Mémoire présenté devant le jury de l'EURIA en vue de l'obtention du Diplôme d'Actuaire EURIA et de l'admission à l'Institut des Actuaires

le 26 Septembre 2019

Par : EHUI Aman-Yah Andréa

Titre : L'apport des objets connectés dans la définition d'un plan de prévention en assurance santé

Confidentialité : Non

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

Membres présents du jury de l'Institut des Actuaires :
Anthony Nahelou
Alain Moeglin
Dominique Abgrall
Signature :

Entreprise :
SIA PARTNERS
Signature :

Membres présents du jury de l'EURIA :
Brice Franke

Directeur de mémoire en entreprise :
Nicolas SERVAN
Signature :

Invité :

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 :

EURIA
EURo Institut
d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

T +33 (0)2 98 01 66 55
euria@univ-brest.fr

Résumé

La maîtrise des risques et l'optimisation des coûts constituent des problématiques auxquelles les assureurs en protection sociale sont confrontés depuis maintenant plusieurs années. Après avoir exploré les pistes immédiates (meilleure connaissance et anticipation des risques, maîtrise des frais), les assureurs orientent de plus en plus leurs plans stratégiques vers la prévention des risques auprès des assurés. Bien que cohérente, la mise en œuvre de cette démarche reste complexe.

La notion de prévention englobe l'ensemble des actions, attitudes et comportements qui tendent à limiter la survenance de maladies ou de traumatismes ainsi qu'à maintenir / améliorer la santé des individus. Pour ce faire, il est nécessaire de disposer de données nouvelles, de les capter et de les interpréter, afin de pouvoir notamment :

- identifier le type d'actions préventives à effectuer ;
- identifier les individus cibles dans des démarches préventives ;
- observer et mesurer l'impact d'un plan de prévention.

De telles données peuvent être mesurées quotidiennement à l'aide d'objets connectés. Se basant sur ces éléments, l'étude ici menée a pour objectif, en premier lieu, de créer des indicateurs de mode de vie permettant d'expliquer des comportements particuliers ou atypiques. Ces indicateurs viseront à classer les individus afin de cibler les actions à engager, puis à observer l'impact et l'efficacité d'une action de prévention qui serait mise en place.

En outre, le système de protection sociale français porté par la Sécurité Sociale et les assureurs souhaite passer d'un système curatif vers un système préventif. L'objectif de ce mémoire est ainsi de montrer dans quelle mesure les objets connectés peuvent être utilisés afin de définir un programme de prévention santé.

Mots clefs: Assurance santé, Prévention, Activité physique, Objets connectés, Classification, Rentabilité, Fonctions d'utilité, Régression logistique

Abstract

Risk control and cost optimization are issues that social protection insurers have been facing for several years now. After exploring immediate avenues (better knowledge and anticipation of risks, cost control), insurers are increasingly focusing their strategic plans on risk prevention for policyholders. Although consistent, the implementation of this approach remains complex.

The notion of prevention encompasses all actions, attitudes and behaviours that tend to limit the occurrence of diseases or injuries as well as to maintain / improve the health of individuals. To do this, it is necessary to have new data available, to capture and interpret them, in order to be able to

- identify the type of preventive actions to be taken ;
- identify target individuals in preventive approaches ;
- observe and measure the impact of a prevention plan.

Such data can be measured daily using connected objects. Based on these elements, the study conducted here aims, in the first instance, to create lifestyle indicators to explain specific or atypical behaviours. These indicators will aim to classify individuals in order to target the actions to be taken, and then to observe the impact and effectiveness of a preventive action that would be implemented.

In addition, the French social protection system supported by the Social Security and insurers wishes to move from a curative system to a preventive system. The objective of this thesis is to show to what extent connected objects can be used to define a health prevention program.

Keywords: Health insurance, Prevention, Physical activity, IoT, Categorization, Profitability, Utility function, Logistic regression

Remerciements

« Un seul mot, usé, mais qui brille comme une vieille pièce de monnaie : merci ! »
Pablo Neruda

La réalisation de ce mémoire a été possible grâce au concours de plusieurs personnes à qui je voudrais témoigner toute ma gratitude.

Je voudrais, avant tout, remercier mon directeur de mémoire M. Nicolas SERVAN, senior manager de l'UC Actuariat, pour sa patience, sa disponibilité et surtout ses judicieux conseils, qui ont contribué à alimenter ma réflexion. De plus, je souhaiterais témoigner toute ma reconnaissance à mes autres encadrants de mémoire, M. Baptiste ANDRIEU et M. Jordan MARIE-ROSE. Je les remercie de m'avoir encadrée, orientée, aidée et conseillée.

A cette véritable équipe, le seul mot merci ne saurait traduire toute ma reconnaissance à votre égard. Merci pour l'investissement ayant conduit à la réussite de ce mémoire, pour avoir toujours su répondre avec pédagogie et patience à mes différentes questions et pour tout le temps que vous m'avez accordé.

Je remercie M. Michaël DONIO, M. Ronan DAVIT et M. Benoît MENONI directeurs du service Actuariat, de m'avoir accueillie au sein de leur service et pour leurs conseils. Je tiens à témoigner toute ma reconnaissance à l'ensemble des consultants de l'UC Actuariat, pour leur disponibilité à toute épreuve.

Un merci s'adresse à Youssef JEBABLI et tout particulièrement à Sarah DIDO, mes collègues stagiaires, pour leur soutien inconditionnel, leurs encouragements, les fous rires et les pauses "pré-dej".

Je suis également reconnaissante envers les 206 personnes qui, à ce jour, ont pris le temps de répondre à mon formulaire de récolte de données.

Enfin, j'adresse mes plus sincères remerciements à Anaelle LEBERRE pour son suivi durant le stage, à l'ensemble des enseignants de l'EURIA, représenté par M. Franck VERMET, directeur de l'EURIA, de l'INSA, représenté M. James LEDOUX directeur du département Génie Mathématique de l'INSA Rennes, pour la qualité de leurs enseignements au cours de ces années de formation.

Je désire aussi remercier Damien TICHIT, Jean Paul GOMIS, Nicolas NGO et Romain LAILY pour avoir relu et corrigé mon mémoire. Leurs conseils de rédaction ont été très précieux.

Je n'oublie pas mes parents, pour leur soutien constant et leurs encouragements. « *Vous avez tout sacrifié pour vos enfants n'épargnant ni santé ni efforts. Vous m'avez donné un magnifique modèle de labeur et de persévérance. Je suis redevable d'une éducation dont je suis fière.*»

Note de Synthèse

Prévention de l'inactivité physique

Les assurances maladies obligatoire et complémentaire œuvrent pour une protection au quotidien des français contre les risques sociaux comme la maladie. La prise en charge de ces risques engendre des coûts qui nécessitent un financement adapté. Dans ce contexte, les organismes d'assurance doivent trouver, de concert, des leviers permettant de maîtriser, voire de limiter, les risques afin de pérenniser le financement de cette protection, et la rendre la plus complète possible.

La prévention – qui constitue un modèle « gagnant-gagnant » pour les assurés, les assureurs et les organismes sociaux – fait aujourd'hui partie des principaux axes de développement dans ce domaine. Elle correspond en effet à l'ensemble des dispositions permettant de préserver le capital santé des individus. Cependant, cette approche oblige à s'éloigner des modèles économiques et commerciaux traditionnels de l'assurance, qui s'appuient sur les principes de réparation et de remboursement, afin de s'orienter vers des approches de prévisibilité et de prévention, propices à engendrer des réductions ou suppressions du risque (phénomènes qui seront traités en termes de prévention primaire et secondaire). Afin de faire cohabiter les deux modèles, il est donc primordial d'être en mesure de comprendre et de mesurer l'impact, sur la sinistralité, généré par la mise en place d'un plan de prévention.

C'est ainsi cette problématique que nous allons étudier dans le cadre de ce mémoire, en se concentrant sur l'impact, à court terme, que peut avoir le manque d'activité physique sur la santé, car il s'agit en effet d'un facteur majeur de risque de décès dans le monde.

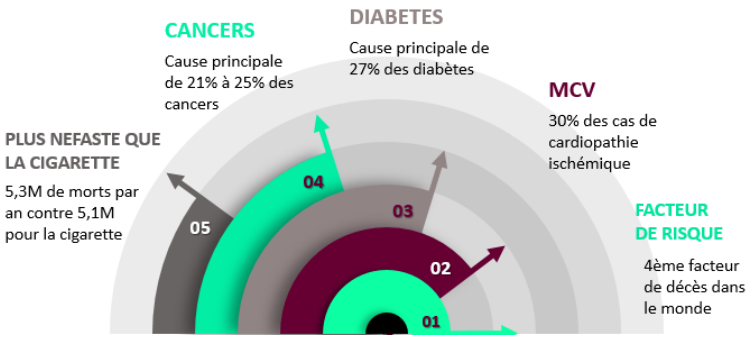


Figure ***1:*** Cinq faits sur l'inactivité physique selon l'OMS

Modèle théorique de la prévention

Un cadre théorique sur la prévention a été défini dans l'objectif de pouvoir en quantifier l'impact. Ce mémoire choisit d'étudier l'impact de la prévention sur l'espérance d'un gain perçu par l'assureur. Pour ce dernier, dans un contexte sans prévention, sur un contrat pour lequel il reçoit une prime w, il aura éventuellement à débourser une somme S_0 si l'individu est sinistré. Ce sinistre arrive avec une probabilité p_0. En introduisant un niveau de prévention e donné, un impact sera observé sur la probabilité d'être sinistré qui deviendra $p(e)$ (en **prévention primaire**) ou sur le coût du sinistre $S(e)$ (en **prévention secondaire**).

Situation	Probabilité d'occurrence	Gain si sinistre	Gain sans sinistre
Sans prévention	p_0	$w - S_0$	w
Prévention primaire	$p(e)$	$w - S_0 - c(e)$	$w - c(e)$
Prévention secondaire	p_0	$w - S(e) - c(e)$	$w - c(e)$

Table *1:* Récapitulatif des changements générés en fonction du type de prévention

Le calcul de l'espérance de gain est utilisé en prenant en compte un niveau de prévention e. Par ailleurs, la notion de fonction d'utilité est introduite afin de relever l'importance de la perception de l'individu sur ce qui est réellement gagné. Cette utilité va expliquer, par exemple, pourquoi pour la même quantité d'un bien, deux individus peuvent éprouver des satisfactions différentes. En notant cette fonction U, l'espérance de gain perçu pour un contrat est donnée par :

- en prévention primaire :

$$E[G]_U = p(e) \times U\{w - S_0 - c(e)\} + (1 - p(e)) \times U\{w - c(e)\}$$

- en prévention secondaire :

$$E[G]_U = p_0 \times U\{w - S(e) - c(e)\} + (1 - p_0) \times U\{w - c(e)\}$$

Les objets connectés en assurance

L'accès à l'information des objets connectés et leur utilisation pour appuyer l'assuré et guider ses décisions offrent un avantage certain. En effet, ces informations issues des objets connectés offrent la possibilité d'identifier le risque pour mieux l'atténuer. Ces objets ont une popularité et un impact sur la vie quotidienne qui ne cesse de croître. Ils s'implantent dans plusieurs secteurs d'activités et l'assurance n'y échappe pas. En assurance non-vie, des offres ont déjà vu le jour dans les domaines de l'assurance automobile et de l'habitation. Néanmoins, en santé, le contexte réglementaire français étant plus délicat, seule la prévention laisse entrevoir des possibilités.

Dans cette étude il a été entrepris de récolter des données issues de smartphones et de montres connectées afin d'illustrer, sur des données concrètes, des procédures d'**identification des risques** et de **mesure d'impacts de stratégie de prévention connectée**.

Identification des risques

À partir des données récoltées et issues d'objets connectés, plusieurs indicateurs de performances ont été construits afin d'établir des profils d'individus. Ce sont des indicateurs tels que un score sportif, un score de marche, un indicateur de stabilité journalière ou encore un indice de fréquence d'activité physique. Ceux-ci ont pour objectif de rendre compte des habitudes sportives des individus en matière de performance et de régularité.

Tout d'abord, une Analyse en Composante Principale - ACP - est réalisée afin de visualiser les individus dans un espace de projection. Ensuite, une classification non-supervisée est effectuée sur la base de ces indicateurs. La méthode des *kmeans* et la Classification Ascendante Hiérarchique - CAH - ont été mises en place. Puis, afin de décrire les groupes observés, deux approches ont été retenues :

- il a été d'abord entrepris de projeter les groupes retenus sur les axes de l'ACP ;
- ensuite, une classification supervisée par arbre de décision a été réalisée afin d'affiner cette description.

Figure *2:* Résultat de l'identification des risques par méthode de classification *kmeans*

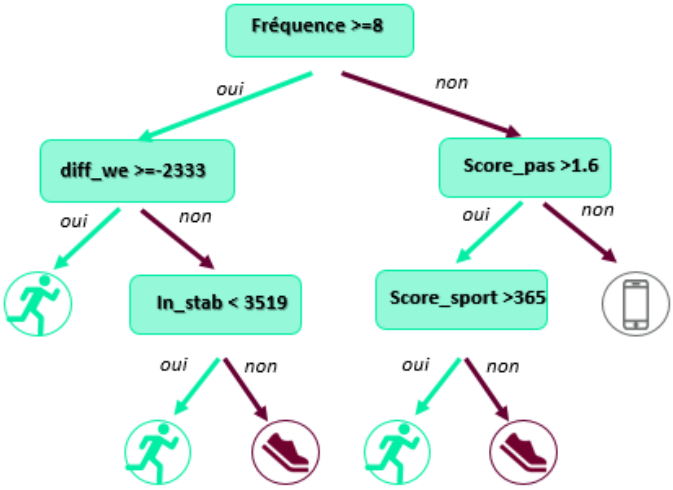


Figure *3:* Description des groupes obtenus par arbre de décision en fonction des indicateurs créés

Il a été décidé de retenir la classification issue des *kmeans* car, contrairement à la CAH, les groupes ainsi obtenus sont plus interpretables et plus fins. Ils sont catégorisés comme **Inactifs**, **Marcheurs** et **Sportifs**.

Impact de la prévention primaire

Dans l'intention de créer une base illustrative, une réplication des groupes sportifs est réalisée sur une base réelle de prestations santé. Cette réplication utilise un tirage aléatoire du groupe sportif, en fonction des proportions observées par type de profil, sur la base recueillie.
Seule l'information sur le nombre de consultations chez le généraliste est disponible. Ainsi, dans cette étude, la notion de sinistre est associée au sous-poste de soins que constituent les consultations généralistes . Afin de mettre en pratique la théorie de la prévention primaire, il est

important de pouvoir évaluer la probabilité d'être sinistré. Pour ce faire, un modèle de régression logistique permettant d'attribuer une probabilité à chacun des profils présents en portefeuille a été mis en place. Ce modèle, dont la qualité s'est révélée satisfaisante, a ainsi été utilisé pour réaliser des applications effectives dans le cadre de ce mémoire.

Des études de cas

Deux études de cas ont été mises en place. La première application pose un cadre dans lequel, pour deux entreprises identiques en matière de composition salariale, le gain obtenu suite à la mise en place d'une stratégie de prévention, devrait être du même ordre de grandeur. Le gain obtenu suite à une stratégie de prévention est plus important pour des individus « **Inactifs** ». Les proportions d'individus par groupe sportif n'étant pas les mêmes sur les deux entreprises, l'entreprise sur laquelle il y a moins d'« **Inactifs** » aura un gain moins important.

Ainsi, la simple connaissance des proportions par groupes sportifs permet d'anticiper les gains possibles suite à la mise en place d'un plan de prévention. La méconnaissance des groupes sportifs issus des objets connectés, empêche, dans ce cas ci, de correctement quantifier le gain obtenu grâce au type de prévention mis en place.

	Gains	Gains en % des primes	Inactifs	Marcheurs	Sportifs
Entreprise 1	**1097 €**	**0,64%**	**37,1%**	**30,4%**	**32,4%**
Entreprise 2	**822 €**	**0,46%**	**12,4%**	**42,8%**	**44,8%**

Table ***2:*** Gain réel sur les consultations chez le généraliste en fonction des proportions sportives

Dans la seconde application, il a été entrepris d'appliquer la théorie de la prévention primaire dans une étude de cas concrète. Celle-ci évalue la rentabilité perçue par deux assureurs sur deux types de prévention. Les deux stratégies de prévention sont basées sur des données issues des objets connectés. La première repose sur des objectifs fixés de nombre de pas annuel et la seconde sur des objectifs sportifs annuels.
Pour chacun des plans de prévention proposés, un coût et un impact en matière de changement de groupe sportif est défini. L'impact en matière de probabilité d'être sinistré est retranscrit au travers du modèle de régression logistique mentionné ci-avant.

Les fonctions d'utilité des deux assureurs ont été construites et ajustées par des fonctions de type racine carrée. Un profil témoin représentant la vision neutre ($U\{x\} = x$) a également été considéré. Ce profil permet d'apprécier la perception de l'assureur, par rapport à la réalité.

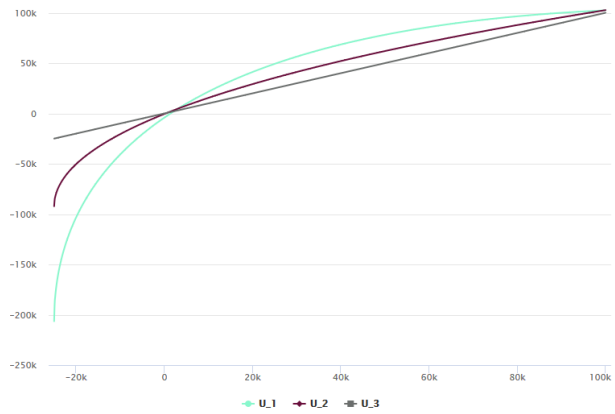


Figure *4:* Fonctions d'utilité considérées

La rentabilité est évaluée en tenant compte de la perception de chaque assureur. Cette évaluation qui s'appuie sur la théorie de la prévention est réalisée sur des niveaux de prévention définis. Il a été observé que la proportion d'individus qui change de classe constitue un indicateur de la performance et de la rentabilité de la stratégie de prévention. De plus, une incitation des individus à devenir « Sportifs » plutôt que « Marcheurs », engendre une plus grande rentabilité pour l'assureur.

Enfin, les outils théoriques présentés dans ce mémoire peuvent également être utilisés dans le cadre de la mise en place d'indicateurs d'aide à la décision. Ce sont :

- la **proportion seuil** : la proportion d'individus qui atteint les objectifs du niveau de prévention considérée, à partir de laquelle mettre en place des stratégies de prévention connectée est rentable ;
- le **coût fixe maximal - CFM** : le point d'équilibre qui permet d'obtenir l'information du coût fixe maximal à investir pour la mise en place de stratégie de prévention. Ce coût permet d'avoir au moins la même rentabilité que si la prévention n'avait pas eu lieu.

À partir de ces indicateurs, l'assureur dont la proportion seuil sera la plus basse et dont le CFM sera le plus élevé aura un avantage certain.

Limites et améliorations

Une des limites de cette étude réside dans l'utilisation d'une base illustrative. La non-exhaustivité des données empêche une généralisation des résultats de l'étude. Cependant, bien que les résultats ne soient pas transposables, la méthodologie est cohérente et reste elle transposable.

Plusieurs améliorations peuvent être apportées à cette étude. L'étude des impacts peut être étendue à d'autres sous-postes de soins. De plus, la prévention secondaire pourrait être appliquée en utilisant le modèle coût $\times$ fréquence utilisé en tarification non-vie, pour faire varier la dépense engendrée par un sinistre en fonction du groupe sportif. Enfin, il serait également possible d'étudier les interactions résultant de la mise en place de plans de prévention simultanés ou encore de modéliser les changements de groupes par des modèles multi-états.

Conclusion

Dans ce mémoire, seront présentées des méthodes de traitements et d'utilisation des données issues des objets connectés. Ces données ont eu un réel apport dans l'identification du risque que constitue l'inactivité physique. Appuyée par la théorie de la prévention, cette identification a permis de mesurer l'impact de stratégies de prévention sur une base illustrative, en estimant la rentabilité *a priori* de celles-ci pour un assureur.

Preventing physical inactivity

Compulsory and supplementary health insurance provide daily protection for French people against social risks such as illness. The management of these risks generates costs that require appropriate financing. In this context, insurance organisations must work together to find levers to control or even limit risks in order to ensure that the financing of this protection is sustainable and as complete as possible.

Prevention - which is a "win-win" model for policyholders, insurers and social agencies - is now one of the main areas of development in this field. It corresponds to all the provisions making it possible to preserve the health capital of individuals. However, this approach requires a shift away from traditional economic and business models of insurance, which are based on the principles of repair and reimbursement, towards approaches of predictability and prevention that are conducive to risk reduction or elimination (phenomena that will be addressed in terms of primary and secondary prevention). In order for the two models to coexist, it is therefore essential to be able to understand and measure the impact on the loss experience generated by the implementation of a prevention plan.

This is the issue that we will study in this paper, focusing on the short-term impact that lack of physical activity can have on health, as it is a major risk factor for death worldwide.

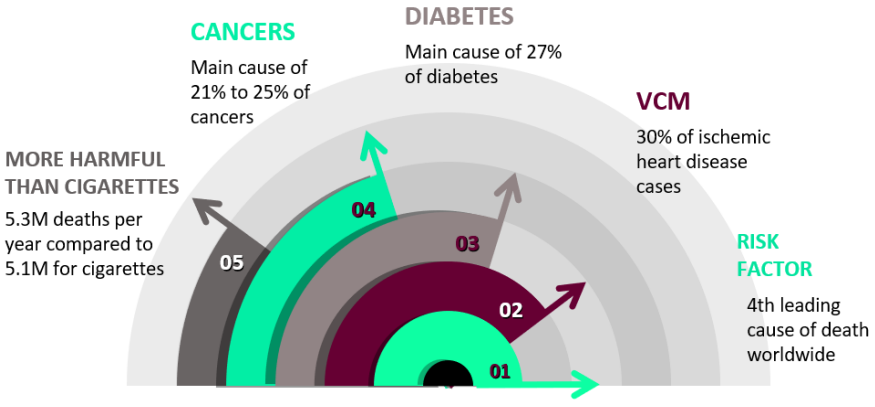


Figure ***5:*** Five facts on physical inactivity according to WHO

Theoretical model of prevention

A theoretical framework on prevention has been defined in order to quantify its impact. This thesis chooses to study the impact of prevention on the expectation of a gain received by the insurer. For the latter, in a context without prevention, on a contract for which he receives a w premium, he will eventually have to pay an amount of S_0 if the individual is a victim. This loss occurs with a probability of p_0. By introducing a given e level of prevention, an impact will be observed on the probability of being affected which will become $p(e)$ (in **primary prevention**) or on the cost of the loss $S(e)$ (in **secondary prevention**).

Situation	Probability of occurrence	Gain if claim	Claim-free profit
Without prevention	p_0	$w - S_0$	w
Primary prevention	$p(e)$	$w - S_0 - c(e)$	$w - c(e)$
Secondary prevention	p_0	$w - S(e) - c(e)$	$w - c(e)$

Table *3:* Summary of changes generated by type of prevention

The calculation of the gain expectation is used taking into account a e prevention level. In addition, the notion of utility function is introduced in order to highlight the importance of the individual's perception of what is actually gained. This utility will explain, for example, why for the same quantity of a good, two individuals may experience different satisfactions. By noting this U function, the expectation of gain received for a contract is given by :

- in primary prevention :

$$E[G]_U = p(e) \times U\{w - S_0 - c(e)\} + (1 - p(e)) \times U\{w - c(e)\$$

- in secondary prevention :

$$E[G]_U = p_0 \times U\{w - S(e) - c(e)\} + (1 - p_0) \times U\{w - c(e)\}$$

Objects connected in insurance

Access to the information of the connected objects and their use to support the insured and guide his decisions offer a definite advantage. Indeed, this information from connected objects offers the possibility of identifying the risk in order to better mitigate it. These objects have a popularity and an impact on daily life that continues to grow. They are setting up in several sectors of activity and insurance is no exception. In non-life insurance, offers have already been made in the areas of automobile and home insurance. Nevertheless, in health, as the French regulatory context is more delicate, only prevention can suggest possibilities.

In this study, it was undertaken to collect data from smartphones and connected watches in order to illustrate, on concrete data, procedures for **risk identification** and **impact measurement of connected prevention strategy**.

Risk identification

From the data collected and derived from connected objects, several performance indicators were constructed to establish profiles of individuals. These are indicators such as a sports score, a walking score, a daily stability indicator or a physical activity frequency index. The objective of these reports is to report on the performance and regularity of individuals' sporting habits.

First, a Main Component Analysis - MCA - is performed to visualize individuals in a projection space. Then, an unsupervised classification is carried out on the basis of these indicators. The *kmeans* method and the Ascending Hierarchical Classification - CAH - have been implemented. Then, in order to describe the observed groups, two approaches were chosen :

- it was first undertaken to project the selected groups on the CPA axes ;
- then, a supervised classification by decision tree was carried out in order to refine this description.

Figure *6:* Result of risk identification by classification method *kmeans*

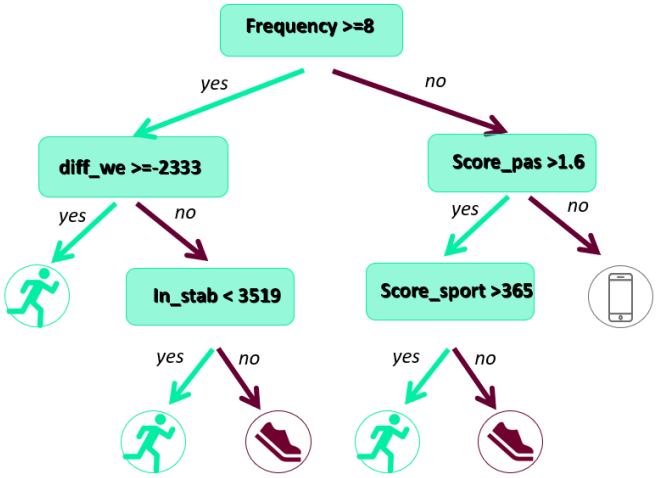


Figure *7:* Description of the groups obtained by decision tree according to the indicators created

It was decided to use the classification from the *kmeans* because, unlike the CAH, the groups thus obtained are more interpretable and refined. They are categorized as **Inactive**, **Walkers** and **Athletic**.

Impact of primary prevention

In order to create an illustrative database, a replication of sports groups is carried out on a real basis of health services. This replication uses a random draw of the sport group, based on the proportions observed by profile type, based on the collected data.

Only information on the number of consultations with the general practitioner is available. Thus, in this study, the notion of claim is associated with the sub-station of care that constitutes general consultations. In order to put primary prevention theory into practice, it is important to be able to assess the probability of being affected. To do this, a logistic regression model was implemented to assign a probability to each of the profiles in the portfolio. This model, the

quality of which proved to be satisfactory, was thus used to make effective applications in the context of this thesis.

Case studies

Two case studies have been implemented. The first application sets out a framework in which, for two identical companies in terms of wage composition, the gain obtained following the implementation of a prevention strategy should be of the same order of magnitude. The gain from a prevention strategy is greater for individuals who are « **Inactive** ». Since the proportions of individuals per sports group are not the same for both companies, the company with fewer « **Inactive** » will have a smaller gain.

Thus, simply knowing the proportions by sport group makes it possible to anticipate the possible gains following the implementation of a prevention plan. The lack of knowledge of the sports groups from the connected objects prevents, in this case, to correctly quantify the gain obtained thanks to the type of prevention implemented.

	Gains	Gain in % of premiums	Inactives	Walkers	Athletic
Company 1	1097 €	0,64%	37,1%	30,4%	32,4%
Company 2	822 €	0,46%	12,4%	42,8%	44,8%

Table *4:* Real gain on general practitioner consultations according to sports proportions

In the second application, it was undertaken to apply the theory of primary prevention in a concrete case study. It evaluates the perceived profitability of two insurers on two types of prevention. Both prevention strategies are based on data from connected objects. The first is based on objectives set in terms of the number of annual steps and the second on annual sporting objectives.
For each of the proposed prevention plans, a cost and impact of changing sport groups is defined. The impact in terms of the probability of being affected is transcribed through the logistic regression model mentioned above.

The utility functions of the two insurers were constructed and adjusted by square root functions. A control profile representing neutral vision ($U\{x\} = x$) was also considered. This profile makes it possible to assess the insurer's perception in relation to reality.

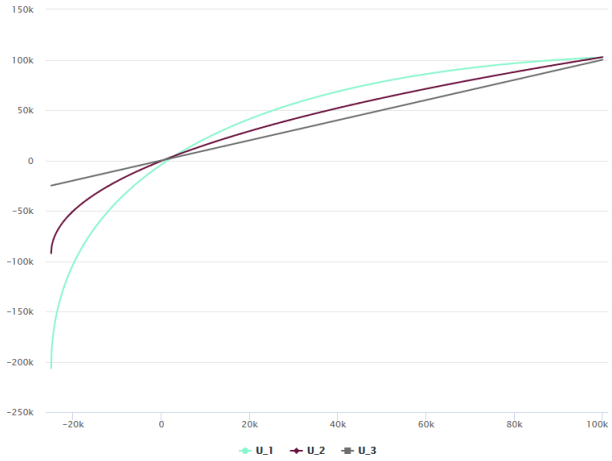


Figure *8:* Utility functions considered

Profitability is assessed by taking into account the perception of each insurer. This evaluation, which is based on prevention theory, is carried out at defined levels of prevention. It has been observed that the proportion of individuals who change classes is an indicator of the performance and profitability of the prevention strategy. In addition, an incentive for individuals to become « Athletics » rather than « Walkers », results in greater profitability for the insurer.

Finally, the theoretical tools presented in this paper can also be used in the context of the implementation of decision support indicators. These are :

- the **threshold proportion** : the proportion of individuals who achieve the objectives of the level of prevention considered, from which to set up connected prevention strategies is profitable ;
- the **maximum fixed cost - MFC** : the equilibrium point that provides information on the maximum fixed cost to be invested for the implementation of a prevention strategy. This cost makes it possible to have at least the same profitability as if prevention had not taken place.

Based on these indicators, the insurer with the lowest threshold proportion and the highest MFC will have a definite advantage.

Limits and improvements

One of the limitations of this study is the use of an illustrative base. The lack of completeness of the data prevents a generalization of the study results. However, although the results cannot be transposed, the methodology is consistent and remains transposable.

Several improvements can be made to this study. The impact study can be extended to other sub-stations of care. In addition, secondary prevention could be applied using the $\times$ frequency cost model used in non-life underwriting, to vary the expense generated by a loss according to the sport group. Finally, it would also be possible to study the interactions resulting from the implementation of simultaneous prevention plans or to model group changes using multi-state models.

Conclusion

In this thesis, methods of processing and using data from connected objects will be presented. These data have made a real contribution to identifying the risk of physical inactivity. Supported by prevention theory, this identification made it possible to measure the impact of prevention strategies on an illustrative basis, by estimating their profitability *a priori* for an insurer.

Table des matières

Introduction

Vitality, un programme de prévention santé créée par Generali, voit le jour le 1er janvier 2017. Il introduit une stratégie de prévention inédite en France, avec les objets connectés au cœur de la promotion d'une bonne hygiène de vie. Bien que l'usage des objets connectés en assurance soit déjà une pratique courante dans de nombreux pays, en France, la réglementation en assurance santé ne laisse une opportunité qu'en prévention.

Dans un contexte d'assurance, deux types de prévention sont présentés dans ce mémoire. La prévention primaire qui réduit la probabilité d'avoir un sinistre et la prévention secondaire qui réduit les coûts associés aux sinistres. Dans le cadre de cette étude, une modélisation de la prévention primaire est réalisée. Elle permet d'apporter un cadre d'application assurantiel aux objets connectés.

L'objectif de ce mémoire est de développer une méthode de quantification de l'impact, à court terme, de stratégies de prévention, basées sur les objets connectés. Les plans de préventions considérés dans cette étude concernent l'inactivité physique. Cette dernière constitue l'un des facteurs de risque les plus importants en santé.

Ce mémoire est structuré en quatre parties. Avant tout, le contexte de l'étude soulignant les notions d'assurance maladie, de prévention de l'inactivité physique et d'enjeux des objets connectés dans l'assurance sera présenté. Un cadre théorique de la prévention sera ensuite défini. En outre, des modélisations sont effectuées, d'une part dans l'intention d'identifier le risque lié à l'inactivité physique grâce aux informations recueillies à partir des objets connectés et, d'autre part, elles permettront d'évaluer l'impact d'une stratégie de prévention connectée. Enfin, des études de cas seront utilisées afin de montrer l'apport des objets connectés dans l'évaluation *a priori* d'une stratégie de prévention connectée.

1 Contexte de l'étude

« Qui prévient le moment l'empêche d'arriver ;
qui le laisse échapper ne peut le retrouver.» ***André Chénier***

Ce chapitre introductif a pour objectif de présenter le contexte de prévention dans lequel se place cette étude. Il permettra au lecteur de se familiariser d'une part avec les notions de **protection sociale**, d'**Assurance Maladie** et de **prévention**. D'autre part, il sera question de détailler la **notion d'activité physique** afin de comprendre l'importance de sa promotion dans le cadre de la **prévention santé**.

1.1 Protection sociale et Assurance Maladie

La protection sociale désigne tous les mécanismes de prévoyance permettant aux individus de faire face aux conséquences financières des risques sociaux pouvant entraîner l'accroissement des dépenses (ex : hospitalisation) ou la diminution des ressources (ex : arrêt de travail). Il s'agit de situations susceptibles de compromettre la sécurité économique de l'individu ou de sa famille, en provoquant une baisse de ses ressources ou une hausse de ses dépenses. Le concept de protection sociale recouvre les notions d'assurance et de solidarité et permet de se prémunir contre tous les aléas de la vie liés à la personne, qu'ils soient d'origine privée ou professionnelle.

1.1.1 Généralités de la protection sociale en France

1.1.1.1 Acteurs de la protection sociale en France

La France dispose d'un système de protection sociale très complet qui associe de nombreux acteurs publics et privés. Les principaux acteurs du marché de la protection sociale en France sont la Sécurité Sociale, les assureurs, les mutuelles et les institutions de prévoyance.

La Sécurité Sociale

La Sécurité Sociale est un organisme parapublic axé sur le service public et relevant du Code de la Sécurité Sociale. Elle est constituée de deux principaux régimes :

1. le régime général, géré par l'Assurance Maladie, assureur solidaire de quatre personnes sur cinq en France ;
2. le régime agricole, géré par la caisse centrale de la Mutualité Sociale Agricole (MSA), couvre les exploitants et les salariés agricoles.

Il existe également d'autres régimes dits spéciaux comme, par exemple, le régime de la SNCF. Les régimes de la Sécurité Sociale fournissent une couverture obligatoire. Ils sont financés par des cotisations retenues sur le salaire, des impôts et des taxes. Les prestations peuvent être assurées par des services publics, des assurances ou des services privés.

Les sociétés d'assurance

Les sociétés d'assurance se présentent sous plusieurs formes juridiques. Elles peuvent être des sociétés anonymes à but lucratif ou des sociétés d'assurance mutuelles à but non lucratif. Juridiquement, une société d'assurance se caractérise par le fait qu'elle est régie par le Code des Assurances.

Les mutuelles

Selon le Code de la mutualité, les mutuelles sont des personnes morales de droit privé à but non lucratif. Soumises au Code de la mutualité, elles mènent, notamment au moyen des cotisations versées par leurs membres nommés "sociétaires", et dans l'intérêt de ces derniers et de leurs ayants droit, des actions de prévoyance, de solidarité et d'entraide, dans les conditions prévues par leurs statuts.

Les institutions de prévoyance

Les institutions de prévoyance -IP-, sont des organismes paritaires à but non lucratif exerçant uniquement dans le champ des risques sociaux. Elles sont juridiquement régies par le Code de la Sécurité Sociale et sont dirigées par un conseil d'administration qui est constitué à parts égales de représentants de salariés et de représentants d'employeurs. Elles sont historiquement créées par des entreprises, groupes d'entreprises ou branches professionnelles.

1.1.1.2 Branches de la protection sociale

Les risques sociaux couverts par la Sécurité Sociale sont :

- le risque **vieillesse**, comprenant par exemple les retraites ;
- le risque **famille**, comprenant par exemple la maternité ;
- le risque **emploi** qui comprend le chômage, l'insertion et la réinsertion professionnelle ;
- le risque **logement** qui comprend les allocations de logement ;
- le risque **pauvreté et exclusion sociale** qui comprend les prestations diverses de l'assistance sociale en faveur des personnes démunies.

Même s'il ne constitue pas une branche, le risque de **dépendance** qui correspond à la prise en charge des personnes âgées ou handicapées en situation de dépendance [1] peut tout de même être évoqué.

1. État d'une personne qui a besoin d'être aidée pour l'accomplissement des actes de la vie quotidienne ou qui nécessite une surveillance régulière.

Selon le risque social couvert, les parts de marché des différents acteurs de la protection sociale se présentent comme sur la figure 1.1.

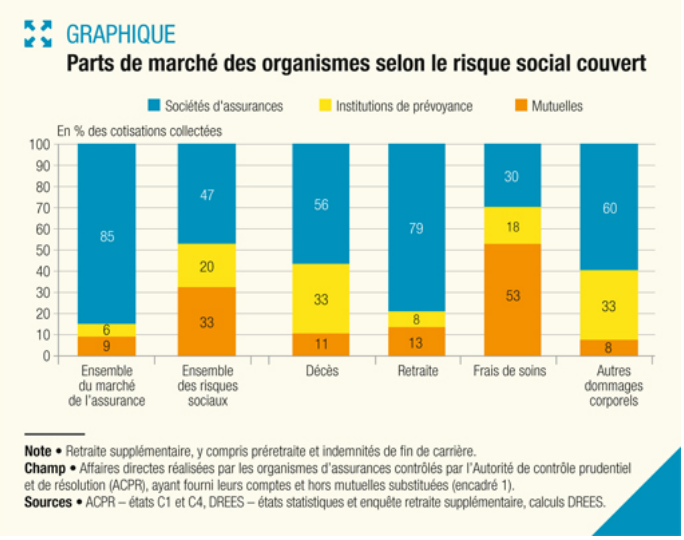


Figure *1.1:* Part de marché des organismes selon le risque social couvert

Sur l'ensemble du marché de l'assurance, les sociétés d'assurance occupent une place majoritaire. Cependant, les autres acteurs sont plus spécialisés dans certains risques sociaux.

1.1.1.3 Types de contrats

Dans une première classification, les contrats collectifs sont distingués des contrats individuels.

Les contrats **collectifs** permettent à une entreprise de faire bénéficier, à ses salariés, d'une couverture de protection sociale. Cette protection complète le régime obligatoire de la Sécurité Sociale et garantit le niveau de vie des assurés et de leur famille en cas de réalisation d'un évènement imprévu impactant leurs revenus. Les cotisations sont réparties entre l'entreprise et le salarié. Le contrat sera constitué sans discrimination d'âge, de rémunération, d'état de santé ou encore de type de contrat de travail. C'est en cela que réside l'avantage d'une assurance prévoyance collective puisqu'elle permet la mutualisation des risques matérialisée par une baisse des tarifs. L'inconvénient majeur de ce type de contrat est le manque de flexibilité car les garanties sont imposées aux salariés.

Les contrats **individuels** permettent aux assurés d'effectuer leurs propres choix de niveaux de garantie et de cotisations. Ce choix impacte les niveaux de remboursement et permet de moduler le lot d'options sur différents postes comme l'hospitalisation, le dentaire, etc. Cependant, la souscription à un contrat individuel met l'assuré face à une sélection plus rigoureuse car, l'âge et l'état de santé au moment de la souscription influent fortement sur l'acceptation du dossier ou, tout du moins, sur le montant des cotisations. De plus, l'effet de mutualisation étant moins présent, cela engendre une hausse des tarifs.

Tous les acteurs du marché de la protection sociale possèdent des contrats collectifs et des contrats individuels.

Figure *1.2:* Part des contrats individuels et collectifs chez les principaux acteurs de la protection sociale en 2013

De manière globale, sur l'ensemble du marché, les contrats collectifs et individuels sont quasiment en proportion égale. Toutefois, les mutuelles et les institutions de prévoyance arborent respectivement des spécificités dans l'individuel et dans le collectif.

1.1.2 Assurance Maladie

L'Assurance Maladie est une organisation chargée d'assurer les français face à des risques financiers de soins en cas de maladie [21], ainsi qu'un revenu minimal lorsque l'affection prive la personne de travail. Son fonctionnement est basé sur la mutualisation du risque. En effet, chaque personne cotise, en échange de quoi, lorsque survient la maladie, elle est remboursée selon un barème défini.
Elle repose sur trois principes fondamentaux : l'égalité d'accès aux soins, la qualité des soins et la solidarité.

1.1.2.1 Assurance Maladie obligatoire

Le principal rôle de l'Assurance Maladie obligatoire est d'assurer la prise en charge des dépenses de santé des assurés et de garantir l'accès aux soins pour tous.
Le système de remboursement des soins de l'Assurance Maladie obligatoire fonctionne sur les notions suivantes :

- la **dépense engagée**, qui correspond au montant effectivement dépensé par l'assuré pour l'accès au soin ;
- la **base de remboursement** de la Sécurité Sociale - BRSS - qui correspond au tarif conventionnel servant de base au calcul de la somme qui sera remboursée après un acte médical ;
- la **participation forfaitaire** - PF - qui correspond à un montant fixe de 1€ restant à la charge de l'assuré ;
- le **ticket modérateur** - TM - est une partie de la base de remboursement qui reste à la charge de l'assuré après le remboursement de l'Assurance Maladie et la participation forfaitaire. Il est défini comme un pourcentage de la BRSS et dépend du type de dépense

et du respect de certaines règles telles que disposer d'une ordonnance ou encore suivre le parcours de soins coordonnés ;

- le **reste à charge** - RAC - qui correspond à la part de la dépense engagée qui reste à la charge de l'assuré.

