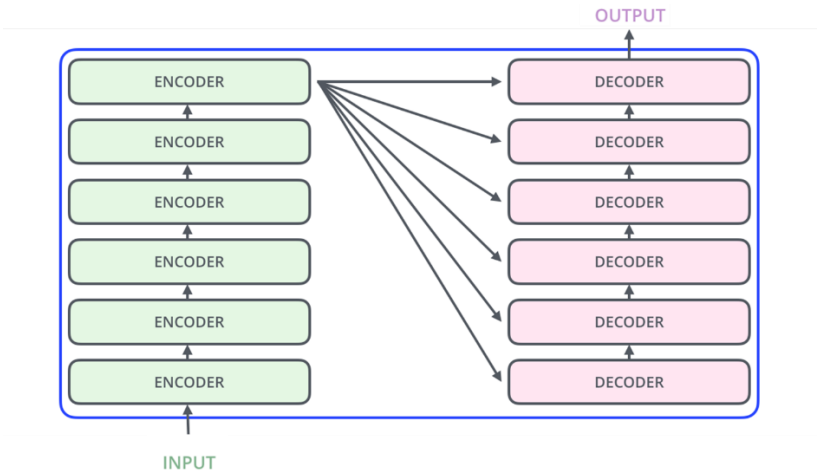


Figure 3.5 – Schéma illustratif modèle *Transformer* [Bourdois, 2019]

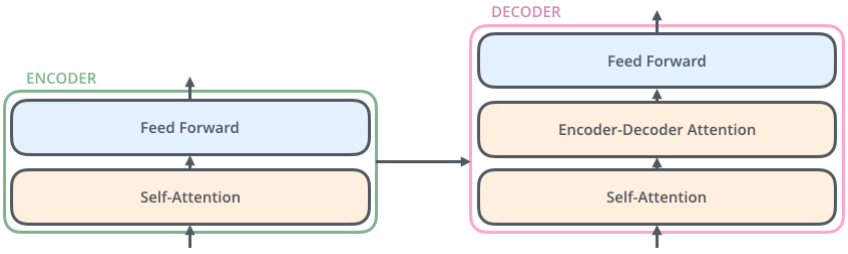


Figure 3.6 – Schéma illustratif bloc *Encoder* et bloc *Decoder* [Bourdois, 2019]

Le processus de *Self-attention* permet de comprendre le contexte de la phrase en *input*, voici ci-dessous un exemple :

[11] Pour une brève définition du réseau de neurones à propagation avant et du réseau de neurones récurrent, le lecteur pourra se rapporter à l'annexe a) et b).

« L'actuaire ne ménage pas son temps car il a un dossier à rendre. »

Le processus de *Self-attention* permet dans cet exemple de déterminer le sens du mot « il » en comprenant le contexte de la phrase (lien entre « actuaire » et « il »).

b. Un modèle BERT pré-entrainé sur deux tâches

De plus, BERT, du fait du nombre de paramètres (110 millions), est un modèle nécessitant une très grosse quantité de données afin d'être entrainé. La puissance de calcul nécessaire à cela n'est donc pas accessible par un utilisateur classique. Le modèle utilisé sera ainsi pré-entrainé (« *transfer learning* »). Autrement dit, Google met à disposition un modèle entrainé en amont sur un jeu de données de très grande taille et l'utilisateur ne fera qu'affiner les poids du modèle afin d'adapter celui-ci à son propre jeu de données. Cette dernière étape, essentielle pour s'adapter au contexte de l'étude, sera appelée entrainement dans la suite du mémoire.

Le pré-entrainement de BERT est non supervisé et est réalisé via des textes de pages Wikipédia (2 500 millions de mots) et via un ensemble de livres (800 millions de mots). Ce pré-entrainement non supervisé est réalisé sur deux tâches :

- **MLM** (*Masked Language Modeling*)

  Des mots vont être masqués aléatoirement et le modèle sera entrainé à les prédire.

  Exemple :

  Séquence initiale : « L'actuaire ne ménage pas son temps car il a un dossier à rendre. »

  Séquence avec des mots masqués : « L'actuaire ne [1] pas son temps car il a un [2] à rendre. »

  Prédiction du modèle : [1] = « ménage » ; [2] = « dossier »

- **NSP** (*Next Sentence Prediction*)

  Le modèle sera entrainé à prédire si deux phrases se suivent ou non.

  Exemple 1 :

  Séquence A : « L'actuariat s'adapte sans cesse à l'évolution de la société. »

  Séquence B : « Ainsi, pour garder un niveau d'expertise élevé, l'actuaire se forme tout au long de sa vie professionnelle ».

  Prédiction du modèle : « Vrai » (elles se suivent)

  Exemple 2 :

  Séquence A : « L'actuariat s'adapte sans cesse à l'évolution de la société. »

  Séquence B : « Il va pleuvoir demain. »

  Prédiction du modèle : « Faux »

c. Un modèle innovant et performant

Et enfin, il est à noter que le modèle BERT est performant et innovant grâce à sa bidirectionnalité. En effet, avant BERT, les modèles parcouraient les séquences de gauche à droite ou de droite à gauche afin de prédire le mot suivant. BERT, quant à lui, grâce au MLM, utilise le contexte complet de la séquence. Il s'agit du premier modèle à utiliser cette approche.

Dans [Sudhir & al, 2021], différents modèles de classification pour l'analyse de sentiments sont comparés, y compris le modèle BERT. Les résultats de l'étude montrent que BERT est l'un des meilleurs pour l'analyse de sentiment. Voici ci-dessous les scores de précision de chacun des modèles sur le jeu de données de classification de sentiments IMDB :

Classification model	Accuracy
[3] Naive Bayes	54.77%
[3] Decision Tree	87.53%
[5] Support Vector Machine (SVM)	88.3%
[3] K-Nearest Neighbor (KNN)	88.86%
[5] LSTM	88.5%
[5] GRU	89.0%
[6] BERT	89.5%
[8] BERT Large with UDA	95.22%

Figure 3.7 – Scores de précision : classification de sentiments IMDB [Sudhir & al, 2021]

Le score obtenu avec BERT est le plus élevé (95,22%). Ainsi, cet article démontre la supériorité de BERT par rapport aux autres modèles dans un cadre de classification de texte. De nombreux articles similaires (sur d'autres jeux de données) ont démontrés que BERT est une référence dans la classification de texte. De plus, [Korpusik & al, 2019] compare différentes méthodes pour la compréhension du langage naturel et les résultats montrent que BERT est supérieur aux autres méthodes sur plusieurs tâches.

Pour cette raison, certaines méthodes développées en III.3 reposeront sur ce modèle BERT. Il sera utilisé comme méthode de classification afin de prédire la suite d'un parcours de classes de risque.

III.3 Présentation des méthodes

Dans cette partie, les méthodes développées et testées afin de prédire la répartition future des entreprises par classes de risque seront présentées.

III.3.1 Les méthodes statistiques

a. Matrice de transition

Tout d'abord, pour faire suite aux travaux présentés en partie II.5, une première méthode reposant sur la matrice de transition a été testée. Les résultats obtenus motivaient l'usage de cette méthode. De plus, il s'agit d'une méthode assez simple et intuitive.

Pour rappel, en II.5, les probabilités de passage d'une classe à une autre par individu avaient été calculées. La matrice de transition suivante avait été obtenue :

$$P = \begin{pmatrix} P_{1\to1} & \cdots & P_{1\to M} \\ \vdots & \ddots & \vdots \\ P_{M\to1} & \cdots & P_{M\to M} \end{pmatrix}$$

Avec M = Nombre de classes de risque

$P_{i\to j}$ = Probabilité de passage de la classe *i* à la classe *j*

Les prédictions par entreprise sont ainsi facilement et rapidement accessibles en utilisant une **approche markovienne**. Ci-dessous le calcul markovien utilisé :

$$\begin{pmatrix} N_{1,1}^{t+1} & \cdots & N_{1,M}^{t+1} \\ \vdots & \ddots & \vdots \\ N_{N,1}^{t+1} & \cdots & N_{N,M}^{t+1} \end{pmatrix} = \begin{pmatrix} N_{1,1}^{t} & \cdots & N_{1,M}^{t} \\ \vdots & \ddots & \vdots \\ N_{N,1}^{t} & \cdots & N_{N,M}^{t} \end{pmatrix} \begin{pmatrix} P_{1\to1} & \cdots & P_{1\to M} \\ \vdots & \ddots & \vdots \\ P_{M\to1} & \cdots & P_{M\to M} \end{pmatrix} \text{ noté } \boldsymbol{N^{t+1} = N^{t}.P}$$

Avec : $N_{i,j}^{t+1}$ = Nombre de personnes de l'entreprise *i* classées en classe *j* à l'instant *t+1*

Il s'agit d'une approche simple et intuitive. En effet, il s'agit encore une fois de se baser sur la loi des grands nombres. Pour un groupe d'assurés donné de taille *x*, si la probabilité de passage dans la classe c au trimestre suivant est p alors il est estimé un effectif de *x*p* assurés dans la classe *c* au trimestre suivant. Ce raisonnement est appliqué par paire de classes.

Cette méthode devrait présenter de **bons résultats sous l'hypothèse que le portefeuille n'évolue pas à travers le temps**. En effet, elle ne permet pas de capter l'évolution temporelle. Les probabilités de passage sont lissées sur tout l'historique. Cependant, **en réalité, les probabilités et les effectifs peuvent évoluer** entre le trimestre 1 et le trimestre 5. Cette évolution doit être prise en compte. La méthode s'avérera donc inefficace pour capter une déformation rapide de la consommation santé comme observée par le passé : pandémie de Covid-19, réforme 100% santé, etc. Les deux méthodes suivantes, détaillées ci-dessous, viseront à prendre en compte cette évolution temporelle assez simplement.

b. Tendance (Régression linéaire)

La modélisation la plus connue et la plus utilisée afin d'évaluer l'évolution temporelle est la **régression linéaire**. Par volonté d'interprétabilité, il sera testé une simple régression linéaire sur les effectifs par classe et par entreprise. Les coefficients seront estimés grâce à l'historique (à partir des six premiers trimestres) et il s'agira ensuite de prédire les effectifs au trimestre 7.

La méthode se décompose en trois étapes :

- **Détermination des coefficients de régression** a et b (e_t = at+b avec e_t : effectif au trimestre *t*) et calcul du coefficient de détermination R^2
- **Prédiction sous conditions**

$$e_{T7} = \begin{cases} 7a + b \; si \; R^2 > 1\% \\ e_{T6} \; sinon \end{cases}$$

  Si $R^2 \leq 1\%$, il est considéré qu'il n'y a pas d'évolution[12].

- **Réajustement par entreprise** : si l'effectif total prédit est différent de l'effectif total réel, des assurés seront ajoutés ou enlevés en classe Consommation nulle.

c. Matrice de transition et tendance

Cette 3ème méthode est une agrégation des deux premières. Elle vise à corriger la non prise en compte de l'évolution temporelle par la méthode présentée en a). Il s'agit des prédictions obtenues par la méthode a) avec prise en compte de la tendance linéaire.

$$Effectif\ T7 = P_m + a.\mathbb{1}_{\{R^2>1\%\}}$$

P_m : prédiction obtenue par la méthode « Matrice de transition ».

Comme précédemment, si besoin, on réajuste si l'effectif total prédit est différent de l'effectif total réel.

Ces 3 méthodes statistiques, très simples, ne prennent pas réellement en compte les parcours de consommation santé, fil rouge de ce mémoire. Elles seront néanmoins utilisées par la suite.

III.3.2 Les méthodes issues du traitement automatique du langage naturel

III.3.2.1 BERT appliqué aux actes

Cette méthode repose sur le modèle BERT introduit précédemment. Il s'agit ici d'utiliser le parcours de soins (la succession des prestations consommées) afin le prédire la classe de risque future. Les prédictions seront données par assuré et l'agrégation par entreprise sera ensuite réalisée. L'agrégation permettra de lisser les erreurs et de retrouver un approche « loi des grands nombres ».

[12] Ce critère, obtenu empiriquement, a été retenu car permettait d'assurer des résultats satisfaisants

Afin d'exploiter le **potentiel de prédiction de BERT** (cf. II.2.2.c), la **base de prestations santé sera assimilée à un corpus de texte** du fait des analogies suivantes peuvent être faites.

- Corpus = base de données des prestations de l'ensemble des assurés
- Texte = succession de l'ensemble des prestations santé d'un assuré
- Phrase = parcours de soins
- Mot = **acte de soins**

Il s'agit ici d'un cadre non standard d'utilisation de BERT, à savoir la prédiction de la consommation santé future (classe de risque future), mais qui est convertie en un problème de *NLP* via ces analogies.

En pratique, **il s'agira de prédire la classe de risque de l'assuré en T7 à partir des prestations du trimestre précédent (T6).**

A noter : il s'agit d'apprentissage supervisé, le modèle est donc entrainé (affinement des poids afin de s'adapter au contexte) sur T5 (variables explicatives) →T6 (variables à prédire).

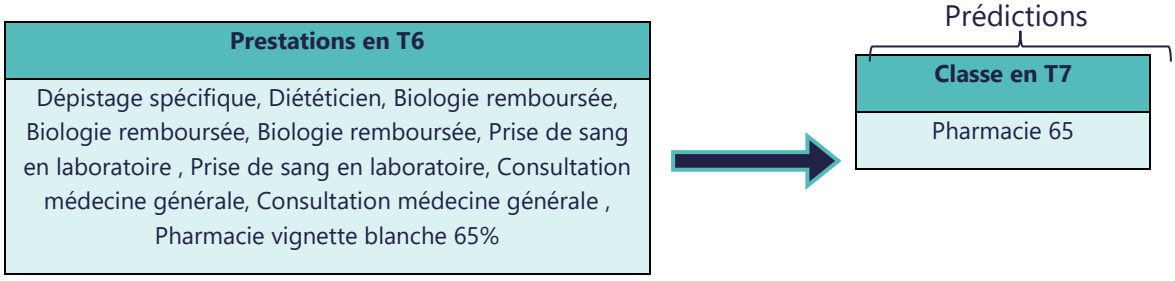


Figure 3.8 – Principe méthode BERT appliqué aux actes

Pour ce faire et pour ensuite prédire la répartition par classe de risque à la maille entreprise, plusieurs étapes sont nécessaires et vont être présentées.

a. Préparation de la base prestations

Il est nécessaire de rendre « le parcours de soins » (les prestations au trimestre précédent) compréhensible par le modèle. Cette étape, complexe, nécessite des recherches poussées et de nombreux tests, le tout réalisé en collaboration avec des experts. Dans le cadre de ce mémoire, ce travail sera seulement ébauché afin de pouvoir répondre à la problématique.

L'entrainement du modèle est réalisé sur T5 → T6. La prédiction est réalisée sur T6 → T7. Il est ainsi important qu'un unique acte soit nommé de la même manière en T5 et en T6. Or, en 2019, de nouveaux libellés d'actes sont apparus avec la réforme 100% santé. Afin de répondre à ce changement, les libellés d'actes ont été uniformisés sur tous les trimestres.

Et enfin, les libellés « honoraires » redondants avec les prestations correspondantes seront supprimés (cf. partie I.3.2c).

b. Choix du modèle et des paramètres

Le modèle pré-entrainé choisi est BERT *base uncased*[13]. Après une étape chronophage de tests de paramètres, les paramètres suivants ont été fixés :

Max length of tokens	33
Optimizer	*Adam*
Loss	*Categorical Crossentropy*
Metric	*Accuracy*
Batch_size	15
Learning rate	9.375e-06
Number of epochs	*Early Stopping*

Figure 3.9 – Paramètres fixés BERT - BERT appliqué aux actes

c. Transformation des scores BERT

Pour chaque assuré, BERT renvoie un score par classe. La matrice des scores suivante est ainsi obtenue :

$$\begin{pmatrix} a_{1,1} & \cdots & a_{1,M} \\ \vdots & \ddots & \vdots \\ a_{N,1} & \cdots & a_{N,M} \end{pmatrix}$$

Avec :
N : effectif total du portefeuille
M : nombre de classes de risque

Les scores ne sont pas des probabilités. Il s'agit de nombres réels centrés en 0 et généralement inférieurs à 5 en valeur absolue. Usuellement, la prédiction par individu est obtenue en prenant la classe correspondant au score maximal.

