DAUPHINE UNIVERSITÉ PARIS

Master Actuariat de Dauphine

DAUPHINE UNIVERSITÉ PARIS

Mémoire présenté devant l'Université Paris Dauphine pour l'obtention du diplôme du Master Actuariat et l'admission à l'Institut des Actuaires

le 21/01/2016

Par : Ivan HERBOCH

Titre: Predictive Analytics en actuariat : application à la modélisation de la résiliation non-vie

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

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

Membre présent du jury de l'Institut des Actuaires :

Emmanuel Berthelé

Signature :

Entreprise : Deloitte Conseil

Nom : Claude Chassain

Signature :

Directeur de mémoire en entreprise :

Nom : Charlotte CHOQUET

Signature :

Membres présents du jury du Master Actuariat de Dauphine :

Marc Hoffmann

Vincent Rivoirard

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

Signature du responsable entreprise :

Secrétariat :

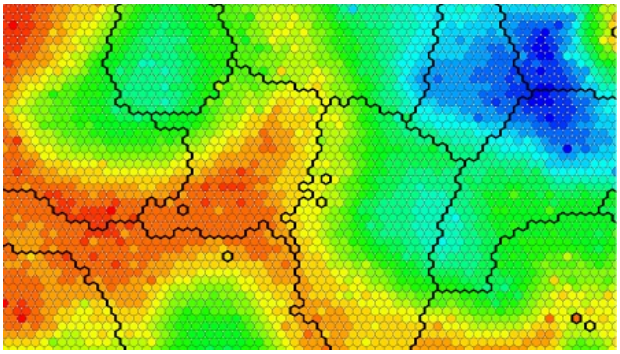
Bibliothèque :

Signature du candidat :

Université Paris-Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 PARIS Cedex 16

Ivan Herboch

Predictive Analytics en actuariat : application à la modélisation de la résiliation non-vie



Directeur de mémoire Entreprise : **Jean-Guillaume Zanotti**
Directeur de mémoire Entreprise adjoint : **Charlotte Choquet**
Directeur de mémoire Université Paris Dauphine : **Vincent Rivoirard**
Directeur de mémoire Ecole Centrale Paris : **Laurence Lévêque**

Résumé

Mots-clés : *Predictive Analytics, cartes auto-adaptatives, segmentation, résiliation, loi Hamon, assurance auto, lift, forêts aléatoires, data science*

La révolution du Predictive Analytics consiste à exploiter la masse de données nouvellement disponible afin d'établir des hypothèses prédictives à partir de l'analyse du passé et du présent. Ces analyses se basent sur des techniques d'apprentissage automatique. Leur utilisation est rendue possible par la puissance de calcul à disposition des scientifiques des données (data scientists) et notamment des actuaires. L'objectif du mémoire est d'utiliser ces algorithmes, et plus particulièrement un type de réseaux de neurones appelé carte auto-adaptative, pour traiter un sujet d'actualité : la loi Hamon. Celle-ci donne de nouveaux droits en matière de résiliation pour les assurés non-vie et met une pression supplémentaire sur un secteur déjà très concurrentiel. L'étude se base sur un portefeuille d'assurance auto fourni dans le cadre d'un partenariat avec un assureur. L'approche du mémoire consiste à segmenter le portefeuille en groupes homogènes d'assurés pour lesquels l'intuition permet déjà de savoir si ceux-ci ont ou non une forte propension à résilier. L'intérêt de l'approche consiste à trouver parmi ces groupes des poches clients dont le profil de résiliation est différent de la moyenne du groupe en question. L'apport de la connaissance permet d'optimiser les actions opérationnelles envers les assurés.

Abstract

Keywords : *Predictive Analytics, self-organizing maps, segmentation, lift, churn, loi Hamon Consumer Bill, car insurance, random forests, data science*

The Predictive Analytics revolution aims at using the newly available data in order to make predictive hypotheses from analyzing the present and the past. These analyses use various algorithms. Their execution is possible because of the computing power available for data scientists and particularly, actuaries. The goal of this actuarial thesis is to harness data science algorithms, and more precisely an artificial neural network called self-organizing map, in order to address a current topic : the French loi Hamon Consumer Bill. It offers new cancellation rights for the non-life customers and adds a competitive pressure in an already competitive market. The actuarial study is based on a car insurance portfolio provided by a French insurer. The thesis' approach is to cluster the portfolio into homogeneous groups. The evaluation of the churn rate in these clusters is intuitive since they are easily explainable. The main interest of the approach is to find small customer groups in these clusters that have a churn behavior different from the average churn rate of the cluster. The acquired knowledge is used to focus optimization measures on profitable customers.