La figure 1.3 permet de présenter l'ensemble des points évoqués précédemment.

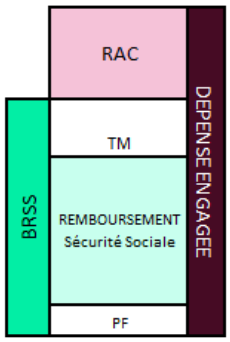


Figure *1.3:* Le remboursement des soins de santé par le régime obligatoire

1.1.2.2 Assurance Maladie complémentaire

Une complémentaire santé permet de compléter la prise en charge de l'Assurance Maladie obligatoire suite à des dépenses pour des soins. En effet, la Sécurité Sociale, par le biais de l'Assurance Maladie obligatoire comme vu précédemment, ne rembourse qu'une partie des dépenses engagées. Ce remboursement peut se montrer, dans certains cas, peu significatif pour l'assuré, lui laissant alors un reste à charge souvent très important. Le niveau de remboursement de la complémentaire dépend du contrat souscrit par l'assuré. La complémentaire a un rôle très important dans le cas de dépenses importantes très peu remboursées par la Sécurité Sociale comme pour les soins dentaires, optiques ou d'hospitalisation.

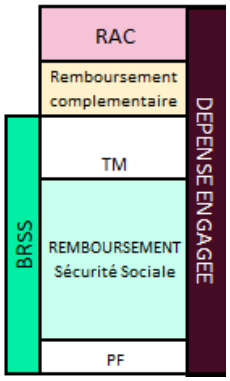


Figure *1.4:* Le remboursement des soins de santé avec une complémentaire

Sur le marché de la complémentaire santé, les assureurs, les mutuelles et les institutions de prévoyance se partagent le marché de l'Assurance Maladie complémentaire tel que présenté dans la figure 1.5.

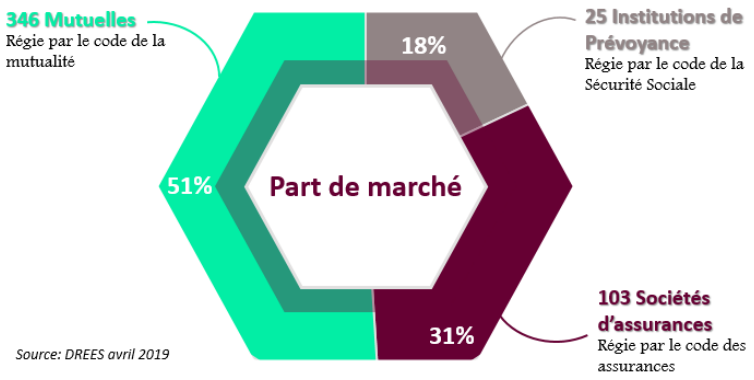


Figure ***1.5:*** Les catégories d'organismes complémentaires d'Assurance Maladie

1.1.2.3 Dépenses en santé

En Assurance Maladie, les dépenses totales par catégorie de maladie sont les suivantes :

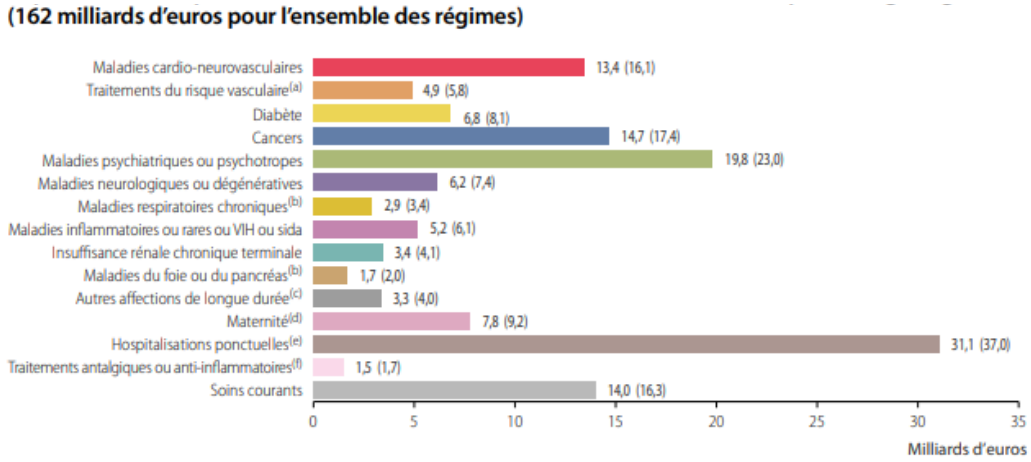


Figure ***1.6:*** La répartition des dépenses de l'Assurance Maladie en 2016 en France selon le rapport des charges et produits de l'Assurance Maladie 2019

Au total, l'Assurance Maladie représente 176,4 milliards d'euros de dépenses engagées en 2018 pour un taux moyen de remboursement de la Sécurité Sociale de 84.72% en 2018. Dans le contexte réglementaire de Solvabilité II, les notions de pilotage des risques, de rentabilité et de pérennité des activités sont encore plus importantes. Ces notions donnent une grande importance à la gestion des coûts. Celle-ci constitue une problématique importante de l'assurance maladie. L'anticipation et la réduction des coûts sont des objectifs majeurs pour tous les acteurs de l'assurance maladie, qu'ils soient privés ou publics. Le succès de la maîtrise des coûts est conditionné par plusieurs facteurs, dont la stratégie de prévention. La notion de prévention en santé sera donc explicitée dans la suite de cette étude.

1.2 Prévention santé

1.2.1 Qu'est ce que la prévention ?

La prévention des risques correspond à l'ensemble des dispositions à mettre en œuvre pour préserver la santé et la sécurité des individus. La réduction des risques par l'accompagnement au quotidien est, aujourd'hui, un des principaux objectifs à court terme pour les compagnies d'assurance. La prévention est un bon outil de réduction des risques et elle fait partie intégrante du métier d'assureur [12]. A ce titre, la prévention peut aussi être définie comme l'ensemble des mesures coûteuses qui sont prises pour réduire le risque auquel est confronté l'assureur. Cependant, la quantification de l'impact de ces mesures préventives reste délicate. En effet, pour cela, un assureur devrait pouvoir mesurer, même sur les risques longs, la variation de sinistralité directement liée à la prévention effectuée. Il doit aussi être en mesure d'exclure l'ensemble des autres facteurs pouvant influer sur cette écart de sinistralité.

La prévention reprend les notions de réduction du risque ou encore de suppression du risque à la source. Deux types de prévention peuvent être identifiées [2].

- **La prévention primaire** correspond à une réduction des probabilités de sinistre. Cela s'exprime aussi souvent par le terme *atténuation*.
- **La prévention secondaire** correspond à la réduction des montants des dommages. Cela s'exprime aussi souvent par le terme *adaptation*.

1.2.2 Prévention santé actuelle

La prévention santé se présente sous plusieurs formes. Qu'elle soit primaire ou secondaire, elle vise à réduire les coûts liés aux pathologies qu'elle permet d'amoindrir ou d'éviter. De plus, la prévention santé actuelle passe aussi par l'information des assurés sur les risques auxquels ils sont soumis. Les assurés sont poussés à agir de manière plus responsable par des politiques de sensibilisation. A titre d'exemple, le dépistage et la vaccination peuvent être cités.

1.2.2.1 Dépistage

Le dépistage est une démarche visant à diagnostiquer tôt certaines maladies. Cette action permet de vérifier la présence, ou non, de signes précurseurs de maladie. Cela correspond par exemple à des examens médicaux effectués en général chez des personnes semblant être en bonne santé. Le dépistage permet, dans certains cas, de reculer l'apparition de la maladie ou d'en amoindrir les impacts.
Dans le cas des cancers par exemple, cela permet souvent de les déceler dans une forme peu avancée et, par conséquent, augmenter les chances de guérison. Le dépistage constitue donc une mesure de **prévention secondaire** car un diagnostic précoce permet de réduire l'impact que pourrait avoir une maladie.

2. La prévention tertiaire, dont le but est de diminuer la prévalence des incapacités chroniques ou des récidives dans une population et de réduire les complications, les invalidités ou les rechutes consécutives à une maladie ne sera pas étudiée ici. L'inclure dans une étude nécessiterait d'être en mesure de capter les conséquences à long terme des mesures de prévention.

Le dépistage peut être prescrit par un médecin et être :

- gratuit comme, par exemple, dans le cas du cancer du sein ou dans le cas du SIDA s'il est réalisé dans un CEGIDD [3] ;
- remboursé en partie par la Sécurité Sociale comme dans le cas du cancer colorectal ;
- ou complètement à la charge des assurés comme le bilan endocrinien [4] s'il est fait sans ordonnance.

1.2.2.2 Vaccination

La vaccination est une action qui consiste à immuniser un individu contre une maladie infectieuse en lui administrant un vaccin. Ce dernier va stimuler le système immunitaire et permettre ainsi de combattre et d'éliminer des maladies infectieuses potentiellement mortelles. C'est l'un des investissements les plus rentables dans le domaine de la santé. De plus, la vaccination constitue une mesure de **prévention primaire** car elle va réduire la probabilité de tomber malade. L'organisation mondiale de la santé - OMS - estime qu'ainsi plus de 2 à 3 millions de décès par an sont évités dans le monde.
Le caractère obligatoire ou recommandé de certains vaccins n'influe pas sur leur remboursement par la Sécurité Sociale. Ils peuvent être en partie, totalement ou pas du tout remboursés par la Sécurité Sociale.

1.2.2.3 Consultations spécialisées

Plusieurs types de consultations chez des médecins spécialisés peuvent constituer des mesures préventives. Par exemple, les consultations chez des diététiciens ou des nutritionnistes en font partie. Ces consultations permettent d'effectuer un état du comportement alimentaire. Cela permet par la suite d'élaborer un plan d'amélioration de la nutrition et recevoir des conseils diététiques avec un suivi plus ou moins régulier. Cela constitue une mesure de **prévention primaire** puisque cette action préventive permet d'améliorer l'hygiène de vie afin d'éviter toutes les maladies liées à la mauvaise alimentation.

1.2.3 Un concept qui s'inscrit dans une orientation politique

Au cours du comité interministériel pour la santé, du 26 mars 2018, le premier ministre français Édouard Philippe, affirme que *"la prévention doit devenir centrale dans toutes les actions qui visent à améliorer la santé de nos concitoyens"*. Partant de cette volonté, l'ensemble du Gouvernement *"s'engage pour que la prévention ne soit plus seulement un concept mais une réalité (. . .) avec une obsession : celle de l'efficacité et des résultats concrets"*. Plusieurs initiatives gouvernementales rentrent dans le cadre de cette stratégie de développement de la prévention[10]. Dans le domaine de la santé, des initiatives comme l'ANI ou la reforme 100% santé s'inscrivent dans cette démarche.

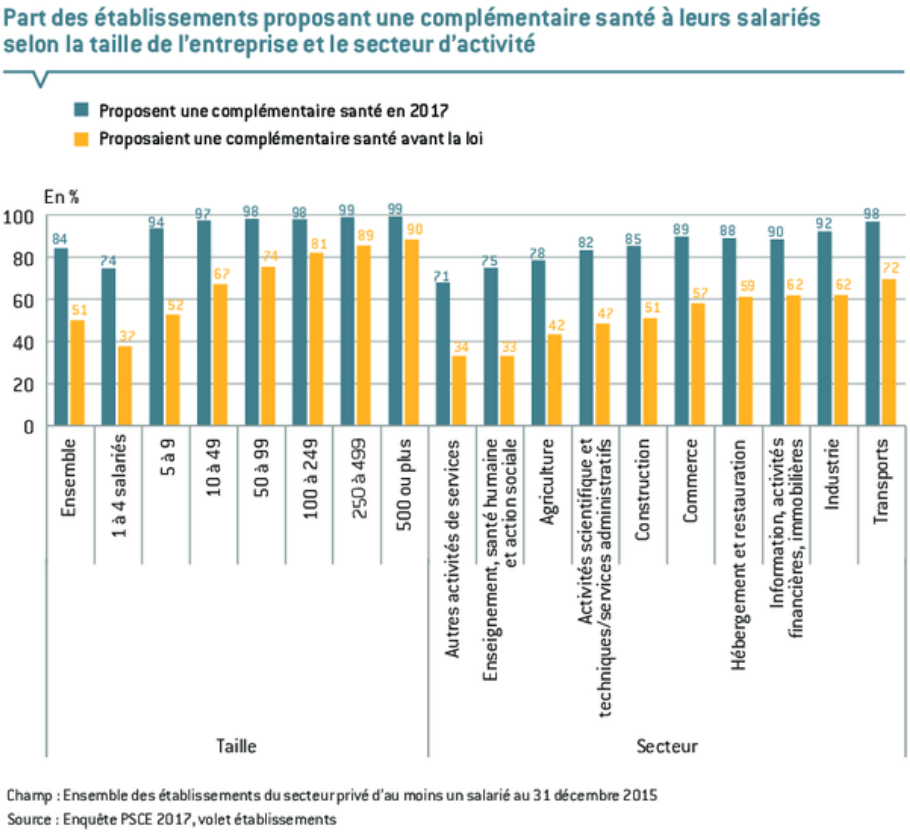
3. Centre Gratuit d'Information, de Dépistage et de Diagnostic des infections par les virus de l'immunodéficience humaine, des hépatites virales et des infections sexuellement transmissibles.

4. Examen biologique qui consiste à doser dans le sang ou les urines, des hormones. L'objectif est le dépistage de certaines atteintes hormonales responsables de maladies ou de troubles plus ou moins importants.

1.2.3.1 Accord National Interprofessionnel

L'Accord National Interprofessionnel - ANI - est un accord négocié et signé par les différents partenaires sociaux au niveau national et qui s'applique à l'ensemble des secteurs d'activités sur le territoire national. Il traite des sujets tels que les conditions de travail et les garanties sociales des salariés. Ces accords sont intégrés dans la législation par le biais d'ordonnances publiées au Journal Officiel.

Le dernier, datant du 11 janvier 2013, comprend comme parties prenantes l'ensemble des organisations patronales et les confédérations syndicales de salariés. Cet accord touche tous les secteurs d'activités et prévoit de modifier les droits sociaux des salariés et des employeurs. Entré en vigueur le 1er janvier 2016, cet accord contient notamment la mise en place obligatoire d'une complémentaire santé d'entreprise. Cette complémentaire serait en partie financée par l'employeur (au moins 50%), quelle que soit la taille de l'entreprise. Elle prévoit aussi des *minima* de couvertures imposés par l'ANI ou par les conventions collectives des secteurs d'activité des entreprises. Comme le montre la figure 1.7, depuis sa mise en place, le nombre d'établissements proposant une complémentaire santé à leurs salariés a augmenté, et cela, quelque soit le secteur d'activité. Cette augmentation est surtout visible pour les petites entreprises, les grosses entreprises étant, pour la plupart, déjà équipées.



Champ : Ensemble des établissements du secteur privé d'au moins un salarié au 31 décembre 2015
Source : Enquête PSCE 2017, volet établissements

Figure *1.7:* Évolution du nombre d'entreprises proposant une complémentaire santé depuis l'ANI

La mise en place de l'ANI permet aux assureurs de concevoir des politiques de prévention plus aisément. Puisque 2% au moins des cotisations seront consacrées à la solidarité, le service de

prévention peut être gratuit et inclus dans certains contrats par un élargissement de la gamme de produits d'assurance complémentaire santé avec une offre de programmes de prévention en lien avec la santé. C'est en cela que cet accord est perçu, par certains acteurs du marché de l'assurance, comme un levier de différenciation et de fidélisation important.

1.2.3.2 100% Santé

Cette réforme du domaine de la santé cherche à réduire, voire annuler, le reste à charge des français lors de dépenses en santé [28], spécifiquement des dépenses liées aux domaines optique, dentaire et auditif. Partant du constat d'une renonciation aux soins importante pour cause de faibles moyens financiers, cette réforme a pour objectif l'accès aux soins pour tous. Elle s'inscrit dans la stratégie globale de prévention de l'État puisque diverses pratiques préventives telles que les bilans de santé seront favorisées (remboursement à 100%). De plus, l'amélioration de l'accès aux soins des français permet aux pouvoirs publics de soutenir la prise de conscience des assurés.

Plusieurs facteurs de risques sont activement ciblés dans le cadre de la prévention. Ce sont des facteurs qui vont influencer l'état de santé de la population. Ainsi, l'objectif des pouvoirs publics et privés est d'améliorer la connaissance et la quantification de ces facteurs de risques afin de mieux les combattre et éventuellement les supprimer. Trois types de facteurs sont identifiables [32] :

- les facteurs **endogènes** tels que l'âge, le sexe, l'hérédité ;
- les facteurs **exogènes** tels que la pollution, le niveau de revenu ;
- les facteurs **comportementaux** tels que fumer, faire du sport ou encore l'alimentation.

Les facteurs endogènes et exogènes n'étant pas contrôlables par l'individu, une stratégie de prévention ne peut être mise en place que sur les facteurs comportementaux, qui eux peuvent varier plus simplement. Par conséquent, l'étude développée dans le cadre de ce mémoire se focalisera sur le facteur de risque comportemental que constitue l'inactivité physique.

1.3 Importance de la promotion de l'activité physique

1.3.1 Définitions

Notion d'activité physique

L'OMS définit l'activité physique comme tout mouvement produit par les muscles squelettiques [5], responsable d'une augmentation de la dépense énergétique (quantité d'énergie dépensée par une personne).

Notion de sédentarité

La sédentarité correspond à une activité physique faible ou nulle avec une dépense énergétique proche de zéro. Elle est caractérisée par une position assise tenue de manière prolongée. Le temps passé devant un écran [6] constitue, par exemple, un bon indicateur de sédentarité.

5. Muscles qui assurent la motricité du squelette.
6. Ici, un écran désigne, de manière générale, un ordinateur ou une télévision.

1.3.2 Des impacts sur la santé

La sédentarité est considérée comme le quatrième facteur de risque de décès dans le monde. En effet, les maladies non transmissibles, comme le diabète, le cancer et les cardiopathies, provoquent plus de 70% des décès, soit 41 millions de décès annuels, dont 15 millions de décès prématurés entre 30 et 69 ans. Les cinq facteurs majeurs favorisant ces maladies sont le tabagisme, l'inactivité physique, la consommation excessive d'alcool, la mauvaise alimentation et la pollution de l'air. Aussi, selon l'OMS, si les quatre facteurs de risque principaux que sont la sédentarité, la mauvaise alimentation, le tabac et l'alcool étaient mieux appréhendés par les populations, près de 80% des crises cardiaques ou des AVC prématurés seraient évités [25][24].

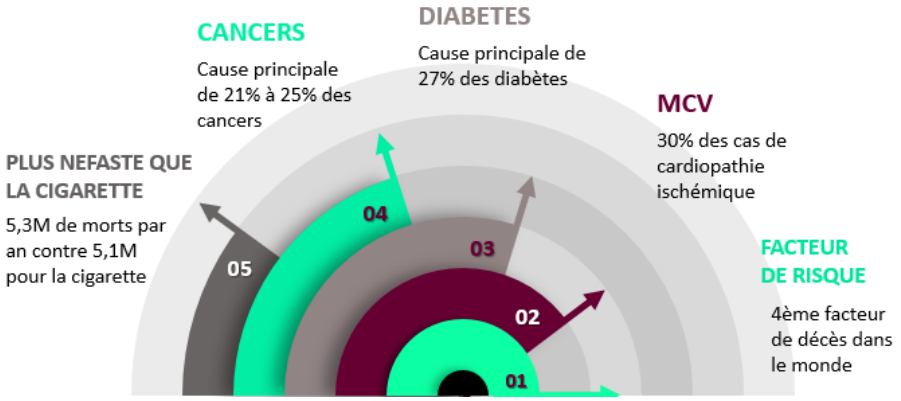


Figure *1.8:* Cinq faits sur l'inactivité physique

La sédentarité augmente considérablement le risque de développer de nombreuses pathologies et représente un problème de santé publique majeur. Par ailleurs, l'OMS estime que la sédentarité est la cause principale de 21 à 25% des cancers du sein ou du colon, de 27% des cas de diabète et d'environ 30% des cas de cardiopathie ischémique [7]. De plus, selon la Chaire de recherche internationale sur le risque cardiométabolique (ICCR), l'inactivité physique tuerait 5,3 millions de personnes par an dans le monde, contre 5,1 millions pour le tabac.

Selon le rapport d'activité 2017 de l'Assurance Maladie (dont une synthèse est affichée en figure 1.9), les maladies non transmissibles représentaient en 2016, 31% des dépenses du régime général de l'Assurance Maladie. Une réduction des facteurs de risques de ces maladies constituerait donc un levier relativement important de réduction des dépenses de santé.

L'OMS recommande la pratique d'au moins une demi-heure de marche par jour pour entretenir sa santé. Cependant, moins de 40% de la population mondiale s'adonne à cette activité régulière. Adopter un mode de vie en deçà des recommandations est nuisible à la santé physique car c'est un facteur aggravant de l'obésité et de la fatigue. Ce mode de vie est également psychiquement néfaste : il favorise la dépression, les troubles du comportements alimentaires, etc.

7. Ensemble des troubles dus à l'insuffisance des apports en oxygène au muscle cardiaque.

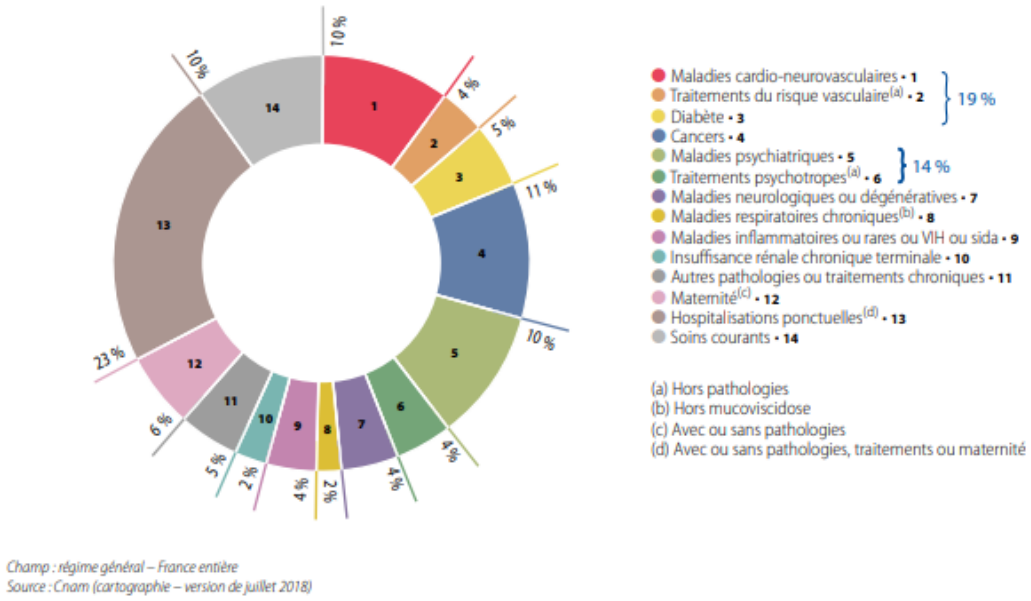


Figure *1.9:* Répartition des dépenses en santé de l'Assurance Maladie du régime général en 2016

Chez l'adulte, la pratique d'une activité physique régulière et adaptée permet, en plus de réduire les risques liés aux maladies non transmissibles précédemment citées, certains effets bénéfiques tels que le renforcement des os. L'activité physique est aussi un facteur clé de la dépense énergétique. Elle est donc fondamentale pour l'équilibre énergétique et le contrôle du poids.

1.3.3 Promotion de l'activité physique dans le cadre de la prévention santé

La promotion de l'activité physique est une préoccupation croissante dans les débats sociaux. Par exemple, l'OMS a pour objectif, en 2019, d'aider les gouvernements à atteindre la cible mondiale d'une réduction de 15% de l'inactivité physique d'ici à 2030. Cela se fera par exemple par la mise en œuvre du module pratique nommé "SOYONS ACTIFS" visant à promouvoir l'activité physique quotidienne de chacun.

Les recommandations de l'OMS en matière d'activité physique varient en fonction de l'âge. L'activité physique englobe notamment les activités récréatives ou les loisirs, les déplacements (par exemple la marche ou le vélo), les activités professionnelles, les tâches ménagères, le jeu, les sports ou l'exercice planifié, dans le contexte quotidien, familial ou communautaire. Pour réduire le risque de maladies non transmissibles et de dépression, certaines recommandations de l'OMS pour les adultes entre 18 et 64 ans ont été formulées :

- pratiquer au moins, au cours de la semaine, 150 minutes d'activité d'endurance d'intensité modérée ou au moins 75 minutes d'activité d'endurance d'intensité soutenue, ou une combinaison équivalente d'activité d'intensité modérée et soutenue ;
- pratiquer par périodes d'au moins 10 minutes des activités d'endurance.

Ces recommandations peuvent être dépassées pour augmenter les bénéfices pour la santé. Pour cela, il faut par exemple :

- augmenter la durée de l'activité d'endurance d'intensité modérée de façon à atteindre 300 minutes par semaine ou pratiquer 150 minutes par semaine d'activité d'intensité soutenue, ou une combinaison équivalente d'activité d'intensité modérée et soutenue ;
- pratiquer au moins deux jours par semaine des exercices de renforcement musculaire faisant intervenir les principaux groupes musculaires.

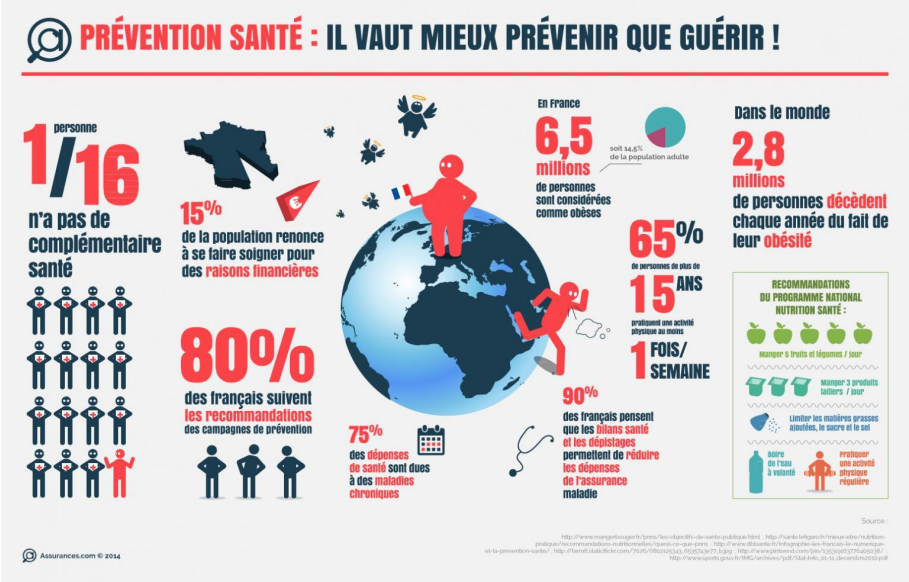


Figure *1.10:* Chiffres clés de la prévention de l'obésité. Assurance.com 2014

Conclusion

Les acteurs de l'assurance maladie garantissent une protection aux français face aux risques sociaux. Cette prise en charge génère des coûts qui nécessitent d'être maîtrisés. Dans ce contexte, la prévention se présente comme un dispositif permettant d'atténuer les risques et les coûts associés. De plus, elle est appuyée par de multiples réformes de l'État. En santé, un des facteurs de risques les plus importants est l'inactivité physique.

L'activité physique est un enjeu important dans la promotion d'une bonne hygiène de vie. Cette pratique, malgré tous les bénéfices qu'il en ressort, est peu ou pas du tout appliquée. Afin de ré-hausser le niveau d'activité des individus, une grande stratégie de prévention est nécessaire. Cela passe notamment par le biais de recommandations en termes d'habitudes sportives. Afin de rétablir un bon équilibre de vie, il est important de diminuer le temps de sédentarité. Pour cela, l'environnement quotidien doit être présenté comme un parcours de santé. Cependant, si des gestes simples peuvent être pratiqués par tous, sportifs et non sportifs, la réelle implication de ceux-ci est peu maîtrisable. Cela est en partie dû au fait que la quantification de la réalisation des objectifs n'est pas toujours accessible. C'est en cela que les objets connectés peuvent se présenter comme un excellent outil de prévention santé.

Partie I

Les objets connectés en Assurance

2 État de l'art et enjeux des objets connectés

« L'Internet des Objets exprime la profondeur et le caractère incertain de l'évolution anthropologique et collective qui s'actualise sous nos yeux ». ***Jean-Max Noyer, IMSIC***

L'objectif de ce chapitre, composé de deux parties, est d'éclairer le lecteur quant au concept d'objets connectés. Il s'agira, tout d'abord, de définir clairement le périmètre de ce qui est considéré comme objet connecté, puis, de mettre en exergue l'enjeu du contexte réglementaire lié à ces derniers. Le lecteur pourra dès lors comprendre la mécanique et l'intérêt des données issues de ces objets.

2.1 Généralités

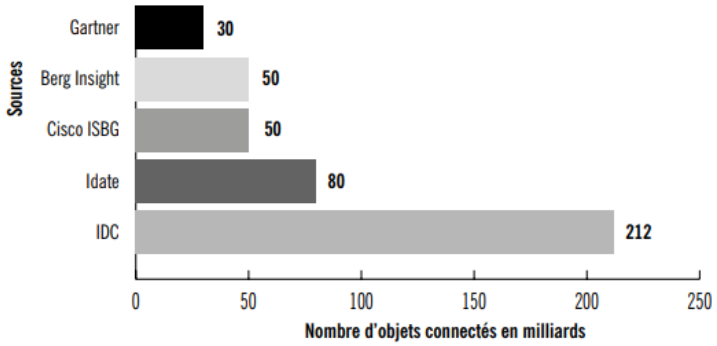
Selon le guide de la santé connectée, tout objet composé de capteurs envoyant des informations vers une application mobile ou un service web est qualifié d'objet connecté. Plus généralement il s'agit d'un objet, souvent électronique, capable de communiquer des informations vers un autre objet, vers un serveur informatique ou vers un ordinateur ou une tablette, grâce à un réseau dédié (le plus souvent Internet). L'objet devient intelligent et intègre un réseau d'autres objets appelé « Internet des objets » (*Internet Of Things* – IOT -), capables d'interagir entre eux.

2.1.1 Des usages en croissance

Les objets connectés connaissent un essor depuis la fin des années 90. Entraînés par la vague internet, ces outils entrent dans les moeurs dans l'ère du *big data*. Le smartphone constitue le premier des objets connectés en termes d'utilisation et d'ancrage dans les moeurs. Sa notoriété n'étant plus à présenter, dans ce chapitre, la notion d'objets connectés exclura le smartphone.

Figure *2.1:* Anticipation de *Cisco* en 2011 sur l'utilisation des objets connectés

En 2011, Cisco prévoyait qu'en 2015 dans le monde, 25 milliards d'objets seraient connectés et IBM en annonçait 1000 milliards. Cependant leur nombre n'a atteint que 15 milliards cette année-là selon Cisco et selon l'Idate [1] et tout juste 6 milliards selon Gartner. De même, les chiffres pour 2020 varient énormément d'une estimation à l'autre [23]. Ces dernières sont présentées dans la figure 2.2.



Source : G9+, *Les nouveaux eldorados de l'économie connectée*, 2013.

Figure ***2.2:*** Estimation du nombre d'objets connectés en 2020 selon différentes entreprises

Bien qu'une croissance beaucoup plus forte avait été anticipée, depuis leurs premières apparitions en 2003, le nombre d'objets connectés utilisés par l'Homme a tout de même été multiplié par trente en 2015. Cela continue de nourrir les anticipations de croissance de ce phénomène. L'essor de ces outils est aussi lié à la diversité croissante présente dans les objets connectés mis à disposition. Dans la figure 2.3, il est possible d'observer le panel d'objets connectés les plus connus en France.

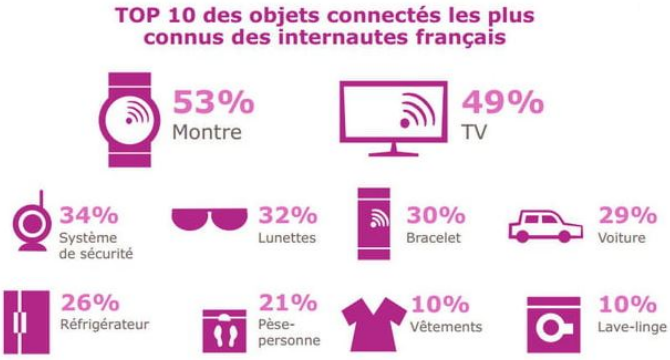


Figure ***2.3:*** Les principaux objets connectés connus en France en 2014 selon une étude de Médiametrie

Ce panel d'objets connectés connus constitue un exemple de la diversité présente au sein des objets connectés.
Les objets connectés connus du grand public sont ceux réservés aux usages individuels. Ce sont par exemple les montres, lunettes et vêtements connectés. Ils sont évoqués souvent sous le terme de "*wearables*" qui fait référence aux objets connectés pouvant être "portés". Dans le domaine du bien-être, il est possible de citer par exemple la brosse à dents connectée, la fourchette

1. Institut de l'audiovisuel et des télécommunications en Europe.

connectée et autre pèse-personne. Ces objets permettent d'améliorer le quotidien des personnes en leur apportant un suivi régulier. En santé particulièrement ceux-ci permettent un meilleur suivi des individus souffrant d'affections particulières et nécessitant un suivi régulier. Ce sont, par exemple, les lentilles de contact pour diabétiques, le pilulier intelligent, la montre connectée pour identifier la crise cardiaque avant qu'elle ne se manifeste, le patch connecté pour les patients atteints de la maladie d'Alzheimer ou encore le tensiomètre sans fil.
En dépit de leur apport effectif dans certains secteurs, beaucoup de ces objets ont une utilisation de type gadget. De ce fait, ceux-ci n'apportent pas de réelle valeur ajoutée sur le long terme. Ainsi, l'intérêt actuel de ce type d'objet connecté réside dans l'usage qu'il va apporter à son utilisateur en matière de "*quantified self*[2]".

2.1.2 Des freins à cette croissance

L'essor des objets connectés se heurte tout de même à quelques freins dont les principaux sont détaillés sur la figure 2.4.

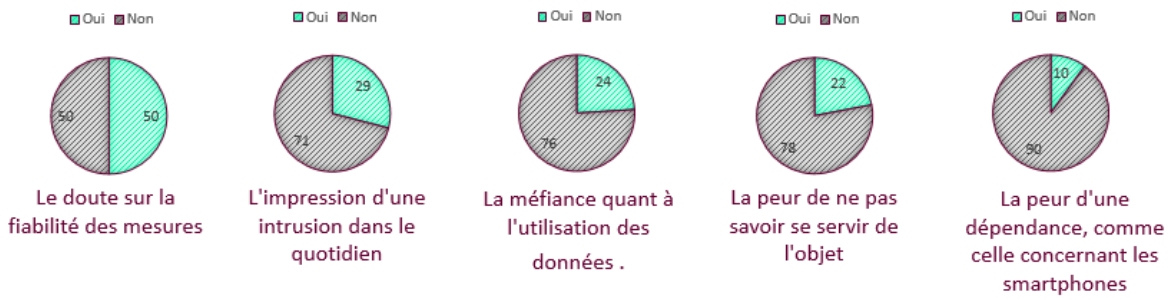


Figure ***2.4:*** Les principaux freins à l'essor des objets connectés selon un article de 2018 de la Mutuelle générale

Selon cette étude de la Mutuelle générale, la principale cause de non achat d'un objet connecté reste le doute quant à la fiabilité des mesures avant même le sentiment d'intrusion dans le quotidien. De plus, en 2017, selon les entreprises Gartner et l'Idate, pour 53% des français, les équipements connectés représentent le futur, dans des domaines comme la sécurité, la santé et la domotique. 47% mettent l'accent sur l'aspect pratique et utile alors que pour 29%, il s'agit d'un simple phénomène de mode.

De manière générale, l'expansion des objets connectés est ralentie par leur coût d'acquisition relativement élevé mais également par le sentiment d'atteinte à la vie privée qu'ils peuvent engendrer chez leurs utilisateurs.

2.1.3 Des catalyseurs pour un nouvel essor

Plusieurs apparitions technologiques permettent d'anticiper une grande croissance dans l'utilisation des objets connectés. Ces vecteurs de croissance sont par exemple la 5G, l'*Edge Computing* ou le *machine learning*.

2. Principes et méthodes permettant à chacun de mesurer ses données personnelles, de les analyser et de les partager.

La cinquième génération de standards pour la téléphonie mobile - **5G** -, est une innovation qui permet des débits de télécommunication mobile de plusieurs gigabits de données par seconde [3]. Son déploiement constituera un élément favorable à la croissance de l'IOT. En effet, celle-ci est développée afin de contribuer à l'essor des objets connectés en améliorant considérablement les débits de communications.

L'***Edge Computing*** est une technologie idéale dans le cadre des objets connectés car elle permet de réduire les temps de latence entre la collecte et le traitement des données. Cette analyse en temps réel offre de multiples possibilités pour les objets connectés.

Les algorithmes de type ***machine learning*** préalablement entraînés permettront aux objets connectés d'acquérir une meilleure puissance de calcul. Cela introduit la possibilité de récupérer des données et de les assigner à un modèle prédéfini. Depuis une caméra, un objet connecté pourrait, par exemple, détecter par le biais d'un modèle de *machine learning* les signes précurseurs d'un dégât des eaux pour enclencher l'alerte et le suivi.

2.1.4 Des secteurs d'application variés

Les objets connectés offrent des opportunités de développement dans plusieurs secteurs. Aussi, plusieurs domaines voient leurs activités évoluer avec la démocratisation des objets connectés.

Dans le B2C [4], l'intégration de ces objets connectés dans une stratégie d'entreprise est une garantie de différenciation pour les entreprises et permet d'apporter une connaissance client beaucoup plus fine. Les secteurs d'activités suivants peuvent être cités.

- Dans la **santé** connectée par exemple, les objets connectés analysent l'historique des données et peuvent envoyer des alertes à celui qui les porte et au médecin s'ils détectent un comportement anormal, tel que des pulsions cardiaques anormalement rapides.
- Dans la **domotique**, un exemple d'application destinée aux consommateurs est le thermostat connecté de l'entreprise Netatmo. Il permet aux consommateurs de réaliser des économies d'énergie. En effet, la programmation "intelligente" du chauffage intérieur et le pilotage à distance évitent un trop grand gaspillage énergétique.

Dans le B2B [5], les applications sont diverses et variées. Des exemples d'application des objets connectés dans le B2B sont :

- la **maintenance prédictive** : anticiper les pannes des machines et donc les réparer avant qu'elles ne tombent en panne, tel est l'objectif de l'entreprise CityTaps. Cette entreprise propose des compteurs d'eau connectés qui collectent des informations sur la consommation et la pression d'eau, la température des boîtiers etc. Ces informations sont croisées avec des

3. Cela correspond à 42 secondes afin de télécharger une vidéo de 1 Go contre 7 min en 4G.

4. *Business to Consumer.*. Appellation qui désigne l'ensemble des activités commerciales qui mettent en relation une entreprise et un consommateur

5. *Business to Business.*. Appellation qui désigne l'ensemble des activités commerciales nouées entre deux entreprises

données externes de température extérieure et de pression atmosphérique afin de réaliser une anticipation de la surchauffe des boîtiers, et à *priori* d'éviter des pannes ;

- les **flottes automobiles connectées** : piloter un parc automobile et découvrir des *scorings* de sécurité ou d'éconduite attribués à chaque conducteur, il s'agit là des possibilités offertes par Total Fleet Connect. Cela est possible avec l'aide d'un boîtier connecté, installé dans les véhicules de la flotte et qui transmet toutes les données de conduite en temps réel. D'après l'entreprise Total, ce boîtier connecté permet de réaliser des économies concernant la consommation de carburant, la sinistralité des véhicules ainsi que le coût pour la compagnie d'assurance ;

Selon l'entreprise Gartner, le B2C engendre le plus gros volume de ventes, mais le B2B engendre plus de revenus.

2.2 Limites réglementaires françaises actuelles

Dans ce paysage d'objets connectés en expansion, il est important de relever des points d'attention particuliers comme la sécurité des données. Il est ainsi important de se demander dans quelles mesures les données produites par ces objets sont sécurisées. Pour toutes les entités qui comptent faire usage de ces données, la confidentialité et la protection des données personnelles est un enjeu de taille qu'il est nécessaire de prendre en compte dans cette explosion du paysage connecté. Plusieurs règles [30] et codes de bonne conduite concernant la collecte, le traitement et le stockage des données sont importants à souligner.

2.2.1 CNIL et RGPD

Sont considérées comme données personnelles, toutes informations relatives à une personne physique identifiée ou identifiable, directement ou indirectement, par référence à un numéro d'identification comme le numéro de sécurité sociale ou à un ou plusieurs éléments qui lui sont propres (ex. : nom et prénom, date de naissance, éléments biométriques, empreinte digitale, ADN ...). Ces données bénéficient d'une forte protection du fait de leur caractère personnel. A ce titre, des organismes tels que la CNIL assurent le rôle de gardien des libertés et des droits des individus sur les données qu'ils produisent [4].

2.2.1.1 Commission Nationale de l'Informatique et des Libertés

La Commission Nationale de l'Informatique et des Libertés - CNIL - est un organisme public qui agit au nom de l'Etat et qui a été créée par la loi Informatique et Libertés du 6 janvier 1978. Elle est chargée de veiller à la protection des données personnelles contenues dans les fichiers et traitements informatiques ou papiers, aussi bien publics que privés. Ses missions en France se résument en quatre points :

1. informer, protéger les droits ;
2. accompagner la conformité / conseiller ;
3. anticiper et innover ;
4. contrôler et sanctionner.