Ici, il n'est pas possible de prédire directement par l'argument maximum. En effet, les données sont déséquilibrées comme l'illustre cet extrait des effectifs au trimestre 6 :

	Effectif au trimestre 6
Analyses	1333
Analyses+Imagerie	255
Appareillage	179
ATM	904
Consommation nulle	4852
Dentaire	317
Généraliste	367
Hospitalisation	451
…	…

Figure 3.10 – Illustration données déséquilibrées

[13] Le modèle utilisé a été téléchargé à l'aide de la librairie open source « Transformers » disponible sur huggingface.co. L'API Tensorflow.keras a ensuite été ensuite utilisée afin de créer et entrainer un modèle propre à la prédiction de classes de risque.

L'effectif en classe Consommation nulle est nettement plus élevé que celui en classe Appareillage (4852 > 179). Les données sont déséquilibrées. Le modèle aura donc tendance à prédire que la classe Consommation nulle (ou plus généralement les classes prépondérantes).

Il s'agit d'un problème très connu qui est souvent résolu par du rééchantillonnage (suppression ou duplication aléatoire de certaines lignes de la base d'entrainement afin d'avoir des données rééquilibrées).

Plusieurs tests ont été réalisés ici afin de résoudre ce problème dont les deux suivants :

- Rééchantillonnage et prédiction par score maximal
- Rééchantillonnage, normalisation min-max et simulations stochastiques

Aucun de ces tests, ni aucune méthode présente dans la littérature n'a permis d'obtenir des résultats satisfaisants. Ce mémoire propose une nouvelle méthode empirique (qui ne sera pas justifiée théoriquement) qui consiste à transformer les scores sans rééchantillonnage de la manière suivante :

On notera $A = (a_{i,j})_{i,j}$, la matrice des scores et $\tilde{A} = (\tilde{a}_{i,j})_{i,j}$ la matrice des scores transformés.

$$\tilde{A} = f(A) \text{ avec } f \text{ telle que } f : a_{i,j} \mapsto \frac{rang_j(a_{i,j})}{N_j^{t-1}}$$

Avec :

N_j^t : nombre de personnes classées en classe j à l'instant t

$rang_j(a_{i,j})$: rang de $a_{i,j}$ dans la liste triée par ordre croissant des valeurs de la colonne j de A

Exemple calcul de $rang_j(a_{i,j})$:

$$a = \begin{pmatrix} 1 & 9 & 4 \\ 5 & 8 & 6 \\ 3 & 5 & 2 \end{pmatrix} \qquad rang_2(a_{1,2}) = 3 \quad \text{car } a_{1,2} = 9 \text{ et la liste triée est : (5,8,9)}$$

Cette transformation suivie d'une prédiction par la classe correspondant au score minimal (et non pas maximal) donne des résultats très satisfaisants une fois les prédictions agrégées par entreprise et a donc été choisie. Il s'agit d'une solution analogue au rééchantillonnage. Cette méthode a été construite afin de maximiser la performance du modèle et en vérifiant empiriquement l'adéquation entre prédictions et données réelles.

L'agrégation par entreprise est ensuite obtenue en comptant les effectifs prédits par classe et par entreprise.

III.3.2.2 BERT appliqué aux classes

Cette méthode est assez similaire à la précédente or il ne s'agit plus de parcours de soins mais de parcours de classes de risque.

Les analogies suivantes peuvent être faites :

- Corpus = base de données des classes de risque de l'ensemble des assurés sur tout l'historique
- Texte = succession de l'ensemble des classes de risque d'un assuré
- Phrase = parcours de classes de risque de T2 à T6
- Mot = **classe de risque**

En pratique, il s'agira de prédire la classe de risque de l'assuré en T7 à partir des classes de risques précédentes (T2 à T6).

A noter : il s'agit d'apprentissage supervisé, le modèle est entrainé sur (T1->T5) → T6.

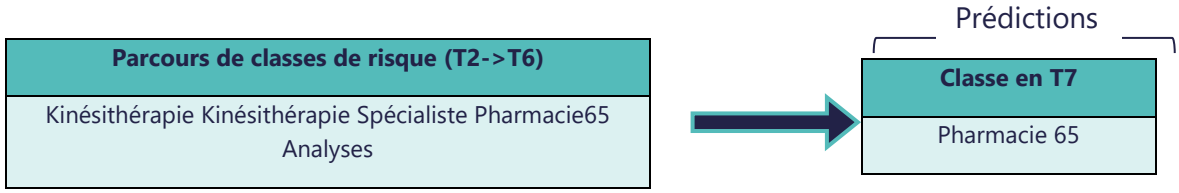


Figure 3.11 – Principe méthode BERT appliqué aux classes

La méthodologie est la même que pour la méthode précédente (même modèle, même transformation des scores, etc.). Seul le paramètre « *Learning rate* » a légèrement changé mais il présente le même ordre de grandeur :

Learning rate	8.75e-06

Des modèles plus classiques (*CART, Random Forest*) peuvent aussi être utilisés afin de prédire à l'aide des classes de risque qui précèdent. Ils n'utilisent pas la notion de parcours mais présentent l'avantage d'être plus simples. Ces modèles ont aussi été testés et les méthodes correspondantes sont les suivantes.

III.3.3 Les méthodes traditionnelles de Data Science

a. CART (classes)

Cette méthode repose sur le modèle *Classification And Regression Trees* (CART)[14]. Il s'agit ici d'utiliser les classes de risque qui précèdent sous la forme de variables (et non plus sous la forme de parcours).

Il s'agira de prédire la classe de risque en T7 à l'aide des classes de risque observées en T2, T3, T4, T5 et T6.

A noter : il s'agit d'apprentissage supervisé, le modèle est donc entrainé sur (T1, T2, T3, T4, et T5) → T6.

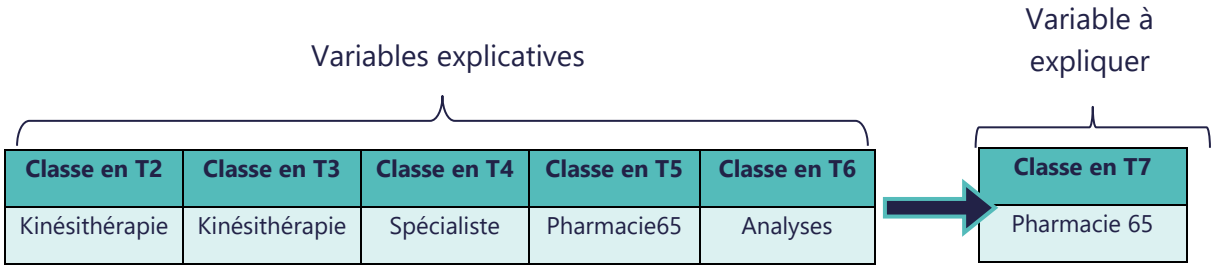


Figure 3.12 – Principe méthode CART

Pour ce faire, la méthodologie se déroule en plusieurs étapes :

- **Préparation de la base**

Comme précédemment, les données sont déséquilibrées. Pour cette méthode, elles seront rééchantillonnées afin d'obtenir 1000 assurés par classe de risque. Autrement dit, du suréchantillonnage sera fait sur les classes présentes moins de 1000 fois et du sous-échantillonnage sur les classes présentes plus de 1000 fois.

- **Construction de l'arbre CART**

Tout d'abord, l'arbre CART[15] sera construit avec une profondeur maximale fixée à 8. Ensuite, l'arbre obtenu sera élagué[16] avec comme paramètre de complexité : 0,001. Cela est réalisé afin de contrôler le sur-apprentissage et d'obtenir des résultats cohérents.

- **Prédiction CART**

[14] Le lecteur pourra se rapporter à l'annexe e) pour une présentation du CART.
[15] Le modèle est construit à l'aide de la librairie R *rpart*. Ce package ne permet pas de construire l'arbre maximal, car trop profond. Cela explique la profondeur choisie.
[16] Elaguer consiste à désactiver de manière récursive les fractionnements les moins importants en fonction du paramètre de complexité.

Comme dans le cas de BERT, des scores sont obtenus par individu et par classe. Cependant, pour CART, les scores sont des probabilités. La matrice des probabilités suivante est ainsi obtenue :

$$\begin{pmatrix} \alpha_{1,1} & \cdots & \alpha_{1,M} \\ \vdots & \ddots & \vdots \\ \alpha_{N,1} & \cdots & \alpha_{N,M} \end{pmatrix}$$

Avec :

N : effectif total du portefeuille
M : nombre de classes de risque
$\alpha_{i,j}$ = probabilité que l'individu i soit classé en j au trimestre suivant

- **Agrégation par entreprise**

Pour BERT, les prédictions étaient réalisées par individu. Il fallait ensuite agréger ces prédictions par entreprise en comptant les effectifs prédits par classe. Ici, les prédictions sont des probabilités, il suffit donc de les sommer par entreprise afin d'obtenir les effectifs prédits.

Exemple avec une entreprise de deux assurés pour un trimestre donné :

Probabilités	Classe a	Classe b	Classe c
Individu 1	35%	5%	60%
Individu 2	70%	30%	0%

	Classe a	Classe b	Classe c
Effectif prédit	1,05	0,35	0,60
	(0,35 + 0,70)	(0,05 + 0,30)	(0,60 +0,00)

Figure 3.13 – Exemple agrégation par entreprise méthode CART

b. *Random Forest* (classes)

Il s'agit du même principe que la méthode précédente. La modélisation par ***Random Forest***[17] nécessite la même base de données en entrée. **Il s'agit donc encore de prédire la classe de risque en T7 à l'aide des classes précédentes** (T2, T3, T4, T5 et T6).

[17] Le modèle est construit à l'aide du *package* Python *scikit-learn*. Le lecteur pourra se rapporter à l'annexe e) pour une présentation du *Random Forest*.

De plus, comme précédemment, la base d'entrainement est rééchantillonnée avec 1000 assurés par classe. Et enfin, les prédictions sont présentées sous la même forme que pour le modèle CART, les scores sont obtenus sous forme de probabilités.

1000 arbres aléatoires sont construits. Chacun d'entre eux est construit sans fixer de paramètres de complexité ou de profondeur.

III.3.4 Les méthodes en deux étapes

Dans cette partie, il sera présenté des méthodes en 2 étapes **complétant et améliorant les méthodes** : BERT appliqué aux actes ; BERT appliqué aux classes ; *Random Forest* (classes).

Les **deux étapes** sont les suivantes **:**

- Prédiction de la **classe *Consommation nulle*** par un CART ''binaire''
- **Prédiction des autres classes de risque** par l'une des les trois méthodes **:** BERT appliqué aux actes ; BERT appliqué aux classes ; Random Forest (classes)

a. Motivations

Ces trois méthodes distinguent mal les assurés consommateurs des non-consommateurs. En effet, il s'agit de classifications en classes multiples (ie. non binaire). Les modèles ne sont donc pas focalisés sur la distinction Consommation nulle/Consommation non nulle. Il semble donc nécessaire de **prédire en amont la classe Consommation nulle (1ère étape).**

Pour ce faire, il semble judicieux de créer un modèle adapté à cette problématique et qui exploitera les bonnes variables pour distinguer *Consommation nulle* et *Consommation non nulle*. D'après l'analyse des données, les montants de dépenses engagées sont plus adaptés que les fréquences/quantité d'actes pour répondre à cette classification binaire.

b. Principe

Ci-dessous un schéma résumant la méthodologie :

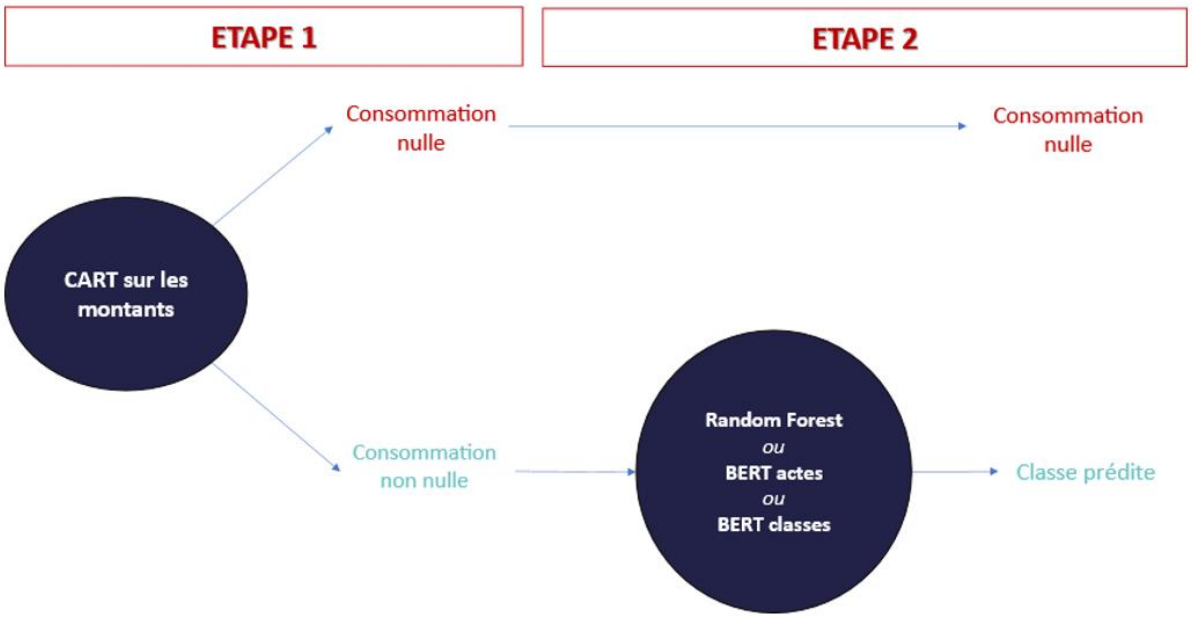


Figure 3.14 – Principe méthode en 2 étapes

Etape 1 : prédiction des « Consommation nulle »

Comme mentionné précédemment, il existe de meilleurs indicateurs que les prestations consommées afin de distinguer Consommation nulle/Consommation non nulle. **La prédiction des « Consommation nulle » sera ici réalisée à partir des frais réels et du nombre d'actes sur les trimestres précédents.**

Plus précisément, la prédiction en T7 est réalisée à partir des variables explicatives suivantes :

T2 montants	…	T6 montants	Montants moyens	T2 nombre d'actes	…	T6 nombre d'actes	Nombre d'actes moyen
	…				…		

Figure 3.15 – Variables explicatives CART Consommation nulle

Avec : $\text{Montants moyens} = \frac{\text{Frais réels de l'individu sur les cinq trimestres}}{5}$

$\text{Nombre d'actes moyen} = \frac{\text{Nombre d'actes de l'individu sur les cinq trimestres}}{5}$

$T_i \text{ montants} = \frac{\text{Frais réels de l'individu en } T_i}{\text{Montants moyens}}$

$T_i \text{ nb actes} = \frac{\text{Nombre d'actes de l'individu en } T_i}{\text{Nombre d'actes moyen}}$

A noter : comme précédemment, l'entrainement sera réalisé sur (T1, T2, T3, T4, et T5)→T6 et l'arbre maximal sera construit avec une profondeur maximale fixée à 8 puis élagué ensuite avec comme paramètre de complexité : 0,001. De plus, la base d'entrainement n'aura pas été rééchantillonnée.

Etape 2 : prédiction des « Consommation non nulle »

A la sortie de l'étape 1, les assurés sont soit classés en classe Consommation nulle soit classés en classe Consommation non nulle.

Pour les assurés qui ont été prédits en *Consommation non nulle*, la 2ème étape consistera à leur assigner une classe de risque en utilisant l'un des modèles BERT ou le *Random Forest* entrainés en III.3.2 et III.3.3. La prédiction est ensuite la plus petite valeur[18] ne correspondant pas à la classe Consommation nulle.

[18] A noter que ces modèles peuvent prédire Consommation nulle, dans ce cas, la prédiction sera la classe correspondant à la deuxième plus petite valeur (et non la première).

III.4 Synthèse et limites

Plusieurs modèles ont été introduits : des modèles dits **statistiques** fondés autour des matrices de transition (avant ou sans prise en compte de la tendance), des modèles usuels de ***Data Science*** (CART et *Random Forest*) ainsi qu'un **modèle de *NLP* : BERT**.

Une dualité est introduite entre les libellés **d'actes de soins** et les **classes de risque** et certains modèles peuvent être utilisés avec l'une ou l'autre des deux approches.

De plus, une **méthodologie dite « en deux étapes »** a été introduite et permet d'utiliser les performances des modèles plus complexes là où ils sont le plus attendus grâce à un pré-modèle CART qui permet de pré-classifier les non consommants.