Synthèse

Mots-clés : *Predictive Analytics, cartes auto-adaptatives, segmentation, résiliation, loi Hamon, assurance auto, lift, forêts aléatoires, data science*

Le *Predictive Analytics* est l'utilisation des différentes méthodes statistiques comme l'apprentissage automatique (machine learning), l'analyse exploratoire de données (data mining) ou la théorie des jeux afin d'analyser le présent et le passé pour faire des hypothèses prédictives sur des évènements futurs. Ces modèles cherchent à établir des schémas dans les données pour identifier les risques et les opportunités. Le *Predictive Analytics* peut notamment être utilisé en actuariat et ceci est exposé dans cette étude.

L'objectif du mémoire est de donner une meilleure connaissance de la résiliation en assurance non-vie. En effet, la loi Hamon, entrée en vigueur le 1er janvier 2015, autorise les assurés non-vie à résilier à l'expiration d'un délai d'un an à compter de la première souscription, sans frais ni pénalités les contrats tacitement reconductibles. Cette loi met sous pression les assureurs dans un marché très concurrentiel et à marges réduites. Elle les oblige à mieux sélectionner leurs assurés, notamment en fonction de leur comportement de résiliation, afin de ne conserver que des clients qui résilient peu. Cela permet d'amortir les frais d'acquisition du client sur une période plus longue et ainsi augmenter la rentabilité de l'assureur, dès lors que la sinistralité du client n'est pas trop importante.

L'étude de la résiliation se base une portefeuille de contrats d'assurance auto fourni par un assureur français avec lequel un partenariat a été établi pour le mémoire. Améliorer la compréhension du portefeuille vis-à-vis de la résiliation passe par trois étapes :

— Segmentation du portefeuille hétérogène en clusters homogènes et analyse des taux de résiliation

— Choix d'un cluster d'intérêt et obtention de poches clients dans le cluster qui sont très différentes du comportement moyen des assurés du cluster

— Compréhension des facteurs de la résiliation au sein de poches clients d'intérêt

L'utilité de cette modélisation est d'identifier des profils de clients dont le comportement de résiliation est contraire à l'intuition. Par exemple, identifier parmi les jeunes conducteurs, qui à priori ont une forte propension à résilier, une poche client constituée de jeunes qui résilient peu est très intéressant pour l'assureur. En effet, il peut par exemple cibler ses campagnes marketing vis-à-vis des jeunes conducteurs en priorité sur ce type de profil identifié. De plus, les actions de rétention ou même de réduction tarifaire peuvent être accordées de préférence aux profils identifiés comme intéressants plutôt qu'à ceux qui le sont moins. L'apport de cette connaissance permet d'augmenter la rentabilité de l'assureur.

Le mémoire commence par la définition du contexte de l'assurance non-vie en décrivant les principales garanties et chiffres clés de l'assurance auto et habitation. Le loi Hamon est ensuite abordée et les problématiques soulevées par cette nouvelle mesure sont exposées. Le *Predictive Analytics* est ensuite défini et relié à la révolution du Big Data. Quelles sont les applications du *Predictive Analytics* pour créer de la valeur ? Le

mémoire détaille ensuite les phénomènes sous-jacents qui motivent l'analyse de données. Le Big Data n'est alors qu'une extension du *Predictive Analytics* dans le cas où les volumes de données traités sont importants, où les données utilisées proviennent de sources très variées et où elles doivent être traitées très rapidement. Ces trois points constituent les 3 V du Big Data, pour Volume, Variété et Vitesse. La variété des sources de données traitées permet d'apporter un nouveau regard sur l'activité de l'entreprise, par exemple en donnant de nouvelles informations sur les clients. Le mouvement de l'Open Data vise à mettre à disposition de tous des données publiques. La France participe à cet essor et a mis en place un portail regroupant les données venant de plusieurs entités du secteur public, notamment de l'INSEE. Les différentes types de données sont évoqués ainsi que les limites légales quant à leur utilisation.