Les données personnelles sont au coeur des préoccupations de la CNIL. Son pouvoir de contrôle et de sanction lui permet d'assurer la mise en place réelle et efficace de procédures de protection des données personnelles. Aussi, pour toutes initiatives ou projets d'entreprise qui utilisent ou traitent des données personnelles des citoyens de l'Union européenne, l'accord préalable de la CNIL est requis. A ce titre, une collaboration avec la CNIL peut s'avérer appropriée afin de créer une offre légale.
L'arrivée du règlement général sur la protection des données personnelles renforce la position de cet organisme.

2.2.1.2 Règlement général sur la protection des données personnelles

Le règlement général sur la protection des données personnelles - RGPD - est entré en application le 25 mai 2018. Il a pour objectif de permettre à l'Europe de mieux s'adapter aux réalités du numérique. Il renforce les droits des citoyens européens en leur donnant plus de contrôle sur leurs données personnelles.
Cette réforme vise trois principaux objectifs :

- renforcer les droits des personnes en créant par exemple le droit à la portabilité des données personnelles ;
- responsabiliser les acteurs traitant des données [6] ;
- crédibiliser la régulation grâce à une coopération renforcée entre les autorités de protection des données.

Pour un citoyen européen, ses droits se résument comme suit :

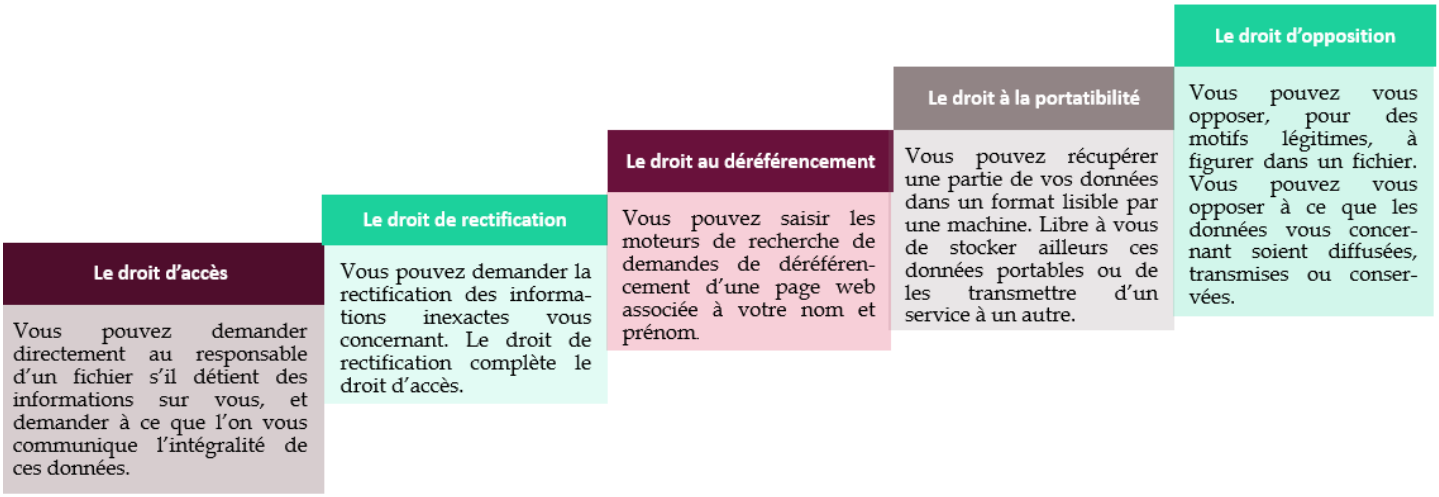


Figure ***2.5:*** Les droits des citoyens européen en matière de protection de données

La prise en compte de ce règlement est importante dans la suite de cette étude car les modèles développés dans le cadre de ce mémoire sont construits à partir de données personnelles. En effet, les données produites par les objets connectés peuvent être, soit directement, soit indirectement, révélatrices d'informations sensibles relatives notamment à l'état de santé actuel ou futur. C'est pour cela que tout au long de cette étude des dispositions ont été prises afin de toujours être dans le respect de ce règlement.

6. Responsables de traitement et sous-traitants.

Les étapes de la gestion du RGPD peuvent se résumer sur la figure 2.6

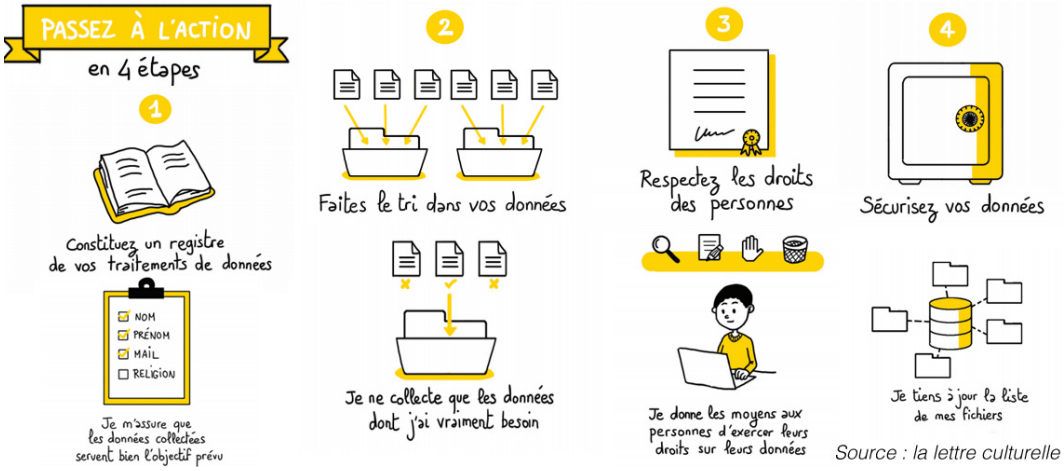


Figure ***2.6:*** La mise en conformité RGPD en quelques étapes

De manière plus détaillée l'application de ce règlement a été respectée notamment dans les étapes suivantes :

- **la collecte des données**. Une demande de consentement spécifique et une mention d'information ont été nécessaires dans le cadre de la récolte des données personnelles ;
- **le respect des droits des individus**. Lors de la collecte, une messagerie de contact est mise en place afin de permettre aux propriétaires des données de faire part de leur volonté d'exercer leurs droits énoncés dans la figure 2.5 ;
- **le stockage des données**. Les données sont stockées sur un serveur sécurisé et ne sont accessibles que par les personnes participant à cette étude.

2.2.2 L'assureur et les données de santé

En France, en assurance collective, la loi Evin du 31 décembre 1989 interdit la tarification basée sur des données de santé. Cette loi constitue une vraie barrière de sécurité contre les discriminations individuelles. Le recueil et le traitement de ces données sont plus ou moins autorisés juridiquement en fonction du type d'assurance. La table 2.1 fait un état des lieux de cette autorisation.

	Recueil des données personnelles	Exploitation de données personnelles
Assurances sociales	Aucune	Pas de sélection
Assurances privées collectives	Recueil de données personnelles non interdit	Sélection interdite (loi Evin)
Assurances privées individuelles	Recueil de données personnelles non interdit (sauf données de santé pour les mutuelles)	Liberté de sélection des risques (Incitation fiscale à ne pas le faire)

Table ***2.1:*** Autorisation/ interdiction du recueil et du traitement des données personnelles [30]

3 Enjeux des objets connectés dans l'assurance de demain

« Il faut trouver un juste équilibre entre la segmentation et le partage du risque entre assurés - ce sont les principes de solidarité et de mutualisation consubstantiels à l'assurance qui sont en jeu ; et il ne faut pas perdre de vue nos valeurs collectives à l'égard du respect de la vie privée ». ***François Villeroy de Galhau***

Dans ce chapitre, les enjeux pour le domaine de l'assurance liés à l'usage des objets connectés seront abordés. Le lecteur pourra, mieux se familiariser d'une part avec les pratiques existantes d'utilisation des objets connectés en assurance et d'autre part avec la prévention basée sur des données issues d'objets connectés .

3.1 Pratiques assurantielles existantes

En assurance, trois secteurs ont déjà été investis par les objets connectés. Les compagnies d'assurance se sont penchées sur des produits pour lesquels la présence d'objets connectés est déjà effective. Comme détaillé sur la figure 3.1, il s'agit des secteurs de la santé, de l'habitation et de l'automobile.

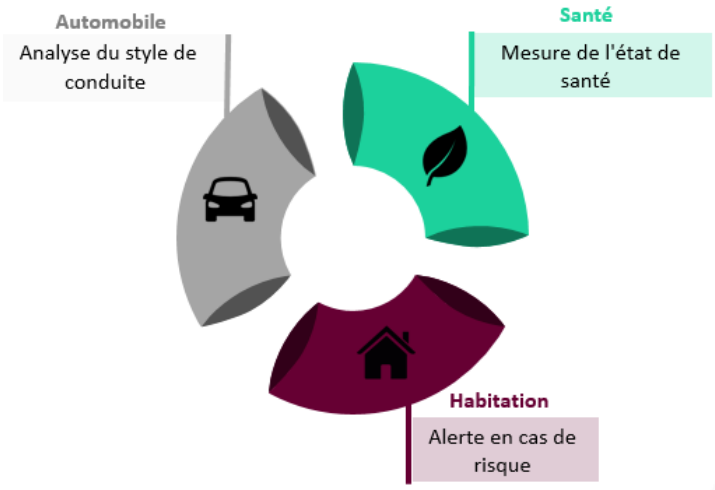


Figure ***3.1:*** Les principaux secteurs d'utilisation des objets connectés

- **La santé :** les bracelets/montres connectés sont capables de mesurer des indicateurs physiques du porteur. Cette connaissance peut l'inciter à avoir une meilleure hygiène de vie (activité physique, alimentation, sommeil. . .) ;

- **L'habitation :** une alerte peut être envoyée à l'aide des objets connectés en cas de vol, fuite ou encore court-circuit ;
- **L'automobile :** correspond historiquement au premier marché des objets connectés. Les boîtiers installés dans les véhicules sont dotés d'une technologie capable d'analyser le style de conduite de l'assuré.

Les objets connectés offrent, quelque soit le secteur d'activité, l'avantage à l'assureur de maintenir un lien quasi-permanent avec les assurés. Traditionnellement, assuré et assureur n'échangent qu'au moment du renouvellement du contrat ou lors d'un sinistre. Avec les objets connectés, les assureurs peuvent renforcer l'engagement de leurs assurés, et ce au travers d'une approche personnalisée.

3.1.1 Le benchmark français

Il est important de noter que sur le marché français de l'assurance, les objets connectés s'inscrivent dans une dynamique de prévention. En effet, à l'image du programme "*Vitality*" de Generali ou encore de "*Pay how you drive*" - PHYD - de AXA en France, les problématiques d'objets connectés sont utilisées pour accompagner l'assuré dans sa gestion quotidienne du risque. Cela peut être désigné par le terme **assurance comportementale**.

- **Generali Vitality** : cette initiative de Generali en collaboration avec les montres connectées GARMIN incite ses adhérents à avoir une bonne hygiène de vie grâce à des récompenses. Cela se déroule en plusieurs étapes pour l'assuré. Il est amené à remplir un questionnaire détaillé en ligne pour faire un bilan de santé. En fonction des réponses, des objectifs personnalisés sont proposés. Aussi, en les atteignant, les candidats gagnent des points qui se convertissent en cadeaux de marques partenaires. Un an après la mise en place, sur les 100 000 entreprises clientes de Generali France, 1 800 ont choisi d'entrer dans le programme Vitality, mais seulement 2 700 collaborateurs, ont accepté d'y participer. La mise en place prend du temps. Ce programme a aussi suscité une controverse dans les phases de lancement du programme, la place de l'assureur dans la prévention des assurés ayant été remise en question.
- **Pay how you drive :** cette offre, connue sous le nom de "YouDrive" chez AXA permet de récompenser les « bons conducteurs » à travers l'abaissement de leurs primes. Grâce aux boîtiers connectés installés dans les véhicules, les assurés voient leurs tarifs modulés en fonction de leur conduite. Les points de conduite varient selon les assureurs mais regroupent généralement le nombre de kilomètres parcourus, les horaires de conduite, le temps de repos lors d'un trajet, la façon de prendre les virages, les accélérations, les freinages ou encore la vitesse en fonction du trafic. Le conducteur peut analyser son comportement au volant sur la base d'un score de conduite entre 0 et 100 points à la fin de chaque trajet. La somme de tous ces scores est pondérée par la distance parcourue sur le mois. Elle permet de prouver à son assureur la prudence de sa conduite et ainsi de bénéficier d'une réduction sur la prime de son assurance pouvant aller jusqu'à 50%. A ce jour, environ 90% des conducteurs ont bénéficié d'une baisse.
- **Pay as you drive :** ce type d'assurance proposé par plusieurs assureurs du marché

français comme Groupama, AXA ou encore la MAAF constitue l'assurance au kilomètre. Une cotisation directement dépendante du nombre de kilomètres parcourus est versée par l'assuré. Cependant, le tarif au kilomètre dépend du profil du conducteur : risque, antécédents, bonus/malus etc ...

- **Protection 24 et Somfy Protect** : depuis 2015, AXA propose des objets connectés afin d'assurer le suivi de la protection domestique. Pilotable depuis une application mobile, une alerte est envoyée en cas, par exemple, de détection d'intrusion, de départ d'incendie ou de fuite d'eau. A partir des caméras disposées dans leur habitation, les assurés peuvent se rendre compte visuellement de la situation, et selon le cas, prévenir un voisin, alerter les secours ou faire intervenir les équipes d'AXA Assistance pour vérification, gardiennage, ou effectuer des réparations. L'adhésion à cette protection permet aussi à l'assuré, en fonction du programme souscrit (Protection 24 et Somfy Protect), de voir baisser, sa franchise ou sa cotisation sur les garanties vol et vandalisme, ou encore d'avoir des bons plans et des avantages.

3.1.2 Sur le marché international

La principale problématique concernant l'usage des objets connectés dans la pratique assurantielle en France correspond au contexte juridique et réglementaire qui est encore en construction. Aussi, à l'échelle planétaire, de nombreuses applications actuarielles sont d'ores et déjà proposées aux assurés. Parmi elles, peuvent notamment être citées :

- **John Hancock Financial - Pay how you behave** est une entreprise américaine qui propose à ses clients, à la place des contrats traditionnels, des polices accordant une place importante aux données portant sur leur activité physique et leur état de santé. Le cumul de points acquis sur la base d'habitudes saines enregistrées permet de réduire les cotisations d'assurance. En fonction du programme souscrit et des points cumulés des avantages variés sont à gagner. Parmi les avantages, peuvent être cités, une réduction de 15% des primes, des remises dans des magasins ou encore des montres connectées Apple. Cependant des pénalités qui vont jusqu'à 15$ par mois sont aussi à prévoir lorsque le niveau d'activité n'est pas jugé suffisant ;
- **UnitedHealthcare** est une entreprise américaine qui incite les employés couverts par l'assurance *UnitedHealthcare* à recevoir des récompenses financières s'ils atteignent leurs objectifs quotidiens, mesuré via leur montre connectée. Les utilisateurs peuvent obtenir une Apple Watch par le biais du programme et peuvent ensuite utiliser l'argent qu'ils gagnent grâce au programme pour rembourser le reste de l'appareil. Après remboursement, l'argent est mis sur un compte d'épargne santé ou de remboursement santé. Les employés qui participent à ce programme peuvent gagner quelques dollars par jour pour atteindre trois objectifs :
  - ✔ un objectif de **fréquence** pour atteindre 500 pas en sept minutes, six fois par jour, à au moins une heure d'intervalle (1,50$ par jour) ;
  - ✔ un objectif d'**intensité** pour atteindre 3 000 pas en 30 minutes (1,25$ par jour) ;

- ✔ un objectif de **ténacité** pour atteindre 10 000 pas au cours de la journée en tout (1,25$ par jour) .

- **Pay as you live - PAYL** est un concept d'assurance commercialisé par l'entreprise *Ernst and Young* qui propose aux assurés d'utiliser des *wearables* pour gérer leur santé [8]. Les données sont transmises à l'assureur par une interface portable. Les données comportementales sont ensuite incluses dans des programmes de fidélisation et dans la tarification des polices. De plus, à travers l'interface de l'utilisateur, ce dernier peut faire le lien entre un comportement positif et l'impact sur le prix de la police d'assurance. Ce concept permettrait d'inciter les titulaires de ce type de contrat à améliorer leur mode de vie pour améliorer leur état de santé et réduire les primes.

Plusieurs pratiques assurantielles utilisant les objets connectés sont déjà mises en place en France et dans le monde dans des domaines d'applications variés. En santé, les assurés sont incités à réadapter leurs comportements afin d'être plus actifs. Cette pratique s'appuie sur le fait que le comportement des assurés peut être capté par des données issues de leurs objets connectés. Par cette connaissance plus concrète de l'assuré, il serait possible d'évaluer la performance des actions préventives effectuées.

3.2 Notion de prévention de demain

Les objets connectés sont parfois présentés comme les outils d'une « prévention de demain ». Il n'existe cependant actuellement aucune étude certifiant de leur efficacité en tant qu'outil de prévention. La connaissance de soi et de sa santé, amenée par la collecte et la mesure des différentes variables physiques, permet d'objectiver les comportements et de modeler les habitudes. Cet atout peut justifier l'usage des objets connectés dans le cadre d'une stratégie de prévention.

3.2.1 L'impact de la prévention connectée chez l'assureur

Le principe de la prévention connectée [1] est très proche de l'assurance comportementale, notamment par son principe de modulation de l'offre en fonction de l'assuré. Dans les exemples d'applications en assurance cités précédemment, il est important de noter que l'assurance comportementale impacte les éléments principaux de la théorie des contrats que sont : l'asymétrie d'information [2], l'aléa moral [3] et l'anti-sélection [4]. En ce sens, la prévention peut être complémentaire à l'activité d'assurance.

L'**asymétrie d'information** peut être réduite par le biais des objets connectés puisque l'assureur a une meilleure connaissance de l'assuré. Un phénomène inverse peut apparaître puisque l'analyse des informations récoltées peut permettre de connaître l'assuré mieux que lui-même.

1. Appellation désignant ici un programme de prévention utilisant des objets connectés.
2. Répartition inégale de l'information entre les deux acteurs.
3. Absence d'incitation à se prémunir contre le risque lorsque l'on est protégé contre ses conséquences, par exemple par une assurance.
4. Sélection des mauvais risques.

Pour éviter le risque d'**aléa moral**, l'assureur a pour volonté de responsabiliser ses assurés. Cette responsabilisation est incluse dans un processus de prévention. De plus, adhérant de son plein gré aux offres de type assurances comportementales ou prévention connectée, l'assuré est conscient des enjeux de son comportement. Cette conscience crée un aléa moral "inversé" chez l'assuré, puisqu'il sera enclin à changer son comportement afin d'améliorer sa notation. L'assuré est donc encouragé à effectuer des actions d'auto-protection. Le domaine de la santé est particulièrement soumis à des problématiques d'**anti-sélection**. Ainsi, les offres de prévention connectée peuvent être utilisées en tant que levier afin de diminuer le phénomène de sélection adverse, puisque, comme observé pour les programmes tels que PAYD de AXA, les bons profils sont attirés par les offres comportementales.

Ehrlich et Becker mettent en lumière des liens complexes entre activité d'assurance et prévention[5], notamment leur caractère **substituable**. En effet, lorsque la prime d'assurance augmente, les assurés peuvent être tentés de limiter leur risque en se tournant plus vers la prévention, au détriment de leur assurance. De plus, la prévention secondaire est nommée "auto-assurance" par Ehrlich et Becker car, pour l'assuré, elle a le même rôle de diminution du coût d'un sinistre.

3.2.2 L'adhésion aux programmes de prévention connectée

Comme vu dans la section 3.1, des offres d'assurance prenant en compte les données issues des objets connectés existent déjà en France et ailleurs. Dans le cadre purement préventif, seul le programme *Vitality* de Generali propose des offres basées sur les données issues d'objets connectés en France. Plusieurs caractéristiques individuelles impactent la décision d'adhésion à un programme de prévention santé [19]. Une étude d'Harmonie Mutuelle en collaboration avec l'université de Lyon [18] s'est intéressée aux facteurs explicatifs de l'adhésion des individus à ce type de prévention.

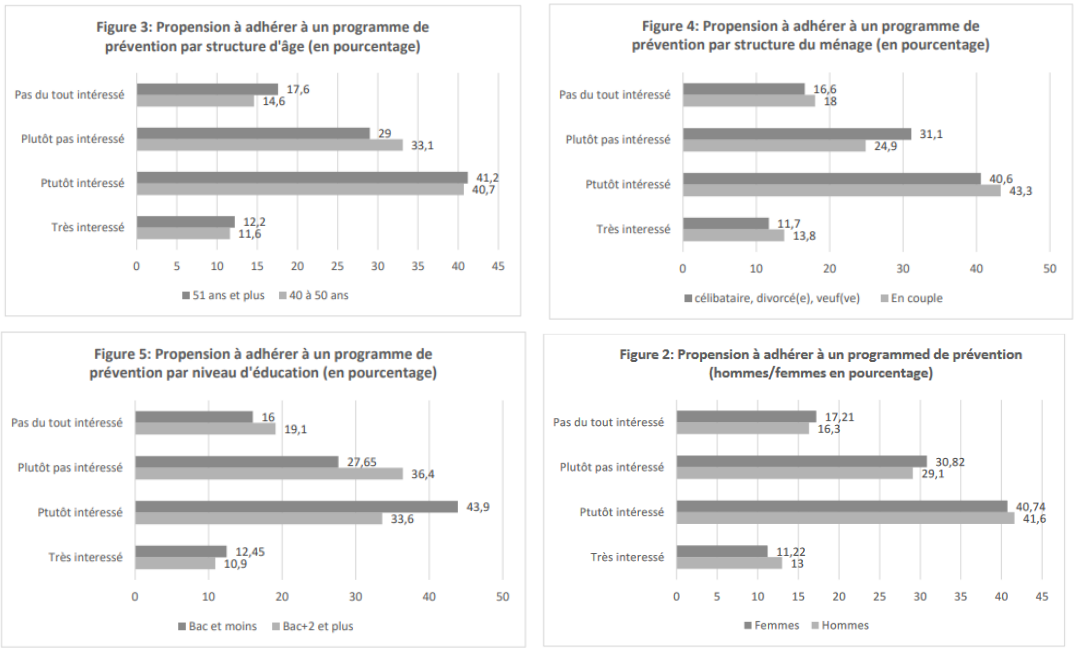


Figure ***3.2:*** La propension à adhérer à un programme de prévention en fonction de différents paramètres

Cette étude met dans un premier temps en exergue le fait que ce type de programme est envisagé par les acteurs de l'assurance. De plus, ces résultats constituent une étude de marché plutôt optimiste. En effet, pour la majorité des profils interrogés, plus de la moitié étaient intéressés par un programme de prévention connectée, moyennant l'apport de leurs données connectées. Il faut cependant noter que la limite principale de cette étude réside dans le fait qu'elle n'ait été réalisée que sur des individus de plus de 40 ans.
Une seconde enquête de 2015 de HubTday Insurance/ Gn research a interrogé 2215 personnes de plus de 18 ans. Les résultats de cette enquête sont présentés sur les figures 3.3 et 3.4.

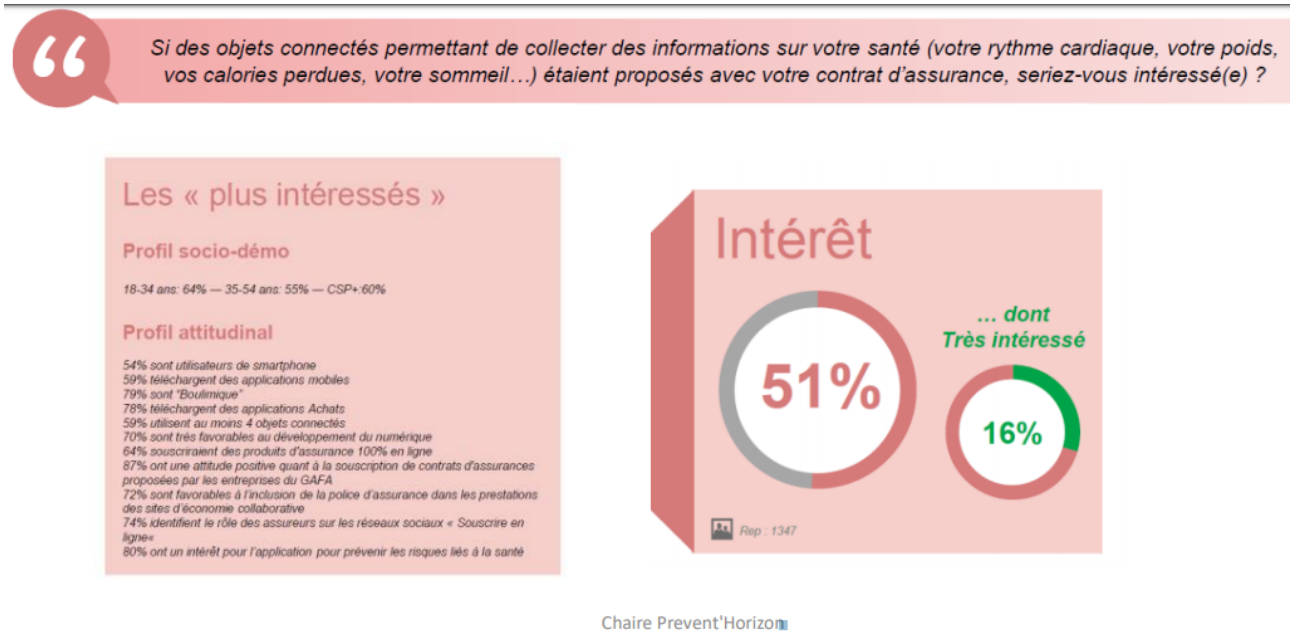


Figure *3.3:* Résultat de l'enquête de 2015 "Révolution numérique et assurance" de HubTday Insurance

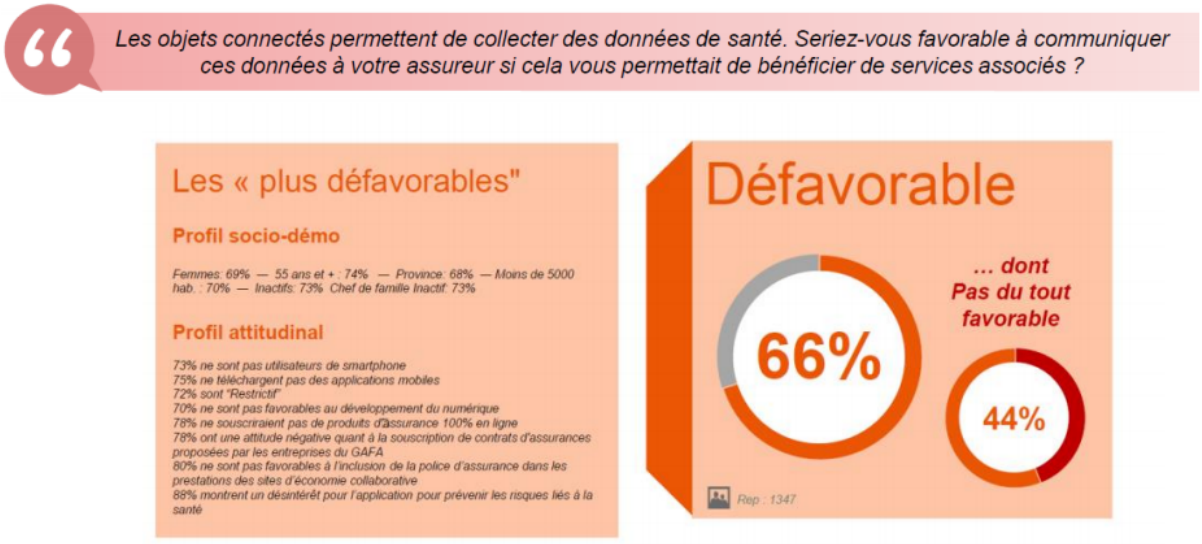


Figure *3.4:* Résultat de l'enquête de 2015 "Révolution numérique et assurance" de HubTday insurance

L'adhésion aux programmes de prévention connectée incluse dans des contrats d'assurance est mitigée mais séduit particulièrement les profils jeunes et les cadres. La communication des don-

nées issues des objets connectés à l'assureur moyennant des services est cependant très mal accueillie selon l'étude, surtout par les profils de femmes de plus de 55 ans et professionnellement inactives.

Conclusion

Les objets connectés sont des outils qui connaissent une grande expansion dans tous les domaines, dont celui de l'assurance. Des pratiques assurantielles sont déjà mises en place en France et dans le monde. Dans le domaine de la santé, le marché des objets connectés peut apporter un renouveau dans l'encadrement de la prévention pour l'assureur. Cependant, il est important, dans le développement d'offres de prévention connectée, de prendre en compte les exigences réglementaires liées à l'utilisation des données issues des objets connectés. Les assureurs se doivent donc d'être transparents dans leurs utilisations des données personnelles pour lever les craintes de leurs assurés. Ces nouveaux outils peuvent permettre de quantifier l'impact des mesures de prévention. Pour cela, il est important de poser un cadre théorique dans la quantification de la performance de la prévention.

Partie II

Le cadre théorique de la prévention et l'apport des méthodes classiques d'apprentissage statistique dans l'analyse des données connectées

4 Modèles classiques d'apprentissage statistique

« Le modèle est une représentation réaliste et pertinente dans le contexte des utilisations possibles ». ***NPA2 3.1.1***

Dans ce chapitre, les méthodes et algorithmes utilisés dans la suite de cette étude sont développés.

4.1 Apprentissage non-supervisé

L'objectif de l'apprentissage non-supervisé est de comprendre, représenter ou classer des données multi-dimensionnelles. Une exploration des données est réalisée et permet de révéler des structures naturelles dans les données. Les principaux résultats de cette exploration sont :

- **la segmentation** : construction de classes de manière automatique en fonction de l'échantillon disponible ;
- **les règles d'association** : analyse des relations entre les variables ou détections d'associations ;
- **la réduction de dimension** : projection des individus et des variables sur des axes.

4.1.1 Analyse en Composantes Principales

Principe

L'analyse en composantes principales - ACP - est une méthode d'analyse de données quantitatives [13]. Elle permet d'explorer des jeux de données sur plusieurs dimensions en recherchant les liaisons entre p variables et les ressemblances entre n individus. Pour cela, elle introduit une notion de distance entre individus et la proximité des variables est constatée par leurs corrélations au sens de Pearson (annexe A.2). Dans sa finalité, cette méthode apporte une représentation des n individus [1], dans un sous-espace de dimension $k < p$. L'objectif est donc de définir k nouvelles variables [2], combinaisons linéaires des p variables initiales et qui feront perdre le moins d'informations possibles.

Notion d'axes, de composantes principales, de valeurs propres et de contributions

L'ACP fournit une représentation synthétique et visuelle du jeu de données en projetant les nuages de points représentant les individus sur des axes (également appelés dimensions). De

1. p et $n \in \mathbb{N}^*$
2. $k \in \mathbb{N}^*$

plus, à chaque axe est associé une variable appelée composante principale. Celle-ci correspond au vecteur renfermant les cordonnées des projections des individus sur les axes. Ces composantes constituent une combinaison linéaire de la dimension initiale du nuage de données et sont non-corrélées les unes des autres. Chacune des composantes contient donc une partie de l'information initiale.

La réduction de la dimension consiste à conserver un nombre restreint d'axes tout en gardant un maximum de la variance initiale. De manière générale, l'ACP est ici réalisée sur des données normées. Cette mise à l'échelle permet d'assimiler directement la part de variance expliquée par l'axe, à la notion de valeurs propres issues de la décomposition en valeurs singulières de la matrice résumant les données.
Puisqu'ils sont représentatifs de la variance des composantes auxquelles ils sont associés, l'observation des valeurs propres permet de choisir le nombre d'axes à retenir. L'éboulis des valeurs propres représente la décroissance des valeurs propres. Il existe trois critères empiriques généralement utilisés pour sélectionner les axes.

- Le **critère du coude** : le choix du nombre d'axes se fait à partir de l'éboulis des valeurs propres. Lorsqu'un décrochement (coude) suivi d'une décroissance régulière est observée, les axes avant le décrochement sont conservés.
- Le **critère de Kaiser** : les axes dont les valeurs propres sont supérieures à 1 sont sélectionnés [3].
- La **règle de l'inertie minimale** : les axes dont le cumul représente 80% de la variance totale sont retenus.

Afin de réaliser une bonne interprétation de la structure observée par l'ACP, il est nécessaire de prendre en compte la notion de contribution aux axes. Cela correspond au poids que la variable ou que l'individu a eu dans la constitution d'un axe. L'étude de la contribution d'un individu ou d'une variable permet de déceler les observations influant le plus sur la constitution d'un axe.

Les avantages et les inconvénients de la méthode

L'avantage de l'ACP réside principalement dans sa **simplicité**. C'est une méthode simple à la fois sur le plan mathématique et sur le plan de l'interprétabilité des résultats. Cela est dû au fait que, mathématiquement, elle ne fait appel qu'à des outils de calcul de valeurs/vecteurs propres d'une matrice et de changement de base. De plus, le caractère visuel des résultats permet d'appréhender de manière directe les résultats. Cette méthode a aussi l'avantage d'être **flexible**, puisqu'elle peut s'appliquer directement sur un ensemble de données quantitatives de contenu et de taille quelconques. Étant une méthode d'analyse de données, l'ACP n'a pas réellement d'inconvénients. La perte d'information liée à cette méthode est nécessaire à l'obtention d'un résultat.

3. Cela est valable dans le cas d'une ACP normée. Dans le cas contraire, les axes dont la variance est supérieure à la variance moyenne sont retenus.

4.1.2 Classification Ascendante Hiérarchique

Principe

La classification ascendante hiérarchique - CAH - est une méthode de classification automatique qui constitue des groupes homogènes en se basant sur une notion de distance [34]. Cette méthode cherche à regrouper les individus les plus semblables possibles dans des classes qui sont les plus dissemblables possibles. Cela introduit les notions d'homogénéité intraclasse et d'hétérogénéité interclasse. Cette classification est dite ascendante car elle part des observations individuelles, tandis qu'elle est dite hiérarchique car elle produit de moins en moins de classe, en incluant des sous-classes en leur sein.

Notion de distance

La proximité entre deux observations est calculée par une distance. Puisque les regroupements seront effectués en fonction de cette distance, la définition de celle-ci est importante. Plusieurs indices de distances existent (distance moyenne, distance maximum ou minimum...) et la classification retenue peut être sensiblement différente d'un indice à l'autre si les classes ne sont pas bien discriminées dans l'espace dans lequel les individus sont projetés.

La méthode de calcul de la distance utilisée dans cette étude correspond à la méthode de « variance » ou de « Ward ». Celle-ci définit la distance entre classes comme suit :

$$d_w(h_1, h_2) = \frac{card(h_1)card(h_2)}{card(h_1) + card(h_2)} d(g_{h_1}, g_{h_2})^2$$

avec g_{h_i} le centre de gravité de la classe h_i, et $d(g_{h_1}, g_{h_2})$ la distance euclidienne entre les deux centres de gravité, telle que

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Cette méthode permet de regrouper les classes de telle sorte que la variance interclasse soit maximale et que la variance intraclasse soit minimale.

Algorithme

À partir de l'indice de distance retenu, les distances entre tous les individus sont calculées. Puis, à chaque itération, un regroupement de classes d'individus est effectué en minimisant la distance intraclasse et en maximisant la distance interclasse. Le résultat de la classification hiérarchique est un arbre représentant les relations entre les différentes classes et sous-classes. Cet arbre appelé dendrogramme, permet de synthétiser les agrégations effectuées en rendant compte des classes obtenues à chaque itération. Celui-ci permet de définir les classes obtenues en fonction du nombre de classes souhaité.

Les avantages et les inconvénients de la méthode

Cette méthode présente comme avantage principal le fait de permettre, sans *a priori* de classe, d'exprimer la proximité entre individus. À partir du dendrogramme obtenu, il est possible

d'observer, pour n'importe quel nombre de classes, la segmentation associée. Cependant, le temps exécution de cet algorithme augmente avec le nombre d'individus.

4.1.3 *Kmeans*

Principe

Le regroupement *Kmeans* est un type d'apprentissage non-supervisé dont l'objectif est de trouver K groupes dans les données [34]. Une classe est représentée par son centre de gravité, et une observation appartient à la classe dont le centre de gravité lui est le plus proche. La distance euclidienne est utilisée comme mesure de proximité.

Algorithme

L'algorithme utilise une seule information *a priori* : le nombre de classes K. Une fois ce dernier fixé, l'algorithme fonctionne comme suit :

1. une initialisation des centres est réalisée de manière aléatoire. Ainsi, K individus sont tirés au hasard et représentent le centre de gravité d'une classe ;
2. pour chaque observation, une affectation à la classe qui minimise la distance entre le centre de gravité et l'observation est effectuée ;
3. de nouveaux centres de gravité sont calculés à partir de ces classes ;
4. les étapes **2** et **3** sont réitérées jusqu'à obtenir une stabilité des groupes.

Les avantages et les inconvénients de la méthode

L'avantage principal de cette méthode est sa capacité à traiter des volumétries importantes. En effet, puisqu'il n'y a pas de calcul des distances deux à deux des individus, la méthode est plus rapide. Cet algorithme est cependant utilisable uniquement lorsque la notion de moyenne existe. De plus, il est sensible à l'initialisation aléatoire des centres de gravité.

4.2 Apprentissage supervisé : cas de la régression binomiale

4.2.1 Notations et hypothèses

La régression binomiale est adoptée dans un contexte de classification binaire. Elle suppose donc qu'il existe seulement deux issues à une variable catégorielle. Les notations ci-après seront utilisées dans cette étude pour n individus et p variables explicatives (n et $p \in \mathbb{N}^*$).

Y	La variable expliquée qui ne prend que deux modalités (0 et 1)
X_i	La $i$ème variable explicative ($i \in \{1, \ldots, p\}$)
π	La probabilité d'un succès telle que $\pi = P(Y = 1)$
h	La fonction de lien
β_i	le coefficient réel associé à la variable explicative X_i
ϵ	La composante aléatoire du modèle aussi appelée résidu

Le modèle linéaire généralisé s'écrit comme suit pour p variables explicatives :

$$h(E[Y|X]) = \beta_0 + \beta_1 \times X_1 + \cdots + \beta_p \times X_p + \epsilon$$

Dans le cas du modèle de régression binomiale, l'hypothèse principale est que la variable à expliquer Y sachant les variables explicatives X suit une loi binomiale $\mathbb{B}(\pi)$. Ainsi, puisque :

$$E[Y|X] = 1 \times P(Y = 1) + 0 \times P(Y = 0) = \pi$$

le modèle de régression binomial est le suivant :

$$h(\pi) = \beta_0 + \beta_1 \times X_1 + \cdots + \beta_p \times X_p\epsilon$$

Le choix de la fonction lien h est donc important. Dans ce cas ci, il doit s'agir d'une fonction telle que :

$$h : [0, 1] \to \mathbb{R}$$

Les fonctions de lien h utilisées en général dans le cadre de la régression binomiale présentent une forme sigmoïdale. Les plus connues et utilisées sont les suivantes :

Lien	*Fonction*	*Spécificité*
probit	$h(\pi) = \phi^{-1}(\pi)$ avec $\phi(\pi) = \int_{-\infty}^{\pi} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt$	Expression non explicite
log-log	$h(\pi) = \ln(-\ln(1-\pi))$	Fonction dissymétrique
logit	$h(\pi) = \ln \frac{\pi}{1-\pi}$	Expression explicite

Table *4.1:* Présentation des principales fonctions de lien utilisées dans la régression binomiale

Les liens **logit** et **probit** étant très proches dans les prédictions, l'utilisation du lien **logit** est surtout motivée par son caractère explicite. En effet, les coefficients estimés sont d'une part, directement interprétables, et d'autre part, permettent de calculer directement l'estimation $\hat{\pi}$ de la probabilité de succès, à partir des estimations $\hat{\beta}_i$ de β_i. La quantité suivante :

$$\frac{\pi}{1-\pi} = \exp(\beta_0 + \beta_1 \times X_1 + \cdots + \beta_p)$$

exprime un rapport de chances aussi appelé *odds*. Par exemple, lorsqu'un individu présente un *odds* de trois, cela signifie qu'il a trois fois plus de chances d'avoir un succès. C'est en cela que chaque coefficient est directement interprétable.
De plus, la probabilité estimée $\hat{\pi}$ se calcule comme suit :

$$\hat{\pi} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 \times X_1 + \cdots + \hat{\beta}_p \times X_p)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 \times X_1 + \cdots + \hat{\beta}_p)}$$

Les paramètres du modèle binomial sont généralement estimés en maximisant la log-vraissemblance du modèle. Soit $\log(L_M)$ la log-vraissemblance du modèle.

$$\log(L_M) = \sum_{i=1}^{n} Y_i \times \ln(\pi_i) + (1 - Y_i) \times \ln(1 - \pi_i)$$

où n désigne le nombre d'individus de l'échantillon de modélisation, Y_i la réponse binaire de l'individu i et $\pi_i = P(Y_i = 1)$.

4.2.2 Qualités du modèle

Il existe plusieurs manières de juger de la qualité d'un modèle binomial. Les critères retenus dans cette étude sont développés ci-après.

AIC

Le critère d'information d'Akaike - AIC - s'applique aux modèles estimés par une méthode du maximum de vraisemblance. Il est défini par la quantité :

$$AIC = 2 \times \log(L_M) + 2 \times k$$

avec L_M la vraisemblance du modèle et k le nombre de paramètres dans le modèle. Plus elle est faible, meilleur est le modèle. Cette valeur représente un compromis entre le biais du modèle qui diminue avec le nombre de paramètres et la simplicité du modèle (le plus petit nombre de paramètres possible).