Ces méthodes peuvent présenter certaines limites qu'il est bon de mentionner :

- **Délai de réception des données**

Les prestations consommées au dernier trimestre sont utilisées afin de prédire la consommation santé future. Il s'agit d'une méthodologie idéale et non réaliste. En effet, il existe un délai entre la date de survenance d'une prestation et la date de réception des données par l'assureur. En pratique, il ne s'agira donc pas exactement des données au dernier trimestre mais plutôt des dernières données disponibles (dernier trimestre glissant).

- **La modélisation est réalisée indépendamment des garanties**

La modélisation est réalisée sur tout le portefeuille. Or, la fréquence de consommation santé visible par l'assureur risque d'être impactée par les garanties.

Exemple :

Supposons deux personnes ayant exactement le même profil et la même consommation santé. Supposons qu'elles consomment toutes les deux une séance chez l'ostéopathe en janvier et une autre en juin. La première personne a un contrat remboursant deux séances d'ostéopathie par an. La deuxième, quant à elle, a un contrat remboursant qu'une seule séance par an. Cette deuxième personne ne déclarera donc pas sa deuxième séance à l'assureur (aucun intérêt à le faire). Ces deux personnes ont donc des consommations santé réelles identiques mais deux consommations santé différentes selon l'assureur. Cet exemple illustre bien que deux personnes ayant la même consommation santé peuvent être différemment perçues par l'assureur selon les garanties qu'elles ont.

A noter qu'en pratique cette limite peut être contournée en retenant une approche segmentée appliquée à chacun des niveaux de garanties.

IV Prédictions du risque santé par entreprise : résultats et proposition d'une méthode mixte

Précédemment, les méthodes développées et testées dans ce mémoire afin de prédire le risque santé par entreprise ont été présentées. **Il s'agit maintenant d'analyser et de comparer les résultats (prédictions obtenues) des différentes méthodes.**

Tout d'abord, des critères seront proposés afin d'évaluer les méthodes. Ensuite, une étude de la **robustesse des méthodes** sera menée. Cela permettra d'obtenir une synthèse des résultats obtenus et une comparaison des méthodes. Et enfin, une **méthode mixte** sera proposée comme méthode finale.

IV.1 Critères d'évaluation des méthodes

Dans cette partie, les critères d'évaluation des méthodes seront fixés et expliqués. D'abord, une métrique sera proposée afin d'évaluer la capacité de prédiction des méthodes. Ensuite, d'autres indicateurs, tout aussi importants, seront présentés.

IV.1.1 Capacité de prédiction

A des fins de mesure de capacité de prédiction (performance) des méthodes, les prédictions ont été réalisées sur le dernier trimestre (T7). Il s'agit d'un choix délibéré afin de pouvoir valider les prédictions (comparer les prédictions avec les données réelles) avec les dernières classes de risque connues (i.e. en T7).

La variable à prédire étant quantitative (effectif), la capacité de prédiction sera mesurée par les deux métriques suivantes :

- Le **RMSE** (*Root mean square error*) qui correspond à la racine carrée de l'**erreur quadratique moyenne**. Il est donné par la formule suivante :

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Avec : y_i = valeur réelle observée ; $\hat{y}_i$ = valeur prédite

- Le **Q^2** qui correspond au **pouvoir prédictif du modèle**. Il est donné par :

$$Q^2 = 1 - \frac{RMSE^2}{Var(Y)}$$

Avec : Y = variable correspond aux données réelles observées

Afin d'obtenir de bonnes prédictions, il est nécessaire de minimiser le RMSE et ainsi d'avoir un Q^2 proche de 1, le plus grand possible. Il s'agira du premier critère d'évaluation.

IV.1.2 Stabilité et complexité du modèle

Au-delà de la capacité de prédiction de la méthode, d'autres critères sont à prendre en compte afin que celle-ci soit utilisable :

- **La stabilité**

  Certaines méthodes présentées en III.3 utilisent des modèles de *Machine Learning*. Dans certains cas, ces modèles peuvent présenter un aléa. Autrement dit, à paramètres fixés, les modèles peuvent présenter des résultats différents si on les exécute plusieurs fois.

  Une méthode présentant des résultats stables est une méthode plus facilement utilisable. On parle de **robustesse intrinsèque**.

  De plus, certaines méthodes peuvent être sensibles à la base de données en entrée. Autrement dit, un petit changement de la base en entrée pourrait complétement changer les résultats. Ceci n'est bien-sûr pas souhaitable. Lorsqu'une méthode n'est pas trop sensible aux données entrées, on parle de **robustesse par rapport à l'échantillonnage**.

  Ces deux types de robustesse seront mesurés en IV.2.1 et IV.2.2 à l'aide de la variance du Q^2.

- **La complexité du modèle**

  Afin de maximiser la capacité de prédiction, il est tentant d'utiliser des méthodes de plus en plus sophistiquées. Cependant, il est important de ne pas négliger l'**accessibilité**, l'**interprétabilité** et la **complexité en temps** des méthodes.

  D'abord, une méthode beaucoup plus sophistiquée que la précédente et dont l'amélioration des résultats est négligeable n'est pas pertinente. En effet, des méthodes trop sophistiquées sont difficiles à mettre en place (développement et industrialisation). De plus, cela peut aussi être source d'erreurs ou de difficultés lorsque les données sont mises à jour ou lorsque l'utilisateur n'est pas le développeur. Il faut ainsi être capable de mesurer l'apport et l'accessibilité d'une méthode plus complexe.

  Ensuite, il est important de trouver un juste milieu entre performance et interprétabilité. L'interprétabilité d'un modèle est la compréhension des mécanismes mathématiques du modèle, autrement dit, son fonctionnement. Contrairement aux modèles traditionnels comme les GLM (Modèle linéaire généralisé) qui sont facilement interprétables par des tests statistiques, la plupart des modèles de *Machine Learning* sont des « boites noires ». L'usage de ces méthodes peut donc ne pas être pertinente dans certains cas lorsque l'interprétabilité est tout aussi importante que la performance.

Et enfin, malgré la puissance de calculs des ordinateurs récents, la complexité en temps peut être un problème posé par les méthodes les plus sophistiquées. Dans ce mémoire, l'idée est de cibler les actions de prévention au bon moment. Or, il existe déjà un délai de réception des données non négligeable. Il ne faut donc pas rajouter en plus un délai trop important entre la réception des données et le ciblage. A noter qu'il est quand même possible d'investir dans de la puissance de calcul afin de réduire le temps d'exécution mais que cela est couteux.

IV.2 Etude de la robustesse

IV.2.1 Robustesse intrinsèque

Dans cette partie, la robustesse intrinsèque (notion introduite en IV.1.2) des méthodes proposées en III.3 sera étudiée. Certaines d'entre elles utilisent des modèles de *Machine Learning* présentant une part d'aléa. C'est le cas des méthodes utilisant un modèle BERT ou un *Random Forest*. Les autres méthodes (matrice de transition, régression linéaire et CART) utilisent des modèles stables par construction et ne seront donc pas étudiées dans cette partie.

a. Méthodologie

Dans la majorité des cas, la part d'aléa présente dans les modèles de *Machine Learning* peut être maitrisée en fixant la graine aléatoire (nombre utilisé pour l'initialisation d'un générateur de nombres pseudo-aléatoires). La robustesse intrinsèque est alors mesurée en faisant varier la graine.

Pour le modèle BERT, la méthodologie reste la même mais l'aléa ne dépend pas seulement de la graine. En effet, il n'est pas possible d'éliminer la totalité de l'aléa en fixant la graine aléatoire avec l'implémentation choisie (cf. III.3.2). Il s'agit d'un problème connu par les utilisateurs de BERT mais qui ne posera pas de problème ici pour l'étude de la robustesse intrinsèque.

Les modèles seront ici exécutés plusieurs fois en faisant varier la graine aléatoire. La robustesse des modèles BERT, modèles plutôt instables, sera mesurée sur 100 itérations avec 100 graines aléatoires distinctes.

b. Résultats des méthodes simples

La robustesse intrinsèque des différentes méthodes simples (méthodes en une étape) est observée à l'aide des valeurs de Q^2. Les résultats obtenus, présentés sous forme de distribution, sont les suivants :

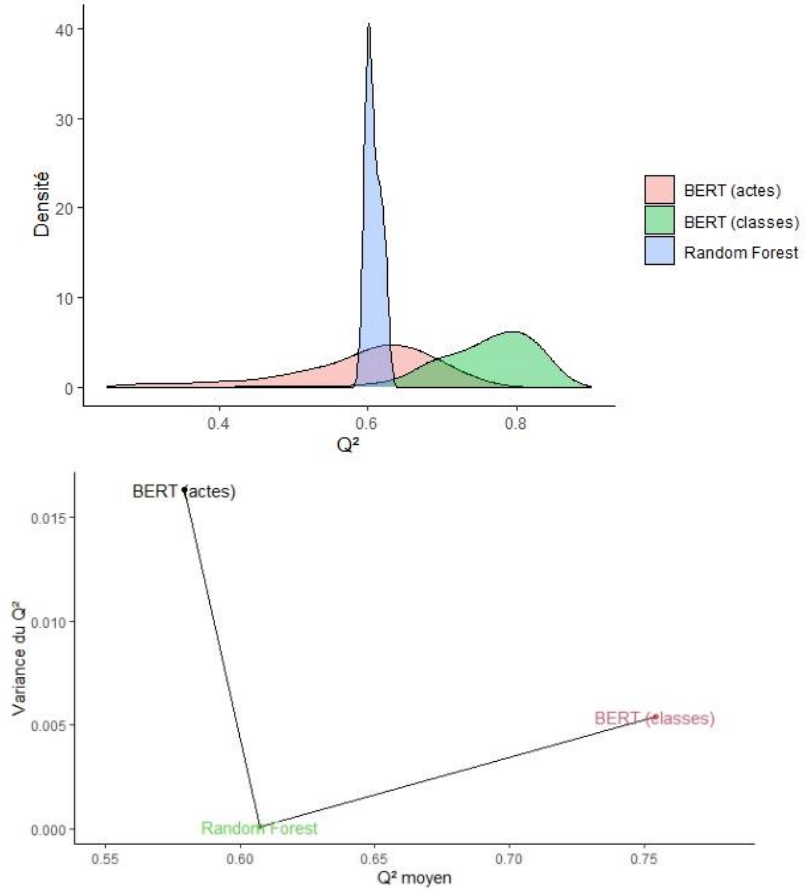


Figure 4.1 – Robustesse intrinsèque : distribution (haut) et front de Pareto (bas) des Q^2

D'après la distribution, les méthodes utilisant **BERT semblent moins stables** que la méthode utilisant le modèle *Random Forest* (du fait d'une plus grande dispersion des Q^2 mesurés). De plus, la méthode présentant en moyenne **les meilleurs résultats est la méthode BERT utilisant les classes de risque** (méthode III.3.2.2).

Il est ainsi possible de conclure que parmi ces méthodes, la méthode la plus robuste est celle utilisant le *Random Forest*. Cependant, BERT avec les classes de risque (méthode III.3.2.2) semble être un bon compromis entre robustesse intrinsèque et performance.

Remarque : le CART n'est pas positionné ici du fait de sa stabilité, il n'est pas sensible à ce type de test.

c. Proposition d'amélioration de la robustesse

La partie précédente, la variabilité non-négligeable des résultats obtenus par les modèles BERT. Cette variabilité implique une non-représentativité des résultats obtenus par une unique exécution. Il s'agira ici de répondre à ce problème. Pour ce faire, il a été retenu de mesurer une moyenne sur plusieurs exécutions.

Afin d'optimiser ces mesures, ce nombre d'exécutions doit être réduit tout en conservant la représentativité de la métrique. Il s'agit ici **de déterminer le nombre minimal d'exécutions de BERT pour obtenir une stabilité suffisante.**

Voici ci-dessous les résultats moyennisés (la valeur correspondant à l'abscisse i est la moyenne des i premières exécutions) d'abord pour la modèle BERT (actes) et ensuite pour le modèle BERT (classes) :

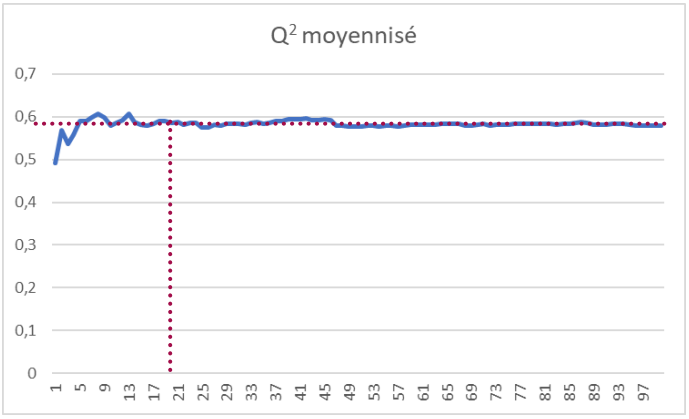


Figure 4.2 –Q^2 moyennisé – BERT appliqué aux actes

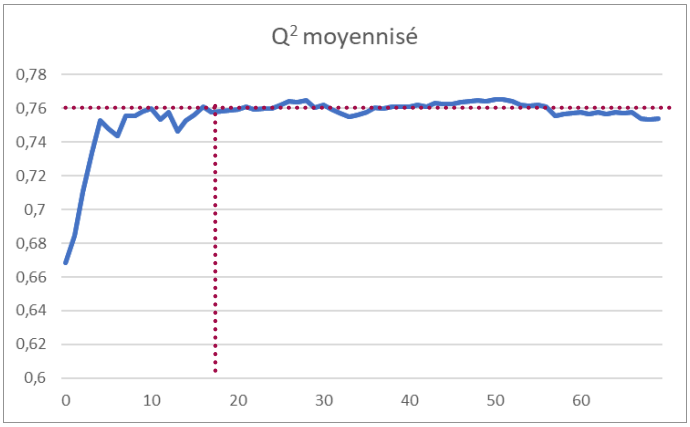


Figure 4.3 –Q^2 moyennisé – BERT appliqué aux classes de risque

Par l'étude des graphes ci-dessus, le choix du nombre d'exécutions sera fixé à 20. En effet, à 20, pour les deux modèles, le résultat est bien représentatif du point de convergence.

Avec cette méthode, on obtient :

	Q^2 moyen	**Variance du Q^2**
BERT (actes)	58%	1,13%
BERT (classes)	75%	0,44%

Figure 4.4 – Résultats modèles moyennisés sur plusieurs exécutions

d. Résultats méthodes en deux étapes

Et enfin, pour terminer rigoureusement l'étude de robustesse intrinsèque, il est important d'étudier les méthodes en deux étapes (i.e. avec pré-traitements des données par un CART

binaire). En effet, les modèles BERT ou *Random Forest* entrainés en III.3.2 et III.3.3 et dont la robustesse a été étudiée précédemment sont aussi utilisés dans les méthodes en deux étapes. Celles-ci présentent donc une instabilité qui doit être étudiée. Ci-dessous un tableau permettant de comparer la robustesse intrinsèque entre les méthodes simples et les méthodes en 2 étapes :

	Méthodes simples		**Méthodes en deux étapes**	
	Q^2 moyen	**Variance du Q^2**	**Q^2 moyen**	**Variance du Q^2**
BERT (actes)	58%	1,13%	63%	0,83%
BERT (classes)	75%	0,44%	79%	0,32%
Random Forest	61%	0,06%	85%	0,04%

Figure 4.5 – Résultats Robustesse intrinsèque

Il est notable que les méthodes en deux étapes sont dans les trois cas plus performantes mais aussi plus robustes que les méthodes simples. Cela était prévisible car le modèle CART - utilisé comme 1ère étape- est parfaitement stable.

IV.2.2 Robustesse par rapport à l'échantillonnage

Dans cette partie, la robustesse par rapport à l'échantillonnage (notion introduite en IV.1.2) des méthodes proposées en III.3 sera étudiée. Certaines de ces méthodes peuvent en effet être sensibles à un léger changement de la base de données en entrée, ce qui n'est pas souhaitable.

a. Méthodologie

Les modèles seront entrainés à l'aide de 20 échantillons (E_1, …, E_{20}) issus de la base complète. Les vingt modèles entrainés seront ensuite testés sur la base de test complète à l'aide des indicateurs RMSE et Q^2.

Les 20 échantillons (E_1, …, E_{20}) sont créés de la manière suivante :

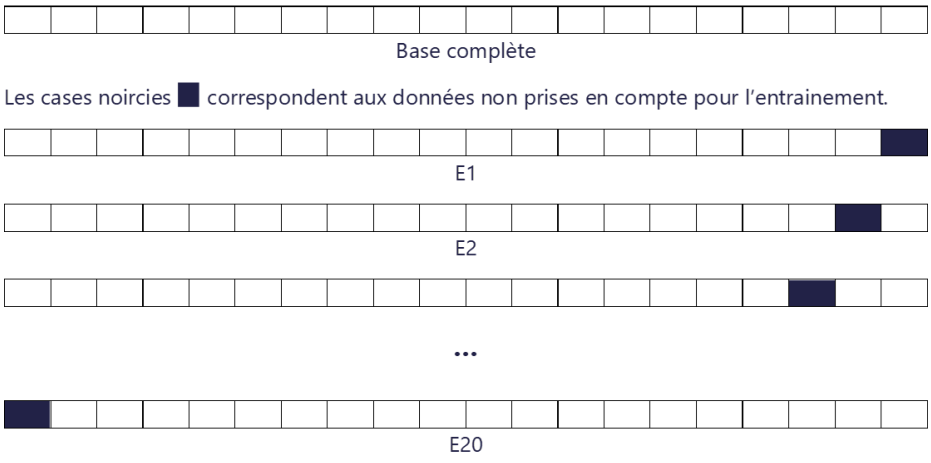


Figure 4.6 – Méthodologie de création des échantillons pour l'étude de la robustesse

Chaque échantillon est la base complète à laquelle on a enlevé 1/20ème des données. Ce sous-échantillonnage est le même (et s'inspire) de celui utilisé dans une méthode couramment utilisée en *Machine Learning* : la validation croisée *k-fold*.

b. Résultats des méthodes simples

La robustesse par échantillonnage des différentes méthodes simples est observée à l'aide des valeurs de Q^2. Ci-dessous la distribution des Q^2 obtenue :

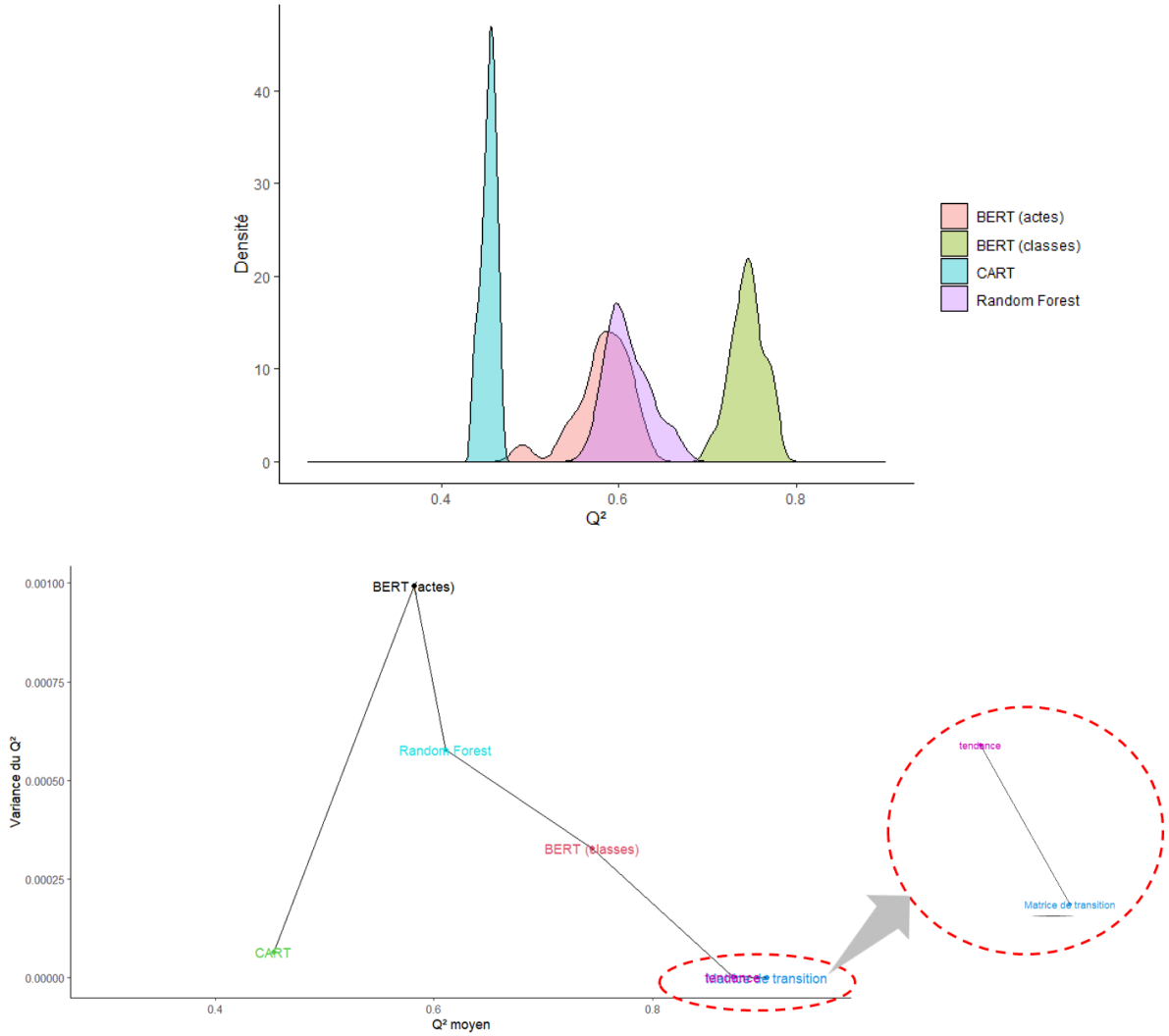


Figure 4.7 – Robustesse par rapport à l'échantillonnage : distribution et front de Pareto des Q^2

A la lecture des résultats, il apparaît que **les méthodes utilisant BERT soient aussi stables que la méthode utilisant le modèle *Random Forest***. De plus et comme attendu, la méthode **CART s'avère très stable**.

Il est ainsi remarquable que les méthodes *Matrice de transition* et *Tendance* ne présentent pas de problème de stabilité par rapport à l'échantillonnage (et ont donc été retirées du graphique de la distribution des Q^2). De plus, la **même conclusion que pour la robustesse intrinsèque** peut être dressée : parmi les méthodes présentant de l'aléa, **BERT (classes) présente un bon compromis entre performance et robustesse.**

c. Résultats méthodes en deux étapes

Et enfin, pour terminer l'étude de la robustesse par rapport à l'échantillonnage, ci-dessous sont comparés les résultats entre les méthodes simples et les méthodes en deux étapes :

	Méthodes simples	Méthodes en deux étapes
	Variance du Q^2	Variance du Q^2
BERT (actes)	0,10%	0,07%
BERT (classes)	0,03%	0,02%
Random Forest	0,06%	0,04%

Figure 4.8 – Résultats Robustesse par rapport à l'échantillonnage

Il est notable que les méthodes en deux étapes sont plus robustes que les méthodes simples par rapport à l'échantillonnage. Cela était prévisible car les modèles CART sont très stables par rapport à l'échantillonnage et la première étape est réalisée avec un modèle CART.

IV.3 Synthèse des résultats

Dans cette partie, une synthèse des résultats obtenus sera présentée afin de pouvoir comparer les méthodes entre elles. Les méthodes seront séparées en deux catégories :

- **Les méthodes statistiques**
  - Matrice de transition
  - Tendance
  - Matrice de transition et tendance

Les méthodes statistiques sont des méthodes modélisant l'ensemble du portefeuille sans étudier réellement le comportement des assurés à l'échelle individuelle. Elles présenteront donc des **résultats bons dans l'ensemble** (Q^2 élevé) **mais ne seront pas capables d'anticiper des changements** dans la consommation santé d'une entreprise ou d'un sous-groupe d'assurés.

Méthodes statistiques			
	Q^2	Robustesse intrinsèque et robustesse par rapport à l'échantillonnage	Temps d'exécution (apprentissage)
Matrice de transition	**90%**	**Parfaitement robuste**	**Moins de 1sec**
Régression linéaire	**89%**	**Parfaitement robuste**	**~ 5sec**
Matrice de transition + régression linéaire	**90%**	**Parfaitement robuste**	**~ 5sec**

Figure 4.9 – Synthèse des résultats des méthodes statistiques

- **Les méthodes de Data Science**
    - BERT appliqué aux actes
    - BERT appliqué aux actes
    - CART
    - *Random Forest*
    - Les méthodes en deux étapes

Les méthodes de Data Science quant à elles se focalisent sur le **comportement individuel** des assurés. Elles ne présenteront pas des résultats excellents dans l'ensemble (**Q^2 inférieurs**) mais **capteront des situations particulières et intéressantes dans le cadre de la prévention.**

Méthodes de Data Science				
	Q^2	Robustesse intrinsèque	Robustesse par rapport à l'échantillonnage	Temps d'exécution (apprentissage)
BERT (actes)	**58%**	**Peu robuste**	**Très robuste**	**~ 20min**
BERT (classes)	**75%**	**Robuste**	**Très robuste**	**~ 20min**
CART	**45%**	**Parfaitement robuste**	**Très robuste**	**~ 1min**
Random Forest	**61%**	**Très robuste**	**Très robuste**	**~ 1min**
BERT (actes) en deux étapes	**63%**	**Robuste**	**Très robuste**	**~ 20min**
BERT (classes) en deux étapes	**79%**	**Robuste**	**Très robuste**	**~ 20min**
Random Forest en deux étapes	**85%**	**Très robuste**	**Très robuste**	**~ 1min**

Peu robuste : $1\% < Var(Q^2)$ Robuste : $0{,}1\% < Var(Q^2) \leq 1\%$
Très robuste : $0{,}01\% < Var(Q^2) \leq 0{,}1\%$ Parfaitement robuste : $Var(Q^2) \leq 0{,}01\%$

Figure 4.10 –Synthèse des résultats des méthodes *Data Science* et *NLP*

Ainsi, il n'est ainsi pas judicieux de comparer les Q^2 entre les méthodes statistiques et les méthodes de Data Science. **Ces deux types de méthodes sont complémentaires**.

IV.4 Proposition d'amélioration via une méthode mixte

Dans cette partie, afin d'améliorer les prédictions du risque santé par entreprise obtenues précédemment, il sera proposé de **combiner les méthodes**. Cette idée repose sur les deux constats suivants :

- Il est nécessaire de combiner des méthodes statistiques avec des méthodes de Data Science afin de capter la **tendance générale du portefeuille et en même temps les situations particulières intéressantes.**
- Les méthodes de Data Science n'utilisent pas toutes les mêmes informations et peuvent ainsi **être complémentaires**.

Il sera d'abord présenté la méthodologie afin de combiner les méthodes et ensuite les résultats obtenus seront étudiés.

IV.4.1 Méthodologie

Cette méthode qui sera nommée méthode mixte utilise un modèle *Random Forest* avec 100 arbres aléatoires. Il s'agit en fait d'utiliser les prédictions obtenues par les trois méthodes en deux étapes et la méthode Matrice de transition sous la forme de variables en entrée du *Random Forest*. A ces quatre variables sera ajoutée la variable classe correspondant à la classe dont on souhaite prédire l'effectif. Ci-dessous une illustration de la méthodologie :

Prédicteurs					Variable à prédire
	Effectifs prédits				Effectif Méthode Mixte
Classe	Matrice de transition	BERT actes+[15]	BERT classes+	*Random Forest+*	
Infirmiers	5	9	3	7	6

Figure 4.11 – Principe méthode mixte illustré par un exemple pour la classe Infirmiers

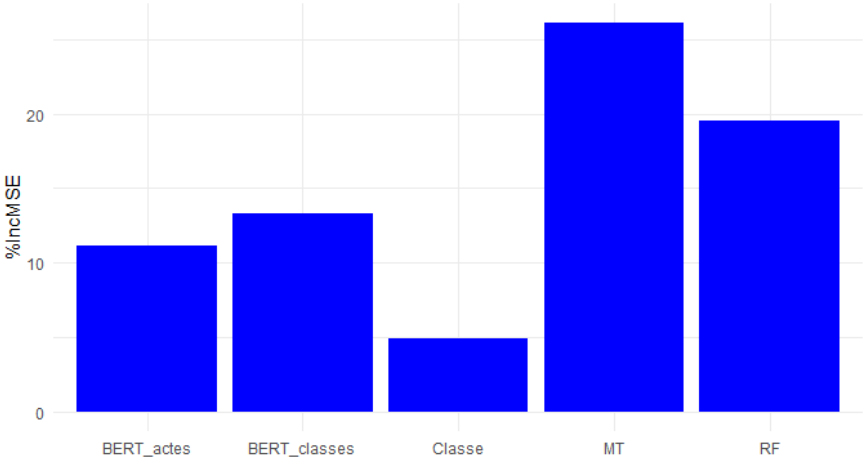
A noter : Comme précédemment, le modèle doit d'abord être entrainé sur les données en T6 afin de pouvoir prédire les effectifs en T7.

IV.4.2 Résultats

Dans un premier temps, avant de présenter la performance de la méthode mixte, il convient de rappeler que le *Random Forest* ne permet pas d'expliciter des règles de calcul ou de décision. En effet, les prédictions sont obtenues en agrégeant les prédictions d'un nombre important d'arbres CART. Cependant, pour apporter un peu d'explicabilité, l'importance de chaque variable dans la prédiction sera ici étudiée. Pour ce faire, l'indicateur choisi est le

%IncMSE (*Increase in MSE*) qui correspond à la perte de précision du modèle si la variable est omise.

Ci-dessous, les résultats obtenus :



MT : prédictions obtenues par la méthode matrice de transition.

RF : Prédictions obtenues par la méthode *Random Forest+*.

BERT_actes/BERT_classes : prédictions obtenues par la méthode *BERT actes+/BERT classes+*.

Classe* : classe correspondant à la classe dont on souhaite prédire l'effectif.

*La variable Classe a son importance car selon la classe, les poids des différentes méthodes ne seront pas les mêmes. Par exemple, on peut supposer que les prédictions obtenues avec la méthode MT seront plus importantes pour une classe très absorbante.

Figure 4.12 – Variables d'importance méthode mixte

Il est remarquable, sans surprise, que la variable la plus importante est la variable MT correspondant aux prédictions obtenues par la matrice de transition. De plus, les variables correspondant aux 3 méthodes en deux étapes sont aussi toutes les trois discriminantes. L'hypothèse de complémentarité des méthodes est donc confirmée.

De plus, cela se vérifie par la performance de la méthode mixte :

	Q^2	Robustesse intrinsèque	Robustesse par rapport à l'échantillonnage	Temps d'exécution (apprentissage)
BERT (actes)	**58%**	**Peu robuste**	**Très robuste**	**~ 20min**
BERT (classes)	**75%**	**Robuste**	**Très robuste**	**~ 20min**
CART	**45%**	**Parfaitement robuste**	**Très robuste**	**~ 1min**
Random Forest	**61%**	**Très robuste**	**Très robuste**	**~ 1min**
BERT (actes) en deux étapes	**63%**	**Robuste**	**Très robuste**	**~ 20min**
BERT (classes) en deux étapes	**79%**	**Robuste**	**Très robuste**	**~ 20min**
***Random Forest* en deux étapes**	**85%**	**Très robuste**	**Très robuste**	**~ 1min**
Méthode mixte	**95%**	**Très robuste**	**Très robuste**	**~ 45min**

Peu robuste : $1\% < Var(Q^2)$ Robuste : $0{,}1\% < Var(Q^2) \leq 1\%$

Très robuste : $0{,}01\% < Var(Q^2) \leq 0{,}1\%$ Parfaitement robuste : $Var(Q^2) \leq 0{,}01\%$

Figure 4.13 – Synthèse des résultats : méthode mixte vs autre méthodes

Comme on peut le constater ci-dessus, **le score Q^2 est de 95%**. Cela dépasse largement les Q^2 obtenus précédemment quelque soit la méthode. La méthode mixte sera donc la méthode retenue pour la suite du mémoire et le ciblage des actions de prévention.