Pseudo R^2 de MacFadden

Le pseudo R^2 de MacFadden a pour objectif de mettre en évidence une régression inadaptée, c'est-à-dire lorsque les variables explicatives n'expliquent rien. Cet indicateur prend la valeur 1 lorsque la régression est parfaite et 0 dans le cas contraire. Les ouvrages de littérature [27] suggèrent que le R^2 de MacFadden est le plus proche, dans son concept, du coefficient de détermination de la régression linéaire simple. Il se calcule comme suit :

$$R^2_{MF} = 1 - \frac{\log(L_M)}{\log(L_0)}$$

avec $\log(L_0)$ la log-vraissemblance du modèle trivial [4] et $\log(L_M)$ la log-vraissemblance du modèle à qualifier. Lorsque $\log(L_0) = \log(L_M)$ alors $R^2_{MF} = 0$, le modèle n'améliore pas le modèle trivial. Lorsque $\log(L_M) = 0$, alors $R^2_{MF} = 1$, le modèle est parfait.

Courbe ROC et AUC

Une courbe ROC (*Receiver Operating Characteristic*) est une représentation graphique qui représente les performances d'un modèle de classification pour tous les seuils de classification. Elle trace la sensibilité (taux de vrais positifs - TVP -) en fonction de la spécificité (taux de faux positifs - TFP -).

$$TVP = \frac{\text{Nombre de bien classé 1}}{\text{Nombre de 1}} \text{ et } TFP = \frac{\text{Nombre de mal classé 1}}{\text{Nombre de 0}}$$

l'AUC correspond à l'aire sous la courbe ROC. Plus cette aire est proche de 1, meilleur est le modèle. Plus cette valeur est proche de 0.5, plus les performances du modèle peuvent être

4. Modèle tel que $\beta_i = 0 \; \forall i \neq 0$

assimilées au hasard. L'AUC mesure les qualités prédictives du modèle, quel que soit le seuil de classification sélectionné.

Score de Brier

Le score de Brier est une mesure de la performance globale du modèle. Pour cela, il compare les probabilités estimées à la réponse réelle. Ce score se définit comme suit :

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\pi}_i - Y_i)^2$$

Un score de Brier proche de 0 signifie que les probabilités sont correctement calibrées.

5 Théorie de la prévention

« Les individus investissent dans la santé et la prévention pour rentabiliser leur capital humai ». ***Becker***

Ce chapitre pose le cadre théorique dans l'étude de la prévention[1][32][7][31]. L'objectif sera de modéliser le gain utile [1] de l'assureur et de l'optimiser en fonction du niveau de prévention. Pour cela, la modélisation du gain utile est faite en fonction :

- du type de prévention ;
- de l'utilité de l'assureur ;
- des niveaux de prévention retenus.

5.1 Notations, notions et hypothèses

5.1.1 Notations

Dans cette étude, les notations ci-après seront utilisées. À noter que la notation «$'$» fera référence à la dérivée d'une fonction.

w	le montant de la prime à valeurs dans $\mathbb{R}^+$;
e	le niveau de prévention retenu à valeurs dans $\mathbb{N}$ tel que $e = 0$ correspond à l'absence de préventi
$p(e)$	la probabilité d'occurrence d'un sinistre en fonction du niveau de prévention e ;
$p_0 = p(0)$	la probabilité d'occurrence d'un sinistre sans mise en place de prévention ;
$S(e)$	la fonction de coût d'un sinistre en fonction du niveau e de prévention ;
$S_0 = S(0)$	le coût d'un sinistre sans mise en place de prévention ;
$c(e)$	le coût de la mise en place de la prévention de niveau e.

5.1.2 Hypothèses

La modélisation du gain utile de l'assureur sera réalisée selon les hypothèses suivantes [5] :

- Le coût d'un sinistre est réduit par la prévention et diminue avec le niveau de prévention, de sorte que :

$$S(e) < S_0 \ \forall e \in \mathbb{N}^* \text{ et } S'(e) < 0$$

1. Gain qui prend en compte la fonction d'utilité.

- Le coût de la mise en place de la prévention de niveau e n'est pas nul, augmente avec le niveau de prévention, et les coûts de la prévention augmentent rapidement, de sorte que :

$$c(e) > 0 \; \forall e \in \mathbb{N}^* \; , \; c'(e) > 0 \text{ et } c''(e) > 0$$

- La probabilité d'occurrence d'un sinistre en fonction du niveau de prévention e diminue avec la prévention, de sorte que :

$$p(e) < p_0 \; \forall e \in \mathbb{N}^* \text{ et } p'(e) < 0$$

5.2 Notion de fonction d'utilité

En théorie économique, la fonction d'utilité mesure la satisfaction du consommateur qui possède une quantité x d'un bien. On notera $U\{x\}$ l'utilité perçue par le consommateur du bien. L'utilité est associée aux notions :

- d'**ordinalité** : car elle permet d'ordonner les préférences dans un panier de biens ;
- de **cardinalité** : car elle permet de quantifier les préférences dans un panier de biens ;
- de **marginalité** : car elle permet d'évaluer comment évolue la satisfaction globale pour une unité de bien.

L'utilité doit être en mesure de conserver la structure de préférences transitive de l'individu appelé dans ce contexte le **décideur**. Pour ce faire, les équivalences suivantes doivent être respectées :

$$U\{a\} - U\{b\} > 0 \iff a \text{ est préférée à } b$$

$$U\{a\} - U\{b\} = 0 \iff a \text{ est indifférente à } b$$

Les fonctions d'utilité utilisées dans le cadre de cette étude sont croissantes car elles cherchent à modéliser l'utilité perçue du gain de l'assureur. Cette utilité croît forcément puisque l'objectif de l'assureur reste le profit et plus la quantité d'un bien est importante, plus la satisfaction de l'individu sera grande. Dans notre cas, on distinguera l'indice d'utilité $IU\{x\}$ de la fonction d'utilité $U\{x\}$. L'indice et la fonction d'utilité sont tels que, pour une quantité maximale Q d'un bien,

$$IU :]-\infty, Q] \to]-\infty, 1]$$

$$\begin{cases} \text{IU}\{0\} &= 0 \\ \text{IU}\{\text{Q}\} &= 1 \end{cases} \text{ et } U\{x\} = Q \times IU\{x\}$$

La quantité Q est à déterminer et correspond à la quantité de bien attendu par le décideur. Elle sera appelée la **valeur attendue** par le décideur.

5.2.1 Typologie

Il existe trois formes d'indice d'utilité. Ils peuvent être de type concave, convexe ou linéaire. La figure 5.1 résume les caractéristiques de ces formes.

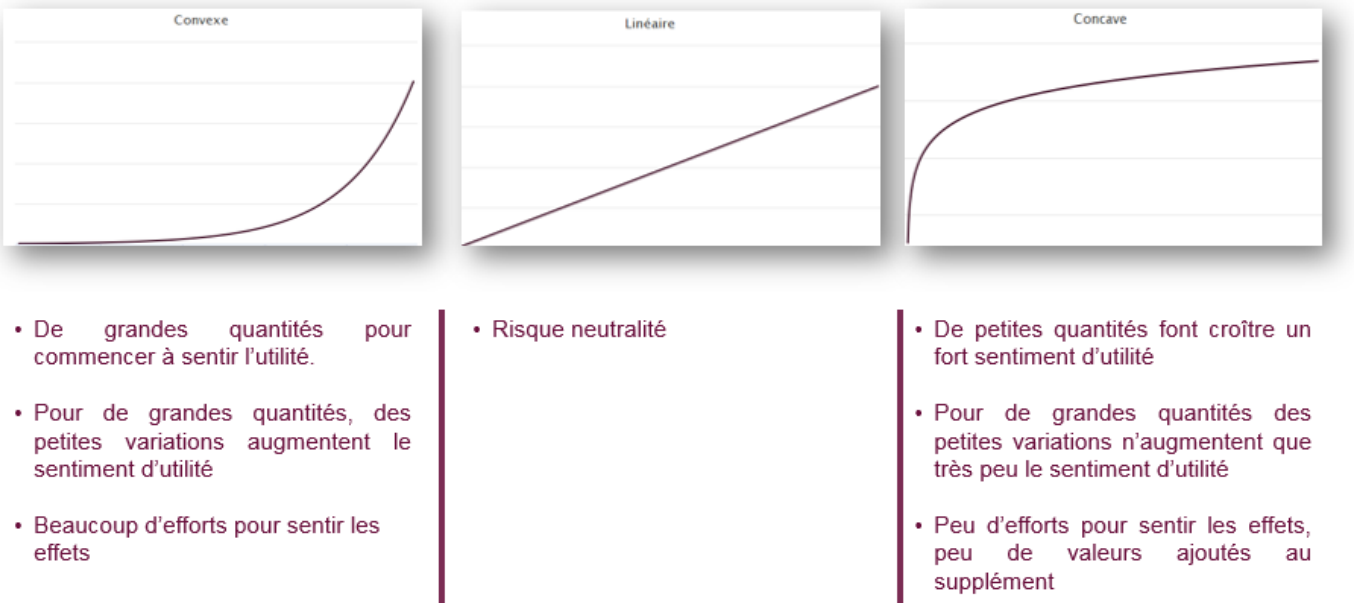


Figure *5.1:* Les trois types de d'indice d'utilité

Sur les fonctions d'utilité convexes, un phénomène d'**explosion** est observable. Cela correspond au fait que l'utilité marginale augmente fortement après un certain niveau. Ce type d'indice d'utilité correspond à un caractère irrationnel marqué par une forte appétence au risque.

Sur les fonctions d'utilité neutres au risque ($U\{x\} = x$), l'utilité marginale est constante.

Sur les fonctions d'utilité concaves, un phénomène de **saturation** est observable. Cela correspond au fait que l'utilité marginale est très faible après un certain niveau. Ce type d'indice d'utilité rend compte d'un comportement prudent, dit averse au risque.

La fonction d'utilité est cependant généralement supposée croissante et concave en chacun de ses arguments. Dans le cadre de cette étude, cela est d'autant plus vrai car elle met en évidence le caractère prudent de l'assureur.

5.2.2 Construction d'une fonction d'utilité

Von Neumann et Morgenstern définissent un axiome selon lequel un individu peut toujours exprimer ses préférences. La construction de fonctions d'utilité va utiliser cet axiome d'ordonnancement des préférences en supposant que l'individu est rationnel. Pour cela, celui-ci répond à une série de questions sur ses choix concernant une quantité donnée d'un bien. Ces questions consistent à faire définir par l'individu les scénarios équivalents pour en déduire des points d'équilibre. Plus concrètement, une série de questions s'en suit pour déterminer l'utilité. Par exemple, les questions suivantes seront posées pour comprendre l'utilité d'un individu sur un jeu simple.

QUESTION 1 : *Un jeu a une chance sur deux de gagner 100 sinon, il ne rapporte rien. En contrepartie de quelle somme ne participeriez-vous pas au jeu ?* La quantité 100 correspond à la valeur attendue **Q**.

Un individu neutre au risque répondrait 50 puisqu'il s'agit de l'espérance mathématique du jeu. Pour un individu plus averse au risque qui répondrait 30 par exemple, cela revient à poser l'équation :

$$\begin{aligned} \text{IU}\{0\} \times 0.5 + \text{IU}\{100\} \times 0.5 &= \text{IU}\{30\} \\ 0 \times 0.5 + 1 \times 0.5 &= \text{IU}\{30\} \\ 0.5 &= \text{IU}\{30\} \end{aligned}$$

QUESTION 2 : *Cette fois, le jeu a une chance sur deux de gagner 30 sinon, il ne rapporte rien. En contrepartie de quelle somme ne participeriez-vous pas au jeu ?*

Pour une réponse de 10,

$$\begin{aligned} \text{IU}\{0\} \times 0.5 + \text{IU}\{30\} \times 0.5 &= \text{IU}\{10\} \\ 0 \times 0.5 + 0.5 \times 0.5 &= \text{IU}\{10\} \\ 0.25 &= \text{IU}\{10\} \end{aligned}$$

QUESTION 3 : *Le jeu a une chance sur deux de gagner 100 sinon, il rapporte 30. En contrepartie de quelle somme ne participeriez-vous pas au jeu ?*

Pour une réponse de 65,

$$\begin{aligned} \text{IU}\{100\} \times 0.5 + \text{IU}\{30\} \times 0.5 &= \text{IU}\{65\} \\ 1 \times 0.5 + 0.5 \times 0.5 &= \text{IU}\{65\} \\ 0.75 &= \text{IU}\{65\} \end{aligned}$$

QUESTION 4 : *Le jeu permet de gagner 10 sinon, il fait perdre 10. Quelle doit être la probabilité de gagner pour que vous participiez à ce jeu ?*

La probabilité énoncée donne la situation d'équilibre (une espérance nulle) perçue pour l'individu. Pour une réponse de 80%,

$$\begin{aligned} \text{IU}\{\text{-}10\} \times 0.2 + \text{IU}\{10\} \times 0.8 &= 0 \\ \text{IU}\{\text{-}10\} \times 0.2 + 0.25 \times 0.8 &= 0 \\ \text{IU}\{\text{-}10\} &= \text{-}\ 1 \end{aligned}$$

À partir des réponses déjà fournies, l'indice d'utilité sera celui de la figure 5.2.

Les questions peuvent être multipliées afin d'affiner l'allure de la fonction. Cet individu dont l'indice d'utilité semble concave pour de faibles quantités est prudent, il sera dit averse au risque. En effet, il sera prêt à perdre statistiquement, c'est à dire qu'il acceptera un montant certain moins élevé que l'espérance statistique. Pour de grandes valeurs, avec les questions posées, il est neutre au risque puisqu'il acceptera un montant certain égal à l'espérance mathématique pour ne pas jouer.

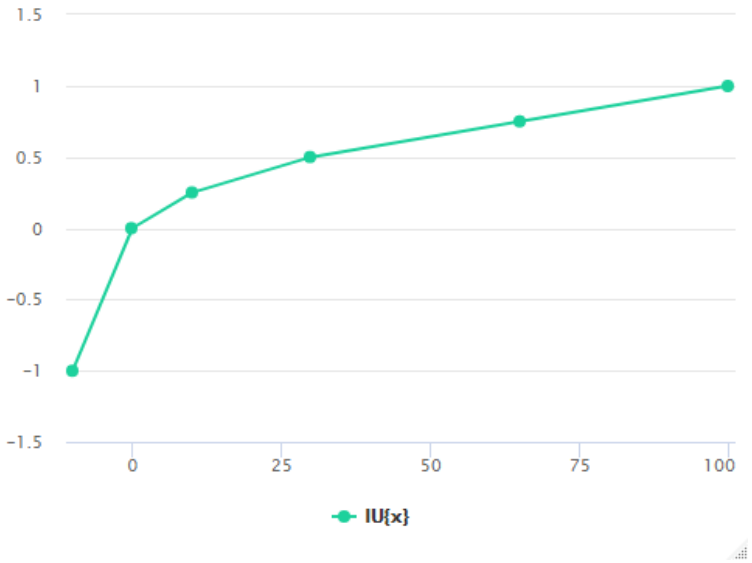


Figure ***5.2:*** Indice d'utilité de l'individu

5.3 Modélisation du gain utile

Selon le type de prévention (primaire ou secondaire) défini dans la section 1.2.1, pour un individu donné, trois espaces probabilisés peuvent être considérés :

- $(\Omega_0, \mathcal{F}_0, \mathbb{P}_0)$, l'espace probabilisé qui décrit les états de l'univers sans prévention avec :

$$\begin{array}{rcl} \Omega_0 & = & \{\text{Sinistre } S_0, \text{ Pas de sinistre }\} \\ \mathcal{F}_0 & = & \{\Omega_0, \emptyset, \text{ Sinistre } S_0, \text{ Pas de sinistre }\} \\ \mathbb{P}_0 & : & \Omega_0 \rightarrow [0,1] \text{ tel que P(Sinistre } \mathrm{S_0}) = p_0 \text{ et P(Pas de sinistre)} = 1 - p_0 \end{array}$$

- $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$, l'espace probabilisé qui décrit les états de l'univers après une prévention primaire avec :

$$\begin{array}{rcl} \Omega_1 & = & \Omega_0 \\ \mathcal{F}_1 & = & \mathcal{F}_0 \\ \mathbb{P}_1 & : & \Omega_1 \rightarrow [0,1] \text{ tel que P(Sinistre } \mathrm{S_0}) = p(e) \text{ et P(Pas de sinistre)} = 1 - p(e) \end{array}$$

- $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$, l'espace probabilisé qui décrit les états de l'univers après une prévention secondaire avec :

$$\begin{array}{rcl} \Omega_2 & = & \{\text{Sinistre } S(e), \text{ Pas de sinistre }\} \\ \mathcal{F}_2 & = & \{\Omega_2, \emptyset, \text{ Sinistre } S(e), \text{ Pas de sinistre }\} \\ \mathbb{P}_2 & : & \Omega_1 \rightarrow [0,1] \text{ tel que P(Sinistre S(e))} = \mathrm{p_0} \text{ et P(Pas de sinistre)} = 1 - p_0 \end{array}$$

Cette modélisation peut être résumée pour un niveau de prévention e donné par les paramètres suivants :

Situation	Probabilité d'occurrence	Gain si sinistre	Gain sans sinistre
Sans prévention	p_0	$w - S_0$	w
Prévention primaire	$p(e)$	$w - S_0 - c(e)$	$w - c(e)$
Prévention secondaire	p_0	$w - S(e) - c(e)$	$w - c(e)$

Table *5.1:* Récapitulatif des changements générés en fonction du type de prévention

5.3.1 En prévention secondaire

La prévention secondaire correspond à la réduction des montants des dommages, elle est également appelée adaptation. En fonction du niveau de prévention, l'espérance de gain utile $E[G]_U$ de l'assureur s'écrit comme suit :

$$E[G]_U = f(e) = p_0 \times U\{w - S(e) - c(e)\} + (1 - p_0) \times U\{w - c(e)\} \tag{5.1}$$

À partir de cette formulation de l'espérance de gain utile de l'assureur, il est possible de trouver le niveau de prévention e^* qui maximise ce gain. Le niveau de prévention optimal est tel que :

$$f'(e^*) = 0$$

En développant cette équation on obtient l'expression suivante :

$$c'(e^*)\Big[p_0 \times U'\{w - S(e^*) - c(e^*)\} + (1 - p_0) \times U'\{w - c(e^*)\}\Big] = -S'(e^*) \times p_0 \times U'\{w - S(e^*) - c(e^*)\} \tag{5.2}$$

Cet optimum est un maximum si $f''(e^*) < 0$

Propriété 1

Pour un individu « neutre au risque », sa fonction d'utilité est telle que $\forall x$, $U\{x\} = x$. Dans ce cas, nous avons $U'\{x\} = 1$, l'équation 5.4 devient donc :

$$\begin{aligned} c'(e^*)\Big[p_0 + (1 - p_0)\Big] &= S'(e^*) \times p_0 \\ c'(e^*) &= -S'(e^*) \times p_0 \end{aligned}$$

Propriété 2 (Jullien et al 1999)[16]

Soit U_1 et U_2 deux fonctions d'utilité concaves et croissantes telles que $U_2 = \phi(U_1)$ avec ϕ une fonction concave et croissante. Alors

$$e_1^* < e_2^*$$

5.3.2 En prévention primaire

La prévention primaire correspond à une réduction des probabilités de sinistre, elle est également appelée atténuation. En fonction du niveau de prévention, l'espérance de gain utile $E[G]_U$

de l'assureur s'écrit comme suit :

$$E[G]_U = f(e) = p(e) \times U\{w - S - c(e)\} + (1 - p(e)) \times U\{w - c(e)\} \tag{5.3}$$

À partir de cette formulation de l'espérance de gain utile de l'assureur, il est possible de trouver le niveau de prévention e^* qui maximise ce gain. Le niveau de prévention optimal est tel que :

$$f'(e^*) = 0$$

En développant cette équation on obtient l'expression suivante :

$$c'(e^*)\Big[p(e^*)\times U'\{w-S-c(e^*)\}+(1-p(e^*))\times U'\{w-c(e^*)\}\Big] = -p'(e^*)\times\Big[U\{w-c(e^*))\}-U\{w-S-c(e^*))\}\Big] \tag{5.4}$$

Cet optimum est un maximum si $f''(e^*) < 0$

Propriété 1

Pour un individu « neutre au risque », sa fonction d'utilité est telle que $U\{x\} = x \ \forall x \in \mathbb{R}$. Dans ce cas, puisque $U'\{x\} = 1$, l'équation 5.4 devient ;

$$\begin{aligned} c'(e^*)\Big[p(e^*) + (1 - p(e^*))\Big] &= -p'(e^*) \times \Big[w - c(e^*) - (w - S - c(e^*))\Big] \\ c'(e^*) &= -p'(e^*) \times S \end{aligned}$$

Propriété 2 :(Jullien et al 1999)

Soit U_1 et U_2 deux fonctions d'utilité concaves et croissantes telles que $U_2 = \phi(U_1)$ avec ϕ une fonction concave et croissante. Alors

$$e_1^* < e_2^* \iff p(e_1^*) < p_0$$

5.3.3 Modèle de perception des risques

Le modèle de perception des risques permet d'introduire des déformations du modèle précédent en introduisant des notions de caractère pessimiste ou optimiste [15]. Pour cela, une fonction ψ de transformation des probabilités est introduite. Cette fonction est telle que :

$$\psi : [0,1] \to [0,1] \text{ avec } \begin{cases} \psi(0) &= 0 \\ \psi(1) &= 1 \end{cases}$$

La fonction de transformation des probabilités peut dépendre de facteurs comme le type de risque considéré ou encore des variables émotionnelles de l'assuré. Elle est appliquée afin de déformer la probabilité d'éviter un sinistre $1-p$ par une nouvelle probabilité $\psi(1-p)$ telle que :

- pour un acteur **pessimiste**, $\psi(1-p) < 1-p$, les chances d'éviter le sinistre sont plus faibles que prévues ;

- pour un acteur **optimiste**, $\psi(1-p) > 1-p$, les chances d'éviter le sinistre sont meilleures que prévues ;
- pour un acteur **neutre** dans le, $\psi(1-p) = 1-p$, les chances d'éviter le sinistre sont identiques aux prévisions.

Les nouvelles espérances de gain sont données par :

- en **prévention primaire**

$$E[G]_U = (1 - \psi(1 - p(e))) \times U\{w - S - c(e)\} + (\psi(1 - p(e))) \times U\{w - c(e)\}$$

- en **prévention secondaire**

$$E[G]_U = (1 - \psi(1 - p_0)) \times U\{w - S(e) - c(e)\} + (\psi(1 - p_0)) \times U\{w - c(e)\}$$

Ces déformations permettent d'obtenir de nouveaux niveaux de prévention optimaux qui prennent en compte des scénarii optimistes ou pessimistes.

Conclusion

La théorie de la prévention présentée dans ce chapitre permet de poser un cadre théorique dans lequel il est possible pour l'assureur de mesurer d'une part le gain inhérent à un programme de prévention spécifique, afin de savoir si celui-ci est rentable. D'autre part, cette théorie lui permet d'optimiser son gain en choisissant le niveau de prévention qui lui fournirait un gain optimal au vu de son appétence au risque. Si le coût associé à un niveau de prévention est en général connu, d'autres variables comme l'impact de la prévention sur la sinistralité et la probabilité d'occurrence ne le sont pas. C'est à ce titre que l'introduction des objets connectés peut être utile. En effet, captant le comportement de l'assuré avant et après la prévention, il est possible de percevoir la variation de la sinistralité et de la probabilité d'occurrence en fonction de l'hygiène de vie de l'assuré. Le challenge de cette approche sera de pouvoir associer un comportement-type de l'assuré à un niveau de sinistralité.

Partie III

Traitement des données connectées, et modélisations préliminaires

Dans cette partie, le lecteur pourra se saisir du contexte factuel de cette étude. Il s'agira, dans une approche en deux temps, de se familiariser avec les données utilisées et de comprendre la création et la modélisation des indicateurs qui serviront de base à cette étude. Puis, dans une seconde partie, une modélisation du niveau de sinistralité en fonction du caractère sportif sera réalisée. Toutes les conclusions ne seront applicables que pour la base utilisée et ne pourront en aucun cas être généralisée.
Afin de réaliser cette étude, il est nécessaire de bénéficier d'une base de données prenant en compte :

- des données individuelles issues d'objets connectés ;
- des données de prestations santé associées aux mêmes individus ;
- des données personnelles non mesurables par les objets connectés ;

Face à la difficulté d'obtenir une telle base de données auprès d'organismes qui en dispose, il a été entrepris, dans un cadre illustratif, de construire une telle base. Constituée d'éléments concrets et réels, cette base a été construite de telle sorte que les résultats, bien qu'illustratifs, restent cohérents avec ce qu'il est possible d'obtenir dans un contexte réel. Tout au long de cette étude, il sera important de distinguer deux sources de données qui sont utilisées : la base personnelle de pratique sportive et la base de prestations santé.

L'objectif de cette partie décomposée en deux chapitres sera de pouvoir réaliser toutes les modélisations nécessaires à :

- l'identification des individus cibles dans des démarches de prévention d'une part ;
- et à la mesure de l'impact d'un plan de prévention, d'autre part.

6 Objets connectés : outils de classification des risques

« Nous n'avons pas seulement besoin de données brutes, il nous faut également disposer d'une formulation adéquate. Ceux qui révolutionnent la pensée humaine ne sont pas ceux qui collectionnent le plus d'informations, mais ceux qui conçoivent la trame de nouvelles structures intellectuelles ». ***Stephen Jay Gould***

Dans ce chapitre, l'objectif est l'identification des individus cibles dans des démarches de prévention. Pour cela, la démarche adoptée est de construire des groupes reflétant le caractère sportif des individus du portefeuille. Ces groupes sportifs sont issus de la base personnelle de pratique sportive, appelée **base sport**. Aussi, dans ce chapitre, une analyse de la base à disposition sera effectuée dans un premier temps. Puis, des indicateurs de performance seront créés afin de caractériser les comportements sportifs. Enfin, une classification des individus sera effectuée sur la base des indicateurs précédemment calculés.

6.1 Analyses des données

Cette étape d'analyse de données est une étape importante avant toute modélisation. Elle permet de comprendre les données par leur origine, leur traitement et leur structure.

6.1.1 Origine des données

La **base sport** a été construite à partir d'un formulaire disponible en ligne dont les questions sont présentées en annexe B.1.2. Au 30/06/2019, 201 participations ont été recensée. Il est important de préciser que dans des considérations de protection des données personnelles, la collecte des données a été réalisée de manière anonyme. Le participant au formulaire apporte des réponses personnelles telles que le sexe, l'âge et le poids et importe sur le formulaire des données récupérées depuis son objet connecté. Les données récupérées sont issues à la fois de smartphones, via les applications de santé, et de montres connectées. Il est important de préciser qu'il existe une grande variété dans la provenance des données récupérées. Cela implique des formats et spécificités propres à chaque appareils en termes de nomenclature et de stockage de données et donc requiert un traitement distinct par type d'appareil. La figure 6.1 met en valeur la variété dans l'origine des données.

Enfin, il est à souligner que, par souci de cohérence entre les bases Sport et Sinistre (décrite par la suite), des restrictions ont été réalisées. Ainsi, seuls les individus résidant en France, possédant une complémentaire santé et ayant un historique de données sur l'année 2018 ont été retenus.

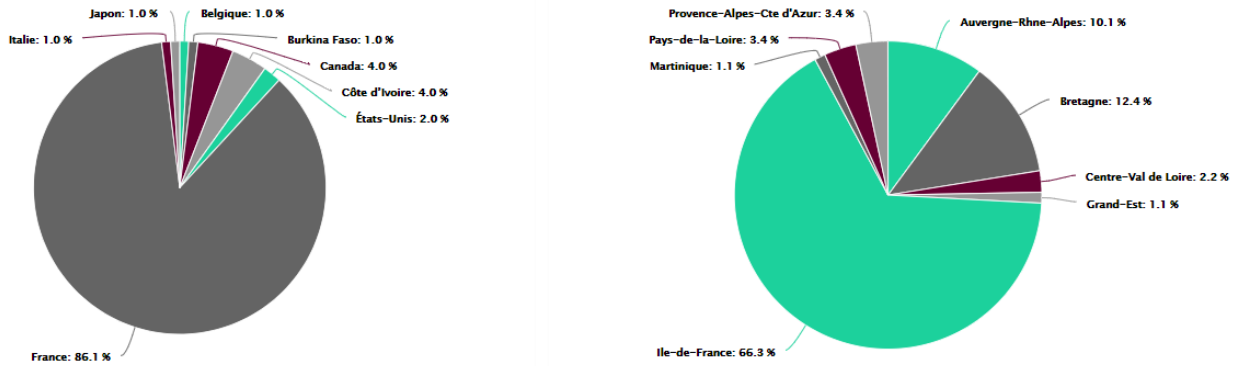


Figure *6.1:* Origine géographique des données

Le formulaire distribué a permis de récupérer deux types d'informations, d'une part les informations génériques du répondant et, d'autre part, les informations issues des données sportives extraites.

- *Les informations génériques du répondant* sont décomposées ainsi :
  - les caractéristiques **physiques** : le sexe, l'âge, le poids et la taille ;
  - les caractéristiques **professionnelles** : la catégorie socioprofessionnelle - CSP - et le secteur d'activité professionnel ;
  - les caractéristiques **familiales** : le statut marital et les enfants à charge ;
  - les caractéristiques **géographiques** : le code postal, le pays [1] ;
  - les caractéristiques de **santé** : le nombre de consultations annuelles chez le généraliste, la possession d'une complémentaire santé et une note d'hygiène de vie attribuée par le répondant lui même ;
  - les caractéristiques **sportives** : les activités sportives réalisées et la fréquence d'activité physique ;
  - les caractéristiques **des appareils d'extraction** : le type d'appareil de récupération des données et la marque de l'appareil.
- *Les informations issues des données sportives extraites* sont les suivantes :
  - le jour considéré ;
  - le nombre de pas ;
  - les calories dépensées ;
  - le type d'activité pratiquée ;
  - la distance parcourue.

Dans l'objectif d'adapter l'étude à n'importe quel jeu de données, une chaîne d'automatisation a été mise en place. Cette chaîne d'automatisation a été matérialisée à l'aide d'une plateforme web : `http://showroom-actuariat.sia-partners.com/shiny/Andrea2/` sur laquelle les principaux indicateurs de suivi sont disponibles. Dans un soucis de respect de la réglementation, notamment le RGPD, les résultats publiés sont agrégés.

1. si hors de France

6.1.2 Contrôles des données et retraitements

Les données de cette étude présentent la particularité d'être peu exhaustives puisque, par exemple, les 23 - 27ans sont sur-représentés. Ceci impacte l'âge moyen qui est donc faible. De plus, le secteur banque/assurance est aussi sur-représenté.
Certaines variables telles que le nombre de calories dépensées et la distance parcourue ont été retirées. Ce retrait est principalement dû à la corrélation trop importante existante entre ces variables et le nombre de pas. De plus, la variable liée au nombre de calories dépensées est très subjective puisqu'elle dépend du métabolisme de l'individu que le constructeur mobile n'a pas nécessairement.
Il est aussi important de noter que, bien que les données issues des smartphones présentent généralement une faible marge d'erreur, les montres connectées se montrent plus fiables.

Les limites et restrictions imposées lors du remplissage des données permettent de limiter le contrôle des données sans pour autant l'exclure. Cette standardisation des entrées permet de focaliser le contrôle sur la cohérence des informations. Un contrôle sur le type d'activité et la fréquence d'activité a été mis en place tel que :

$$\text{activité } \textit{aucune} \iff \sum(\text{autres activités}) = 0 \text{ et Fréquence} = 0$$

De plus, la variable représentant l'Indice de Masse Corporelle - IMC - a été créée.

$$IMC = \frac{Poids(kg)}{Taille(m)^2}$$

6.1.2.1 Des retraitements

Certaines variables telles que l'âge, l'IMC et la variable $Famille$, ont fait l'objet de retraitements. La variable âge est retraitée en quatres catégories.

Figure *6.2:* Histogramme des âges avant et après retraitements

La variable IMC est retraitée selon les paliers Maigre - Normal - Surpoids - Obèse définis par l'OMS. Cette distinction permet de mettre en évidence les individus éventuellement à risque dans le portefeuille.

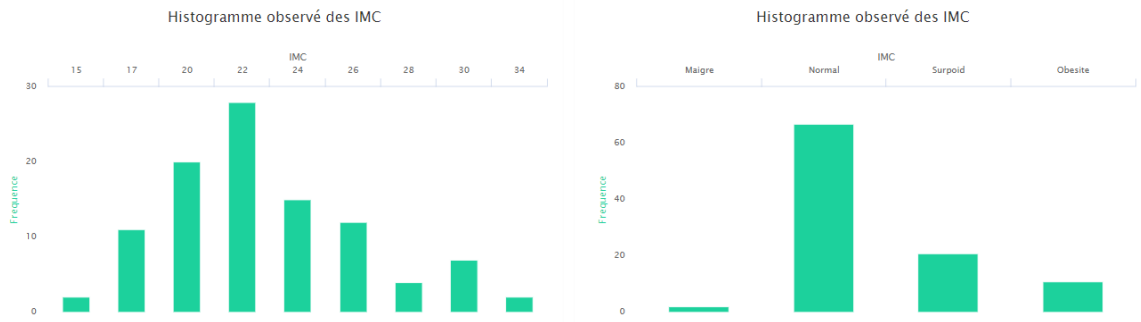


Figure *6.3:* Histogramme des IMC avant et après retraitement

La variable *Famille* est retraitée en deux variables booléennes [2] Célibataire et Enfant.

6.1.2.2 Autres variables

Les individus participants sont issus de catégories socioprofessionnelles et de secteurs d'activités variés. Cependant, ce portefeuille est caractérisé par une forte proportion d'étudiants et d'individus du secteur "Banque/Assurance".

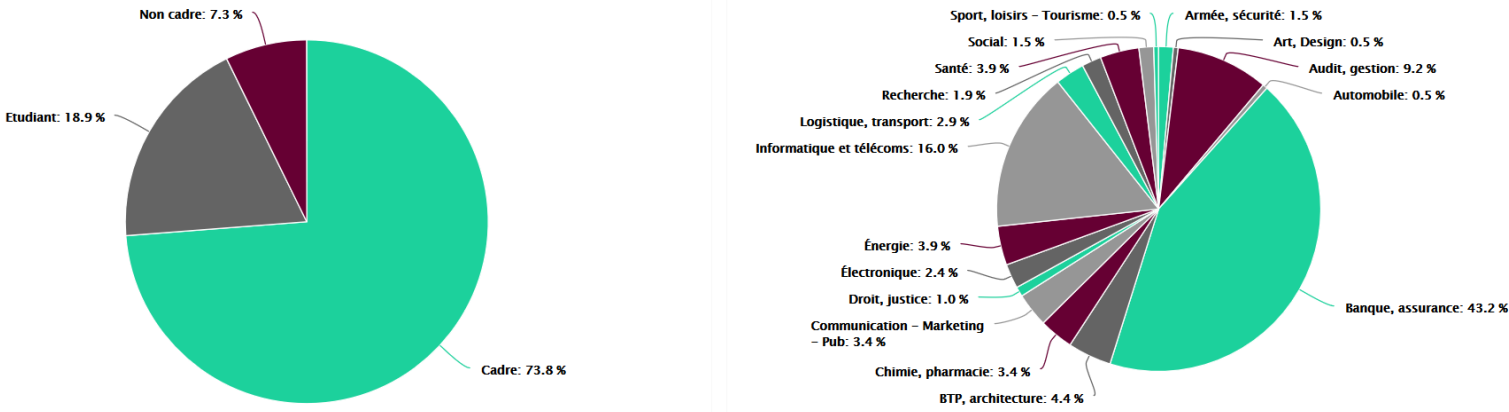


Figure *6.4:* Répartition des CSP et des secteurs d'activités professionnelles

Il faut aussi noter que ce portefeuille possède une répartition quasi-paritaire des sexes avec une proportion d'hommes à 52.7% et de femmes à 47.3%

6.1.3 Analyses bivariées

La figure 6.5 met en évidence que ce portefeuille est constitué de femmes globalement plus jeunes que les hommes. De plus, celles ci s'octroient une meilleure hygiène de vie, pour un nombre de consultations légèrement plus élevé et un IMC plus dispersé.

La figure 6.6 ne peut laisser place à une interprétation sur la modalité "Non cadre" puisqu'elle est mal représentée dans ce portefeuille. Cependant, les "Cadres" s'octroient tout de même une hygiène de vie plus saine que celle des "Étudiants".

2. Variables qui ont comme modalités : *TRUE* et *FALSE*.

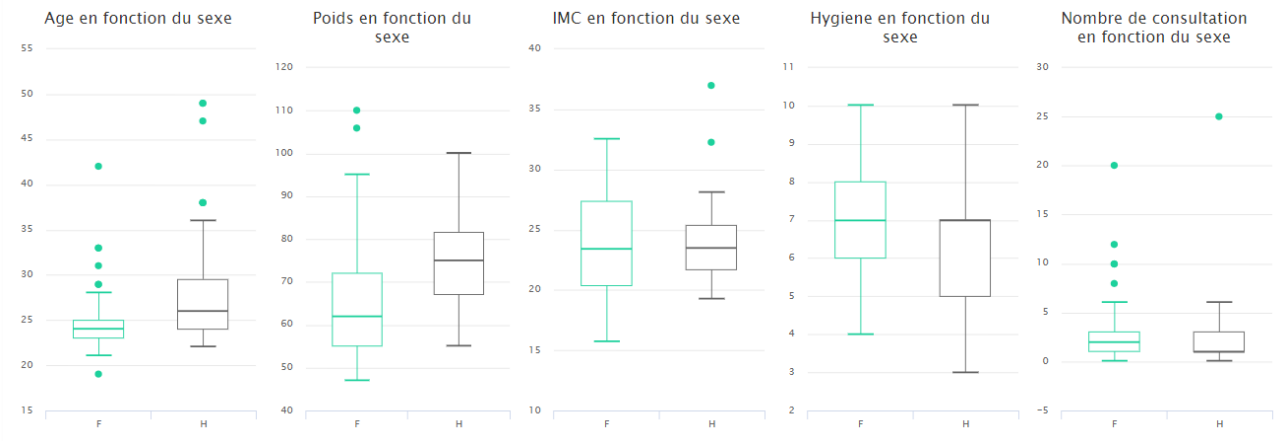


Figure *6.5:* Analyse bivariée des sexes

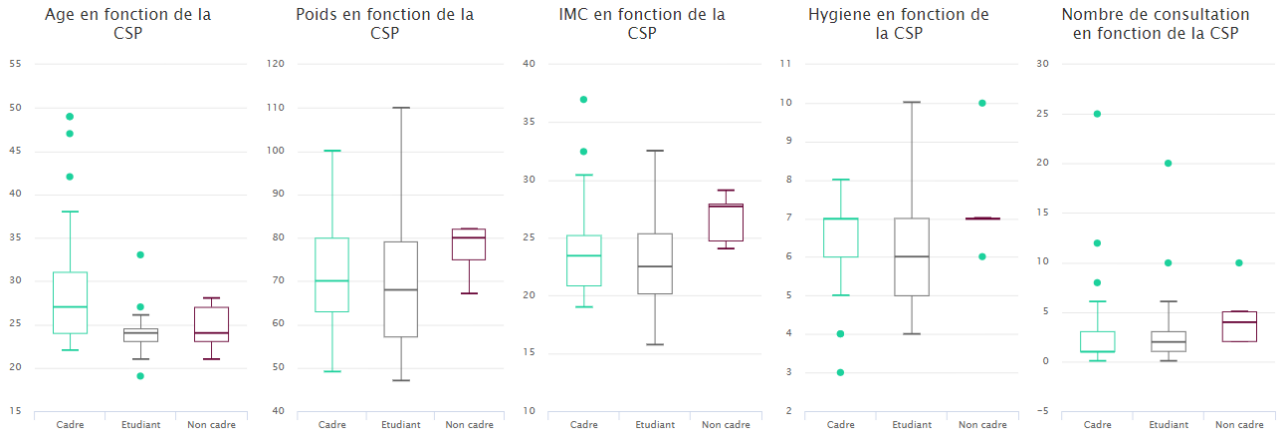


Figure *6.6:* Analyse bivariée des CSP

6.1.4 Corrélations

Il sera important ici de distinguer le cas des variables qualitatives et quantitatives.

Variables quantitatives

La figure 6.7 reprend les résultats de corrélations linéaires observées sur les variables quantitatives de la base de données sportives. Les corrélations de Spearman, Kendall et Pearson, explicitées en annexe A.2 sont ici représentées.
Cette analyse permet de mettre en évidence :

- des corrélations évidentes comme celles existant entre le poids, la taille et l'IMC ;
- des corrélations négatives faibles selon Pearson entre la fréquence d'activité et l'âge. Dans ce groupe d'individus, plus l'individu est âgé, plus il exercerait une activité physique à une fréquence plus faible ;
- des corrélations positives faibles entre la fréquence d'activité et la note d'hygiène de vie. Dans ce groupe, l'attribution d'une note d'hygiène de vie prendrait en compte la régularité de la pratique sportive.