V Application : création d'un outil d'aide au ciblage d'actions de prévention

Dans les parties précédentes, une méthodologie a été présentée afin d'étudier et prédire la déformation de la consommation santé par entreprise. **Dans cette partie, il s'agira de présenter et démontrer l'utilisation de ces prédictions dans le cadre du ciblage d'actions de prévention.**

Après avoir défini et illustré deux termes fondamentaux de cette partie le *groupe d'assurés ciblé (GAC)* et le *potentiel de prévention (PP)*, l'applicabilité des modèles précédents au ciblage sera démontrée en comparant les prédictions obtenues à la réalité. Enfin, la démarche d'utilisation de l'outil sera illustrée par un cas concret.

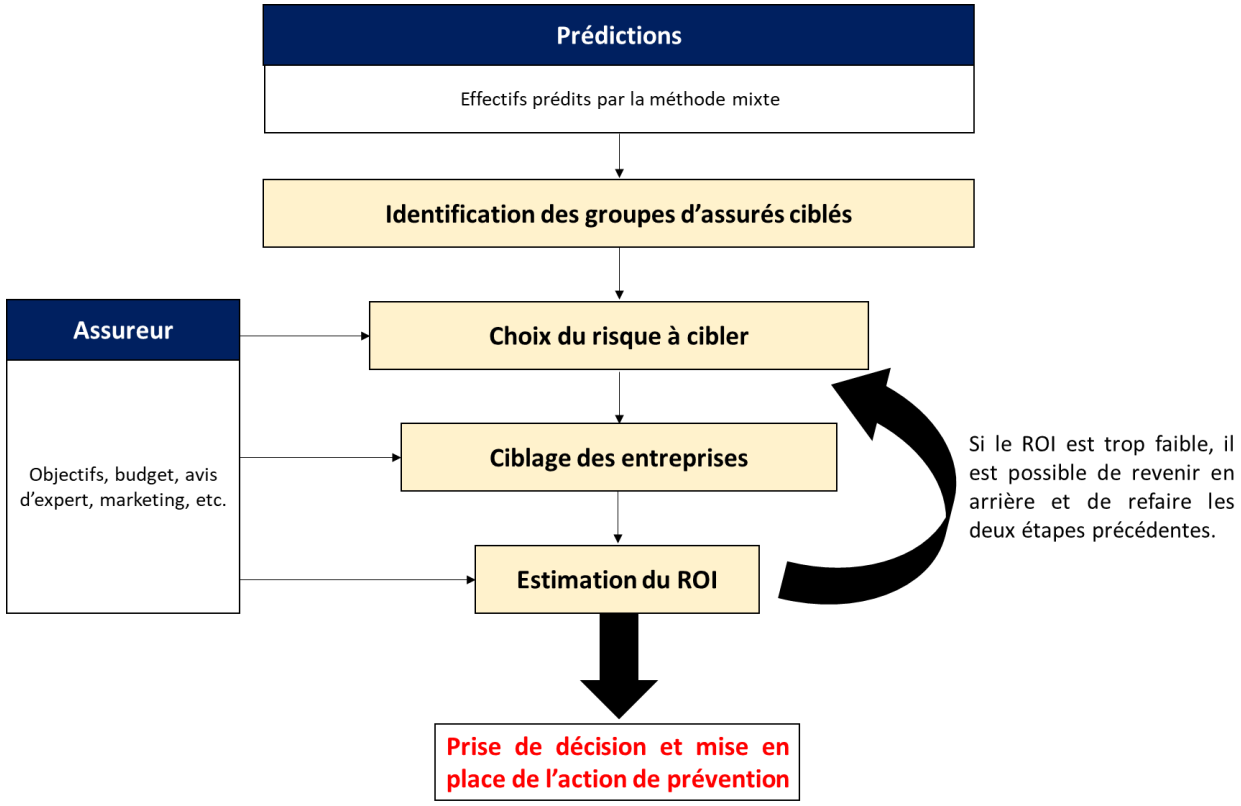


Figure 5.1 – Schéma illustratif de la démarche de ciblage d'actions de prévention

V.1 Présentation de la démarche

Afin de pouvoir proposer une méthodologie de ciblage d'actions de prévention, il est nécessaire dans un premier temps de préciser la définition choisie d'un « groupe d'assurés ciblé ». En effet, intuitivement, il s'agit d'un groupe d'assurés pour lequel l'impact d'une action de prévention sera important (recherche d'un ROI élevé). Or, cette définition est assez floue et nécessite donc d'être clarifiée.

V.1.1 Le potentiel de prévention

Dans ce mémoire, il sera choisi de proposer une action de prévention (qui reste à être définie à ce stade) à un groupe d'assurés[19] dont le RC moyen est largement supérieur au RC moyen du portefeuille. En effet, on supposera ici qu'il est possible de ramener ces individus à la moyenne du portefeuille à l'aide d'actions de prévention.

Le **potentiel de prévention** correspond à la baisse de RC qui pourrait être observée en modifiant le comportement de consommation santé d'un groupe d'assurés (à l'aide d'actions de prévention).

A noter : le potentiel de prévention défini ci-dessus présente l'inconvénient de biaiser l'analyse en mettant en avant les classes les plus coûteuses.

a. Ecart au portefeuille

Afin de définir mathématiquement le « potentiel de prévention », il est d'abord nécessaire d'introduire « **l'écart au portefeuille** ». Celui-ci est calculé de la manière suivante :

$$\Delta RC_{etp,\,cl}^{T7} = R_{etp} * \left(\overline{RC}_{cl}^{T7} - \overline{RC}^{T7}\right)$$

Avec :

$\Delta RC_{etp,\,cl}^{T7}$: écart au portefeuille au trimestre T7 des assurés de l'entreprise *etp* classés en classe *cl*.

$\overline{RC}^{T7}$: RC moyen par assuré du portefeuille en T7

$\overline{RC}_{cl}^{T7}$: RC moyen par assuré du portefeuille classé en classe de risque *cl* en T7

$R_{etp} = \frac{\sum_{i=1}^{6} \overline{RC}_{etp}^{Ti}}{\sum_{i=1}^{6} \overline{RC}^{Ti}}$: moyenne (de T1 à T6) des RC moyens trimestriels de l'entreprise sur la moyenne (de T1 à T6) des RC moyens trimestriels du portefeuille

$\left(\overline{RC}_{cl}^{T7} - \overline{RC}^{T7}\right)$ représente l'écart moyen de RC entre un assuré du portefeuille classé en classe cl et un assuré quelconque du portefeuille.

R_{etp} est un coefficient permettant d'affiner cet écart par entreprise (selon si l'entreprise a une consommation par assuré plus ou moins importante par rapport au portefeuille)

[19] Le groupe d'assurés dont il est question ici est uniquement un **groupe d'assurés ayant une consommation santé homogène**. En pratique, ce groupe d'assurés sera caractérisé par son entreprise et sa classe. Il s'agira donc d'identifier les groupes d'assurés (**entreprise *etp*, classe *cl***).

b. Potentiel de prévention

Le **potentiel de prévention (PP)** des assurés de l'entreprise *etp* classés en classe *cl* est ensuite calculé par la formule suivante :

$$PP_{etp,\,cl}^{T7} = \Delta RC_{etp,\,cl}^{T7} * n_{etp,\,cl}^{T7}$$

Avec $n_{etp,\,cl}^{T7}$: nombre d'assurés de l'entreprise etp classés en classe de risque cl en T7

D'autres métriques (plus ou moins complexes) auraient pu être construites afin de mesurer le PP d'un groupe d'assurés. On aurait pu par exemple utiliser des métriques simples, comme la dépense moyenne par assuré selon l'âge, ou la sur/sous-représentation d'une classe de risque par rapport à celle de l'ensemble du portefeuille. Cependant, ces métriques n'ont pas été retenues du fait de leur manque de précision et parfois des difficultés de mise en œuvre dans cette étude, sur les données exploitées.

V.1.2 Les groupes d'assurés ciblés

Dans ce mémoire, **la définition d'un groupe d'assurés ciblé (GAC) se basera sur les deux critères suivants :**

- **Fort PP** dont le critère « fort » sera détaillé ci-dessous

et/ou

- **Croissance du PP (entre T4 et T7)**

En effet, un groupe d'assurés (entreprise *etp*, classe *cl*) avec un fort PP est un groupe d'assurés anormalement coûteux par rapport au reste du portefeuille. La mise en place d'actions de prévention peut ainsi permettre de ramener ces individus à une consommation moyenne (proche de celle des assurés du portefeuille classés en classe *cl*).

De plus, une situation de croissance du PP est tout aussi problématique car elle correspond à une dégradation du risque. Il est donc souhaitable d'inverser cette tendance au plus tôt par la mise en place d'actions de prévention.

La croissance du PP sera mesurée par l'indicateur suivant :

$$\delta PP_{etp,cl} = PP_{etp,\,cl}^{T7} - \frac{PP_{etp,\,cl}^{T4} + 2*PP_{etp,\,cl}^{T5} + 3*PP_{etp,\,cl}^{T6}}{6}$$

Cette métrique a été choisie afin de prendre en compte la croissance du PP sur différentes intervalles de temps (2, 3 ou 4 trimestres) en mettant davantage de poids sur les données les plus récentes. De plus, les pondérations choisies sont arbitraires (on aurait pu tout autant prendre 2, 3 et 4 au numérateur et 9 au dénominateur).

Finalement, afin de prendre en compte les deux critères mentionnés, nous choisirons dans ce mémoire la **définition** suivante d'un **groupe d'assurés ciblé (GAC)** :

Soit (etp, cl) un groupe d'assurés et $s \in \mathbb{R}^+$ le seuil de ciblage[20].

L'ensemble $\{(etp, cl) \mid d_{etp,cl} > s\}$ est appelé le groupe d'assurés ciblé de seuil s.

$Avec : d_{etp,cl} = \delta PP_{etp,cl} + PP^{T7}_{etp,cl}$ [21]

Cet ensemble sera noté par la suite GAC_s.

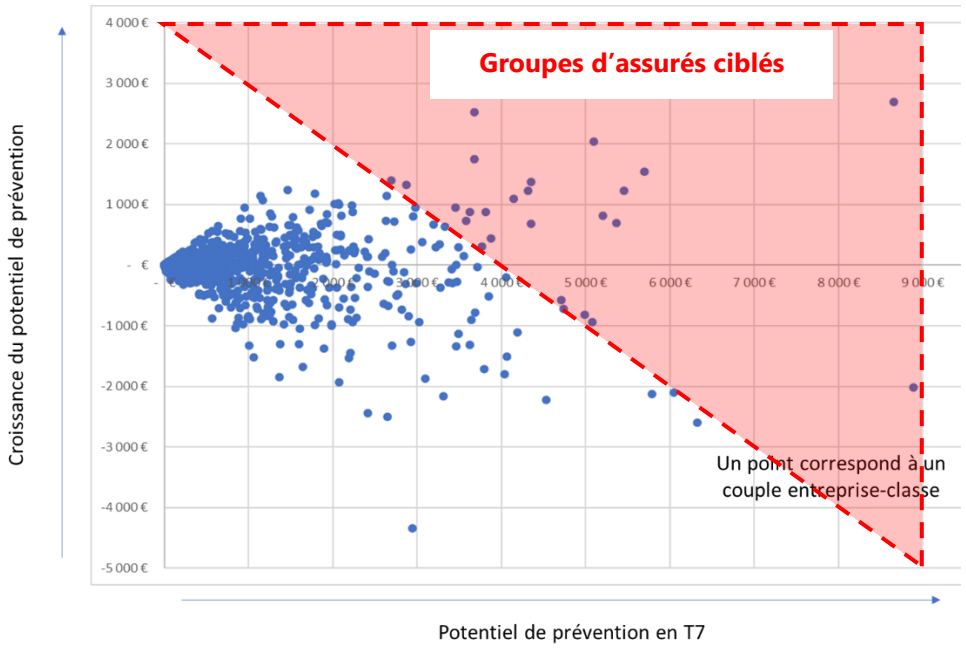


Figure 5.2 – Illustration GAC_s avec s = 4000€

[20] Le seuil de ciblage s est défini en fonction des caractéristiques du portefeuille, du nombre d'assurés ou entreprises à cibler, des coûts, du budget, des avis d'expert, etc.

[21] L'indicateur $d_{etp,cl}$ sera abusivement nommé « **distance** » dans la suite de ce mémoire.

V.2 Applicabilité des prédictions d'effectifs au ciblage d'actions de prévention

Dans la partie IV, une méthode a été présentée afin de prédire les effectifs par classe et par entreprise. Celle-ci présentait des prédictions cohérentes et un Q^2 élevé. Dans cette sous-partie, il s'agira de vérifier l'applicabilité de cette méthode au ciblage d'actions de prévention. Pour ce faire, la bonne prédiction des effectifs pour l'ensemble du portefeuille n'est pas suffisante, il est surtout nécessaire de **bien prédire les PP des GAC** (définis en V.1.2).

V.2.1 Méthodologie

Afin de vérifier la bonne prédiction des PP des GAC, **les PP réels seront comparés avec des PP prédits**.

Pour obtenir ces PP prédits, on calculera d'abord l'écart au portefeuille à l'aide des RC moyens en T6 :

$$\Delta RC_{etp,\,cl}^{T7\ pred} = R_{etp} * \left(\overline{RC}_{cl}^{T6} - \overline{RC}^{T6}\right)$$

$R_{etp} = \frac{\sum_{i=1}^{6} \overline{RC}_{etp}^{Ti}}{\sum_{i=1}^{6} \overline{RC}^{Ti}}$: moyenne (de T1 à T6) des RC moyens trimestriels de l'entreprise sur la moyenne (de T1 à T6) des RC moyens trimestriels du portefeuille

Ici, l'intérêt de multiplier par R_{etp} est de tenir compte du comportement global de consommation de l'entreprise dont le groupe d'assurés est issu.

On utilisera ensuite les prédictions d'effectifs obtenues par la méthode mixte pour obtenir le PP :

$$PP_{etp,\,cl}^{T7\ pred} = \Delta RC_{etp,\,cl}^{T7\ pred} * n_{etp,\,cl}^{T7\ pred}$$

Avec : $n_{etp,\,cl}^{T7\ pred}$: prédiction du nombre d'assurés de l'entreprise *etp* classés en classe de risque *cl* en T7

Et enfin, **l'étude se focalisera sur les groupes d'assurés ayant une distance parmi les plus élevés pour chacune des classes. Le critère retenu sera le suivant :**

$$\boldsymbol{d_{etp,cl} > centile^{95}(d_{cl})}$$

Avec : $centile^{95}(d_{cl})$ le 95ème centile de la distance pour les groupes associés à la classe *cl*

S'il est possible de bien prédire qu'un groupe soit dans les 5% les plus élevés en termes de distance, alors il sera possible de détecter les groupes d'assurés ciblés. Autrement dit, l'applicabilité des prédictions sera démontrée empiriquement.

V.2.2 Résultats

Les PP réels et prédits ont été comparés selon la méthodologie décrite ci-dessus (seulement pour les groupes validant $d_{etp,cl} > centile^{95}(d_{cl})$). La matrice de confusion suivante est ainsi obtenue :

En nombre de groupes d'assurés		REALITE	
		Non GAC	GAC
PREDICTION	Non GAC	2928	29
	GAC	27	49

Figure 5.3 – Matrice de confusion résultant de l'application du modèle mixte

Il s'agit de la matrice de confusion par nombre de groupes d'assurés. On remarque que parmi les 78 GAC réels (29+49), 49 ont été bien prédits (PP prédit supérieur au 95ème centile). Cela correspond à un **taux de vrais positifs de 63%**.

De plus, seulement 76 groupes ont été prédits comme GAC (dont 49 bien prédits), cela correspond à une **précision de 64%**.

Et enfin, l'indicateur ***accuracy***[22] n'est ici pas calculé car pas adapté du fait du déséquilibre GAC/non GAC.

Ces indicateurs permettent d'avoir une première idée de l'applicabilité des prédictions au ciblage. Cependant, comme mentionné précédemment, le seuil de 5% permettant de séparer les groupes en deux catégories (GAC/non GAC) est un seuil arbitraire. Autrement dit, des groupes comptabilisés dans les faux positifs peuvent en réalité être des groupes dont l'impact d'une action de prévention serait important (par exemple, les groupes dont le PP est compris entre le 90ème et le 95ème centile).

Ainsi, par la suite, une étude des montants et une analyse plus fine par classe seront présentées afin d'aller au-delà de cette vision binaire (GAC/non GAC).

a. Montants

Les 78 GAC réels présentent conjointement un PP (par trimestre) de 174 k€. Les 49 GAC bien prédits, quant à eux, présentent un PP de 131 k€. Ainsi, les 63% des GAC bien prédits correspondent à 75% du PP total des GAC. On en déduit que parmi les GAC, ceux présentant les PP les plus élevés sont ceux présentant les meilleures prédictions.