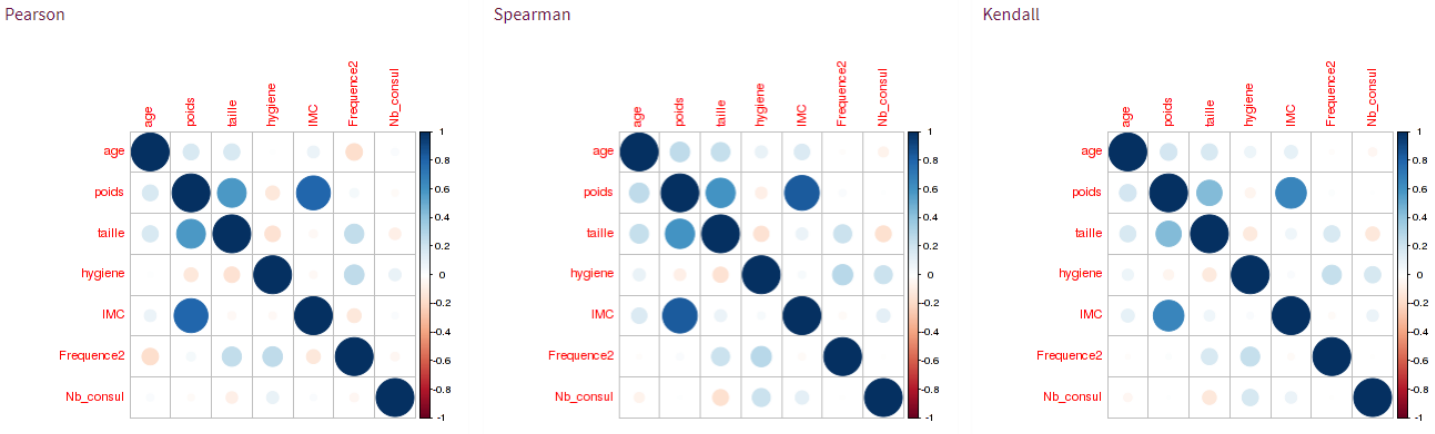


Figure *6.7:* Corrélations observées sur les variables quantitatives

Variables qualitatives

La figure 6.8 reprend les résultats de corrélations linéaires observées sur les variables qualitatives de la base de données sportives. Ces corrélations sont observées en utilisant les méthodes du test du chi-deux et le V de Cramer dont les spécificités sont explicitées en annexe A.2.2,[2].

Le test du chi-deux est un test qui admet une hypothèse d'indépendance entre deux variables. Cette hypothèse est rejetée au niveau 95% lorsque la *p-value* du test est inférieure à 5%. Plus la *p-value* est grande, moins il est possible de rejeter l'hypothèse d'indépendance des deux variables. Ce test statistique est cependant très sensible au nombre d'observations et au nombre de modalités des variables qualitatives à tester. Ainsi, plus grand est le nombre d'observations ou de modalités, plus élevée sera la valeur de la statistique sans qu'il n'y ait nécessairement indépendance entre les variables.

Le V de Cramer quant à lui a un coefficient qui se base sur la statistique du χ^2 et varie entre 0 et 1. Plus il est proche de 0 moins les deux variables sont corrélées ; plus il est proche de 1 plus une forte corrélation entre les deux variables est détectée.

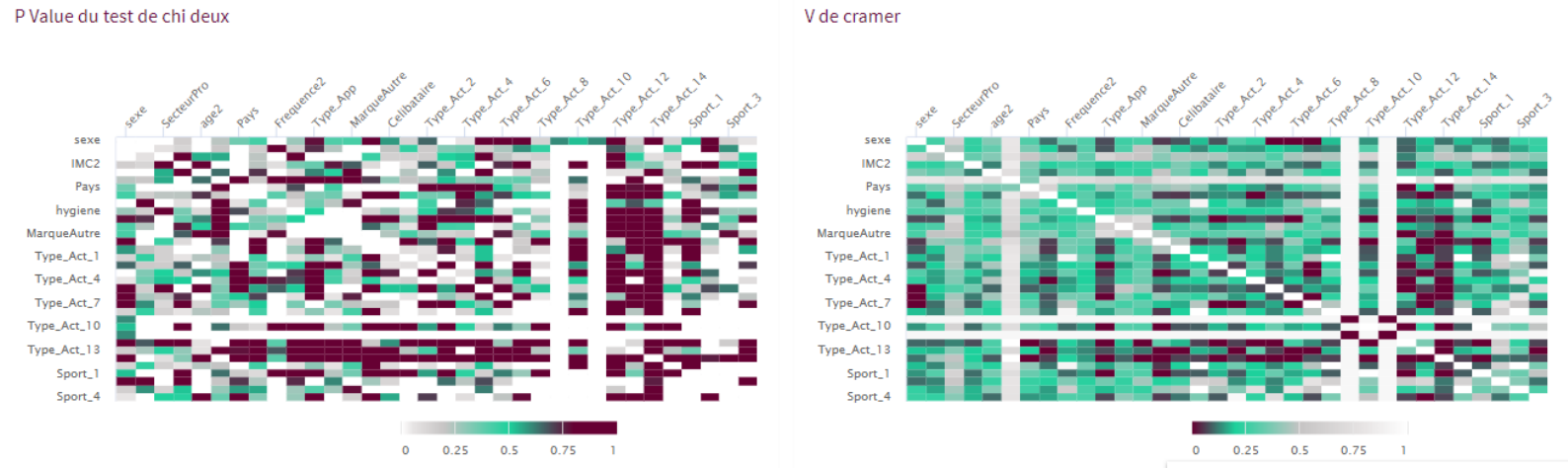


Figure *6.8:* Corrélations observées sur les variables qualitatives

Dans le cas de cette étude des corrélations sont observées entre :

- la fréquence d'activité et l'activité type 7 (musculation) ;
- l'activité type 7 (musculation) et l'activité type 8 (sport de forme/fitness) ;
- le code postal et la complémentaire santé puisque systématiquement, les répondants étrangers (Code postal = 99999) n'ont pas de complémentaire santé ;
- la catégorie d'âge ($age2$) et le type d'activité 4 (sport de raquette).

Il faut cependant relever que certaines variables comme le type d'activité (9 à 15) [3] ont des corrélations apparentes qui ne sont pas significatives puisque ces activités sont peu pratiquées par les participants.

Cette section a permis de se familiariser avec les données récoltées constituant la **base sport**. Fort de cette connaissance, l'étape suivante consistera à créer des indicateurs individuels de performance reflétant le caractère sportif.

6.2 Indicateurs de performance

Dans cette section, plusieurs indicateurs sont créés dans l'objectif de mesurer, à partir des données à disposition, des comportements sportifs.

6.2.1 Indicateur individuel de performance : la note d'hygiène de vie

Cet indicateur est demandé aux participants afin de capter les paramètres en dehors de l'activité physique qui ne seraient pas mesurables par le biais des objets connectés dédiés au sport. Cette note englobe par exemple la nutrition, le sommeil ou encore la consommation d'alcool et de tabac. Cet indicateur reste cependant subjectif puisqu'il va dépendre de la pondération accordée par chaque individu aux différentes composantes de son hygiène de vie. Ainsi, un individu pourrait s'accorder une note de 8 par exemple parce qu'il fait du sport et mange correctement, alors qu'il consomme beaucoup d'alcool et du tabac. Tandis qu'une autre personne ayant le même rythme de vie peut s'accorder une note plus basse parce qu'elle considère que la consommation d'alcool et de tabac est plus importante dans la mesure de l'hygiène de vie.

Pour cela, tout au long de cette étude, l'indicateur "note d'hygiène de vie" ne sera pas utilisé dans la modélisation. Cependant, cet indicateur pourra être utilisé afin de comparer la perception individuelle et la classification factuelle. Un retraitement de cette variable est tout de même effectué (figure 6.9).

Figure *6.9:* Répartition de la variable note d'hygiène de vie avant et après retraitement

3. Les équivalences de notations des activités physiques sont disponibles en annexe à la table B.1.

6.2.2 Indicateur individuel de performance : L'activité sportive

En remplissant le formulaire d'activité physique, les participants peuvent renseigner le type d'activité physique qu'ils pratiquent ainsi que la fréquence. Dans un objectif de hiérarchisation des comportements sportifs, il sera important par la suite de créer des classes d'activités physiques et un indice de fréquence d'activité.

6.2.2.1 Classes d'activités physiques

La construction de ces classes d'activités physiques est réalisée en plusieurs étapes.

Étape 1 : Le regroupement des activités sportives

Une activité sportive est valorisée dans cette étude par la dépense énergétique qu'elle favorise. Ainsi, à partir du tableau en annexe B.2.2, trois groupes d'activités peuvent être formés. Cette équivalence de dépense énergétique a été validée par un médecin du sport.
Ces groupes d'activités sont créés en fixant des paliers de dépense énergétique. Ces paliers sont fixés afin de construire des groupes équirépartis.
Cette classification permet de passer de quatorze activités définies dans le formulaire, à cinq catégories d'activités physiques. Il faut noter que la catégorie $Sport_0$ correspond à une catégorie d'inactivité physique.

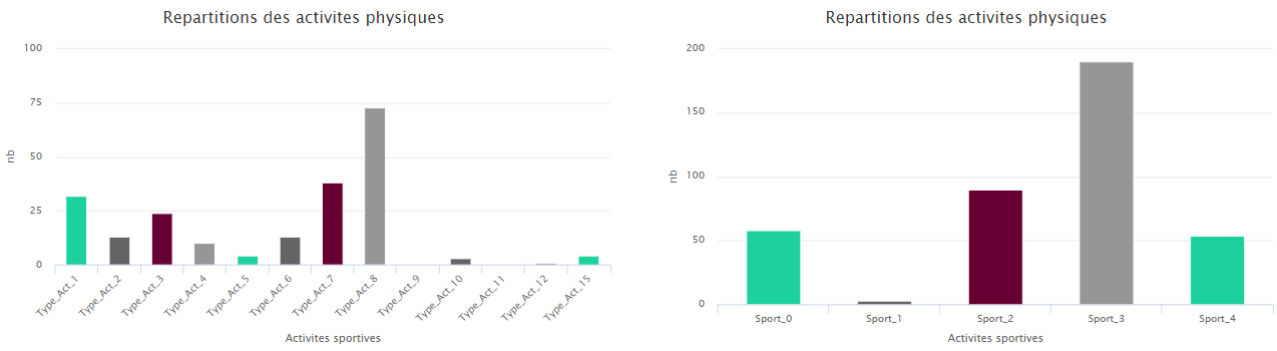


Figure *6.10:* Répartition des différents sports en catégories sportives

Étape 2 : le score d'activité physique

Le score d'activité a pour but de regrouper les informations d'activités sportives réalisées. Pour cela, l'approche retenue de construction du score correspond à une pondération du nombre d'activités d'une catégorie sportive en fonction de la dépense moyenne en calories de cette catégorie. Les coefficients de pondération utilisés sont ceux de la table 6.1. Ils correspondent à la dépense moyenne en quinze minutes pour une activité physique de la catégorie désignée. Ces moyennes sont calculées sur la base du tableau des équivalences de l'annexe B.2.2.

Catégorie sportive	Moyenne de calories dépensés en 15min
Sport_0	0
Sport_1	35
Sport_2	65
Sport_3	105
Sport_4	175

Table *6.1:* Coefficients de pondérations utilisées

Cela permet de constituer le score sportif de la manière suivante :

$$Score\_sport = Sport_1 \times 35 + Sport_2 \times 65 + Sport_3 \times 105 + Sport_4 \times 175$$

Cette formulation permet de décrire le niveau d'activité sportive des individus en hiérarchisant les pratiques sportives en fonction de leurs apports en termes de dépenses énergétiques.

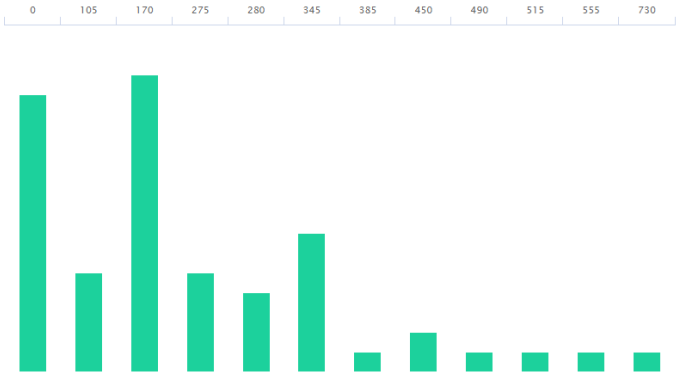


Figure *6.11:* Distribution des *scores_sport*

Étape 3 : La constitution de groupe

La constitution des groupes en fonction de l'activité physique est réalisée en constituant des paliers d'activité. Pour cela, à partir de la distribution des *scores_sport* de la figure 6.11, des paliers permettant de capter les différents types de comportement sont fixés. Les paliers utilisés sont ceux du tableau 6.2. Ces paliers ont été choisis afin d'obtenir des classes ayant des proportions les plus homogènes possible.

Seuil du palier	Groupe	Nombre de personne dans le groupe
[0,170]	Grp_0	38
]170,300]	Grp_1	55
]300,500]	Grp_2	26
>500	Grp_3	38

Table *6.2:* Tableau des correspondances entre seuil de palier et groupe actif

Dans les groupes formés uniquement sur la base de l'activité physique, il est possible d'observer des comportements spécifiques en matière de nombre de pas.

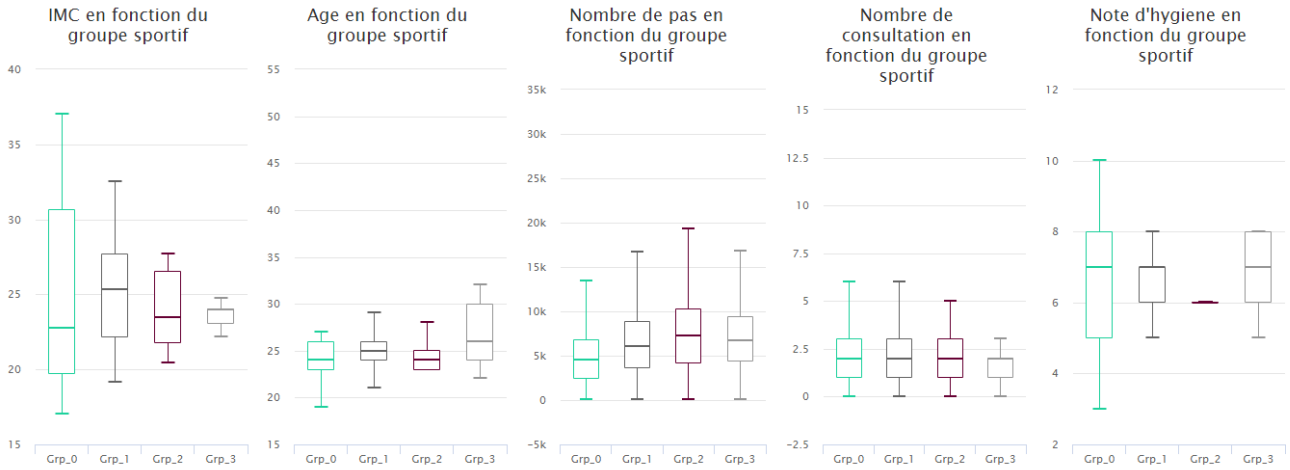


Figure *6.12:* Box-plot de différentes variables quantitatives en fonction du groupe

Cette représentation permet de mettre en évidence les points suivants :

- les groupes Grp_2 et Grp_3 ont des IMC moins dispersés et un IMC en moyenne plus faible ;
- l'âge moyen des personnes du groupe Grp_3 est significativement supérieur aux autres ;
- le nombre de pas augmente en fonction de l'activité physique. Cependant, cette évolution est moins marquée entre les groupes Grp_2 et Grp_3 ;
- le nombre de consultations est moins dispersé et plus faible pour les individus du groupe Grp_3 ;
- la note d'hygiène de vie est en moyenne plus élevée chez les individus du groupe Grp_3 et identique (hygiène = 6) pour les individus du groupe Grp_2.

6.2.2.2 Indice de fréquence d'activité physique

L'indice de fréquence d'activité physique permet de hiérarchiser la fréquence d'activité physique en fonction de son impact sur la santé. Partant du constat qu'une activité physique régulière a un plus fort impact sur la santé qu'une activité occasionnelle, l'indice de fréquence est calculé comme suit [4].

$$Frequence = 3\times4\times I_{\text{plusieurs fois par semaine}}+1\times4\times I_{\text{une fois par semaine}}+3\times I_{\text{plusieurs fois par mois}}+I_{\text{une fois par mois}}$$

$$I \text{ une indicatrice telle que } I_{\text{plusieurs fois par semaine}} = \begin{cases} 1 & \text{si l'activité est pratiquée plusieurs fois par semaine} \\ 0 & \text{sinon} \end{cases}$$

4. Les pondérations données servent à ramener la fréquence d'activité à une fréquence mensuelle. Une activité marquée comme étant réalisée plusieurs fois par semaine sera considérée comme correspondante à trois fois par semaine et donc $3 \times 4 = 12$ fois par mois

6.2.3 Indicateur individuel de performance : le score de marche

6.2.3.1 Les proportions de temps en régime d'activité

Seront considérés comme des régimes d'activités, les caractères "sédentaire", "peu actif", "modérément actif" , "actif" et "très actif". Ils sont déterminés pour caractériser le niveau d'activité d'une journée en matière de nombre de pas. Pour cela, des paliers d'activités sont définis comme dans le tableau 6.3.

Palier	Caractère
[0 ; 5000]	Sédentaire
]5000 ; 7500]	Peu actif
]7500 ; 10000]	Modérément actif
]10000 ; 12500]	Actif
]12500 ;]	Très actif

Table *6.3:* Paliers d'activités

Ces paliers sont fixés selon l'étude de Tudor-Locke (2004) [33] qui estime le mode de vie d'un individu en fonction de son nombre de pas. Cette étude est souvent reprise dans la littérature pour des études d'impacts de la marche régulière sur la santé.

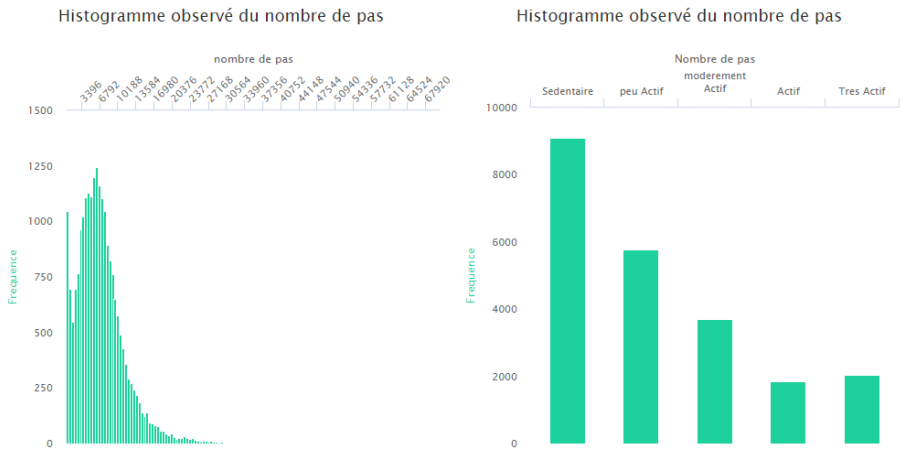


Figure *6.13:* Répartition du nombre de pas et répartition par paliers d'activités

Cette répartition permet de déterminer, pour chaque individu, la proportion du temps passé sur son historique 2018 dans chaque régime d'activité.

6.2.3.2 Score de marche

A partir des proportions de temps passé dans un régime d'activité par individu, il sera déterminé par la suite un score de marche. Ces proportions seront notées :

TActif	Proportion de temps passé en régime "Très actif"
Actif	Proportion de temps passé en régime "Actif"
MActif	Proportion de temps passé en régime "Modérément actif"
PActif	Proportion de temps passé en régime "Peu actif"
Sedentaire	Proportion de temps passé en régime "Sédentaire"

Le score de marche qui est souhaité doit être d'autant plus élevé que le régime d'activité est considéré comme bon. De plus, ce score doit également être pénalisé par une proportion trop importante passée en régime sédentaire. Ce score a donc été construit afin de mettre en valeur les comportements actifs en matière de marche. Pour cela, le score de marche est calculé comme suit :

$$score\_pas = TActif \times 5 + Actif \times 4 + MActif \times 3 + PActif \times 2 - Sedentaire$$

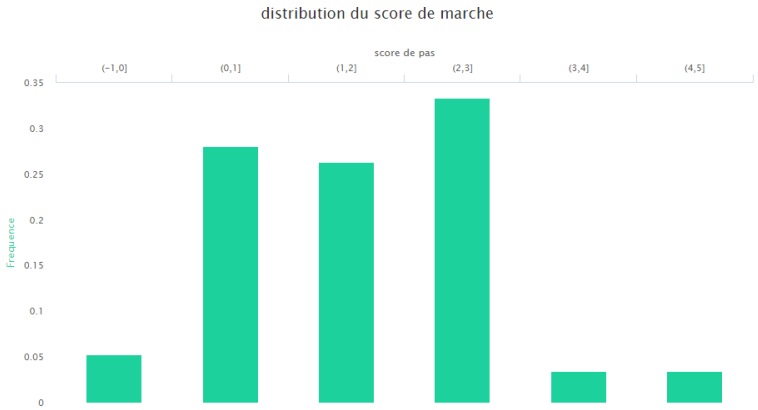


Figure *6.14:* Répartition observée du score de marche

La figure 6.14 met en évidence la répartition du score de marche observée dans le portefeuille.

6.2.4 D'autres indicateurs

Indicateur de stabilité journalière

L'indicateur de stabilité a pour objectif de repérer un comportement stable en matière de marche des individus. Celui-ci sera calculé par la moyenne des écarts entre deux jours consécutifs. Pour un individu j avec n jours d'historique et un nombre de pas Y_i pour le $i^{ème}$ jour d'historique, l'indicateur de stabilité se calcule comme suit :

$$In\_stab_j = \frac{1}{n-1} \times \sum_{i=1}^{n-1} |Y_{i+1} - Y_i|$$

Cet indicateur est calculé sur les jours simples, hors jours fériés et hors week-end. Plus il est faible, plus l'individu est régulier dans son comportement.

Indicateur de dispersion

L'indicateur de dispersion ici retenu est l'écart-type. Cet indicateur mesure la dispersion du nombre de pas journalier autour du nombre de pas moyen. Plus le nombre de pas sera dispersé, c'est-à-dire moins les valeurs seront concentrées autour de la moyenne, plus l'écart-type sera élevé. Il sera noté *ec_type*.

Nombre de pas moyen le week end, en semaine et en jour férié

Ces indicateurs permettent de repérer des comportements éventuellement différents pendant les jours fériés et pendant les week-ends. Ils seront notés :

diff_spe différence entre nombre de pas moyen en jour férié et nombre de pas moyen en jours ouvrés
diff_we différence entre nombre de pas moyen en week end et nombre de pas moyen en jours ouvrés

Avec tous les indicateurs conçus pour décrire le comportement des participants, il est possible de dresser des classes de profils sportifs. Pour cela, il faudra, en premier lieu, étudier les variables retenues afin de comprendre leurs structures et les relations qui les relient [14][20][29].

6.3 Classification

Les variables retenues premièrement dans la classification pour exprimer le comportement sportif sont les variables : *Frequence2, score_sport, In_stab, score_pas, ec_type, diff_spe, diff_we.* Avant tout, une étude de la structure et des liens entre les variables retenues sera réalisée. Ensuite, deux méthodes usuelles de classification seront utilisées afin de créer des regroupements de comportements sportifs. Enfin, il sera important de caractériser les groupes formés.

6.3.1 Les variables de la classification

En étudiant, tout d'abord, les corrélations entre les variables sus-mentionnées, certaines variables peuvent déjà être retirées du processus de classification, lorsqu'elles présentent une corrélation supérieure, en valeur absolue, à 0.5.

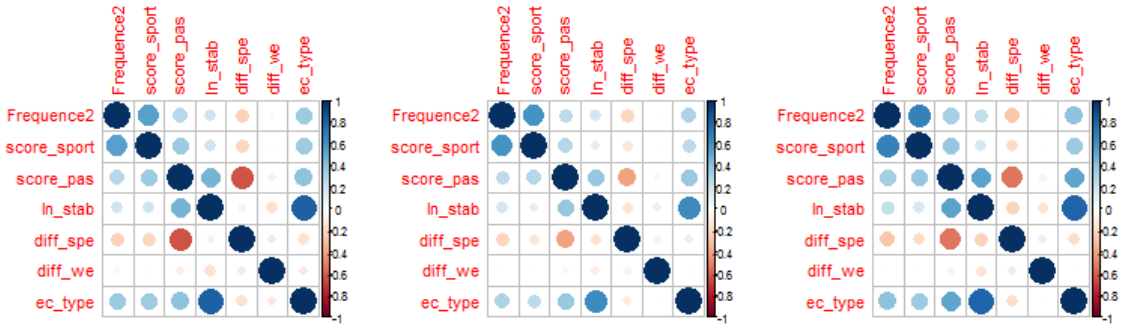


Figure *6.15:* Corrélations de Pearson, Kendall et Spearman

La variable *ec_type* sera considérée comme supplémentaire puisqu'elle est trop corrélée à *Ind_sta.* Afin d'étudier les variables de la classification, une analyse en composante principale - ACP - a été réalisée. Cette analyse permet de condenser l'information quantitative retenue par analyse des corrélations linéaires entre les variables et donne une visualisation graphique des distances entre les individus. Cette analyse permettra de visualiser les liaisons entre variables et la proximité entre individus.

6.3.1.1 Valeurs propres

Dans une ACP, une projection est réalisée sur plusieurs axes. Ici, le choix des axes retenus se fait selon la règle de l'inertie minimale. Cette règle est choisie dans une volonté de conserver au mieux la variance initiale. Les axes sont donc retenus tels que 80% de la variance est expliquée. À partir de la figure 6.16 et au regard de la règle de l'inertie minimale, quatre axes sont donc retenus dans cette étude.

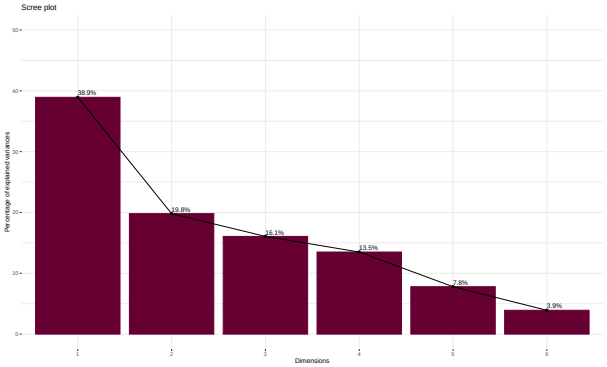


Figure *6.16:* Valeurs propres et pourcentage de variance expliquée

6.3.1.2 Axes et contributions

Sur les 4 axes retenus, la structure de la figure 6.17 est observée.

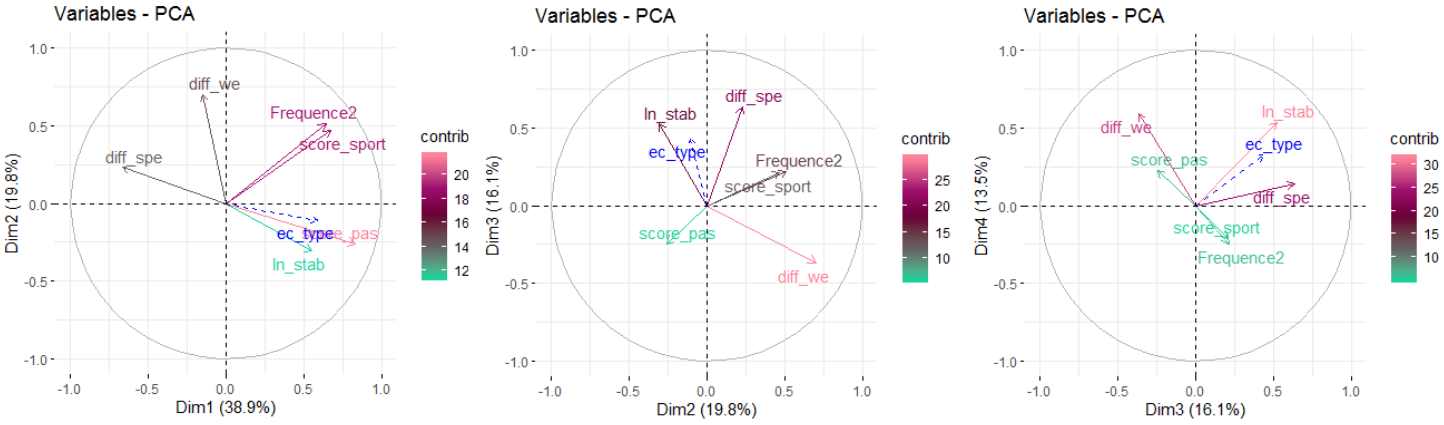


Figure *6.17:* Cercle des corrélations de l'ACP

Axe 1 Cet axe principal explique 39% de la variance totale. Les performances en termes de pas (*score_pas*) et en matière de sport (*score_sport*) contribuent le plus à cet axe.

Axe 2 Cet axe explique 20% de la variance totale. Les écarts entre week end et jours ouvrés en matière de nombre de pas (*diff_we*) contribuent le plus à cet axe.

Axe 3 Cet axe explique 16% de la variance totale. Les écarts entre jours fériés et jours ouvrés en matière de nombre de pas (*diff_spe*) contribuent le plus à cet axe.

Axe 4 Cet axe explique 13% de la variance totale. L'indicateur de stabilité entre jours ouvrés en matière de nombre de pas (*In_stab*) contribue le plus à cet axe.

Il faut noter que, même si l'ACP met en avant une importante corrélation entre les variables *Frequence2* et *score_sport* sur chacun des axes, ces deux variables méritent d'être inclues puisque l'une apporte une information de fréquence d'activité et l'autre d'intensité d'activité. Aussi, cette proximité observée est principalement due au fait qu'une règle d'association existe entre ces deux variables. En effet un individu qui ne pratique aucune activité physique a nécessairement une fréquence nulle. L'analyse par ACP permet de comprendre les structures existantes dans le panel d'individus présents. Une analyse plus approfondie par apprentissage non-supervisé va notamment permettre de délimiter les groupes homogènes grossièrement repérés dans l'analyse par ACP.

6.3.2 *K-means*

Cet algorithme est une méthode d'apprentissage non-supervisé dont les mécanismes sont expliqués dans la section 4.1.3. La base de données utilisée est centrée et réduite avant l'application du modèle. La première étape dans la classification consiste à déterminer le nombre de groupes/*clusters* à former.

6.3.2.1 Choix du nombre de *cluster* et classification

Le nombre de groupes est choisi de telle sorte que la variance intra-classe soit minimisée et que la variance inter-classe soit maximisée. La figure 6.18.a représente l'évolution du paramètre "proportion d'inertie expliquée par la partition" en fonction du nombre de *clusters*. Dans le cas de la base de données utilisée, trois *clusters* seront formés puisqu'en rajoutant des *clusters* supplémentaires, l'inertie expliquée n'augmente pas de manière significative. En effet, constituer plus de *clusters* correspondrait à complexifier la classification tandis que le gain en matière d'inertie expliquée est minime. Pour trois *clusters* sélectionnés, la répartition des individus par *cluster* est celle représentée dans la figure 6.18.b.

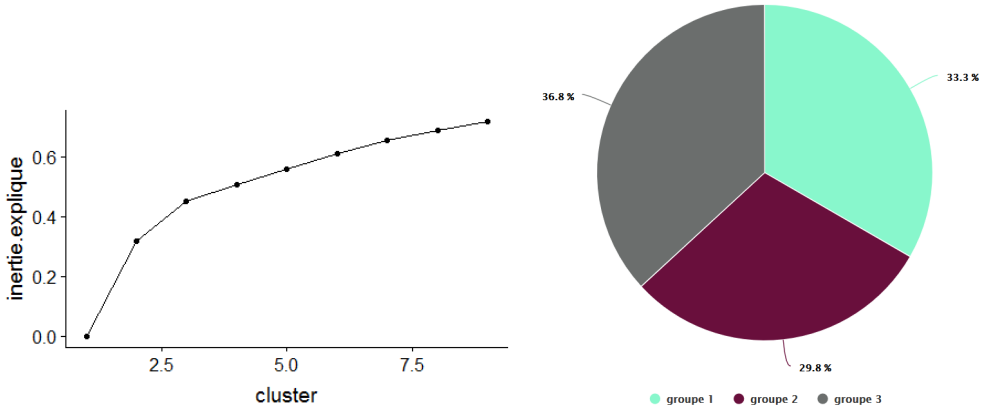


Figure *6.18:* Sélection du nombre de *clusters* (a) et répartition des *clusters* ainsi formés (b)

Pour les différents groupes formés, les caractéristiques moyennes des différents paramètres sont présentées dans la table 6.4. Celle-ci permet de mettre en valeur des caractéristiques principales de chaque *cluster* formé. Les classes semblent avoir, par exemple, des profils distincts en matière de fréquence d'activité sportive (*Frequence_2*) et de score sportif (*score_sport*).

	Frequence2	score_sport	score_pas	In_stab	diff_spe	diff_we
1	10.667	335.714	1.813	3,200.529	-2,110.082	344.022
2	6.471	209.118	2.532	4,068.608	-4,682.601	-2,182.624
3	1.105	50.263	0.766	2,813.595	508.975	-496.489

Table *6.4:* Moyenne des divers paramètres dans les différents *clusters*

6.3.2.2 Caractérisation des groupes

Les *clusters* créés sont à eux seuls peu interprétables. Pour mieux les interpréter, une première approche consiste à projeter les groupes formés sur les axes de l'ACP comme sur la figure 6.19.

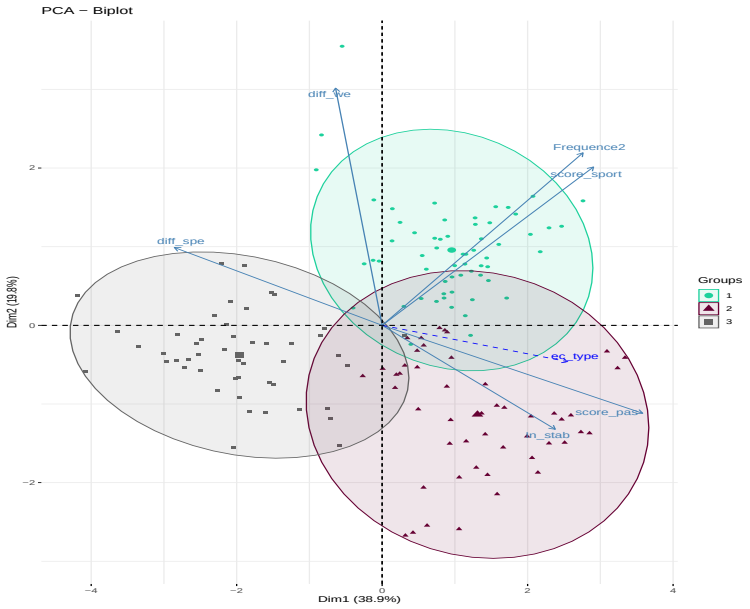


Figure *6.19:* Affichage des groupes formés par le clustering sur les axes 1/2 de l'ACP

Cette visualisation permet de caractériser les groupes tels que :

- le **groupe 1** serait constitué d'individus "sportifs" et ceux ayant un nombre de pas important le week-end ;
- le **groupe 2** serait constitué d'individus "marcheurs" avec un score de pas important ;
- le **groupe 3** serait constitué d'individus sédentaires et sportivement inactifs. Ceux-ci sont aussi caractérisables par leur niveau d'activité qui est significativement plus élevé les jours fériés.

Une autre approche de caractérisation des groupes est une approche par apprentissage supervisé. L'idée première est d'expliquer les *clusters* en fonction des autres paramètres ayant servis à sa création. Pour cela, une approche par arbre de décision permet de retracer le chemin de classification. Grâce à la figure 6.20 la description des groupes peut être affinée.

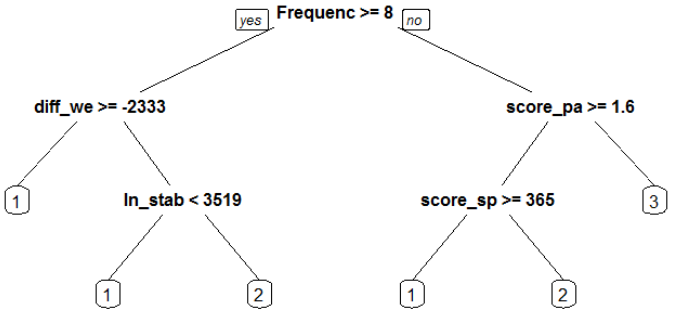


Figure *6.20:* Arbre de décision obtenu pour la classification

- Sont considérés comme des individus du **groupe 1** dans un premier temps les individus "sportifs" avec une fréquence sportive élevée ou un score sportif élevé. Ils ont une meilleure stabilité journalière du nombre de pas.
- Sont considérés comme des individus du **groupe 2** les individus "marcheurs" avec un score de pas important, mais aussi les sportifs avec une moins bonne stabilité journalière du nombre de pas.
- Sont considérés comme des individus du **groupe 3** les individus qui ont une fréquence sportive faible et un score de pas faible. Ce sont des individus fortement sédentaires avec une fréquence d'activité physique faible.

Une approche par *random forest* (annexe A.4) permettant d'obtenir un ordre global d'importance des variables a également été réalisée. Cette importance correspond à une diminution moyenne de l'impureté apportée par chaque variable. Elle est calculée par l'index de Gini [5].

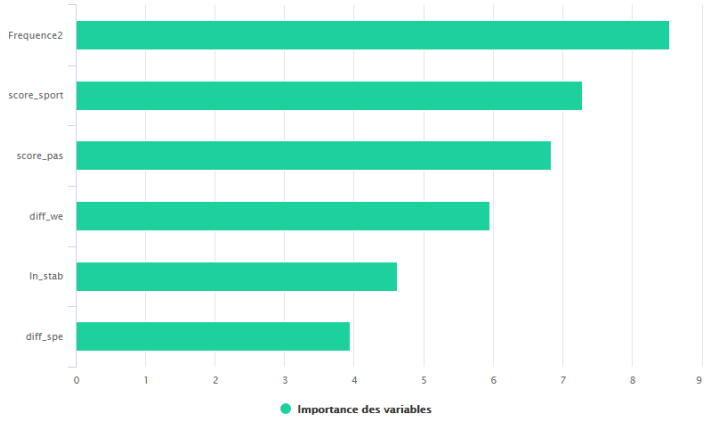


Figure *6.21:* Importance des variables

Cet ordre d'importance présenté dans la figure 6.21 vient confirmer le contenu des groupes et renforcer les conclusions sur la constitution des groupes. Ainsi, l'activité sportive est le premier

5. La diminution cumulée pour chaque noeud est calculée, puis une moyenne sur l'ensemble des arbres est effectuée.

élément retenu dans la classification avant de discriminer sur le comportement en matière de marche.

6.3.3 Classification ascendante hiérarchique

Les mécanismes de cet algorithme d'apprentissage non-supervisé sont détaillés en section 4.1.2. La méthode d'agrégation utilisée pour réaliser les regroupements successifs est la méthode de Ward pour distance euclidienne. Cette méthode consiste à créer des groupes homogènes en minimisant, à chaque étape, la perte d'inertie inter-classe engendrée par le regroupement. La première étape dans cette classification consiste à déterminer le nombre de groupes à former. Dans ce but, la perte d'inertie est minimisée en fonction du nombre de *clusters*.

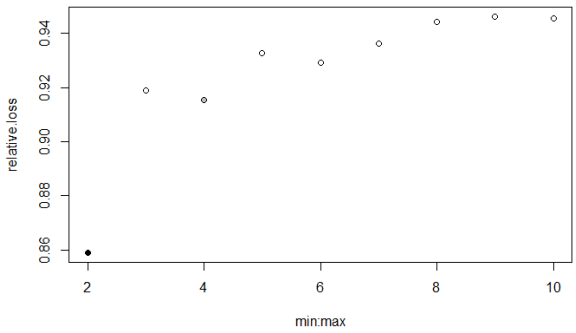


Figure *6.22:* Sélection du nombre de *clusters*

Cette minimisation conduit à sélectionner deux *clusters*. Cependant, afin de pouvoir effectuer une comparaison avec les groupes formés par les *kmeans*, une classification avec trois *clusters* est aussi réalisée.

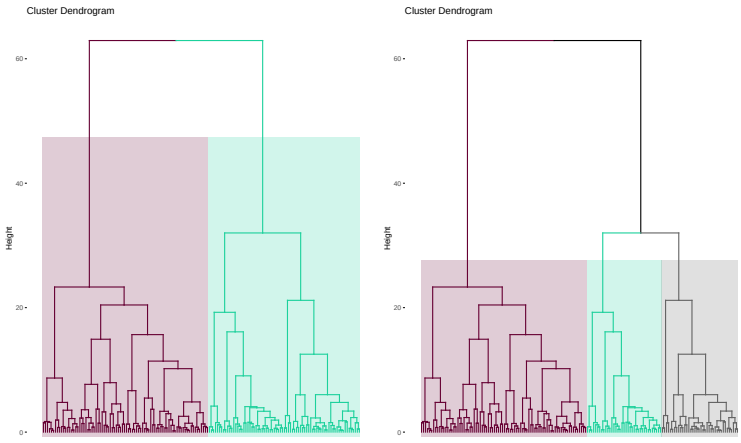


Figure *6.23:* Dendrogrammes pour 2 et 3 clusters sélectionnés

Les groupes formés avec trois *clusters* sont différents de ceux obtenus par *kmeans* et sont moins interprétables sur les axes principaux de l'ACP. En effet, les groupes 1 et 3 sont constitués d'individus ni "sportifs" ni "marcheurs". Ces deux groupes ne sont pas distinguables selon des paramètres et sont donc difficilement interprétables. Cependant, avec deux *clusters*, une

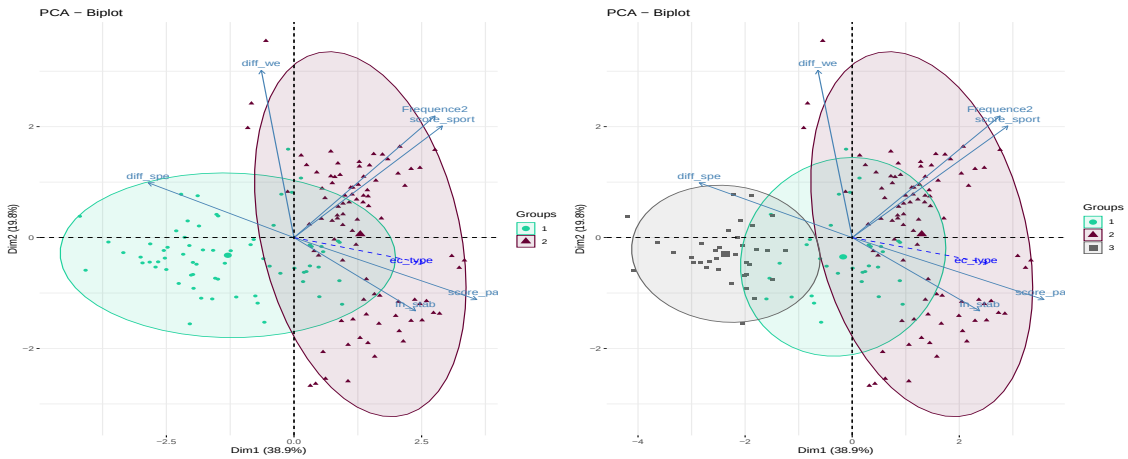


Figure *6.24:* groupes formés avec 2 et 3 *clusters*

distinction nette "Actif"/ "Inactif" est réalisée. Cette classification optimale en classification hiérarchique ascendante ne permet pas d'évaluer la nuance dans les comportements sportifs.

6.3.4 Résultats de classification

La classification retenue correspond à celle réalisée par les *kmeans*. Ce choix de classification a été effectué car les groupes formés sont plus distincts et plus interprétables. Les différents groupes formés correspondent à des profils sportifs. La figure 6.25 expose la répartition des différents profils sportifs en fonction des caractéristiques individuelles non sportives.

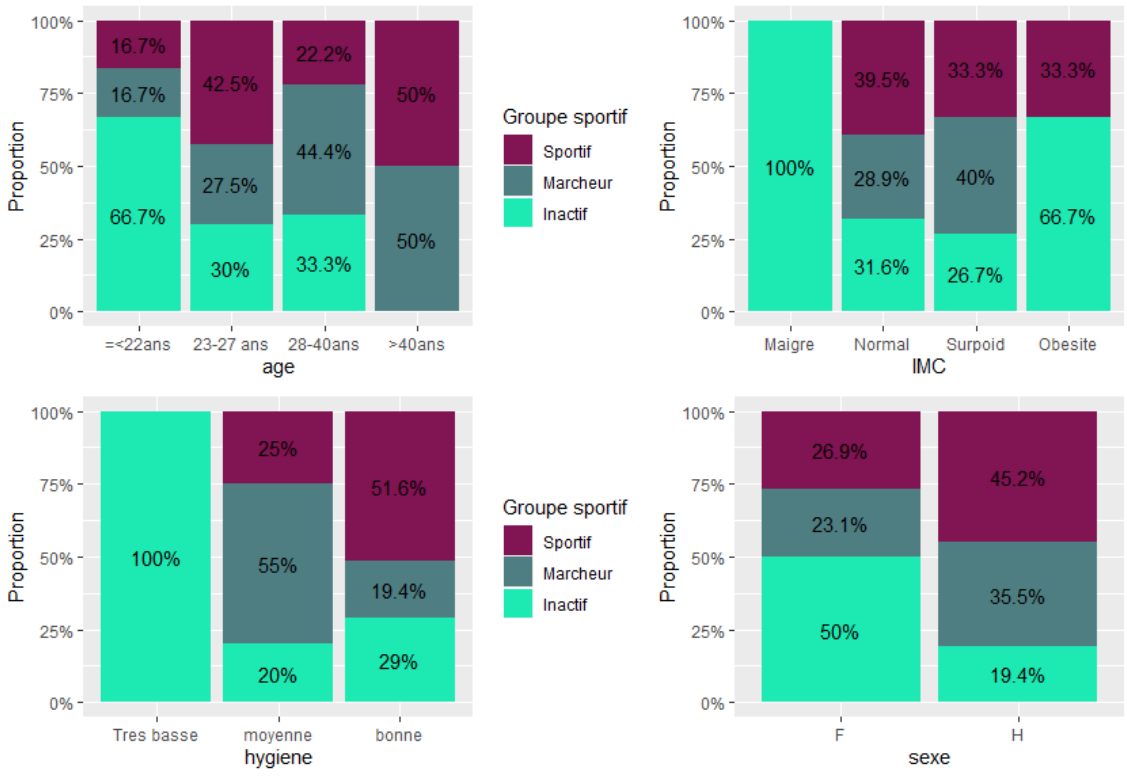


Figure *6.25:* Proportion des profils dans les différents groupes formés

La variable "hygiène de vie", bien que n'ayant pas été utilisée lors de la modélisation, reflète

assez correctement le groupe sportif. En effet, tous les individus ayant déclaré une hygiène de vie très basse sont classés dans le groupe **Inactif**. De plus, les individus ayant déclaré une bonne hygiène de vie restent en majorité classés en tant que **Sportifs**.

L'interprétation de la figure 6.25 doit tout de même tenir compte des quantités initiales des individus de chaque profil. Par exemple, étant donné qu'un seul individu en portefeuille est déclaré "Maigre" par son IMC, il est impossible de généraliser le profil "Inactif" à tous les individus "Maigres".
La figure 6.26 expose la répartition du nombre de consultations en fonction du profil sportif.

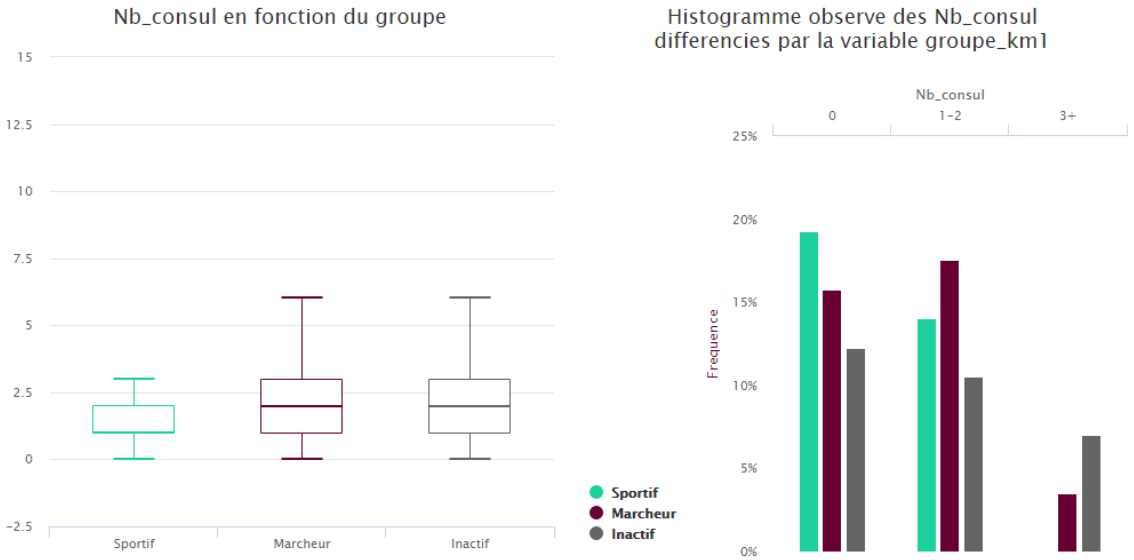


Figure *6.26:* Répartition du nombre de consultations en fonction du groupe

Le nombre de consultations augmente en moyenne avec la qualité du profil sportif. Ainsi, le nombre de consultations du groupe **Sportif** est en moyenne plus faible que celui du groupe **Marcheur** qui est lui même plus faible que celui du groupe **Inactif**.

Les groupes constitués permettent de qualifier les profils sportifs des individus en fonction des mesures de leurs appareils connectés et de leurs déclarations sportives. Ces profils, couplés aux profils individuels, permettent de cartographier les profils du portefeuille étudié. Une telle cartographie peut permettre dans un premier temps de cibler les actions préventives. Par exemple, une action faisant la promotion de l'activité physique uniquement chez les femmes qui sont dans ce portefeuille préventif peut être adéquat puisqu'elles sont en majorité classées comme inactives.

Cette classification, bien que réaliste, se heurte néanmoins à des limites comme la taille de l'échantillon. Les problèmes de qualité des données en matière de véracité des mesures du nombre de pas peuvent aussi être à soulever. En effet, les échantillons récupérés étant constitués de mesures réalisées par des smartphones, celles-ci ne sont correctes que dans la mesure où l'individu est en possession de son téléphone à tout instant. L'hypothèse de possession du smartphone à tout instant reste néanmoins acceptable sur les populations jeunes comme celle du portefeuille

étudié. Dans un cadre réel d'application, des mesures réalisées par des montres connectées seront beaucoup plus fiables.

Conclusion

Dans ce chapitre, les groupes reflétant le caractère sportif des individus du portefeuille sont construits et permettent d'identifier les individus cibles dans la démarche de prévention de l'inactivité physique. Les individus **Inactifs** seront la principale cible des stratégies de prévention. Pour ces derniers, l'objectif sera de les faire évoluer vers des catégories **Marcheurs** ou **Sportifs**. Dans la suite de cette étude, une méthodologie de calcul de l'impact des évolutions de groupes sera développée. L'objectif sera donc de pouvoir associer un groupe sportif à un niveau de sinistralité.

7 Données de prestations santé et modélisation de la probabilité d'occurrence

« La santé n'a pas de prix, mais elle a un coût. » ***André Grimaldi***

Dans ce chapitre, l'objectif est d'apporter un élément de mesure de l'impact d'un plan de prévention primaire. Pour cela, l'approche adoptée sera, tout d'abord, d'associer à la base de prestations santé des groupes reflétant le caractère sportif des individus du portefeuille. Ces groupes sportifs sont issus de la base personnelle de pratique sportive. Aussi, dans ce chapitre, une analyse de la base à disposition sera effectuée avant d'y associer les comportements sportifs. Ensuite, sur la nouvelle base obtenue, une modélisation de la probabilité d'occurrence d'un sinistre sera réalisée.

7.1 Description des données et construction des groupes sur la base de prestations santé

7.1.1 Description de la base de prestations santé

Cette étape de description des données permet de comprendre les données par leur origine, leur traitement et leur structure. La base de prestations santé, appelée **base sinistres** correspond à des données réelles d'un assureur français.

Caractéristiques de la base

Les variables disponibles dans cette base de données sont présentées dans la table 7.1.

Nom de colonne	Description
benefID	Numéro de bénéficiaire
ouvrantDroitID	Numéro d'ouvrant-droits
typeAss	Type d'assuré : ASSURE PRINCIPAL, CONJOINT, ENFANT
CCN	Convention collective à laquelle l'assuré est rattachée (4 différentes)
exercice	Exercice d'assurance : 2014, 2015, 2016, 2017
sousPoste	Sous-poste de soins (Consultations Généralistes, spécialistes, radiologie, analyses, auxiliaires médicaux)
benefCollege	Collège de l'assuré : CAD (cadre), NCA (non-cadre), EP (ensemble du personnel)
exposition	Exposition annuelle
couvObligatoire	Niveau de couverture obligatoire : -99 (pas de couverture obligatoire), 0 (base), 1 (option 1), 2 (option 2)
couvFacultative	Niveau de couverture facultatif : -99 (pas de couverture obligatoire), 0 (base), 1 (option 1), 2 (option 2)
contractFormula	Formule souscrite
structureFoyer	Structure du foyer de l'assuré : isole/couple , Avec/SansEnfant
montantDe	Dépense engagée annuelle
montantRAC	Reste à charge annuel
nbSin	Nombre de sinistres
benefSexe	Sexe de l'assuré
benefAge	Age de l'assuré
benefDept	Département de résidence de l'assuré
prime	Prime versée par l'assuré principal

Table *7.1:* Dictionnaire de variable

Dans ce portefeuille constitué de plus de 70 000 contrats, un retraitement des classes d'âges a dû être effectué. En effet, comme la figure 7.1 le montre, une différence notable en matière de dépense engagée moyenne est visible juste entre les groupes « ≤ 27 » ans et « ≥ 28 ». Dans la suite de cette étude, seules ces deux catégories d'âges seront utilisées.

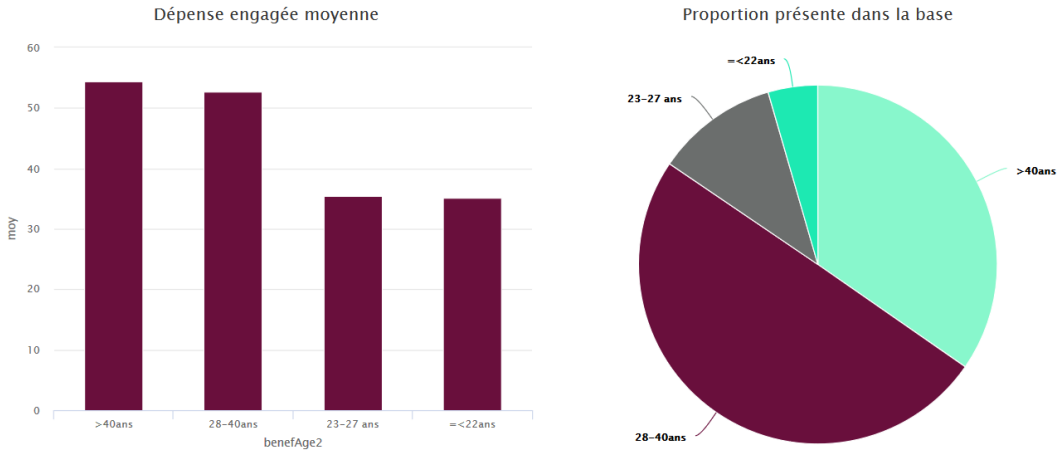


Figure ***7.1:*** Dépenses moyennes par catégories d'âges et proportions des catégories d'âges dans la **base sinistres**

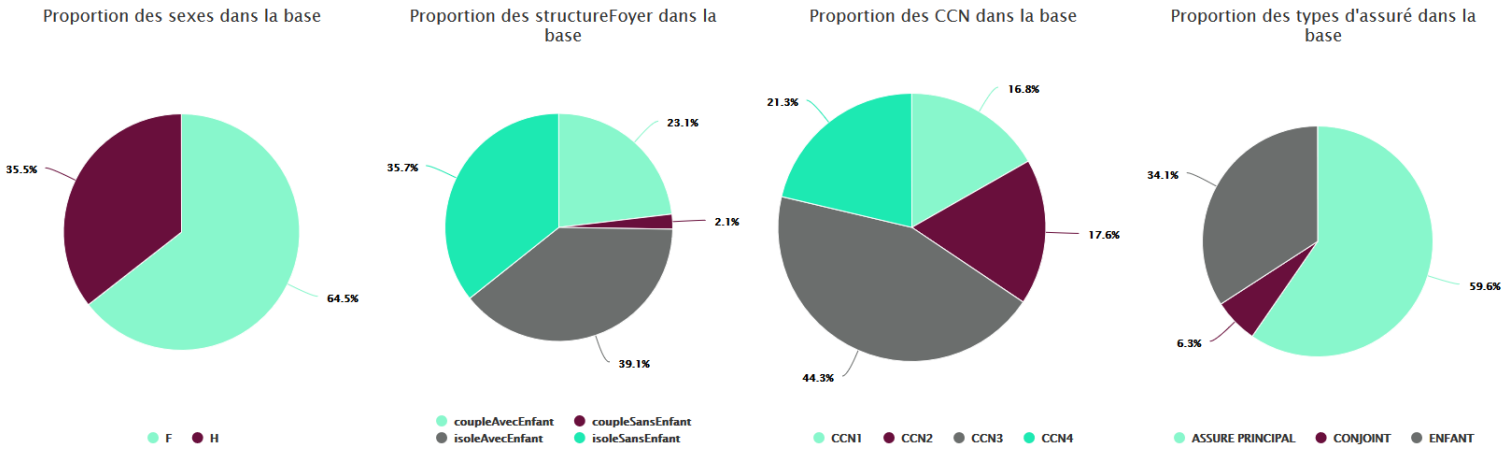


Figure ***7.2:*** Proportions observées dans la base de prestations

Dans la suite de cette étude, seul le sous-poste des consultations chez le généraliste a été conservé. L'étude s'est concentrée sur ce sous-poste car il s'agit de la seule information qui a été à disposition dans la **base sport**. Dans ce portefeuille, des niveaux de dépenses différents avaient été observés par groupe sportif. De plus, dans la **base sinistres** les observations de l'année la plus récente (2017) ont été utilisées pour réaliser les modélisations. Ce portefeuille décompte environ 250 000 consultations chez le généraliste en 2017 pour une dépense totale engagée d'environ 6 millions d'euros.

Création d'un zonier

Afin de réaliser un zonier simplifié, une classification ascendante hiérarchique a été réalisée. L'objectif est de regrouper les départements par zones en fonction des coûts moyens et des fréquences moyennes observées. Les variables retenues pour la mise en place de la CAH sont donc la fréquence moyenne des visites chez le généraliste et les coûts moyens par département. A partir du dendrogramme réalisé, un nombre de trois zones a été retenu. Les coûts moyens et fréquences moyennes par zone sont présentés dans la table 7.2.

Zone	Fréquence moyenne	Coût moyen	Nombre de département
1	3.539	23.986	63
2	3.045	26.027	25
3	4.369	24.033	14

Table *7.2:* Coût moyen et fréquence moyenne observés par zone créée

La zone 2 correspond donc aux départements ayant, en moyenne, des coûts de visites chez le généraliste plus élevés en moyenne. La zone 3, quant à elle, concerne les départements pour lesquels les visites sont les plus fréquentes, La figure 7.3 met en évidence la classification simplifiée des différents départements.

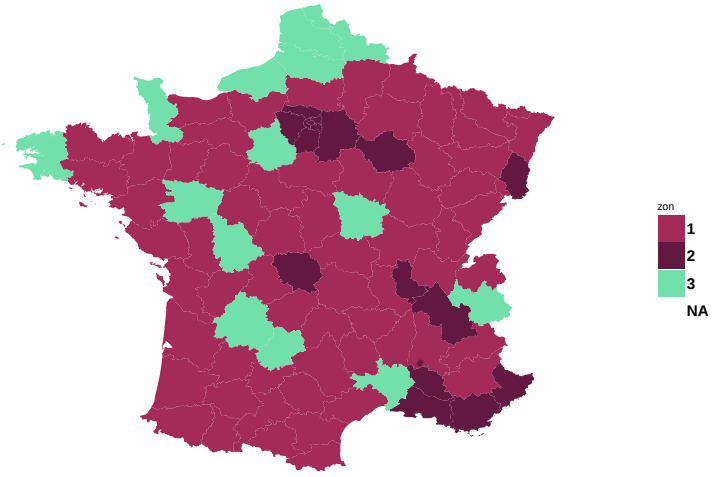


Figure *7.3:* Zonier obtenu par classification ascendante hiérarchique

7.1.2 Construction des groupes sur la base de prestations santé

La **base sport** construite ne possède pas directement d'informations sur la sinistralité des participants, comme elle le pourrait dans un cadre réel d'application par un assureur. Afin de construire une base complète possédant à la fois des données de sinistralité et des indicateurs de performances sportives, il est nécessaire de combiner la **base sinistres** à la **base sport**. Il s'agit de reproduire les groupes définis sur la **base sport** sur la **base sinistres**.

Pour cela, afin de garder une structure cohérente des données, il est important de :

- rester sur les mêmes segments d'âges (18-50 ans) ;
- respecter la répartition hommes/femmes ;
- respecter les niveaux de sinistralité donnés par la connaissance du nombre de consultations annuelles chez le généraliste.

La réplication des groupes est ensuite réalisée en effectuant un tirage aléatoire du groupe sportif, en fonction des proportions observées sur la **base sport**. Une étude d'importance des variables communes aux deux bases (figure 7.4) a été réalisée afin de mettre en évidence les variables les plus discriminantes dans l'explication du groupe sportif. L'importance d'une variable correspond ici à l'écart de deviance normalisée entre le modèle complet et le modèle sans la variable considérée.

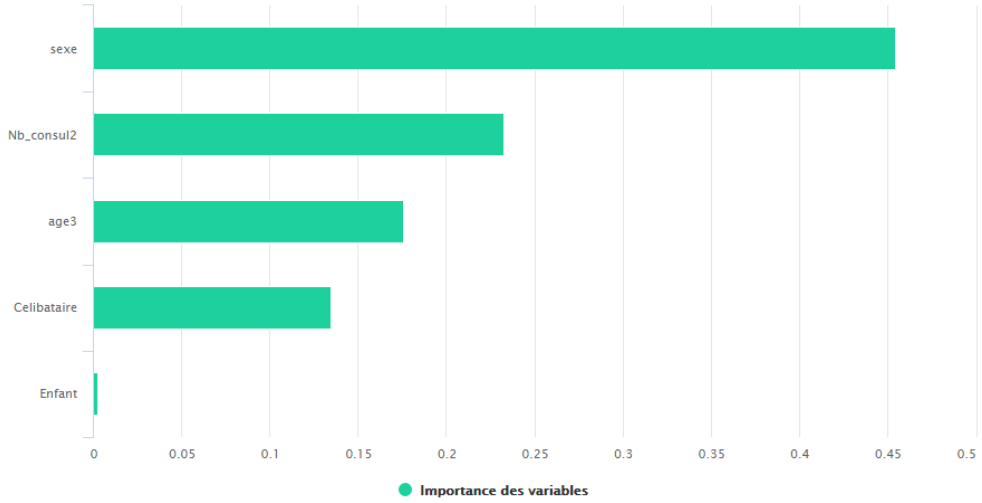


Figure 7.4: Importance des variables communes aux **base sport** et **base sinistres**, dans la définition du groupe sportif

Les variables *Sexe*, *Age*3 et *Nb_consul*2 sont les plus importantes dans la constitution des groupes. Les classes $Sexe \times Age \times Nb\_consul2$ ont donc été utilisées pour répliquer les groupes sportifs. La table 7.3 explicite les proportions utilisées.

Sexe	Age	Nb_consul2	Inactif	Marcheur	Sportif
F	=<27 ans	0	39,4%	30,3%	30,3%
F	=<27 ans	1-2	53,7%	22,0%	24,4%
F	=<27 ans	3+	53,3%	13,3%	33,3%
F	>=28ans	0	63,6%	27,3%	9,1%
F	>=28ans	1-2	23,5%	64,7%	11,8%
F	>=28ans	3+	53,8%	15,4%	30,8%
H	=<27 ans	0	34,5%	29,3%	36,2%
H	=<27 ans	1-2	21,9%	34,4%	43,8%
H	=<27 ans	3+	22,2%	22,2%	55,6%
H	>=28ans	0	30,4%	47,8%	21,7%
H	>=28ans	1-2	36,4%	22,7%	40,9%
H	>=28ans	3+	21,4%	35,7%	42,9%

Table 7.3: Proportions observées sur la **base sport**

Sur la nouvelle base, en plus des informations sur les prestations effectuées, à chaque individu est associé un groupe sportif. Cette base illustrative ainsi créée permettra d'observer un lien entre le groupe sportif et la probabilité d'avoir un sinistre. Ainsi, dans la suite de cette étude, il sera important de relever les méthodologies appliquées et non les résultats obtenus.

7.1.3 Effet de la sélection aléatoire du groupe sportif

La construction des groupes sportifs sur la base de prestations santé entraîne un biais sur l'évaluation des dépenses moyennes par groupe sportif. Les figures 7.5, 7.6 et 7.7 soulèvent la variabilité des résultats au regard des tirages aléatoires effectués.

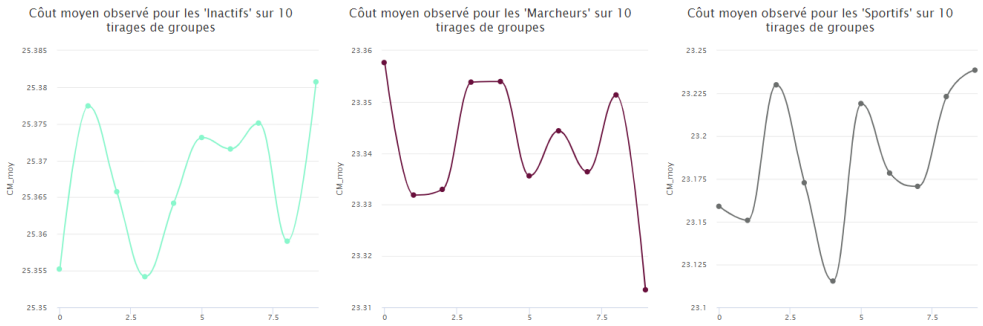


Figure ***7.5:*** Coûts moyens observés des consultations par groupe sportif pour 10 tirages de groupes

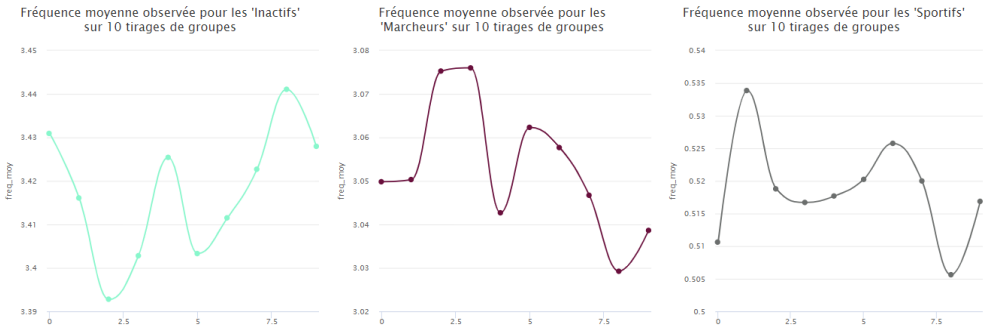


Figure ***7.6:*** Fréquences moyennes observées des consultations par groupe sportif pour 10 tirages de groupes

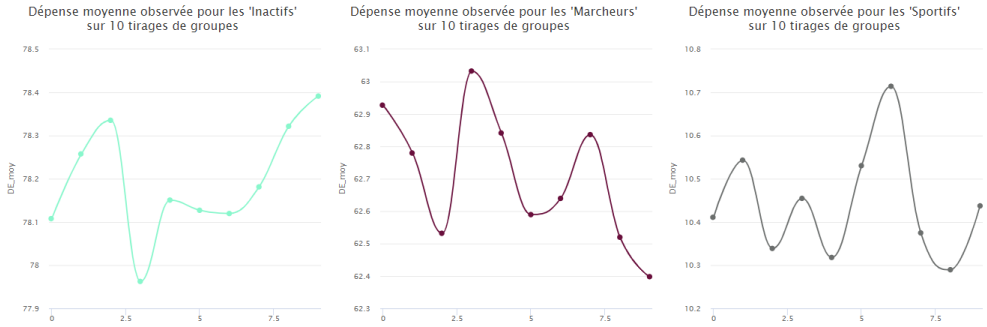


Figure ***7.7:*** Dépenses moyennes observées des consultations par groupe sportif pour 10 tirages de groupes

Une variabilité existe à cause du tirage aléatoire des groupes sportifs. Cependant, la figure 7.8 met en évidence la stabilité relative des observations moyennes. En effet, l'écart-type sur les 10 tirages effectués reste très faible.

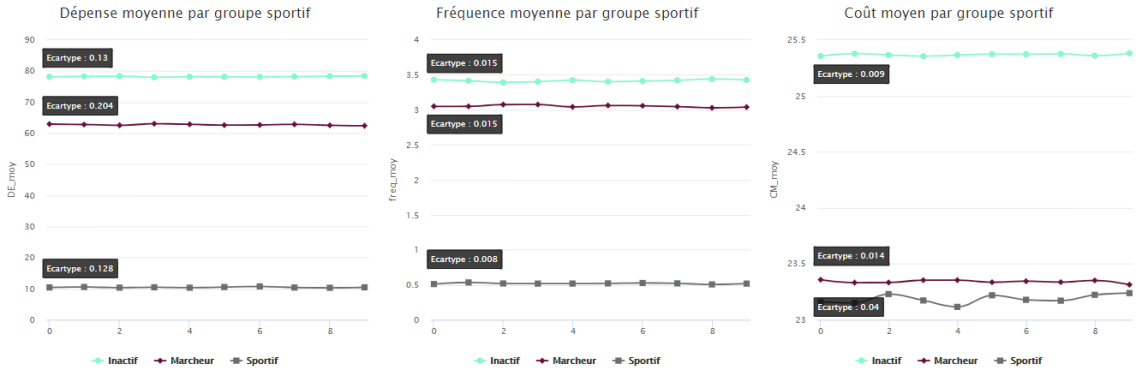


Figure ***7.8:*** Stabilité des dépenses, fréquences et coûts moyens observés des consultations

7.2 Modélisation de la probabilité d'être sinistré

L'objectif de cette section est d'évaluer la probabilité d'avoir un sinistre en fonction des caractéristiques individuelles des individus, dont le groupe sportif. Il sera important de vérifier si la connaissance du groupe sportif impacte la probabilité d'occurrence d'un sinistre. Cette modélisation s'inscrit dans la modélisation d'un cadre de prévention primaire. En effet, l'impact d'une mesure de prévention primaire est évalué en fonction de l'évolution de la probabilité d'occurrence d'un sinistre pour un niveau de prévention donné. Il est important de rappeler qu'ici, un sinistre correspond à une consultation chez le généraliste.

7.2.1 Modèle retenu : la régression logistique

L'approche retenue consiste à réaliser un modèle de régression logistique. Le modèle avec une fonction lien *logit* a été choisi pour sa simplicité dans l'explication des résultats et dans l'interprétation des paramètres. La variable Y à modéliser est telle que :

$$Y = \begin{cases} 1 & \text{si au moins une visite chez le généraliste a eu lieu} \\ 0 & \text{sinon} \end{cases}$$

Il faut noter que dans la modélisation réalisée, le modèle logistique est avant tout utilisé pour l'estimation des probabilités d'occurrence d'un sinistre qu'il fournit. Le modèle retenu pour la modélisation est le suivant :

$$logit(Y) = \beta_0 + \sum_{i=1}^{\text{nb variables}} \left(\sum_{j=1}^{\text{nb modalités}-1} \beta_{ij} * 1_{\{Variable_i = modalite_j\}} \right)$$

Une première étape avant la modélisation consiste à constater les corrélations existantes entre les variables de la modélisation. Les variables retenues sont : le *Groupe sportif*, le *sexe*, la *zone* du département, la catégorie d'*âge*, la *structure du foyer*, la *CCN*, la *formule souscrite* et le *niveau*

de couverture obligatoire. Ces variables étant qualitatives, leurs corrélations sont évaluées par le V de Cramer.

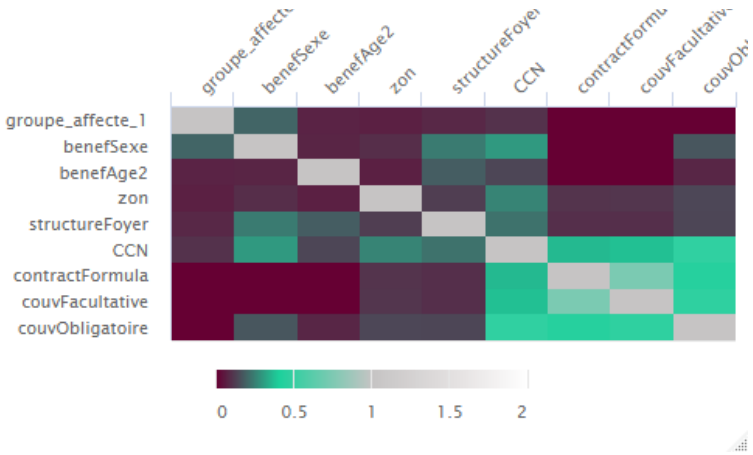


Figure ***7.9:*** Analyse des corrélations par le V de Cramer sur les variables explicatives retenues pour la modélisation

Les variables *couvObligatoire*, *couvFacultative* et *contractFormula* étant assez proches, la connaissance de l'une sera suffisante pour la modélisation. La variable *couvObligatoire* a été choisie au détriment des variables *contractFormula* et *couvFacultative*. En effet, la majorité des modalités de ces dernières n'est pas significative dans le modèle de régression logistique. De plus, la qualité du modèle n'a pas été perturbée par le retrait de ces variables. Une sélection de variables a été réalisée en retirant les variables pour lesquelles la significativité au sens du test de Wald (annexe A.1.1) est mauvaise, et pour lesquelles le retrait n'engendre pas une perte importante en termes de qualité du modèle. Les coefficients de la tables 7.4 sont obtenus après la sélection de variable

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	2,30051	0,06701	34,332	< 2e-16	***
Groupe_1Marcheur	-1,40863	0,02937	-47,966	< 2e-16	***
Groupe_1Sportif	-4,87701	0,03067	-158,995	< 2e-16	***
benefSexeH	-0,0982	0,0227	-4,327	1,51E-05	***
benefAge3>=28ans	0,40708	0,02871	14,18	< 2e-16	***
factor(zon)2	-0,39176	0,02569	-15,25	< 2e-16	***
factor(zon)3	-0,01647	0,02753	-0,598	0,549627	
structureFoyercoupleSansEnfant	-0,5325	0,07244	-7,351	1,97E-13	***
structureFoyerisoleAvecEnfant	0,09916	0,02883	3,44	0,000583	***
structureFoyerisoleSansEnfant	-0,28261	0,02895	-9,762	< 2e-16	***
CCNCCN2	0,35112	0,04156	8,448	< 2e-16	***
CCNCCN3	0,38627	0,0316	12,222	< 2e-16	***
CCNCCN4	-0,20709	0,03731	-5,551	2,84E-08	***
factor(couvObligatoire)0	0,30596	0,05158	5,931	3,00E-09	***
factor(couvObligatoire)1	0,48267	0,06936	6,959	3,42E-12	***
factor(couvObligatoire)2	0,49361	0,06591	7,49	6,91E-14	***

Table ***7.4:*** Coefficients du modèle de régression logistique

Les coefficients d'une régression logistique peuvent être interprétés directement.

Selon la théorie du modèle logistique (voir section 4.2.1), la quantité $\frac{Y}{1-Y}$ correspond à un *odds* ou une côte de succès. Puisque les coefficients correspondent à des $\ln(\frac{Y}{1-Y})$, s'ils sont positifs alors les modalités associées sont plus à risque, et s'ils sont négatifs les modalités associées le sont moins. Ainsi, à partir du signe des coefficients, des interprétations peuvent déjà être réalisées.

- Les coefficients des groupes **Marcheur** et **Sportif** sont négatifs. Les individus de ces groupes présentent, selon le modèle, une probabilité plus faible d'avoir au moins une consultation.
- Les coefficients associés à une structure de foyer sans enfant sont négatifs. Les individus sans enfant ont donc, selon le modèle, une probabilité plus faible d'avoir au moins une consultation.
- Les coefficients des niveaux de couverture obligatoire sont positifs et croissants avec le niveau, la probabilité d'avoir au moins une consultation augmente avec le niveau de couverture.

Ces coefficients permettent aussi de calculer la probabilité associée à un profil donné. Le profil de référence du modèle correspond au profil dont les modalités constituent l'*intercept*. Dans le cas de cette modélisation, le profil de référence est une "*Femme*" de catégorie d'âge "$\leq 27 ans$", "*Inactive*", résidant en *zone 1*, en couple avec enfant, appartenant à la "*CCN1*" et sans couverture obligatoire. Pour ce profil, la probabilité d'obtenir un sinistre se calcule comme suit :

$$\pi = \frac{\exp(\hat{\beta}_0)}{1+\exp(\hat{\beta}_0)} = \frac{\exp(2,30051)}{1+\exp(2,30051)} = 0.9089193,$$

avec $\hat{\beta}_0$ l'estimation du coefficient associé à l'individu de référence dans l'*intercept*. Dans ce portefeuille étudié, l'individu de référence décrit précédemment a une probabilité de 91% d'aller au moins une fois chez le généraliste dans l'année.

Si cet individu de référence devient **Marcheur**, sa probabilité d'aller au moins une fois chez le généraliste dans l'année baisse à 71%

$$\pi = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_{11} * 1_{\{Groupe=\text{"Marcheur"}\}})}{1+\exp(\hat{\beta}_0 + \hat{\beta}_{11} * 1_{\{Groupe=\text{"Marcheur"}\}})} = \frac{\exp(2,30051-1,40863)}{1+\exp(2,30051-1,40863)} = 0.709278$$

Si cet individu de référence devient **Sportif**, sa probabilité d'aller au moins une fois chez le généraliste dans l'année chute à 7%

$$\pi = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_{12} * 1_{\{Groupe=\text{"Sportif"}\}})}{1+\exp(\hat{\beta}_0 + \hat{\beta}_{12} * 1_{\{Groupe=\text{"Sportif"}\}})} = \frac{\exp(2,30051-4,87701)}{1+\exp(2,30051-4,87701)} = 0.07066624$$

Un écart important est observé entre les probabilités d'occurrence d'un sinistre chez les **Sportifs** et chez les **Marcheurs**. Cet écart, trop important pour être réaliste, peut venir du fait de la construction des groupes sportifs sur la base. Il avait déjà été observé un nombre de consultations plus faible chez les individus **Sportifs**. Cette observation se répète, et prend des proportions

plus importantes sur la base construite. Tout de même, ce résultat illustratif de la base utilisée sera employé pour mettre en application la théorie.

7.2.2 Qualité du modèle

Les critères de qualité du modèle peuvent être classés en deux catégories : les critères de qualité intrinsèque du modèle et de qualité en temps que classifieur.

Qualité intrinsèque

Les mesures de la qualité utilisée sont : le Score de Brier, l'AIC et le pseudo R^2 de MacFadden

	Valeur	Commentaire
R^2_{MF}	**0,495**	Cette valeur montre que la régression n'est pas parfaite, mais le modèle se démarque largement du modèle trivial.
Score de brier	**0,097**	Le score Brier étant proche de 0, les probabilités peuvent être considérées comme bien calibrées.
AIC	**62919.57**	La plus faible possible après sélection de variable

Table *7.5:* Mesures et commentaires sur la qualité intrinsèque du modèle

Qualité de classifieur

Ces mesures vont apporter une quantification de la bonne classification que réalise le modèle. Il s'agira ici de l'AUC et du taux d'erreur global sur la base de test.

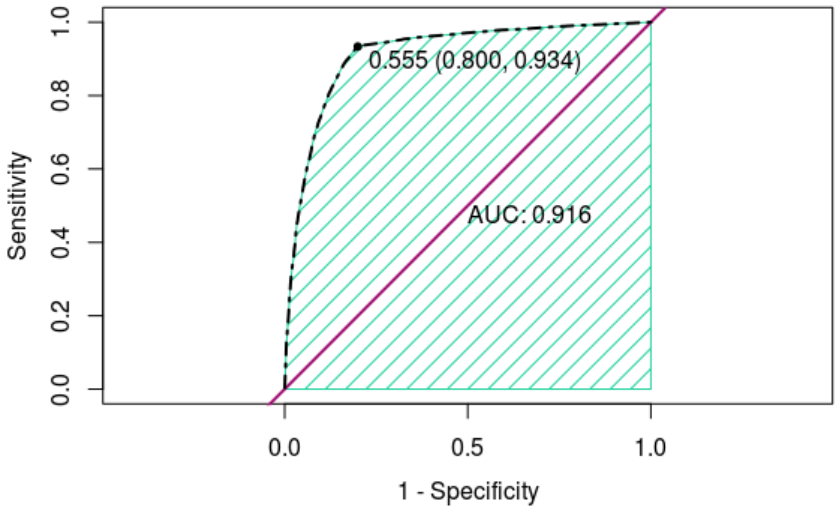


Figure *7.10:* Courbe ROC et AUC

La courbe ROC de la figure 7.10 met en valeur une AUC de 0.916. Celle-ci, étant très proche de 1, le classifieur créé peut être qualifié de satisfaisant. De plus, en évaluant le modèle sur une base de test, avec le seuil optimal de 0.555 proposé, le taux d'erreur global obtenu est de 11%.

7.2.3 Apport du groupe sportif sur la qualité du modèle

Une nouvelle modélisation a été effectuée en retirant la variable *Groupe Sportif*. La table 7.6, résume les mesures de qualité de modèle obtenu.

	Valeur	Commentaire
R^2_{MF}	**0,045**	Cette valeur montre que la régression est loin d'être parfaite, et que le modèle ne se démarque pas vraiment du modèle trivial.
Score de brier	**0,219**	Le score Brier est plus proche de 0 que de 1. Cependant en comparaison avec le modèle précédent, il est très mauvais
AUC	**0,643**	Cet AUC est à peine supérieur à 0.5. Ce classifieur est d'une qualité légèrement supérieure au hasard
Taux d'érreur global	**0,402**	Le modèle échoue 40 % du temps à évaluer si oui ou non au moins une consultation aura lieu
AIC	**119014.3**	Cet AIC est presque deux fois supérieur à celui obtenue avec le modèle précédent

Table *7.6:* Performances du modèle sans les groupes sportifs

Ces résultats mettent en évidence le fait que le modèle créé sans les groupes sportifs est beaucoup moins performant. La connaissance du groupe sportif amenée par les données connectées a donc un réel apport, dans le cas de la base considérée, sur la modélisation de la probabilité d'occurrence d'un sinistre.