[22] $accuracy = \frac{VP+VN}{N}$ avec $VP = vrais\ positifs\ ; VN = vrais\ négatifs\ ; N = nombre\ total\ de\ groupes$

b. Analyse par classe

Afin d'appliquer les prédictions au ciblage, il est souhaitable que la méthode soit performante quelle que soit la classe concernée. En effet, si les prédictions sont mauvaises pour une classe, cela pourrait amener à un mauvais ciblage (mauvais risque ou/et mauvais groupes d'assurés) même si toutes les autres classes sont parfaitement prédites. Voici donc ci-dessous un tableau résumé des résultats par classe :

	REALITE	VRAIS POSITIFS		PREDICTION	
Classe	Nombre de GAC (A)	Nombre de GAC (B)	Taux (B) / (A)	Nombre de GAC (C)	Précision (B) / (C)
Analyses + Imagerie	8	4	50%	8	50%
ATM	8	5	63%	8	63%
Dentaire	8	4	50%	8	50%
Hospitalisation	8	7	88%	8	88%
Infirmiers	8	6	75%	8	75%
Kinésithérapie	8	5	63%	8	63%
Optique	8	6	75%	8	75%
Prothèses dentaires	8	5	63%	8	63%
Psy	7	3	43%	7	43%
Visite généraliste	7	4	57%	5	80%

Les classes Analyses, Appareillage, Consommation nulle, Généraliste, Imagerie + Prophylaxie, Ostéopathie, Pharmacie 30-65, Pharmacie 65, Sous-consommateurs et Spécialiste ne sont pas représentées car ne présentent aucun GAC réel.

Figure 5.4 – Résultats des prédictions de GAC par classe

On remarque ci-dessus que les prédictions sont globalement très bonnes (seize classes bien prédites en prenant en compte les classes non représentées dans le tableau) mais moins bonnes pour quatre classes. Afin de mieux comprendre, ces quatre classes vont être étudiées plus en détails via les montants réels et prédits des PP :

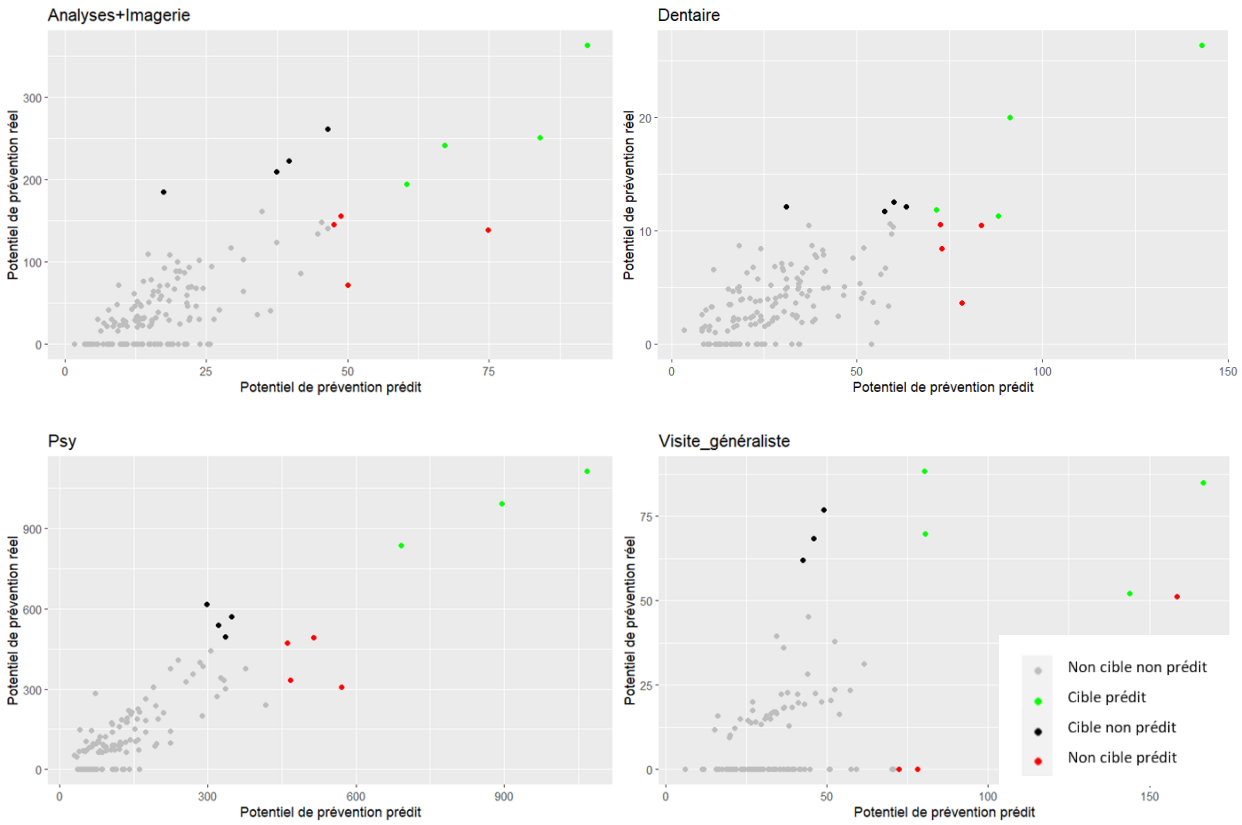


Figure 5.5 – PP prédit vs PP réel pour quatre classes de risque

Il est notable que les groupes non-cibles prédits ont dans la grande majorité des cas un PP réel élevé (mais ne faisant vraisemblablement pas partie des 5% les plus élevés). Les prédictions ne correspondent donc pas à l'attendu réalité vs prédiction mais identifient des groupes qui ont un PP intéressant.

Quant à la classe « Visite Généraliste », ce cas est spécifique. Il s'agit d'une classe très peu représentée (seulement 0.4% des assurés au trimestre T7). Les erreurs viennent du manque de données sur cette classe. Par conséquent, elle ne sera pas étudiée par la suite du fait du manque de fiabilité des prédictions.

Tous les indicateurs valident la bonne prédiction des potentiels de prévention des groupes d'assurés ciblés. Cela démontre que la méthode mixte construite plus haut dans ce mémoire est tout à fait applicable à l'identification des groupes d'assurés ciblés, et, par conséquent, au ciblage d'actions de prévention.

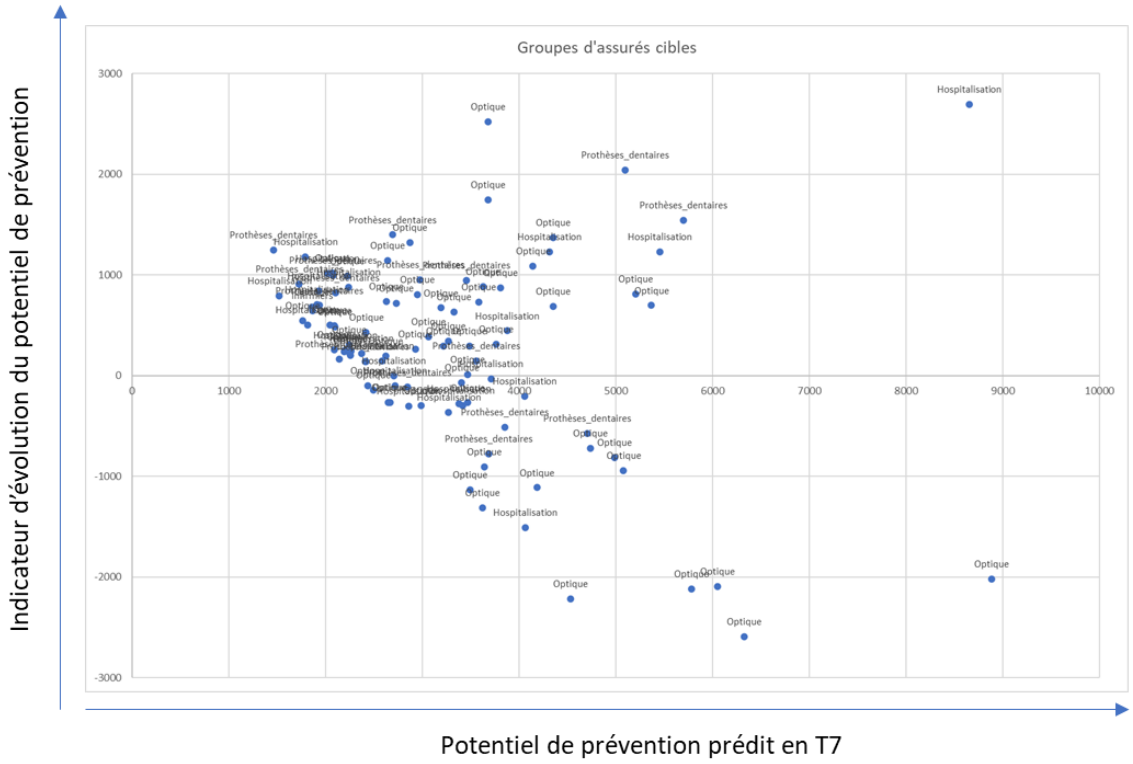
V.3 Présentation de l'outil développé

La notion de ciblage (prédiction de groupes ciblés) étant ainsi bien définie, **il s'agira maintenant de détailler les étapes à suivre afin de mettre en place un plan d'actions de prévention ciblée à l'aide des prédictions d'effectifs.**

V.3.1 Choix du risque à cibler

Dans un premier temps, il est nécessaire de **déterminer quel risque sera visé par l'action de prévention**. Cela peut être défini en amont par avis d'expert mais peut aussi être guidé par la prédiction des PP.

Pour ce faire, il est possible de filtrer les groupes d'assurés par leurs PP et leurs évolutions de PP afin d'identifier les risques majeurs et/ou ceux qui se développent. Ci-dessous le graphe présentant les 100 groupes d'assurés présentant les plus grandes valeurs de $d_{etp,cl}$ ainsi qu'un tableau résumant des statistiques par classe de ces 100 groupes ciblés :



	Nombre de cibles	Effectif	Effectif entreprise	Potentiel de prévention
Hospitalisation	**24**	112	4 559	75 302 €
Infirmiers	**1**	15	320	1 868 €
Optique	**58**	525	10 401	199 317 €
Prothèses_dentaires	**17**	103	3 885	52 429 €
Total	**100**	**755**	**19 165**	**328 916 €**

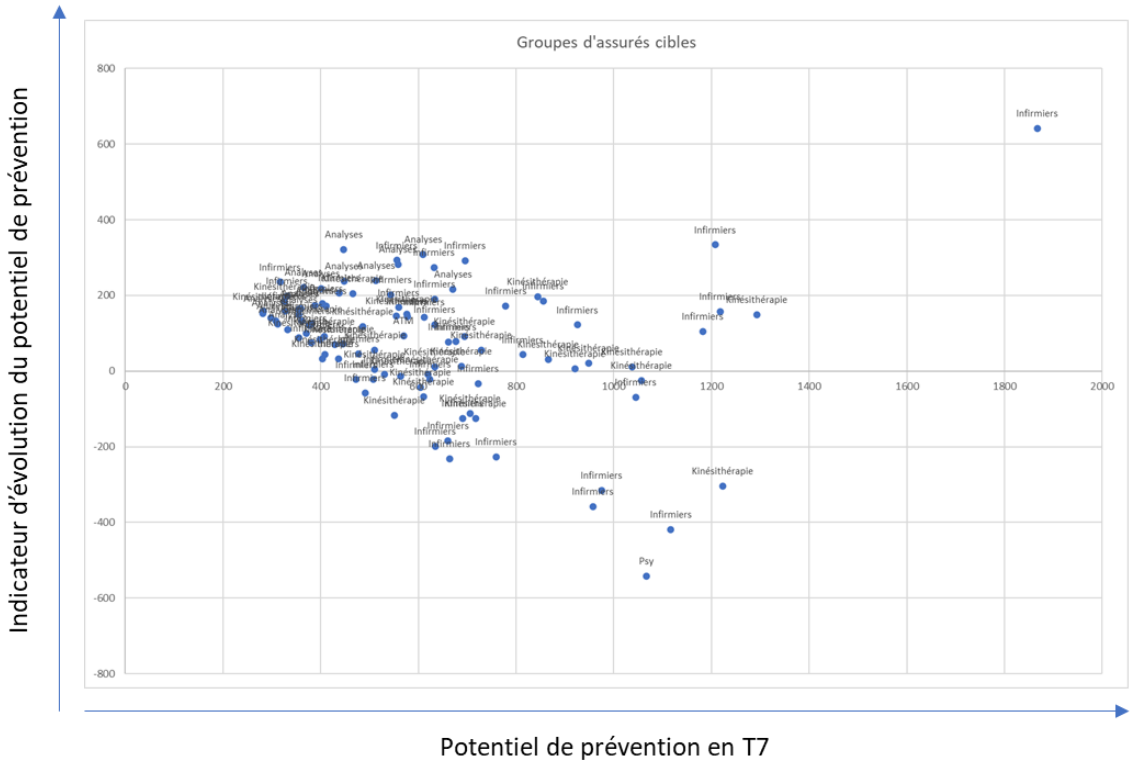
Figure 5.6 – Ciblage du risque : cent groupes ciblés principaux

On observe que les groupes ciblés présentés ci-dessus correspondent aux classes les plus coûteuses (biais mentionné précédemment). Les autres classes sont totalement occultées. Pour résoudre ce problème, il est possible de définir des classes à exclure.

Dans cet exemple, les classes Optique et Prothèses dentaires seront exclues en raison de leur risque considéré comme difficilement réductible ou évitable par des actions de prévention.

La classe Hospitalisation sera elle aussi exclue afin de montrer que la prévention peut également concerner des classes moins coûteuses, telles que la classe Infirmiers.

On obtient ainsi, après les exclusions, 100 nouveaux groupes d'assurés :



	Nombre de cibles	Effectif	Effectif entreprise	Potentiel de prévention
Analyses	**17**	327	4 072	7 003 €
Infirmiers	**47**	336	8 749	31 512 €
Kinésithérapie	**34**	294	6 647	21 850 €
Psy	**1**	7	320	1 067 €
Total	**99**	**964**	**19 788**	**61 432 €**

Figure 5.7 – Ciblage du risque : cent groupes ciblés principaux après suppression des risques non désirés

Parmi les risques restants (après exclusions), la classe Infirmiers présente le PP le plus important et le graphe présente plusieurs groupes d'assurés classés en Infirmiers dans la partie supérieure droite (autrement dit, parmi les groupes les plus réceptifs à une action de prévention).

Il sera ainsi choisi de cibler le risque Infirmiers[23].

[23] Pour rappel, le risque « Infirmiers » ne correspond pas uniquement aux consommations de soins d'infirmiers. Le terme « Infirmiers » correspond seulement à un résumé de la consommation santé (en fréquence) de l'ensemble des assurés de la classe de risque.

V.3.2 Ciblage des entreprises

Dans un second temps, il s'agira de **cibler les entreprises les plus intéressantes**. En effet, parmi les 47 groupes ciblés correspondant à la classe de risque Infirmiers, tous ne doivent pas obligatoirement faire l'objet d'une action de prévention.

Voici ci-dessous un extrait des statistiques par entreprise pour la classe Infirmiers : Les entreprises sont classées par ordre décroissant de distance (les cibles principales sont donc présentées en haut du tableau).