Conclusion

À partir de la base illustratrice créée, il a été possible d'apporter un élément de mesure d'impact du groupe sportif. Cet élément correspond à la probabilité d'occurrence d'un sinistre (consultation chez le généraliste) qui, selon le modèle, peut énormément varier en fonction du groupe sportif. Cependant, des points d'attention sont à mettre en évidence. La base illustrative est créée à partir des proportions observées en matière de groupe sportif sur les profils de la **base sport**. Cette base étant relativement faible en comparaison avec la base de prestations santé, elle n'est pas exhaustive. Ainsi, un biais lié à cette non exhaustivité, est porté à une plus grande échelle sur la base illustrative. De plus, la réplication du caractère sportif sur la base de modélisation à partir du niveau du nombre de consultation, donc de la fréquence, est une limite de cette modélisation. Cette limite peut aussi justifier la très bonne qualité du modèle. La modélisation effectuée reste tout de même valable dans la mesure où elle n'est utilisée qu'à titre illustrative et sur le même type de données.

Dans la suite de cette étude, les modélisations effectuées seront mises au service d'application de la théorie de la prévention dans des contextes assurantiels. Seule la prévention primaire sera abordée, puisque l'étude s'est focalisée sur l'occurrence des sinistres et pas sur le coût des sinistres.

Partie IV

Les objets connectés : outil d'encadrement de programmes de prévention

8 Mise en contexte et application de la théorie de la prévention

« Les lois claires en théorie sont souvent un chaos à l'application ».
Napoléon Bonaparte

Dans ce chapitre, il sera proposé au lecteur deux applications des modèles introduits dans les parties précédentes. Elles lui permettront de se rendre compte d'une part de l'apport des objets connectés dans la quantification *a priori* de l'impact d'un programme de prévention et d'autre part, elles permettront d'appliquer la théorie de la prévention, dans un cadre concret pour l'assureur et l'entreprise.

8.1 Application 1 : apport des objets connectés dans la quantification *a priori* de l'impact d'un programme de prévention

Cette première application, bien qu'immédiate, a pour objectif de mettre en avant l'utilité des informations pouvant être issues des objets connectés.

Contexte

Un assureur A a mis en place un programme de prévention sur une entreprise E_1 de 300 salariés. Celui-ci possède autant de salariés en tarif isolé (400€/an) qu'en tarif famille (750€/an). Ce programme consiste à subventionner la mise en place d'une salle de sport dans l'entreprise E_1. Dans cette entité, l'ensemble des salariés dispose d'objets connectés dont les données sont partagées avec la compagnie d'assurance. Ces informations lui ont permis de classer les salariés en 3 catégories : « Sportif », « Marcheur » et « Inactif ». De plus, sur cette entreprise, l'assureur A a observé des gains par catégories d'individus sur la sinistralité. Ce même assureur souhaite aussi subventionner la mise en place d'une salle de sport dans l'entreprise E_2 qui est strictement similaire à l'entreprise E_1 en matière de type de population. Cependant, dans cette seconde entreprise, les salariés de l'entreprise, réticents quant à la collecte de données privées, refusent de partager les informations issues des objets connectés avec l'assureur.

8.1.1 Construction de l'exemple

Cette étude de cas est construite en utilisant la base de données fournie par un assureur français et présentée en section 7.1. Afin de construire les gains sur la sinistralité observés sur les consultations chez le généraliste de l'entreprise 1 suite la mise en place du plan de prévention, les dépenses moyennes et leurs équivalents en remboursement de l'assureur par groupe sportif

ont été utilisés (voir table 8.1). Il convient également de noter que le montant total de primes reçues pour cette entreprise est de 172 500 € [1].

	Dépense moyenne	Côut pour l'assureur	Gain sur la sinistralité du changement de groupe en sportif
Inactif	78,11 €	23,22 €	**20,13 €**
Marcheur	62,93 €	18,59 €	**15,51 €**
Sportif	10,41 €	3,08 €	**-**

Table *8.1:* Gains moyens observés sur la base illustrative par groupe sportif

Ces gains permettront d'évaluer approximativement le gain total en utilisant les proportions connues des groupes sportifs. Cependant, un programme de prévention n'est efficace que pour ceux qui y adhèrent. Ainsi, dans cette exemple il sera important de considérer une proportion d'individus qui adhère a la salle de sport et devient « **Sportifs** ». Dans cette étude de cas, il a été retenu une proportion de 30% en conformité avec les statistiques fournies par Gymlib, des professionnels de la mise en place du sport en entreprise. Selon ces derniers, pour des entreprises de 200 à 1000 salariés, entre 20% et 40% des salariés utilisent les infrastructures sportives. Par catégorie d'individus avant prévention, les gains observés sur l'entreprise E_1 sont représentés dans la table 8.2.

Sexe	Catégorie d'âge	Nombre de consultation initial	Proportion de la population totale	Gain total	Inactif	Marcheur	Sportif
F	$\leq 27 ans$	0	11%	129 €	39%	30%	30%
F	$\leq 27 ans$	1-2	14%	179 €	54%	22%	24%
F	$\leq 27 ans$	3+	5%	58 €	53%	13%	33%
F	$\geq 28 ans$	0	4%	56 €	64%	27%	9%
F	$\geq 28 ans$	1-2	6%	75 €	24%	65%	12%
F	$\geq 28 ans$	3+	4%	52 €	54%	15%	31%
H	$\leq 27 ans$	0	20%	207 €	34%	29%	36%
H	$\leq 27 ans$	1-2	11%	96 €	22%	34%	44%
H	$\leq 27 ans$	3+	5%	33 €	22%	22%	56%
H	$\geq 28 ans$	0	8%	93 €	30%	48%	22%
H	$\geq 28 ans$	1-2	7%	72 €	36%	23%	41%
H	$\geq 28 ans$	3+	5%	47 €	21%	36%	43%

Table *8.2:* Gains totaux observés par catégorie de personne

1. 750€ $\times 150 + 400$€ $\times 150$

Le gain total observé sur l'entreprise E_1 de 300 salariés et sur les consultations chez le généraliste est de 1097 €. Cela correspond à 0,64% des primes reçues. La proportion d'adhésion au programme sur le gain total a un impact sur le gain obtenu car, comme le montre la figure 8.1, plus il y aura d'individus qui participeront au programme, plus important sera le gain.

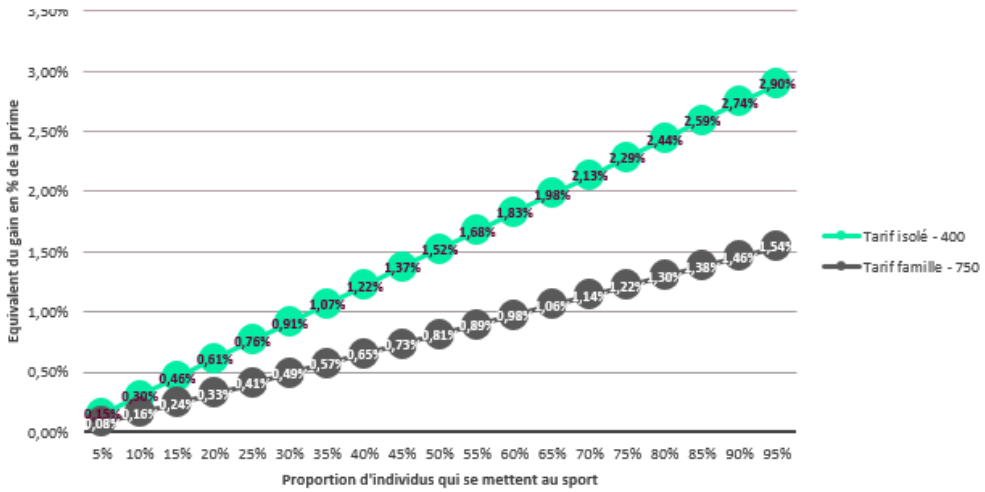


Figure *8.1:* Évolution du gain en pourcentage de la prime en fonction de la proportion d'individus qui se mettent au sport

8.1.2 Apport de la connaissance des groupes sportifs

Dans le cas de l'entreprise 2, en l'absence de données issues des objets connectés, le gain *a priori* devrait être du même ordre de grandeur que celui de l'entreprise 1, puisque les populations sont structurellement identiques. Cependant, lorsque les proportions observées des groupes sportifs varient, le gain n'est plus le même. Comme le montre la table 8.3, bien que la population soit *a priori* identique, l'information des groupes sportifs est cruciale afin de déterminer le gain *a priori* de la prévention mise en place.

	Gains	Gains en % des primes	Proportion d'inactifs	Proportion de marcheurs	Proportion de sportifs
Entreprise 1	1 097 €	0,64%	37,1%	30,4%	32,4%
Entreprise 2	822 €	0,48%	12,4%	42,8%	44,8%

Table *8.3:* Gain réel en fonction des proportions sportives

La méconnaissance des groupes sportifs issus des objets connectés, empêche, dans ce cas-ci, de correctement quantifier le gain obtenu grâce au type de prévention mis en place. L'assureur aurait pu anticiper, avec l'information du groupe sportif, le fait que la prévention soit moins rentable sur l'entreprise 2 puisqu'il s'agit d'une entreprise dans laquelle la proportion d'**Inactifs** est plus faible. De plus, le faible gain en matière de pourcentage des primes rejoint les estimations annoncées par Generali *Vitality*. Ces derniers évaluent l'impact sur la consommation médicale sur un laps de temps aussi court, à moins de 1% [26]. Dans cette application, les coûts engendrés par la mise en place de la stratégie de prévention ne sont pas pris en compte. Il faudrait peut être prévoir une mutualisation sur plusieurs entreprises afin d'améliorer le gain.

À RETENIR

La connaissance des groupes sportifs apportée par les objets connectés permet d'anticiper les gains d'une stratégie de prévention.

8.2 Application 2 : estimation de la rentabilité d'une stratégie de prévention par la théorie de la prévention

Cette application a pour objectif de mettre en pratique la théorie de la prévention afin de quantifier le gain *a priori* de l'assureur, suite à la mise en place d'un plan de prévention.

Contexte

Une CCN [2] souhaite mettre en place un programme de prévention connectée auprès de toutes ses entreprises adhérentes. Elle souhaite notamment que ce programme soit inclus dans ses contrats d'assurance collective et qu'il ne concerne que les assurés principaux. Pour ce faire, elle lance un appel d'offres pour lequel elle retient deux assureurs A_1 et A_2. Ces deux assureurs sont retenus, en premier lieu, pour leurs tarifs de contrats collectifs compétitifs.

Dans la seconde phase de l'appel d'offres, ces derniers doivent formuler des propositions de stratégies sur différents niveaux de prévention proposés par la CCN. Les deux niveaux de prévention abordés par la CCN sont les suivants :

- **Prévention 1** : une prévention connectée qui propose des objectifs en matière de nombre de pas à réaliser sur l'année ;
- **Prévention 2** : une prévention connectée qui propose des objectifs sportifs à réaliser sur l'année.

À partir de la théorie de la prévention et sur la base des données issues des objets connectés transmises par les assurés, les assureurs sont en mesure de faire une proposition du programme le plus rentable, tout en répondant au mieux aux attentes des prestataires sociaux.

L'objectif de cette partie sera, tout d'abord, de mettre en place les éléments nécessaires au calcul du gain utile dans un cadre de prévention primaire. Puisque :

$$E[G]_U = p(e) \times U\{w - S - c(e)\} + (1 - p(e)) \times U\{w - c(e)\}$$

il sera utile de calculer, en amont, les niveaux de prévention ainsi que leurs coûts et leurs impacts sur la probabilité d'obtenir un sinistre. Ensuite, il faudra définir les fonctions d'utilité associées aux deux assureurs. La prime w associée au sous-poste de consultation chez le généraliste est calculée en utilisant les coefficients tarifaires de ce sous-poste qui sont connus. Le coût S du sinistre est supposé connu et correspond au remboursement de l'assureur face à une dépense engagée.

2. Convention Collective Nationale associée à un secteur d'activité. Les décisions prises par la CCN s'appliquent à toutes les entreprises du secteur concerné.

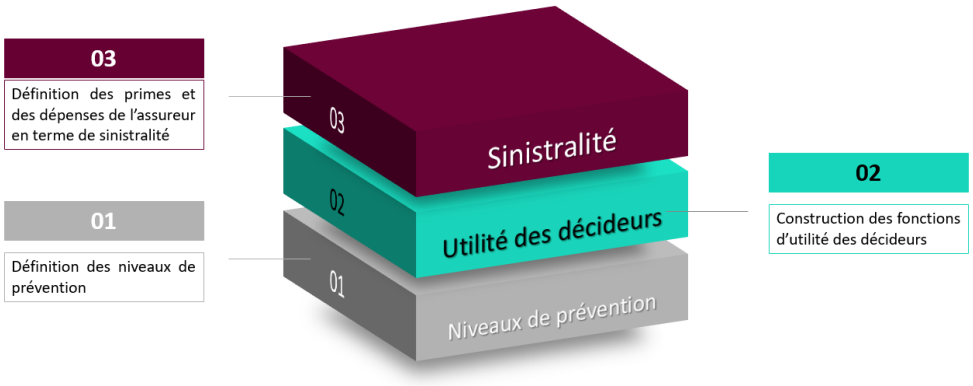


Figure ***8.2:*** Étapes de l'étude de rentabilité d'un programme de prévention

8.2.1 Définition des niveaux de préventions

Afin d'évaluer l'impact des niveaux de prévention, les assureurs conçoivent deux types de programmes de prévention tels que :

- **Prévention 1** : l'objectif de pas défini permet d'augmenter suffisamment le score de marche afin de rendre les individus **Marcheurs** ;
- **Prévention 2** : l'objectif sportif défini permet d'augmenter suffisamment le score de sport afin de rendre les individus **Sportifs**.

Les deux types de prévention mis en place correspondent à des systèmes de récompenses lorsque les objectifs sont accomplis. Ainsi, les coûts liés à ces préventions sont supposés être identiques et divisés en deux types de coûts. Ceux-ci ont des coûts fixes de mise en place de la prévention, valables uniquement sur la première année ainsi que des coûts liés aux récompenses reversées. Les assureurs décident d'utiliser **2%** du volume de primes annuelles afin de récompenser les assurés qui accomplissent les objectifs. Cette proportion a été choisie arbitrairement, par comparaison avec les actions sociales qui bénéficient d'un financement généralement similaire (conformément à l'ANI de 2013 développée en section 1.2.3.1). Des proportions plus grandes n'ont pas été mises en place car, en théorie, plus la récompense est grande plus la proportion d'individus accomplissant les objectifs est élevée. Cependant, dans cette étude, il n'a pas été entrepris de quantifier l'évolution du nombre d'adhérents en fonction de la hauteur de la récompense.

Nombre de personnes	Coûts fixe initial posé
4836	15 000 €

Table ***8.4:*** Caractéristiques du portefeuille utilisé et coûts de prévention associés

Conformément aux données disponibles, les assureurs décident d'évaluer l'impact *a priori* de leurs programmes de prévention connectée sur les consultations chez le généraliste. Un sinistre correspondra donc ici à une visite chez le généraliste. La probabilité d'occurrence d'un sinistre est estimée par le modèle de régression logistique obtenu en section 7.2. Si un individu, de caractéristiques données (*sexe*, *âge* etc.) accomplit les objectifs d'un niveau de prévention, alors il change de groupe sportif et voit sa probabilité d'occurrence de sinistre changer et passer de p_0 à $p(e)$.

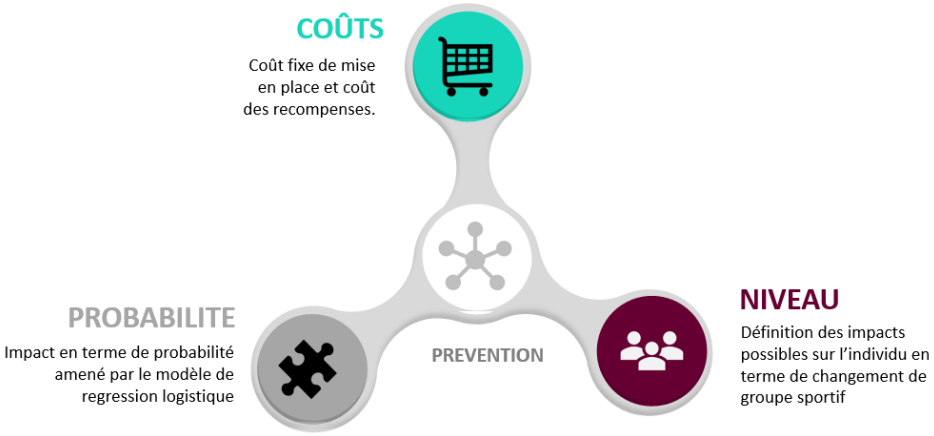


Figure *8.3:* Piliers de la définition d'un niveau de prévention

8.2.2 Construction des fonctions d'utilité des assureurs

La construction des fonctions d'utilité exploitées dans cette étude est basée sur un jeu de questions-réponses comme expliqué dans la section 5.2.2. Deux profils de type « Assureur » ont été interrogés sur des montants allant de $-25k€$ à $100k€$. Ces valeurs ont été choisies en $k€$ afin de mieux situer le décideur interrogé dans un contexte de montant qu'il connaît. Cependant, ces questions ne servent qu'à situer le comportement de celui-ci entre la valeur attendue Q ($IU\{Q\} = 100\%$) et une perte maximale M ($IU\{M\} = -25\%$). Il est important de noter que, l'indice d'utilité tend vers $-\infty$ quand la perte maximale diminue. Même si la théorie autorise cela, en pratique, il est peu concevable de parler d'une perte infinie. Ainsi, la perte maximale de $IU\{x\} = -25\%$ a été choisie afin d'éviter le phénomène d'explosion de la courbe de l'indice d'utilité vers des valeurs infinies.

Profil 1

L'indice d'utilité brut de la figure 8.4 est obtenu en interrogeant le profil assureur 1. Cet indice reflète un décideur qui se contente de peu de gain, pour avoir un fort sentiment d'utilité perçu.

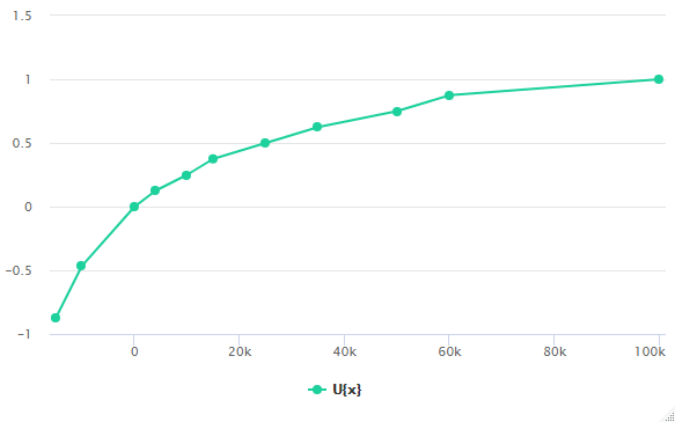


Figure *8.4:* Indice d'utilité brut de l'assureur 1

L'indice d'utilité brut a été ajusté par régression afin d'obtenir une formule fermée de l'utilité de l'assureur. Pour cela, quatre types de fonction ont été ajustées :

- un polynôme d'ordre 2 tel que : $IU\{x\} = ax^2 + bx + c$
- un polynôme d'ordre 3 tel que : $IU\{x\} = ax^3 + bx^2 + cx + d$
- une fonction racine telle que : $IU\{x\} = ax + b\sqrt{x + c} + d$
- une fonction logarithme telle que : $IU\{x\} = ax + blog(x + c) + d$

Ces quatre fonctions ont été retenues car ce sont des fonctions qui peuvent être croissantes et concaves.

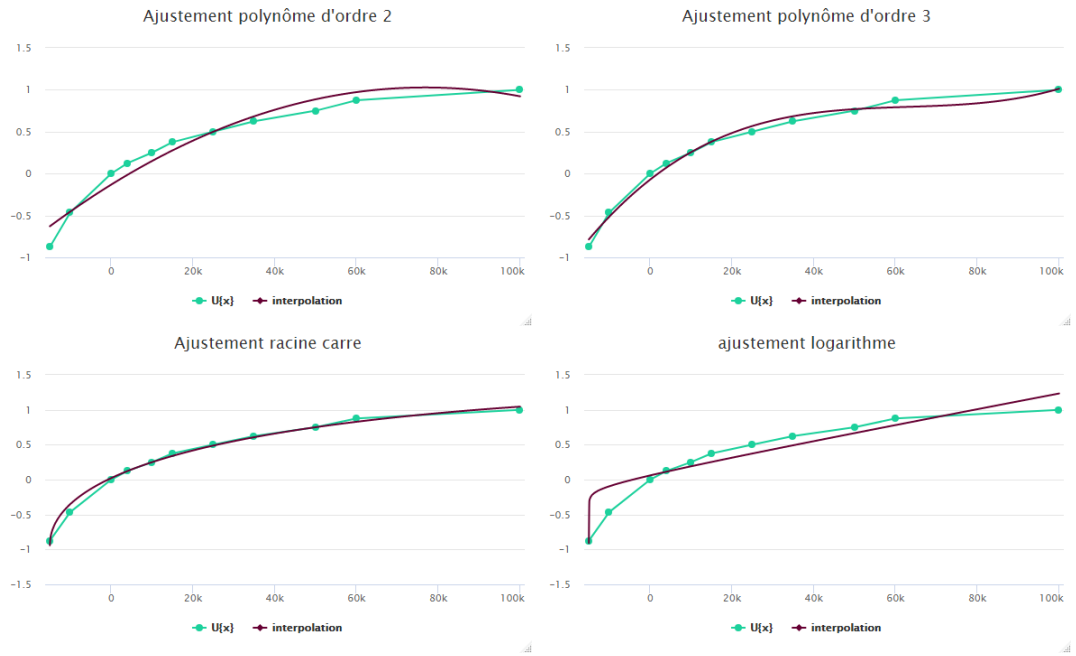


Figure *8.5:* Ajustement de l'indice d'utilité de l'assureur 1

Plusieurs types d'ajustement ont pu être effectués. Cependant, il faut noter que outre les qualités d'adéquation à la courbe, la méthode d'ajustement choisie doit être réalisée avec une fonction concave et croissante afin de respecter les hypothèses retenues de la fonction d'utilité.

Type d'ajustement	R^2	croissante	concave
Polynôme d'ordre 2	0,925	non	oui
Polynôme d'ordre 3	0,985	oui	non
Fonction racine	0,991	oui	oui
Fonction logarithme	0,899	oui	oui

Table *8.5:* Critères de choix du type d'ajustement

À partir des informations du tableau 8.5, l'ajustement choisi est l'ajustement par fonction racine, puisque la qualité d'ajustement en matière de R^2 est la meilleure et les caractéristiques (concavité et croissance) de la fonction d'utilité sont conservées. La fonction d'utilité de l'assureur 1 est donc donnée par une équation de type $IU\{x\} = ax + b\sqrt{x + c} + d$.

Profil 2

L'indice d'utilité du profil 2 a été construit de manière analogue. Les indices d'utilité bruts et ajustés sont présentés à la figure 8.6. Cet indice reflète un décideur qui a plus d'attente en matière de gain que le profil 1 pour avoir un fort sentiment d'utilité.

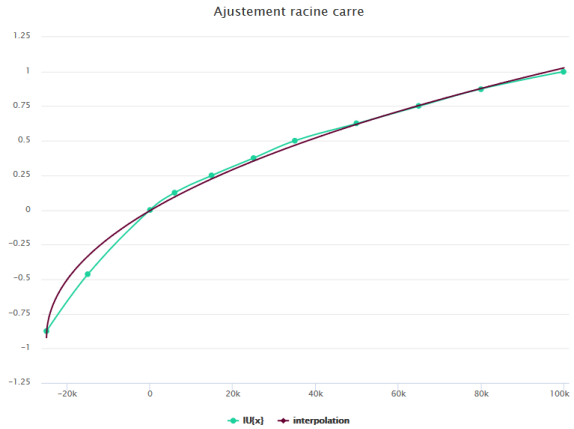


Figure *8.6:* Indice d'utilité de l'assureur 2, brut et ajusté par une fonction de type racine carré

L'indice d'utilité ajusté de l'assureur 2 est donné par une équation de type $IU\{x\} = ax + b\sqrt{x+c} + d$.

Profil 3

L'indice d'utilité du profil 3 correspond à un individu neutre au risque. Il permet de se rendre compte des gains réels espérés. L'équation de sa fonction d'utilité est donnée par $U\{x\} = x$

La figure 8.7 montre que les profils 1 et 2 correspondent à des assureurs qui n'auront pas la même perception des gains qu'ils réalisent puisque leurs fonctions d'utilité sont différentes. Le profil 1 se contentera plus de petits gains alors que le profil 2 en attendra davantage. En termes de pertes, le profil 1 sera plus affecté que le profil 2. Ce dernier est plus proche, dans sa vision, du risque réel intercepté par le profil 3. Ce troisième profil constituera un profil témoin qui exprimera les gains réels observés sans le facteur de la perception individuelle.

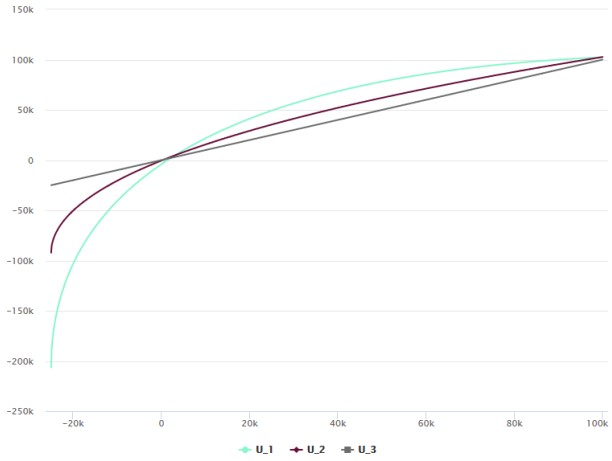


Figure *8.7:* Les fonctions d'utilité des trois profils construits

Ces fonctions sont recalculées pour chaque contrat puisque, pour chacun d'entre eux, la valeur attendue Q est différente. En effet, cela est dû au fait que la prime d'assurance diffère par individu.

8.2.3 Impacts marginaux sur le gain utile

Dans l'objectif, pour les assureurs, d'étudier la rentabilité des programmes de prévention, le calcul des gains utiles de chaque programme de prévention, en incluant la prévention de niveau 0 (aucune prévention) est réalisé. Les gains sont observés sur trois ans en sachant que les frais fixes de mise en place ne sont présents que la première année. De plus, aucune hypothèse d'évolution (entrées et sorties) du portefeuille n'est prise car il sera considéré que le programme sera souscrit sur trois ans par la CCN avec un taux de *turnover*[3] négligeable. Cependant, une évolution de la sinistralité de 1,91% en moyenne a été observée sur les années précédentes de ce portefeuille. Dans la projection sur trois ans des gains obtenus, il sera supposé que la sinistralité augmentera de 2% par an. De plus, les partenaires sociaux étant soucieux de couvrir au mieux leurs assurés, il a été convenu entre assureur et partenaires sociaux que les primes augmentent aussi de 2% par an.

Influence de la proportion d'individus accomplissant les objectifs

Dans l'objectif d'évaluer l'influence de la proportion d'individus accomplissant les objectifs, le gain utile est calculé en considérant des proportions de réussite d'objectif allant de 10 à 90%.

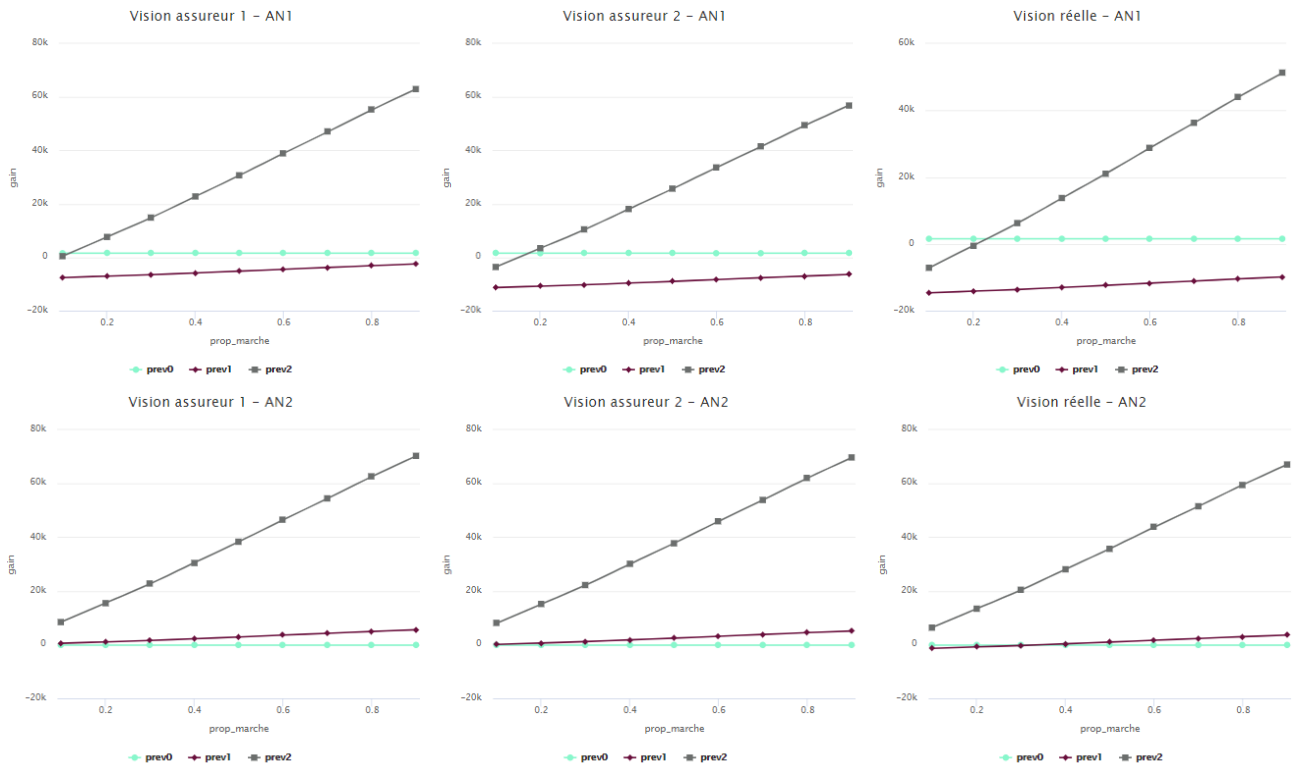


Figure *8.8:* Évolution du gain utile par profil en fonction de la proportion d'individus qui accomplit les objectifs sur les années 1 et 2

3. Taux de renouvellement du personnel d'une entreprise.

La figure 8.8 retranscrit les résultats obtenus et plusieurs conclusions se dégagent.

- La stratégie de prévention 2 (objectifs sportifs) est plus rentable que la stratégie de prévention 1 (objectifs de marche).
- En vision neutre, le programme de prévention 2 n'est rentable la première année qu'à partir de 30% de réussite d'objectifs et la deuxième année à partir de 10% de réussite d'objectifs.
- Au cours de la première année, A_1 perçoit le programme de prévention 2 comme étant rentable dès que 10% des individus deviennent **Sportifs**. Cependant, A_2, dans les mêmes conditions que A_1, ne perçoit ce programme de prévention comme rentable que lorsque 20% des individus deviennent sportifs.
- À partir de la seconde année, il n'y a plus de coûts fixes de mise en place. Seuls les 2% des primes allouées au programme subsistent. Le programme sera rentable dès l'atteinte de 10% de réussite d'objectifs pour A_1 et pour A_2.
- Sur l'année 3, les conclusions sont les mêmes que sur l'année 2.

Figure *8.9:* Évolution du gain utile par profil en fonction de la proportion d'individus qui accomplit les objectifs sur l'année 3

À RETENIR

La proportion d'individus qui change de classe rend compte de la performance et de la rentabilité d'une stratégie de prévention.

Inciter les individus à devenir « Sportifs » plutôt que « Marcheurs », engendre une plus grande rentabilité.

La fonction d'utilité explique la propension de l'assureur à accepter des pertes.

Évolution temporelle et influence du coût fixe initial

Selon un sondage interne de la CCN, ils estiment que 20% des salariés se disent prêts à réaliser les objectifs fixés par la prévention. Partant de cette connaissance, les assureurs anticipent leurs gains (hors variation naturelle de la sinistralité) sur les 3 années d'applications du programme.

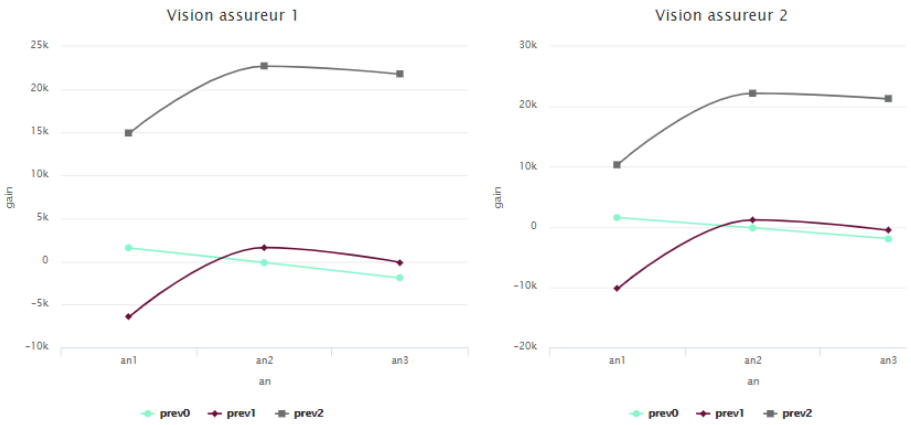


Figure *8.10:* Évolution temporelle anticipée du gain utile de l'assureur

L'écart observé entre l'année 1 et l'année 2 sur les gains utiles après prévention, vient du coût fixe de mise en place de la prévention et de l'évolution de la sinistralité. Cet écart inclut aussi la variation du coût des récompenses induite par l'évolution de 2% de la prime. L'objectif sera de déterminer le point d'équilibre à partir duquel la prévention a la même rentabilité que l'absence de prévention. Ce point d'équilibre permet d'avoir l'information du coût fixe maximal (noté CFM) à investir pour la mise en place de prévention. Ce CFM permet d'avoir au moins la même rentabilité que si la prévention n'avait pas eu lieu. L'intérêt de cet indicateur est de permettre à l'assureur de connaître sa capacité maximale d'investissement si la rentabilité est avérée. Il saura adapter sa proposition s'il veut être compétitif, tout en gardant au moins sa rentabilité initiale.

La recherche du CFM n'est possible que lorsque la rentabilité est avérée au moins sur les années 2 et 3. De plus, son calcul se fera en 3 étapes qui sont :

* Déduction du gain lié à l'évolution de la sinistralité ;

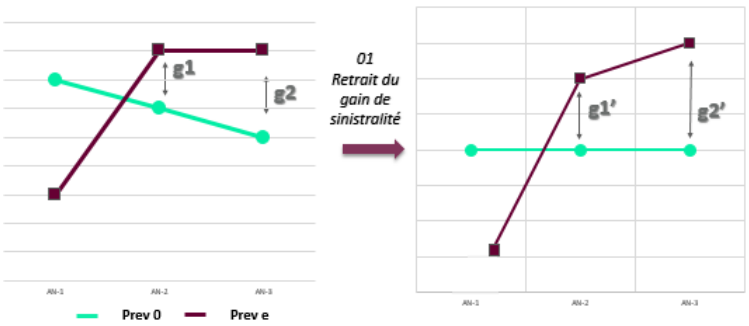


Figure *8.11:* Processus de recherche du Coût Fixe Maximal - CFM - à investir : étape 1

- Recherche de la valeur qui annule le gain sur trois ans ;

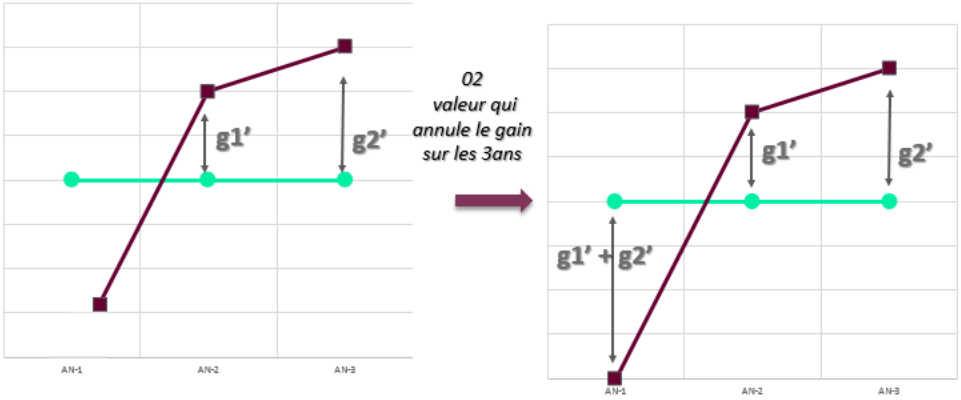


Figure *8.12:* Processus de recherche du Coût Fixe Maximal - CFM - à investir : étape 2

- Déduction du CFM en fonction de la variation du coût des récompenses noté VCV, induite par l'évolution de 2% de la prime par an ;

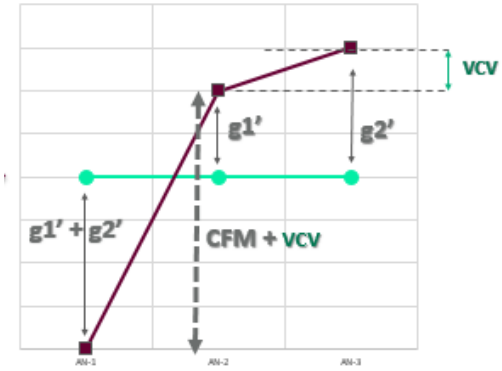


Figure *8.13:* Processus de recherche du Coût Fixe Maximal - CFM - à investir : étape 3

Ainsi, la valeur du CFM est calculée de la manière suivante.

Puisque :

$$VCV = g2' - g1'$$

Et avec :

$$CFM + VCV = 2g1' + g2'$$

on obtient :

$$CFM = 3g1'$$

Le calcul du CFM est réalisé en considérant des proportions d'individus accomplissant leurs objectifs de 10% et de 20%. Il est réalisé pour les deux assureurs et les deux types de prévention lorsque cela est possible (Tables 8.6 et 8.7).

	Profil	g1'	g2'	VCV	CFM
Prevention 1 (Marche)	*Assureur 1*	676,87 €	696,67 €	19,81 €	2 030,60 €
	Assureur 2	249,08 €	259,86 €	10,78 €	747,25 €
Prevention 2 (Sport)	*Assureur 1*	8 608,48 €	8 933,03 €	324,55 €	25 825,44 €
	Assureur 2	8 164,25 €	8 479,47 €	315,22 €	24 492,75 €

Table *8.6:* Récapitulatif pour une proportion de 10%

	Profil	g1'	g2'	VCV	CFM
Prevention 1 (Marche)	*Assureur 1*	1 749,83 €	1 811,12 €	61,29 €	5 249,50 €
	Assureur 2	1 319,24 €	1 371,44 €	52,20 €	3 957,72 €
Prevention 2 (Sport)	*Assureur 1*	22 833,81 €	23 701,19 €	867,38 €	68 501,43 €
	Assureur 2	22 355,49 €	23 212,70 €	857,21 €	67 066,47 €

Table *8.7:* Récapitulatif pour une proportion de 20%

8.2.4 Les résultats

La connaissance des proportions seuils de rentabilité et du CFM permettent aux assureurs d'orienter leur stratégie et de faire une proposition de prévention à la CCN.

- L'**assureur 1** peut proposer à la CCN les stratégies de prévention **1** et **2**. Il n'est cependant prêt à investir que 2 030€ pour mettre en place la stratégie de prévention **1** et jusqu'à 25 835€ pour mettre en place la stratégie de prévention **2**, s'il estime que 10% des individus accompliront leurs objectifs. Quand cette proportion passe à 20%, sa capacité d'investissement sur la stratégie de prévention **2** devient de 68 501€, et de 5 249€ sur la stratégie de prévention **1**.
- L'**assureur 2** peut proposer à la CCN les stratégies de prévention **1** et **2**. Cependant, sa capacité maximale d'investissement est toujours plus faible que celle de l'assureur 1.

Pour chacun des assureurs, un pilotage de la stratégie est réalisé en fonction de son utilité perçue des gains anticipés. Dans cette étude de cas, l'assureur 1 a un avantage certain, puisqu'il offre un investissement initial plus conséquent.

À RETENIR

La capacité maximale d'investissement augmente avec la proportion d'individus qui changent de groupe sportif.

L'estimation de la rentabilité d'un programme de prévention, apportée par la théorie de la prévention, permet de piloter sa stratégie.

Conclusion

Les deux applications développées dans cette étude avaient pour objectif d'illustrer l'apport des objets connectés. En effet, ces derniers apportent une identification des risques en classant les individus par groupe sportif. Cette classification permet une connaissance plus fine du risque et permet d'anticiper les gains à court terme de mesures préventives. La seconde application permet de mettre en pratique la théorie de la prévention, en quantifiant *a priori* la rentabilité perçue par un assureur d'une mesure de prévention.

La théorie de la prévention ici appliquée permet donc :

- d'obtenir une quantification *a priori* de l'impact à court terme d'une prévention. Cette quantification est liée à la connaissance des groupes sportifs amenée par les objets connectés. L'impact se mesure en fonction de la proportion d'individus qui changent de groupe ;
- d'inclure la perception d'un assureur. Cette qualité est importante puisqu'elle permet d'exprimer les différences de comportement en matière de risque entre, par exemple, des gros assureurs **preneurs de risques** et des petites mutuelles plus **prudentes** ;
- de mettre en place un outil de décision quant à l'élaboration des stratégies préventives.

9 Analyse critique des résultats et améliorations de l'approche

« L'humilité n'est pas tant de connaître ses limites que de les admettre ».
Alain Leblay

Dans ce chapitre, le lecteur pourra apprécier les principaux apports et limites de l'étude réalisée. De plus, des recommandations quant au cadre réel d'utilisation seront proposées. Enfin, seront énoncés les apports et les améliorations possibles de l'approche développée.