ID etp	Effectif etp	Age moyen	Effectif classe	Potentiel de prévention	Indicateur croissance
E1	320	42	15	1 868 €	642 €
E2	227	44	9	1 208 €	334 €
E3	208	46	9	1 218 €	157 €
E4	302	47	12	1 183 €	105 €
E5	284	44	11	925 €	122 €
E6	239	46	9	856 €	184 €
E7	239	44	9	696 €	292 €
E8	82	62	5	1 045 €	-69 €
E9	292	41	13	779 €	171 €
E10	231	48	7	632 €	273 €
E11	219	44	7	814 €	43 €
E12	141	39	4	556 €	293 €
			...		
Total	**8749**	**43**	**336**	**31 512 €**	**3 180 €**

Figure 5.8 – Extrait des statistiques des groupes d'assurés (classe Infirmiers)

Plusieurs critères vont être importants pour le ciblage des entreprises. D'abord, on remarque qu'une entreprise (E8) se distingue par un âge moyen élevé. Selon l'action de prévention choisie, il sera nécessaire de vérifier si celle-ci reste pertinente pour un groupe d'assurés âgés. Ensuite, **il faudra prendre en compte la part d'assurés classés en Infirmiers. En effet, une part trop faible impliquerait un ciblage trop large** (tous les salariés de l'entreprise) **pour des retombées bénéfiques trop restreintes** (uniquement pour les salariés de l'entreprise appartenant à cette classe de risque[24]) car il faut supposer que les actions seront proposées à tous les salariés de l'entreprise. Et enfin, comme mentionné précédemment, les groupes d'assurés seront ciblés à l'aide d'un seuil de ciblage ($d_{etp,cl} > s$). Ici, dans notre exemple, il sera choisi $s = s_{15}$ avec s_{15} le seuil permettant de retenir les 15 entreprises présentant les plus grandes valeurs de $d_{etp,cl}$.

On obtient ainsi les entreprises suivantes :

[24] Ceci n'est pas totalement exact : des personnes à l'extérieur de la classe pourraient aussi être impactées positivement par l'action et des personnes appartenant à la classe pourraient être peu réceptives.

Choix des entreprises		
ID	Effectif total	Effectif classe
E1	320	15
E2	227	9
E3	208	9
E4	302	12
E5	284	11
E6	239	9
E7	239	9
E8	82	5
E9	292	13
E10	231	7
E11	219	7
E12	141	4
E13	209	7
E14	191	6
E15	220	9
Total	**3404**	**133**
Portefeuille	**18590**	**643**
Ratio	**18%**	**21%**

Figure 5.9 – Effectifs des entreprises ciblées

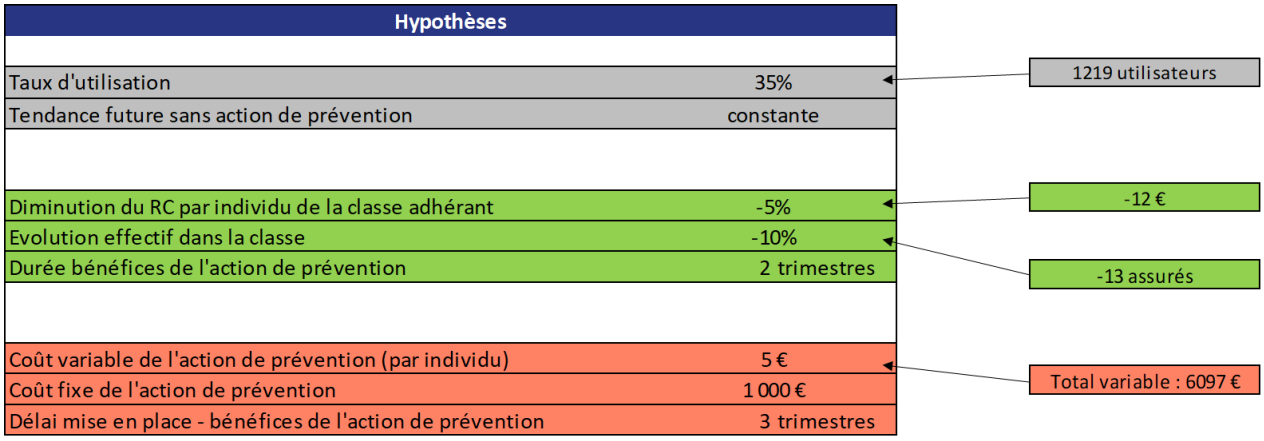
Ainsi :

- **Le risque Infirmiers a été choisi comme cible car il présente le meilleur potentiel de prévention.**
- **Les 15 entreprises présentées ci-dessus ont été ensuite retenues car elles présentent les plus grandes valeurs de $d_{etp,cl}$.**

V.3.3 Estimation du ROI

La dernière étape de la démarche de ciblage d'actions de prévention consiste donc à estimer le ROI associé. Si ce dernier est insuffisant, il sera nécessaire de revenir aux deux étapes précédentes (ciblage du risque et des entreprises). Il est ainsi possible de faire plusieurs allers-retours jusqu'à trouver le risque et les entreprises permettant d'obtenir le meilleur ROI.

Cette étape ne peut pas être basée sur les prédictions. Elle requiert l'avis d'un expert ou de l'expérience pour établir des hypothèses et dépend du design de l'action de prévention. Ci-dessous une illustration d'hypothèses permettant l'estimation du ROI :

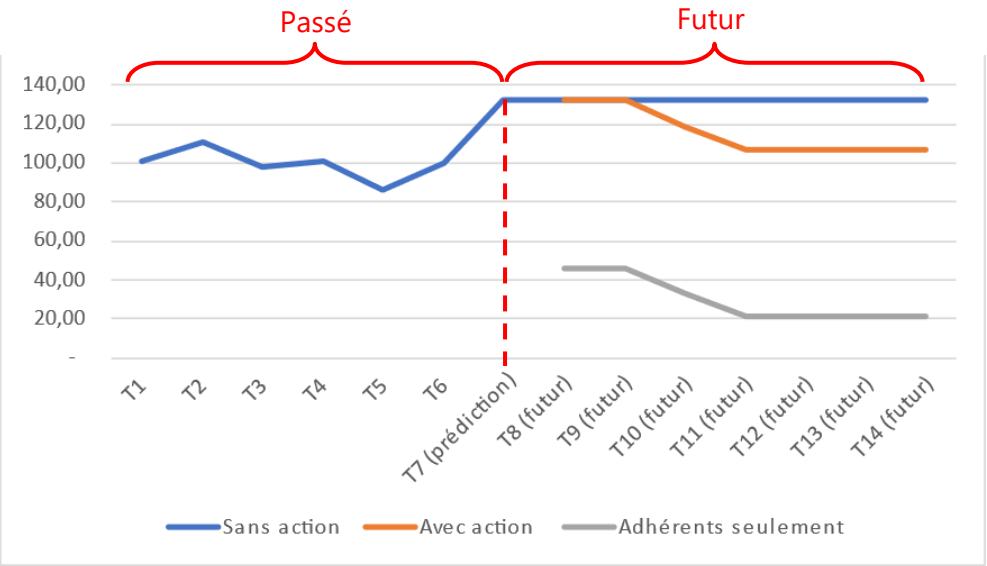


Taux d'utilisation : part des assurés ciblés adhérant à l'action de prévention (utilisant les services/soins proposés)
Tendance future sans action de prévention : tendance future, par trimestre et sans action de prévention, de l'effectif des assurés ciblés classés dans la classe de risque à cibler.
Diminution du RC par individu de la classe adhérant : Diminution du RC (par individu et par trimestre) des individus classés dans la classe de risque à cibler et adhérant à l'action de prévention
Evolution effectif dans la classe : tendance future, par trimestre et avec action de prévention, de l'effectif des assurés ciblés classés dans la classe de risque à cibler.
Durée bénéfices de l'action de prévention : durée pendant laquelle l'action de prévention aura un impact sur le RC et l'effectif dans la classe de risque à cibler.
Coûts : Il est supposé qu'une action de prévention a un coût pour l'assureur, qui se compose d'un coût variable (par assuré) et d'un coût fixe indépendant du nombre d'assurés.
Délai mise en place – bénéfices de l'action de prévention : délai entre l'instant où l'action de prévention est mise en place et l'instant où les bénéfices de l'action deviennent visibles.

Figure 5.10 – Illustration d'hypothèses permettant l'estimation du ROI

A noter : ces hypothèses ne sont qu'à titre illustratif et ne reposent sur aucune donnée réelle ou sur aucun avis d'expert.

Sous ces hypothèses indiquées plus haut, il est ainsi obtenu une diminution de l'effectif en classe Infirmiers illustrée par le graphe ci-dessous :



Sans action : effectif d'assurés classés en classe Infirmiers (estimation en l'absence d'action de prévention).
Avec action : effectif d'assurés classés en classe Infirmiers (estimation si l'action de prévention est mise en place).
Adhérents seulement : effectif d'assurés classés en classe Infirmiers et adhérant à l'action de prévention.
Les trois effectifs présentés concernent seulement les assurés des entreprises ciblées.

Figure 5.11 – Evolution des effectifs sous les hypothèses choisies

Cela se traduit par les résultats suivants :

Bénéfices de l'action de prévention	**14 355 €**
Coût de l'action de prévention	**7 097 €**
ROI	**102%**

Pour résumer :

- **Le risque Infirmiers** a été choisi comme cible car il présente **le meilleur potentiel de prévention.**
- **15 entreprises ont été retenues** (et présentées ci-dessus) car elles présentent les plus grandes valeurs de $d_{etp,cl}$.
- Sous les hypothèses choisies, on estime les **bénéfices de l'action de prévention ciblée** à 14k€ (102% de ROI).

Il a été ainsi présenté une méthode innovante permettant **d'identifier les groupes d'assurés** pour lesquels une action de prévention serait intéressante. De plus, la méthode permet **d'identifier quel est le risque** associé à ces groupes et peut ainsi **orienter le choix de l'action de prévention**. Et enfin, pour terminer, elle permet aussi, une fois le type d'actions fixé, de **mesurer le potentiel retour sur investissement**. Cette méthode, couplée à des tests (afin d'obtenir un retour sur expérience) permet donc de piloter le ciblage d'actions de prévention de A à Z. Dans notre exemple, le ROI est très satisfaisant, l'actuaire pourra donc entamer les discussions avec des professionnels de santé spécialisés afin de préciser le programme de prévention et ensuite le proposer à la direction de sa compagnie d'assurance.

Conclusion

Dans ce mémoire, **une méthodologie a été présentée afin de comprendre et prédire la consommation santé d'un portefeuille d'assurance santé collective**. Ces travaux ont été orientés par la volonté de **proposer un outil aux assureurs afin de cibler leurs actions de prévention**.

D'abord, il a été montré que l'assureur dispose, grâce à ses bases des prestations versées, de données permettant de comprendre et anticiper la consommation santé des assurés.

Ensuite, une méthode a été développée afin d'étudier la consommation santé par assuré d'un portefeuille au cours du temps. Celle-ci repose sur une méthode de réduction des données et de classification des assurés non supervisée développée par ADDACTIS. Dans ce mémoire, cette méthode est complétée par la notion de temporalité (au cours du temps) et de parcours de consommation. On parle ainsi de « **parcours de classes de risque** ».

Ces travaux ont permis d'améliorer la compréhension de la consommation santé par assuré au cours du temps mais ne permettent pas d'anticiper les risques futurs. Pour cette raison, une méthode a été développée dans un second temps afin de prédire les effectifs par classe de risque à la maille entreprise.

Plusieurs méthodes « simples » ont été testées (des méthodes statistiques, des méthodes traditionnelles de *Data science* ou des méthodes issues du traitement automatique du langage naturel). Cela a conduit à l'élaboration d'une méthode plus complexe, nommée « **méthode mixte** », combinant les prédictions de quatre méthodes à l'aide d'un *Random Forest*.

Et enfin, les prédictions, obtenues par la méthode mixte, ont été utilisées pour concevoir un **outil d'aide au ciblage d'actions de prévention**.

Cet outil, combiné avec des avis d'experts (médecins, préventeurs, etc.) permet à l'actuaire de développer des programmes de prévention ciblée de A à Z : de l'identification du risque et des groupes d'assurés à cibler jusqu'à l'estimation du potentiel ROI.

Toutefois, il convient de rappeler qu'il y a certaines limites à ces travaux. Tout d'abord, les techniques utilisées requièrent une grande quantité de données, ce qui peut rendre la mise en œuvre de cet outil complexe. De plus, les prestations consommées au dernier trimestre sont utilisées pour la modélisation, alors qu'en pratique il peut y avoir un délai entre la date de soin et la date de réception des données par l'assureur.

Pour résumer, ces travaux, bien qu'encore difficilement industrialisables, ouvrent la voie à une meilleure compréhension des risques santé par les assureurs. De plus, les méthodes développées dans ce mémoire pourraient être utilisées sur un échantillon plus large, tel qu'un échantillon de la consommation santé française. Cela pourrait permettre de positionner les entreprises/portefeuilles par rapport à la vision du marché français.

Bibliographie

[Bourdois, 2019] : Bourdois, 2019. NLP - Illustration du modèle Transformer de Vaswani et al.
Accessible via : https://lbourdois.github.io/blog/nlp/Transformer/

[Breiman, 1984] : Breiman, 1984. Classification and Regression Trees.

[Breiman, 2001] : Breiman, 2001. Random Forests.

[Capterra, 2021] : Capterra, 2021. 36 % de Français sont inquiets du traitement de leurs données par les applications mobiles de santé.
Accessible via : https://www.capterra.fr/blog/2059/applications-mobiles-sante

[Chen & al, 2012] : Chen, Xu, Weinberger et Sha, 2012. Marginalized Denoising Autoencoders for Domain Adaptation.

[Devlin & al, 2018] : Devlin, Chang, Lee et Toutanova, 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

[Gauchon, 2020] : Gauchon, 2020. Contribution à l'étude de la prévention en assurance santé. (Thèse de doctorat à l'université de Lyon)

[Gauchon & al, 2019] : Gauchon et Hermet , 2019. Individuals tell a fascinating story: using unsupervised text mining methods to cluster policyholders based on their medical history.

[Gordon,1982] : Gordon, 1982. Rapport Flajolet, annexe 1 - La prévention définitions et comparaisons.

[Hermet, 2020] : Hermet, 2020. La psychiatrie en assurance santé. (Mémoire Institut des actuaires)

[IONOS, 2020] : Ionos, 2020. Réseau de neurones artificiels : quelles sont leurs capacités ?
Accessible via : https://www.ionos.fr/digitalguide/web-marketing/search-engine-marketing/quest-ce-quun-reseau-neuronal-artificiel/

[Kohonen, 1990] : Kohonen, 1990. The self-organizing map.

[Korpusik & al, 2019] : Korpusik, Liu et Glass, 2019. A Comparison of Deep Learning Methods for Language Understanding.

[Le Temps, 2023] : Le Temps, 2023. Pourquoi le système de santé français est au bord du gouffre ?
Accessible via : https://www.letemps.ch/monde/europe/systeme-sante-francais-bord-gouffre

[Murtagh, 1995] : Murtagh, 1995. Interpreting the Kohonen self-organizing feature map using contiguity-constrained clustering.
Source : Pattern Recognition Letters, Volume 16, Issue 4, pp. 399–408

[OMS, 1948] : OMS, 1948. Constitution de l'Organisation mondiale de la Santé.

[Parde, 2020] : Parde, 2020. Feedforward Neural Network Basics.

Accessible via :
https://www.natalieparde.com/teaching/cs_421_fall2020/Feedforward%20Neural%20Network%20Basics.pdf
[Sudhir & al, 2021] : Sudhir et Suresh , 2021. Comparative study of various approaches, applications and classifiers for sentiment analysis.

A propos de **BERT** et du **NLP** :

Datascientest, 2020. Natural Language Processing (NLP) : Définition et principes.
Accessible via : https://datascientest.com/introduction-au-nlp-natural-language-processing

Datascientest, 2021. BERT : Un outil de traitement du langage innovant.
Accessible via : https://datascientest.com/bert-un-outil-de-traitement-du-langage-innovant

Ledatascientist, 2020. A la découverte de BERT.
Accessible via : https://ledatascientist.com/a-la-decouverte-de-bert/

A propos des **programmes de prévention mentionnés** dans l'état de l'art :

AESIO mutuelle (dépistage du diabète), 2021.
Accessible via : https://ensemble.aesio.fr/agir-ensemble/agenda/diabete-et-alimentation-webconf-aesio

AXA Prévention, 2010.
Accessible via : https://www.axaprevention.fr/fr

« Relax - bien-être » et « Boost - activité physique » par SwissLife, 2020.
Accessible via : https://www.swisslife.fr/home/des-services-d-accompagnement-pour-prendre-soin-de-votre-sante/Nos-programmes-de-prevention/Boost-et-Relax-deux-programmes-de-coaching-pour-agir-en-faveur-du-bien-etre-et-de-l-activite-physique-en-entreprise.html

Harmonie Prévention, 2013.
Accessible via : https://www.harmonie-prevention.fr

Generali Vitality, 2017.
Accessible via : https://www.generalivitality.com/fr/fr

Bandeau Dreem par Mutuelle Ociane Matmut, 2020.
Accessible via : https://presse.matmut.fr/communique/201265/La-Mutuelle-Ociane-Matmut-Dreem-s-associent-pour-proposer-aux-adherents-de-complementaire-sante-solution-efficace-aux-problematiques-de-sommeil?cm=1

Annexes

a) Réseau de neurones à propagation avant

Le réseau de neurones à propagation avant (*feedforward neural network*) est le type de réseau de neurones le plus simple. Comme son nom l'indique, l'information est propagée en avant, elle ne se déplace que dans un seul sens (de la couche de neurones en entrée à la couche de neurones en sortie). Chaque couche de neurones k reçoit de l'information uniquement de la couche de neurones k-1 and transmet de l'information seulement à la couche k+1.