9.1 Apport de l'approche et cadre réel d'application

Le principal apport de cette étude réside dans la quantification des impacts d'une mesure de prévention. Les données issues des objets connectés permettent d'identifier le risque associé aux individus dans le cadre de l'activité physique et sportive - APS. Les impacts des mesures de prévention sont quantifiés par le biais de la théorie de la prévention et permettent d'anticiper une certaine rentabilité. Ainsi, cette étude fait des objets connectés un outil de mesure de l'impact et de la rentabilité des programmes de prévention liés à l'APS.

Dans un cadre réel d'application, l'association d'une compagnie et d'un intermédiaire (exemple de Generali *Vitality*) peut être adaptée. Cela permettrait de rassurer l'opinion publique en garantissant que les données seront inaccessibles directement par l'assureur et donc que le tarif ne sera pas modulé par la connaissance reçue des objets connectés. En assurance, le cadre réglementaire des objets connectés doit être surveillé. En effet, les pratiques utilisant les objets connectés en assurance étant relativement récentes, le cadre réglementaire n'est pas encore suffisamment établi. De plus, dans le respect du RGPD et dans la continuité des pratiques assurantielles, la protection des données personnelles récupérées doit être au coeur des préoccupations. Ainsi, tout programme de prévention connectée doit être mis en place en considérant, dès la phase de conception, la politique de sécurisation des données.

9.2 Limites de l'approche : une étude théorique avec une illustration

La limite principale de cette étude réside dans l'utilisation d'une **base illustrative**. Cette base est constituée d'éléments éventuellement spécifiques aux individus dont les données ont été recueillies. La non-exhaustivité des données empêche une généralisation des résultats de l'étude, notamment en ce qui concerne les probabilités d'occurrence. Ainsi, même si les résultats sont

non transposables, la méthodologie reste cohérente. À chaque étape de cette étude une base de données plus grande ou plus exhaustive aurait pu modifier les résultats obtenus.

- IDENTIFICATION DES RISQUES : Au niveau de la constitution des groupes, l'approche par classification non supervisée peut varier pour une quantité de données plus importante. Ainsi, les groupes **Sportif**, **Inactif** et **Marcheur** ici formés auraient pu être différents pour d'autres type de données. Une classification plus fine aurait pu être utilisée si des classes spécifiques s'étaient démarquées dans une base de données plus grande.

- IMPACT DES GROUPES SUR LA PROBABILITÉ D'OCCURRENCE : Avec une base plus exhaustive, les probabilités d'occurrence observées par groupe sportifs peuvent être moins éparpillées (71% pour un **Marcheur** contre 7% pour un **Sportif**) et plus réaliste.

Cette étude s'oriente sur l'impact des préventions liées aux consultations chez le généraliste. D'autres sous-postes de soins peuvent être modélisés. Cette méthodologie ne permet cependant que de modéliser des **impacts à court terme**, ce qui correspond à une limite de l'étude. En effet, l'activité physique a aussi des effets à moyen et long termes. C'est le cas des pathologies lourdes (cancers, diabètes etc.) qui peuvent être favorisées par l'inactivité physique. Certains de ces effets sont repris par l'étude du Mouvement des entreprises de France - MEDEF -, du Comité National Olympique et Sportif Français - CNOSF - et AG2R LA MONDIALE [22]. Cette étude quantifie l'impact économique de l'activité physique et sportive sur l'entreprise, le salarié et la société.

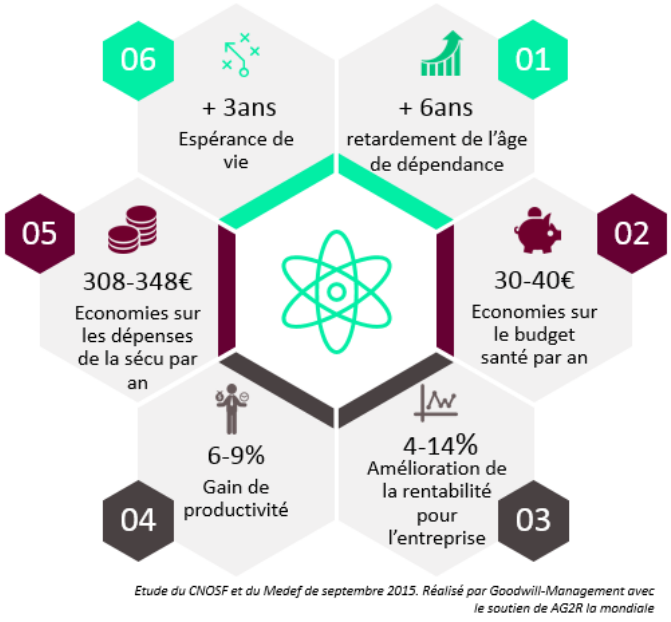


Figure *9.1:* Impact de l'activité physique pour l'individu, l'entreprise et la société civile

9.3 Amélioration de l'approche

Étendre l'approche à d'autres sous-postes de soins

La méthodologie déployée pourrait être étendue aux autres sous-postes de soins comme la pharmacie, les auxiliaires médicaux et les consultations chez un spécialiste. Cette extension permettrait de quantifier le gain général *a priori* après la mise en place d'une prévention, sur la garantie frais médicaux et *a posteriori* sur l'ensemble des garanties. Le parcours de soin ordonné, qui consiste à aller voir le généraliste avant d'autres types de consultation, met tout de même en valeur ce sous-poste qui peut être un bon indicateur du niveau général de la sinistralité en reflétant la propension à consommer.

Évaluer l'impact de la prévention secondaire

L'application et la modélisation réalisées dans cette étude sont attachées à la notion de prévention primaire. Une extension de cette application peut être réalisée en utilisant la théorie de la prévention secondaire. Il s'agira donc de modéliser la dépense engagée en fonction des caractéristiques habituelles en y incorporant les groupes sportifs. Dans le modèle coût $\times$ fréquence généralement utilisé pour modéliser cette dépense engagée, seule la fréquence peut être amenée à être modifiée par le profil sportif. En effet, en santé, les sinistres étant les consultations et autres sous-postes de soins, leurs coûts sont en général fixes. L'impact de la prévention sur la dépense totale par individu serait donc porté par la fréquence.

Phénomène de saturation en prévention

Le phénomène de saturation de la prévention n'est pas pris en compte dans cette modélisation. Ce phénomène consiste à affirmer que, malgré plusieurs stratégies de prévention, la sinistralité ne pourra pas baisser au delà d'un certain seuil. Ce phénomène s'observe dans le cadre de la prévention routière par exemple. En dépit d'une multiplication des actes de prévention, la mortalité sur les routes en Europe est quasiment en stagnation. Une amélioration de cette étude reposerait sur la prise en considération de cette saturation dans le modèle de prévention.

Open Data

Une des améliorations de l'approche consiste à améliorer la connaissance des groupes sportifs en utilisant des données *Open Data*[1] qui pourraient influencer les données mesurées. La principale difficulté rencontrée dans la mise en place de cette amélioration a été le fait que le code postal fourni par le répondant au formulaire n'est pas nécessairement le même sur la durée de l'historique. Il pourrait être envisagé, par exemple, qu'une influence existe entre le climat et le nombre de pas. Cependant, en utilisant le code postal fourni, la figure 9.2 montre qu'aucune structure logique n'existe, par exemple, entre le nombre de pas et le climat en matière de tem-

1. Une donnée ouverte est une information publique brute, qui a vocation à être librement accessible et réutilisable.

pérature maximale dans la journée, de durée d'ensoleillement et de niveau de précipitations [2].

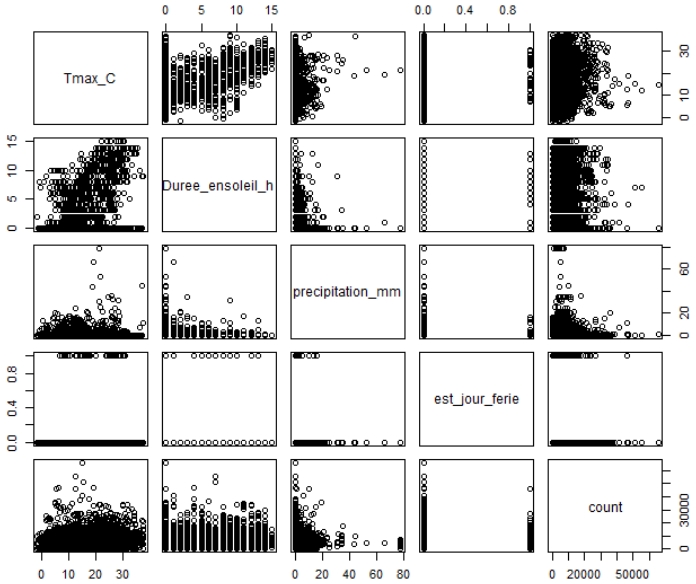


Figure *9.2:* Visualisation du manque de dépendances entre les variables *Open Data* et le nombre de pas (*count*) avec les codes postaux recueillis

Des modèles d'interactions entre actions de prévention

La méthode développée dans cette étude calcule un impact associé à une mesure de prévention. Une amélioration de l'approche consiste à évaluer la sinistralité avant et après mise en place de plusieurs actions de prévention. Le challenge de cette approche réside dans la mesure de l'interaction entre actions de prévention. En effet, lors d'une mise en oeuvre conjointe de plusieurs actions de prévention, les impacts marginaux observés ne sont pas nécessairement additifs. Une mesure de prévention peut démultiplier les impacts d'une autre, l'atténuer ou même l'annuler. Ces interactions peuvent être quantifiées par des outils mathématiques tels que la théorie des jeux et la valeur de Shapley [17][3]. Ces outils permettraient de savoir quel serait le résultat des interactions, quelle serait l'utilité finale des individus et comment serait caractérisé la situation finale suite aux interactions.

Des modèles multi-états de changement de groupes sportifs

Dans cette étude, les changements de groupes sportifs sont supposés connus puisqu'ils sont induits par le type de prévention proposé. Cependant, pour des préventions plus générales proposées, des changements de groupes moins organisés peuvent être réalisés. Dans ce cas, une possibilité d'étude d'impact consiste à utiliser des modèles multi-états qui quantifieraient cet impact en fonction des probabilités de passage d'un groupe sportif à l'autre. Ces probabilités évolueraient en fonction d'un type de profil et de la stratégie de prévention mise en place.

2. Ces données sont récupérées par *web-scrapping*.

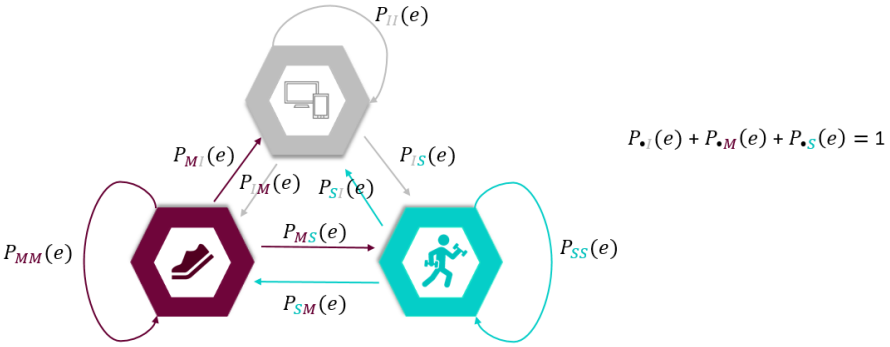


Figure *9.3:* Modèle multi-état de passage d'un groupe sportif à l'autre pour un profil donné

Activité physique extrême

Dans cette étude, la promotion de l'activité physique et sportive - APS - a été prédominante. Cependant, l'activité physique, poussée à l'extrême, conduit à des risques pour le capital santé. L'étude établie ici ne considère pas ce cas spécifique des activités extrêmes. Il serait donc intéressant de savoir quel niveau d'APS détériore la santé, et, de *facto*, la rentabilité d'une mesure de prévention.

Conclusion

L'étude menée tout au long de ce mémoire consistait à développer une méthode de quantification de l'impact, à court terme, de stratégies de prévention basées sur les objets connectés. Le manque d'activité physique étant un facteur important de risque en santé, il a été décidé de focaliser la démarche de prévention sur ce phénomène.

Pour ce faire, des données issues d'un portefeuille d'un assureur français ont été utilisées. De plus, un formulaire a été publié afin de récupérer des données anonymisées de types personnels, professionnels et comportementaux ainsi que des données issues des objets connectés. Un travail conséquent de traitement de ces données et d'automatisation de la récolte a aussi dû être effectué afin de faire face à l'hétérogénéité des données collectées.

Dans l'objectif de répondre à la problématique, une approche méthodologique a été retenue et peut se résumer en deux parties. Tout d'abords, une identification des individus à risques a été nécessaire. Celle-ci a été réalisée par une classification non supervisée sur des indicateurs reflétant les comportements sportifs. Ensuite, appuyée par la théorie de la prévention, une méthodologie de calcul de l'impact a été mise en place. Elle consistait à utiliser un modèle de régression logistique permettant d'estimer la probabilité de survenance des sinistres en fonction du comportement sportif.

Cette approche a pu être mise en application à travers deux études de cas. La première avait pour vocation de montrer que la connaissance des groupes sportifs permettait d'anticiper les gains d'une stratégie de prévention. En outre, la proportion d'individus qui changeaient de classe rendait compte de la performance et de la rentabilité de celle-ci. Il a aussi été observé dans le cadre d'une seconde application que la fonction d'utilité expliquait la propension de l'assureur à accepter des pertes. Enfin, après la mise en place d'indicateurs adaptés, l'estimation de la rentabilité d'un programme de prévention, apportée par le cadre théorique, a permis d'apporter les éléments nécessaires au bon pilotage de la stratégie de l'assureur.

Les travaux réalisés ont aussi consisté à apporter une méthodologie de traitement et de valorisation des données issues des objets connectés. Dans un cadre de prévention, l'identification des profils sportifs rendue possible par les objets connectés a apporté une nouvelle vision comportementale qui a permis de cibler les actions de prévention et ainsi d'optimiser l'équilibre économique des portefeuilles santé.

Le domaine de la prévention est large et cette étude ne développe, de ce fait, qu'une partie du spectre des possibilités. Notamment, par la profondeur des données disponibles, cette étude s'est focalisée sur les impacts court terme et sur la médecine de ville. Dans un processus d'élargissement temporel de l'approche et de prise en compte des liens entre les différents postes de garantie, il serait ainsi possible, sur la base méthodologique déployée, de définir un impact global de la mise en place d'une stratégie de prévention.

Liste des tableaux

Table des figures

Glossaire

APS	Activité physique et sportive. 107, 111, 119
CFM	Le Coût fixe maximal correspond au coût de mise en place d'une prévention maximal à investir pour avoir le même niveau de rentabilité que si la prévention n'avait pas lieu. 103–105, 119
CNIL	Commission Nationale de l'Informatique et des Libertés. 23, 24, 119
CSP	catégorie socio-professionnelles. 58, 119
IOT	Internet of things. Appellation des objets connectés en anglais. 19, 22, 119
Open Data	Une donnée ouverte est une information publique brute, qui a vocation à être librement accessible et réutilisable.. 109, 110, 119
RGPD	Règlement général sur la protection des données. 23–25, 107, 117, 119

A Applications statistiques

A.1 Tests statistiques

A.1.1 Significativité des paramètres : Le test de Wald

Le test de Wald est un test paramétrique dont l'objectif est de déterminer si une variable explicative du modèle est significative. Soient X_i une variable explicative et β_i son coefficient associé dans le modèle considéré. L'hypothèse de ce test est telle que :

$$H_0 : \beta_i = 0 \text{ contre } \beta_i \neq 0$$

Ce test s'appuie sur l'hypothèse de normalité asymptotique des estimateurs du maximum de vraisemblance. Puisque cet estimateur est sans biais et suit une loi normale :

$$W = \frac{\hat{\beta}_i - E[\beta_i]}{\hat{\sigma}(\hat{\beta}_i)} = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}(\hat{\beta}_i)} \approx \chi_1^2$$

Sous l'hypothèse nulle, la statistique de test W devient :

$$W = \frac{\hat{\beta}_i}{\hat{\sigma}(\hat{\beta}_i)} \approx \chi_1^2$$

Le test de Wald va donc évaluer dans quelle mesure cette dernière affirmation est correct, en comparant la statistique W à un quantile X de loi χ_1^2. Ce test se fixe un niveau de confiance α tel que si la quantité $P(W > X_\alpha)$ appelée *p-value* est inférieure à α alors l'hypothèse est rejetée.

A.2 Corrélations

A.2.1 Corrélations en variables quantitatives

Soient X et Y deux variables aléatoires, dont le moment d'ordre 2 est défini.

Corrélations linéaire

La corrélation linéaire du couple (X, Y) est donnée par :

$$\rho(X,Y) = \frac{Cov(X,Y)}{Var(X)Var(Y)}$$

Lorsque X et Y sont indépendants, $\rho(X, Y) = 0$ Cependant, la réciproque n'est correcte que si X et Y sont des vecteurs gaussiens. De plus, ce coefficient ne prend pas en compte tous les types de dépendance. Il ne peut pas être utilisé dans n'importe quelle situation, puisqu'il n'est pertinent qu'en cas de distributions elliptiques ou de dépendance linéaire.

Les tau de Kendall et rho de Sperman sont des mesures de dépendance qui sont appelées corrélations de rang [9].

Tau de Kendall

Soit $(\tilde{X}, \tilde{Y})$ un couple de variable aléatoire avec la même loi jointe que (X, Y) alors :

$$\rho_T(X,Y) = \mathbb{P}[(X-\tilde{X})(Y-\tilde{Y})] > 0] - \mathbb{P}[(X-\tilde{X})(Y-\tilde{Y})] < 0]$$

Rho de Spearman

Soient F_x et F_y les fonctions de répartitions de X et Y.

$$\rho_S = \rho(F_x(X), F_y(Y))$$

A.2.2 Corrélations entre variables qualitatives

Traditionnellement, pour établir s'il existe un effet entre les deux variables qualitatives croisées dans un tableau de contingence, le test du chi-deux (χ^2) est utilisé. Le test V de Cramer permet de comparer l'intensité du lien entre les deux variables étudiées.
Soit le tableau de contingence ci-dessous :

X \ Y	d_1	...	d_k	...	d_s	total
c_1	n_{11}	...	n_{1k}	...	n_{1s}	$n_{1\bullet}$
⋮	⋮		⋮		⋮	⋮
c_h	n_{h1}	...	n_{hk}	...	n_{hs}	$n_{h\bullet}$
⋮	⋮		⋮		⋮	⋮
c_r	n_{r1}	...	n_{rk}	...	n_{rs}	$n_{r\bullet}$
total	$n_{\bullet 1}$	...	$n_{\bullet k}$	...	$n_{\bullet s}$	n

Figure *A.1:* Tableau de contingence

Le χ^2 se calcule par la formule suivante pour $h = 1, \ldots, r$ et $k = 1, \ldots, s$:

$$\chi^2 = \sum_{h,k} \frac{(n_{hk} - \frac{(n_{h.} \times n_{.k})}{n})^2}{\frac{(n_{h.} n_{.k})}{n}}$$

Le V de Cramer est la racine carrée du χ^2 divisé par le χ^2_{max}. Ce χ^2_{max} théorique est égal à l'effectif multiplié par le plus petit côté du tableau [1] moins 1.

$$V = \sqrt{\frac{\chi^2}{\chi^2_{max}}} = \sqrt{\frac{\chi^2}{n \times [min(r,s) - 1}})$$

Plus la statistique V est proche de zéro, moins les variables étudiées sont dépendantes. Il est égal à 1 lorsque les deux variables sont complètement dépendantes, puisque le χ^2 est alors égal au carré de χ_{max} [2]. Donc, plus V est proche de 1, plus la liaison entre les deux variables étudiées, est forte.

1. nombre de lignes ou de colonnes
2. dans un tableau 2 × 2, il prend une valeur comprise entre -1 et 1

A.3 Arbres de décisions

Les arbres de décision sont des algorithmes de prédiction qui fonctionnent en régression et en classification. Ils permettent de trouver une partition qui sépare au mieux les différentes observations. Une fois segmenté, il crée un ensemble de règles (séquences de décision uniques par groupe) en vue de la prédiction d'un résultat ou d'une classe [6].
En théorie des graphes, un arbre est un graphe non orienté, acyclique et connexe. L'ensemble des noeuds se divise en trois catégories :

- **Nœud racine** : l'accès à l'arbre se fait par ce nœud
- **Nœuds internes** : les noeuds qui ont des descendants
- **Nœuds terminaux** (ou feuilles) : nœuds qui n'ont pas de descendants.

Chaque individu, qui doit être attribué à une classe, est décrit par un ensemble de variables qui sont testées dans les noeuds de l'arbre. Les tests s'effectuent dans les nœuds internes, et les décisions sont prises dans les nœuds feuilles.

A.3.1 Apprentissage avec les arbres de décision

Dans cette partie, seul le problème de classification sera traité. Chaque élément x de la base de données est représenté par un vecteur multidimensionnel $(x_1, x_2, \dots x_n)$ correspondant à l'ensemble de variables descriptives du point comme dans le modèle GLM. Chaque noeud interne de l'arbre correspond à un test fait sur une des variables x_i :

- **Variable catégorielle** : génère une branche (descendant) par valeur de l'attribut
- **Variable numérique** : test par intervalles (tranches) de valeurs.

On obtient ainsi que les feuilles de l'arbre spécifient les classes et encodent la règle de décision. Une fois l'arbre construit, classer un nouvel individu se fait par une descente dans l'arbre, de la racine vers une des feuilles. A chaque niveau de la descente on passe un noeud intermédiaire ou une variable est testée pour décider du chemin à choisir pour continuer la descente.

A.3.1.1 Phase 1 : Construction

Au départ, les individus de la base d'apprentissage sont tous placés dans le noeud racine. Chaque noeud est coupé [3] donnant naissance à plusieurs noeuds descendants. Un élément de la base d'apprentissage situé dans un noeud se retrouvera dans un seul de ses descendants. Il existe plusieurs manières de faire des découpes. Il faut donc pouvoir mesurer la qualité de la découpe effectuée. L'arbre est construit par partitions successives de chaque noeud en fonction de la valeur de l'attribut testé à chaque itération. Le critère optimisé est l'homogénéité des descendants par rapport à la variable cible. La variable qui est testée dans un noeud sera celle qui maximise cette homogénéité. Le processus s'arrête quand les éléments d'un noeud ont la même valeur pour la variable cible.

3. opération split ou découpe

Des tests successifs sont effectués sur les $x_1, x_2 \dots x_n$. Chaque noeud feuille est homogène. Cela signifie que ses éléments (points dans chaque région) ont la même valeur pour l'attribut cible. A chaque étape, le but est de couper le noeud en deux régions les plus homogènes possibles. Il existe plusieurs types de séparation ou coupe possibles de l'espace des solutions.

- Une séparation par partition de variables (figure A.2 gauche) qui donne lieu à des découpes orthogonales ;
- Une séparation par combinaison linéaire de plusieurs variables (figure A.2 droite) qui donne lieu à des découpes obliques.

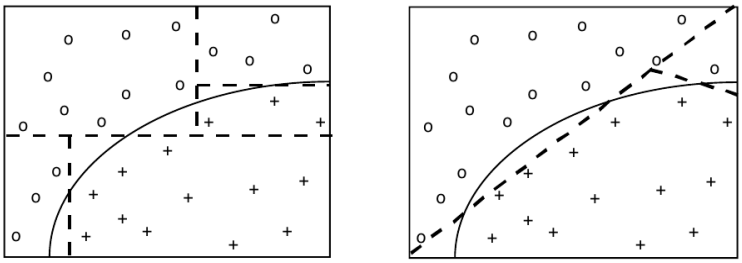


Figure *A.2:* Séparation de classes par partition itérative des variables.

A.3.1.2 Phase 2 : Élagage

Dans cette seconde étape, l'objectif est d'améliorer la qualité du modèle d'apprentissage en agissant sur les paramètres de l'arbre décisionnel. En effet, la performance n'augmente pas forcément avec la profondeur de l'arbre. Après ajustement du modèle, les branches peu représentatives peuvent être supprimées pour garder de bonnes performances prédictives.

A.4 *Random forest*

Les algorithmes de forêt aléatoire pour la régression et la classication sont des algorithmes qui utilisent des stratégies adaptatives (boosting) ou aléatoires (bagging). L'idée principale de cet algorithme est d'utiliser une agrégation d'un grand nombre de modèles tout en évitant le sur-apprentissage[11],[6].

A.4.1 Bagging

Soient Y la variable à expliquer, $X_1, \dots X_p$ les variables explicatives et ϕ le modèle appris sur un échantillon $z = \{(x_1, y_1)...(x_n, y_n)\}$. Le bagging consiste à suivre les étapes suivantes :

1. On considère B échantillons bootstrap $z_1, \dots z_B$ d'individus. Ces échantillons sont issus de z par tirage aléatoire avec remise.
2. Sur chacun des échantillons, on apprend un modèle $\phi(z_i)$.
3. On prédit Y en agrégeant les différentes décisions sur chacun des z_i par

$$Y = \hat{\phi}(x) = \frac{1}{B} \sum_{i=1}^{n} \phi_{z_i}(x) \tag{A.1}$$

Notons que cette manière d'agréger les résultats concerne exclusivement la régression. Pour une classification on prend la décision majoritaire.

A.4.1.1 Lien avec les forêts aléatoires

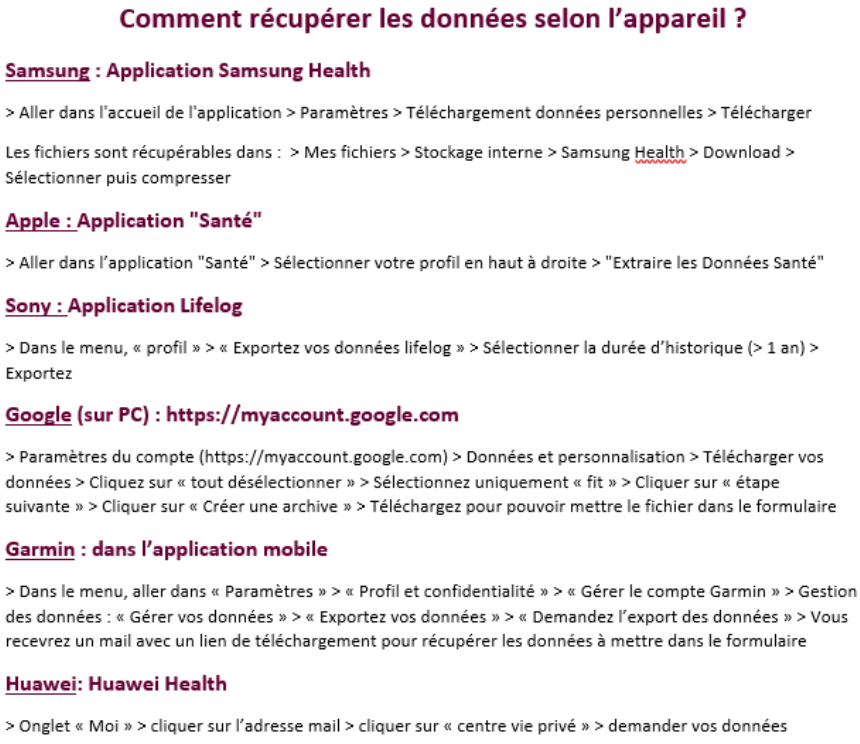
Les Forêts Aléatoires vont permettre l'amélioration du Bagging dans le cas spécifique de l'algorithme CART. L'objectif est de rendre les modèles (arbres) construits plus indépendants entre eux. Cette indépendance va permettre de rendre l'agrégation plus efficace.
Dans ce contexte, les B échantillons bootstrap représentent le nombre d'arbres formés dans la forêt. Le tirage aléatoire des variables explicatives à chaque noeud aboutit à des arbres non corrélés. Pour la construction de chaque noeud de chaque arbre, on tire uniformément q variables [4] parmi les p variables pour former la décision associée au noeud. En fin d'algorithme, on possède B arbres que l'on moyenne pour la régression.

4. En général, un choix optimal pour q est $q = \sqrt{p}$

B Base de données

B.1 Récupération des données

B.1.1 Procédure de récupération des données



Comment récupérer les données selon l'appareil ?

Samsung : Application Samsung Health

> Aller dans l'accueil de l'application > Paramètres > Téléchargement données personnelles > Télécharger

Les fichiers sont récupérables dans : > Mes fichiers > Stockage interne > Samsung Health > Download > Sélectionner puis compresser

Apple : Application "Santé"

> Aller dans l'application "Santé" > Sélectionner votre profil en haut à droite > "Extraire les Données Santé"

Sony : Application Lifelog

> Dans le menu, « profil » > « Exportez vos données lifelog » > Sélectionner la durée d'historique (> 1 an) > Exportez

Google (sur PC) : https://myaccount.google.com

> Paramètres du compte (https://myaccount.google.com) > Données et personnalisation > Télécharger vos données > Cliquez sur « tout désélectionner » > Sélectionnez uniquement « fit » > Cliquer sur « étape suivante » > Cliquer sur « Créer une archive » > Téléchargez pour pouvoir mettre le fichier dans le formulaire

Garmin : dans l'application mobile

> Dans le menu, aller dans « Paramètres » > « Profil et confidentialité » > « Gérer le compte Garmin » > Gestion des données : « Gérer vos données » > « Exportez vos données » > « Demandez l'export des données » > Vous recevrez un mail avec un lien de téléchargement pour récupérer les données à mettre dans le formulaire

Huawei: Huawei Health

> Onglet « Moi » > cliquer sur l'adresse mail > cliquer sur « centre vie privé » > demander vos données

Figure *B.1:* Les méthodes de récupération de données

B.1.2 Questionnaire de collecte de données

Question 1 : Vous êtes : Un homme ; Une femme
Question 2 : Quel est votre âge ?
Question 3 : Quel est votre poids ?

Question 4 : Quelle est votre taille ?
Question 5 : CSP : Cadre ; Non Cadre ; Étudiant
Question 6 : Secteur d'activité
Question 7 : Quel est votre Code postal ?
Question 8 : Pays
Question 9 : Quelle est votre situation familiale ?
Question 10 : Combien de consultations chez le généraliste estimez-vous effectuer par an ?
Question 11 : Possédez-vous une complémentaire santé ?
Question 12 : Quel type d'activité physique pratiquez-vous ?
Question 13 : A quelle fréquence pratiquez-vous cette activité physique ?
Question 14 : Sur une échelle de 1 à 10 quelle note donneriez-vous à votre hygiène de vie ?
Question 15 : De quel type d'appareil provient vos données ?
Question 16 : Quelle est la marque de votre appareil de mesure ?
Question 17 : Si Autre
Question 18 : Données récupérées
Question 19 : Consentez-vous à ce que vos données soient utilisées dans le cadre de cette étude ?

B.2 Données externes

B.2.1 Notations types d'activités

Table *B.1:* Numéro d'activité physique dans la base de donnée

N	sport
1	Aucune
2	Cyclisme
3	Sports collectifs
4	Sports de raquette
5	Sports de combat et arts martiaux
6	Sports nautiques
7	Musculation etc.
8	Sports de forme (fitness
9	Sports de plein air et de nature
10	Sports de cible
11	Épreuves combinées (triathlon etc.).
12	Sports mécaniques
13	Athlétisme
14	Sport équestre
15	Danse

B.2.2 Table d'équivalence de dépenses énergétiques des activités sportives en 15min et 1h

15min	Sport	1h	Groupe
26	Stretching, hatha yoga	102	Groupe 1
34	Voile (loisir)	136	Groupe 1
34	Surf	136	Groupe 1
34	Mini-golf	136	Groupe 1
34	Marche (4 km/h)	136	Groupe 1
34	Bowling	136	Groupe 1
43	Tir à l'arc	170	Groupe 1
43	Canoé-kayak (loisir)	170	Groupe 1
51	Équitation (loisir)	204	Groupe 2
51	Volleyball	204	Groupe 2
51	Tennis de table	204	Groupe 2
51	Tai chi	204	Groupe 2
51	Moto-cross	204	Groupe 2
51	Gymnastique	204	Groupe 2
60	Golf	238	Groupe 2
60	Badminton (loisir)	238	Groupe 2
65	Danse classique	258	Groupe 2
85	Vélo (16 km/h)	340	Groupe 2
85	Ski de piste (descente)	340	Groupe 2
85	Poids, haltérophilie	340	Groupe 2
85	Escrime	340	Groupe 2
85	Course automobile, rallye	340	Groupe 2
85	Boxe (punching ball)	340	Groupe 2
90	Marche (7 km/h)	360	Groupe 3
94	Fitness, aérobic	374	Groupe 3
102	Tennis	408	Groupe 3
102	Squash	408	Groupe 3
102	Skate, roller	408	Groupe 3
102	Rameur	408	Groupe 3
102	Patin à glace (loisir)	408	Groupe 3
102	Natation (brasse)	408	Groupe 3
102	Football (loisir)	408	Groupe 3
102	Bobsleigh, luge	408	Groupe 3

15min	Sport	1h	Groupe
102	Rameur	408	Groupe 3
102	Patin à glace (loisir)	408	Groupe 3
102	Natation (brasse)	408	Groupe 3
102	Football (loisir)	408	Groupe 3
102	Bobsleigh, luge	408	Groupe 3
119	Ski de fond (loisir)	476	Groupe 3
119	Jogging (8 km/h)	476	Groupe 3
119	Hockey sur glace	476	Groupe 3
119	Basket	476	Groupe 3
119	Ski de fond (loisir)	476	Groupe 3
119	Jogging (8 km/h)	476	Groupe 3
119	Hockey sur glace	476	Groupe 3
119	Basket	476	Groupe 3
153	Waterpolo	612	Groupe 4
153	Vélo (24 km/h)	612	Groupe 4
153	Rugby	612	Groupe 4
153	Judo, karate, kick-boxing	612	Groupe 4
153	Football (compétition)	612	Groupe 4
170	Natation (crawl)	680	Groupe 4
170	Escalade	680	Groupe 4
187	Handball	748	Groupe 4
187	Boxe	748	Groupe 4
238	Jogging (15 km/h)	952	Groupe 4
153	Waterpolo	612	Groupe 4
153	Vélo (24 km/h)	612	Groupe 4
153	Rugby	612	Groupe 4
153	Judo, karate, kick-boxing	612	Groupe 4
153	Football (compétition)	612	Groupe 4
170	Natation (crawl)	680	Groupe 4
170	Escalade	680	Groupe 4
187	Handball	748	Groupe 4
187	Boxe	748	Groupe 4
238	Jogging (15 km/h)	952	Groupe 4

Bibliographie

[1] Mehdi Ammi : Analyse économique de la prévention. offre de prévention, incitations et préférences en médecine libérale, 2013. https ://tel.archives-ouvertes.fr/tel-00859358/document.

[2] Sandrine Babin : Tarification en assurance emprunteur : Création de tables de mortalité d'expérience après segmentation du portefeuille par scoring, 2018. http ://www.ressources-actuarielles.net/EXT/ISFA/1226-02.nsf/d512ad5b22d73cc1c1257052003f1aed/acc0cf7001b492a4c1257fb600271f13/$FILE/BABIN.pdf.

[3] Philippe Caillou : Coordination d'agents coopÉratifs :formation de coalitions, 2013-2014. https ://www.lri.fr/ caillou/c6CooperationCoalitionsv1.pdf.

[4] CNIL : Protéger les données personnelles ; accompagner l'innovation ; préserver les libertés individuelles, 2018. https ://www.cnil.fr/sites/default/files/atoms/files/cnil_en_bref_2018.pdf.

[5] Becker Ehrlich : Market insurance, self-insurance and self-protection, 1972. http ://econ2.econ.iastate.edu/classes/econ642/Babcock/ehrlich%20and%20becker.pdf.

[6] Mayoro FALL Ehui Aman-Yah Andréa : Application de méthodes de machine learning au provisionnement non-vie, 2018. https ://euria.univ-brest.fr/digitalAssets/73/73837_BE05.pdf.

[7] Benoît Dervaux et Louis Eeckhoudt : Prévention en économie et en médecine, 2004. https ://www.cairn.info/revue-economique-2004-5-page-849.htm.

[8] EY : Introducing 'pay as you live' (payl) insurance, 2015. https ://www.ey.com/Publication/vwLUAssets/EY-intro ducing-pay-as-you-live-payl-insurance/$FILE/EY-introducing-pay-as-you-live-payl-insurance.pdf.

[9] Brice Franke : Dépendance linéaire et non linéaire, 2018. cours EURIA.

[10] République Française : Comité interministérielle pour la santé - priorité prévention, 2018. https ://solidarites-sante.gouv.fr/IMG/pdf/180326-dossier_de_presse_priorite_prevention.pdf.

[11] Sebastien Gadat : Arbres de classifications et forêts aléatoires, 2015. https ://www.math.univ-toulouse.fr/ gadat/Ens/M2SID/11-m2-Random_Forests.pdf.

[12] Romain Gauchon : Une nouvelle méthode de classification permettant de cibler des actions de prévention, 2018. http ://www.ressources-actuarielles.net/EXT/ISFA/1226-02.nsf/d512ad5b22d73cc1c1257052003f1aed/0ddc88b70ba86b4ec125825500323d0d/$FILE/Memoire%20R
V4.002.pdf/Memoire%20RGA-V4.pdf.

[13] Pierre-Louis GONZALEZ : 'analyse encomposantes principales(a.c.p.), 2016. http ://maths.cnam.fr/IMG/pdf/A-C-P-.pdf.

[14] Simon Hermant : Analyses d'un portefeuille non vie : branche do, 2011. http ://www.ressources-actuarielles.net/EXT/ISFA/1226-02.nsf/d512ad5b22d73cc1c1257052003f1aed/895080339a46e21dc1257a3c0032a322/$FILE/Memoire%20HE

[15] Meglena Jeleva : Économie du risque, 2019. Formation CEA.

[16] Salanié F. Jullien B., Salanié B. : should more risk averse agents exert more effort , geneva papers on risk and insurance theory, 1999. 24, 19-25.

[17] Fabien Lange : Exploration de la valeur de shapley et des indices d'interaction pour les jeux définis sur des ensembles ordonnés, 2008. https ://tel.archives-ouvertes.fr/tel-00274302/document.

[18] Jean-Yves Lesueur : L'adhésion des assurés aux programmes de prévention santé : Quels facteurs explicatifs ? 2018. Université de Lyon, Chaire Prevent Horizon, http ://chaire-prevent-horizon.fr/files/2019/03/JYL-Preprint.pdf.

[19] Jean-Yves Lesueur : Quel impact des programmes de prévention santé sur l'arbitrage « auto-prévention – assurance » ?, 2018. http ://chaire-prevent-horizon.fr/files/2018/02/Pr%C3%A9sentationJYL-esueur_Petit_Dej_PreventHorizon-8_f%C3%A9vrier2018.pdf.

[20] NJOMO NANA YANNICK LIONEL : Segmentation et tarification de la garantie assistance automobile, 2016. http ://www.ressources-actuarielles.net/EXT/ISFA/1226-02.nsf/d512ad5b22d73cc1c1257052003f1aed/391d05ae99c0ccb4c1257fc2002787a2/$FILE/NJOMO%20NAI

[21] L'assurance Maladie : Améliorer la qualité du système de santé et maîtriser les dépenses propositions de l'assurance maladie pour 2019, 2019. https ://assurance-maladie.ameli.fr/sites/default/files/rapport-charges-et-produits-2019-web.pdf.

[22] Goodwill Management : Mesure de l'impact du sport en entreprise - cnosf - medef - ag2r, 2015. http ://www.goodwill-management.com/fr/realisations/mesure-de-l-impact-du-sport-en-entreprise-cnosf-medef-ag2r.

[23] Institut montaigne : Big data et objets connectés faire de la france un champion de la révolution numérique, 2015. https ://www.institutmontaigne.org/ressources/pdfs /publications/rapport%20objets%20connecte%CC%81s(2).pdf.

[24] OMS : Que faire pour éviter une crise cardiaque ou un accident vasculaire cérébral ? 2015. https ://www.who.int/features/qa/27/fr/.

[25] OMS : Maladies cardiovasculaires. 2017. https ://www.who.int/fr/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) https ://www.who.int/features/qa/27/fr/.

[26] GWENDAL PERRIN : Generali vitality : les premiers enseignements du programme. 2017. https ://www.argusdelassurance.com/acteurs/compagnies-bancassureurs/generali-vitality-les-premiers-enseignements-du-programme.124048.

[27] Ricco Rakotomalala : Pratique de la régression logistique - régression logistique binaire et polytomique, 2017. http ://eric.univ-lyon2.fr/ ricco/cours/cours/pratique_regression_logistique.pdf.

[28] Jordan Marie Rose : Modélisation comportementale : impact de la couverture santé sur les dépenses en dentaire, 2018. http ://www.ressources-actuarielles.net/C12574E200674F5B/0/402E86F3DF39EFBFC12583E400284376.

[29] Kevin SADOUN : Apport des télématiques dans la segmentation tarifaire en assurance automobile, 2015. http ://www.ressources-actuarielles.net/C12574E200674F5B/0/2EF09BF7724C8EF3C125806500273FB1.

[30] Marion Del Sol : Enjeux juridiques des objets connectés en matière d'assurance santé. réflexions à partir et au-delà du cadre français, 2018. https ://halshs.archives-ouvertes.fr/halshs-01836170/document.

[31] Augustin Loubatan Tabo : Analyse économique des comportements de prévention face aux risques de santé, year =.

[32] Augustin Loubatan Tabo : Analyse économique des comportements de prévention face aux risques de santé, 2014. https ://tel.archives-ouvertes.fr/tel-00949540/document.

[33] Tudor-Locke : How many steps/day are enough ? preliminary pedometer indices for public health., 2004. https ://www.ncbi.nlm.nih.gov/pubmed/14715035.

[34] Franck Vermet : Apprentissage statistique, une approche connexionniste, 2017. support de cours EURIA.