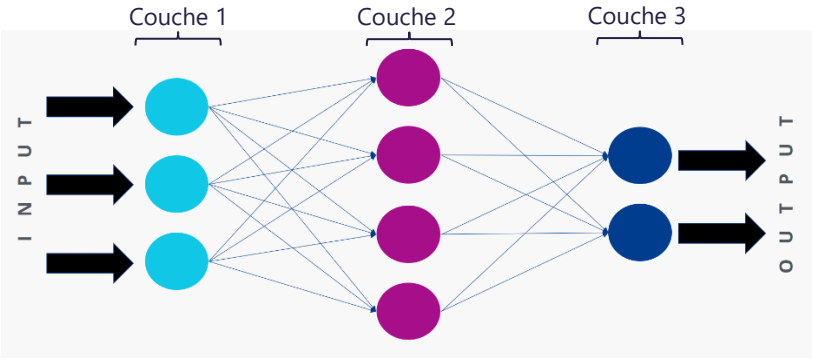


Figure a.1 – Réseau de neurones à propagation avant [IONOS, 2020]

b) Réseau de neurones récurrent

Le réseau de neurones récurrent est un type de réseau de neurones complexe où les connexions entre les neurones peuvent être récurrentes. Autrement dit, l'information n'est plus propagée que dans un seul sens et il existe des cycles dans le réseau.

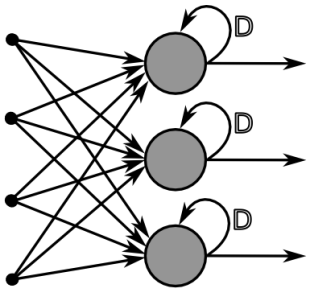


Figure b.1 – Réseau de neurones récurrent : illustration cycle [Parde, 2020]

Ces réseaux de neurones sont particulièrement intéressants en NLP car ils comprennent le contexte des données en entrée grâce aux connexions récurrentes (le réseau a une « mémoire »).

c) mSDA (marginalized Stacked Denoising Autoencoders)

Le mSDA [Chen & al,2012] est une méthode de réduction de dimension permettant dans ce mémoire d'identifier les liens entre les différents actes de santé.

Une matrice de fréquence (figure 2.7) sera donnée en entrée du mSDA et celui-ci retournera une matrice de corrélation des différents actes de santé et une nouvelle matrice de fréquence prenant en compte les corrélations.

Cette méthode se base sur des réseaux de neurones auto-encodeurs débruiteurs :

- Qu'est-ce qu'un auto-encodeur ?

Il s'agit d'un type de réseau de neurones compressant puis reconstruisant une entrée. Cela permet de capter les liens entre les différentes caractéristiques des données en entrée et de créer un espace dimensionnel réduit.

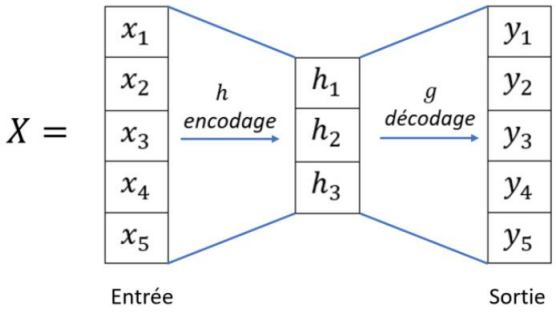


Figure c.1 – Auto-encodeur [Hermet, 2020]

- Qu'est-ce qu'un auto-encodeur débruiteur ?

Les auto-encodeurs débruiteurs fonctionnent de la même manière que les auto-encodeurs classiques, mais l'entrée est délibérément corrompue. Cela permet d'améliorer la correspondance entre l'entrée et la sortie. De cette manière, la sortie est plus précise.

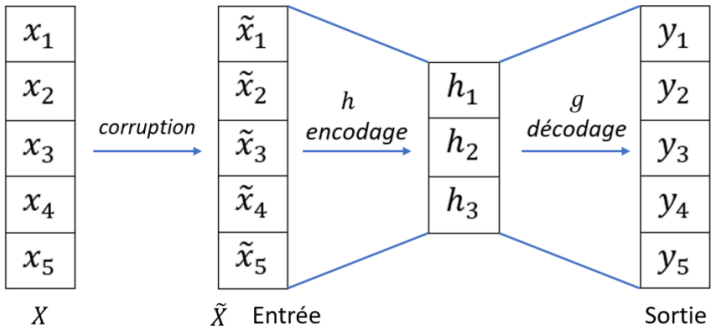


Figure c.2 – Auto-encodeur débruiteur [Hermet, 2020]

Le mSDA est finalement une chaine d'auto-encodeurs débruiteurs (SDA) avec une extension (*marginalized*) signifiant l'utilisation d'une distribution marginale (moyenne des sorties de l'ensemble des auto-encodeurs formant l'architecture) afin d'améliorer la reconstruction des données.

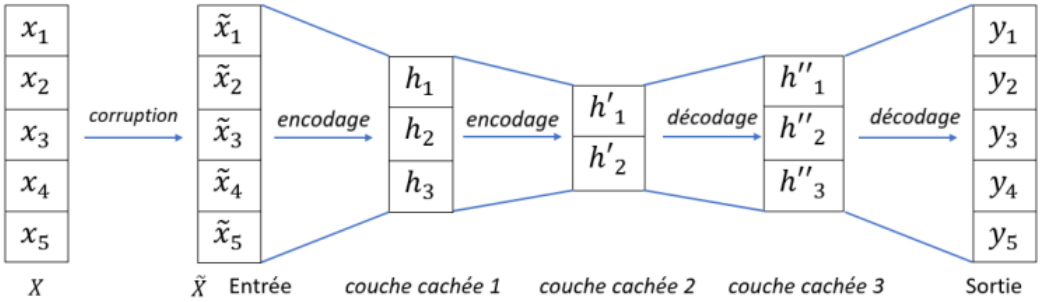


Figure c.3 – SDA [Hermet, 2020]

Pour une description plus complète de l'application du mSDA aux liens entre les différentes actes santé et de la théorie du mSDA, le lecteur pourra se rapporter aux pages 55-67 du mémoire [Hermet, 2020].

d) Carte de Kohonen

La méthode des cartes auto-organisatrices de Kohonen [Kohonen, 1990] est une méthode de classification non supervisée. Elle est utilisée dans ce mémoire pour classer les assurés en classes géographiquement organisées.

Les cartes de Kohonen sont des réseaux de neurones :

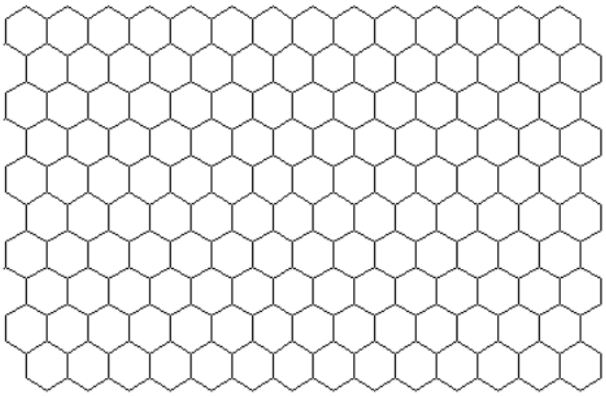


Figure d.1 – Illustration carte de Kohonen

Initialement, chaque neurone est initialisé avec un vecteur de poids aléatoires. Pour chaque assuré, il va être déterminé quel est le neurone qui le représente le mieux et l'assuré sera associé à ce neurone. Ensuite, le poids de ce neurone ainsi que les poids des neurones voisins vont être mis à jour afin de minimiser la distance avec l'assuré. Ce processus sera répété pour tous les assurés. De plus, le processus complet (le passage de tous les assurés) sera répété plusieurs fois (en changeant l'ordre de passage des assurés). Chaque assuré est ainsi, à la fin, associé à un neurone, permettant d'obtenir une classification des assurés en un certain nombre de classes (correspondant au nombre de neurones).

Cependant, le nombre de neurones peut être trop important, c'est pourquoi la méthode des cartes de Kohonen sera couplée avec une classification ascendante hiérarchique selon une approche développée dans [Murtagh, 1995]. Cette méthode semi-supervisée permettra de regrouper les neurones en un nombre de classes décidé par l'utilisateur.

e) CART et *Random Forest*

- CART (*Classification and Regression Trees*)

Il s'agit d'un algorithme d'apprentissage automatique supervisé utilisant un arbre de décision à des fins de classification ou de régression. Dans ce mémoire, il s'agit de classification.

Le principe est le suivant : l'arbre est construit récursivement, en choisissant, à chaque étape, la variable explicative (et le seuil ou modalité) qui donne la meilleure séparation des données en deux sous-ensembles les plus homogènes possibles. Pour commencer, l'ensemble des données en entrée est ainsi divisé en deux sous-ensembles (selon un critère de minimisation de l'erreur de classification). Cette opération est ensuite répétée récursivement pour chaque sous-ensemble obtenu. Un critère d'arrêt (taille minimale des classes, profondeur de l'arbre, etc.) permettra de stopper le processus. De cette manière, on obtient une classification de l'ensemble des données en entrée (données d'entrainement du modèle).

Dans ce mémoire, nous souhaitons prédire. Pour ce faire, l'arbre CART sera parcouru (en fonction des valeurs des données dont on souhaite prédire la classe) et une classe (le sous-ensemble final) sera ainsi déterminée (prédite).

Pour plus de détails, le lecteur pourra se rapporter aux travaux de Breiman [Breiman, 1984].

- *Random Forest*

Il s'agit d'une méthode qui utilise plusieurs arbres CART (« forêt ») afin d'améliorer les performances de prédiction.

Chaque arbre est construit de manière indépendante en utilisant un sous-ensemble aléatoire d'observations issues des données d'entrainement et un sous-ensemble aléatoire de variables explicatives. Les prédictions de chaque arbre sont ensuite agrégées.

Ainsi, pour prédire, chaque arbre renvoie sa prédiction (classe) et la prédiction finale du *Random Forest* est ensuite déterminée par un vote majoritaire.

Pour plus de détails, le lecteur pourra se rapporter aux travaux de Breiman [Breiman, 2001].

f) Caractérisation statistique des classes de risque

Classe (Interprétation de la classe)	Age moyen	% d'hommes	DE moyenne par assuré*	RC moyen par assuré*	Part de l'effectif**
Pharmacie 65	45	54	176 €	73 €	11,89%
Sous-consommateurs	41	67	89 €	38 €	9,09%
Prothèses dentaires	45	64	1034 €	574 €	2,90%
Dentaire	43	66	335 €	146 €	2,29%
Imagerie + Prophylaxie	44	60	204 €	80 €	8,10%
Pharmacie 30-65	44	62	141 €	64 €	13,49%
Hospitalisation	44	59	1585 €	654 €	3,17%
Visite généraliste	43	50	342 €	144 €	0,70%
Appareillage	44	58	203 €	96 €	1,26%
ATM	45	53	302 €	138 €	6,22%
Kinésithérapie	46	54	454 €	190 €	4,59%
Analyses	46	47	312 €	125 €	3,55%
Spécialiste	44	39	235 €	100 €	6,70%
Ostéopathie	44	70	169 €	86 €	2,97%
Analyses + Imagerie	46	53	362 €	137 €	1,78%
Infirmiers	46	56	540 €	214 €	4,52%
Psy	44	45	528 €	244 €	1,38%
Analyses	42	46	307 €	123 €	6,32%
Généraliste	43	49	184 €	80 €	2,41%
Optique	45	59	654 €	483 €	6,67%

*Les montants (DE ou RC) sont exprimés par trimestre (non annualisés).

**La part de l'effectif est calculée sans prendre en compte les assurés non consommants (la classe *Consommation nulle*).

Figure f.1 – Caractérisation statistique des classes de risque

g) Matrice de transition

	Analyses	Analyses + Imagerie	Appareillage	ATM	Consommation nulle	Dentaire	Généraliste	Hospitalisation	Imagerie + Prophylaxie	Infirmiers	Kinésithérapie	Optique	Osthéopathie	Pharmacie 30-65	Pharmacie 65	Prothèses dentaires	Psy	Sous-consommateurs	Spécialiste	Visite généraliste
Analyses	17%	2%	1%	5%	12%	1%	2%	4%	7%	5%	2%	5%	2%	10%	10%	1%	1%	5%	7%	1%
Analyses + Imagerie	11%	2%	2%	6%	12%	2%	2%	3%	6%	5%	4%	5%	2%	8%	11%	5%	1%	7%	7%	0%
Appareillage	7%	1%	3%	5%	24%	1%	2%	3%	7%	4%	3%	6%	3%	8%	6%	2%	0%	6%	7%	0%
ATM	7%	2%	1%	8%	17%	2%	2%	2%	7%	4%	2%	11%	2%	9%	10%	2%	0%	6%	5%	0%
Consommation nulle	4%	1%	1%	3%	51%	2%	1%	1%	5%	2%	1%	4%	2%	7%	3%	1%	0%	8%	3%	0%
Dentaire	6%	2%	1%	4%	23%	5%	2%	1%	5%	3%	2%	5%	2%	7%	7%	15%	0%	6%	4%	0%
Généraliste	6%	1%	1%	6%	15%	6%	9%	2%	9%	3%	3%	5%	2%	8%	8%	6%	1%	5%	6%	0%
Hospitalisation	8%	2%	1%	6%	15%	1%	3%	12%	5%	5%	4%	4%	2%	9%	7%	1%	1%	5%	8%	1%
Imagerie + Prophylaxie	6%	1%	2%	5%	21%	3%	2%	2%	8%	3%	5%	6%	2%	7%	8%	6%	0%	7%	6%	0%
Infirmiers	9%	2%	1%	5%	12%	1%	1%	3%	6%	13%	7%	4%	2%	8%	10%	1%	0%	6%	7%	0%
Kinésithérapie	5%	1%	1%	3%	8%	1%	2%	2%	8%	3%	38%	4%	2%	4%	7%	1%	1%	4%	4%	1%
Optique	7%	2%	1%	5%	28%	2%	2%	2%	7%	3%	3%	0%	3%	11%	10%	2%	1%	7%	5%	1%
Osthéopathie	6%	1%	1%	4%	23%	2%	1%	2%	7%	2%	2%	5%	20%	6%	7%	1%	0%	5%	4%	0%
Pharmacie 30-65	8%	1%	1%	5%	19%	1%	2%	2%	5%	3%	1%	5%	1%	25%	11%	1%	0%	6%	4%	0%
Pharmacie 65	8%	2%	1%	5%	12%	2%	2%	2%	5%	4%	3%	5%	2%	12%	22%	2%	0%	6%	5%	0%
Prothèses dentaires	6%	2%	1%	4%	24%	3%	2%	2%	7%	3%	2%	5%	2%	7%	8%	10%	0%	6%	4%	0%
Psy	5%	1%	0%	2%	5%	1%	1%	2%	2%	1%	1%	3%	1%	3%	4%	1%	59%	2%	7%	0%
Sous-consommateurs	7%	1%	1%	4%	28%	2%	1%	1%	7%	3%	2%	4%	2%	8%	9%	1%	0%	14%	4%	1%
Spécialiste	9%	2%	1%	6%	14%	1%	2%	4%	6%	4%	3%	8%	2%	9%	9%	2%	1%	5%	12%	1%
Visite généraliste	8%	2%	2%	4%	16%	2%	1%	2%	7%	2%	4%	4%	2%	7%	10%	1%	1%	5%	8%	11%

Figure g.1 – Matrice de transition des classes de risque

Pour interpréter ce tableau, voici un exemple :

La cellule située à l'intersection de la ligne « ATM » et de la colonne « Analyses » indique une valeur de 7 %. Cela signifie qu'un individu qui est classé dans la classe de risque « ATM » a une probabilité de 7 % d'être classé dans la classe de risque « Analyses » au trimestre suivant.