Il s'attelle ensuite à l'étude à proprement parler. La première partie consiste en la prise en main des données fournies par le Partenaire. Le portefeuille de contrats d'assurance auto est constitué d'environ 2 millions de polices. Le mémoire explique alors comment les données fournies ont été retraitées pour atteindre l'objectif. Les valeurs manquantes dans les données sont imputées avec des techniques d'apprentissage automatique, afin d'obtenir une complétion des valeurs manquantes plus élaborée que l'approche habituelle consistant à remplacer une valeur manquante par la moyenne ou la médiane.

Le retraitement des valeurs manquantes à l'aide de ces techniques et en utilisant les données de Google Maps permet de ne pas supprimer certaines variables ou observations de la base. Il est donc possible d'utiliser dans une certaine mesure les variables retraitées, dès lors que le biais impliqué par l'utilisation de la variable retraitée n'est pas trop important.

Le portefeuille a ensuite été segmenté afin d'obtenir des clusters homogènes. Pour cela, l'algorithme des cartes auto-adaptatives (*Self-Organizing Maps*) a été employé. Il s'agit d'une technique appartenant aux réseaux de neurones artificiels. Cette méthode vise à modéliser le fonctionnement conjoint des capteurs rétiniens et du cortex cérébral. Des capteurs rétiniens proches sont associées à des zones dans le lobe occipital proches elles aussi. Les cartes auto-adaptatives résument les données en entrée à une grille de neurones aisément compréhensible, de telle sorte que deux vecteurs en entrée soient associés à des neurones voisins dans la carte auto-adaptative. Cette approche permet d'obtenir des clusters aisément interprétables, contrairement aux autres techniques comme la méthode des K-moyennes.

Les 12 clusters obtenus sont par exemple les jeunes conducteurs, les clients âgés et fidèles sans réduction tarifaire ou même les clients multi-équipés. Les taux de résiliation au sein de chaque cluster sont ensuite analysés et l'étude se poursuit en cherchant des poches clients homogènes au sein d'un cluster dont l'étude intéresse particulièrement le Partenaire. Le travail consiste à trouver des poches client ayant une propension à résilier très éloignée de la propension moyenne du cluster, de telle sorte que ces profils auparavant considérés comme peu intéressants pour l'assureur soient en fait très intéressants à prospecter.

Le cluster retenu pour l'étude est celui des clients âgés et fidèles sans réduction tarifaire. Il présente un taux de résiliation et une durée de vie du contrat moyenne du

contrat supérieure à la moyenne. Au sein de ce cluster, un clustering à l'aide d'une carte auto-adaptative a permis d'obtenir 10 poches clients homogènes aux profils de résiliation dont certains ont des profils de résiliation éloignés de la moyenne.

Une fois ces poches obtenues, deux poches en particulier présentent de l'intérêt car elles ont respectivement une très forte et une très faible propension à résilier par rapport à la moyenne des clients âgés et fidèles sans réduction tarifaire. L'approche a alors consisté a mettre en place un modèle explicatif pour connaître les variables qui influent sur la résiliation au sein de chaque poche.

L'intérêt de la modélisation considérée est qu'elle permet, à l'aide des cartes auto-adaptatives, d'avoir une vision claire du portefeuille et d'interpréter aisément les clusters et poches obtenus. L'algorithme employé a l'avantage d'être très robuste et les clusters obtenus sont donc assez stables. Une légère modification des données en entrée ne changera pas le résultat du clustering. Cette approche, contrairement à une approche traditionnelle qui donne des informations sur le comportement moyen des assurés, permet d'apporter de l'information sur le comportement individuel des assurés.

L'approche, basée sur les cartes auto-adaptatives, a l'avantage de pouvoir être mise à jour facilement dans le cas où la structure du portefeuille change légèrement. En effet, l'algorithme est, comme son nom l'indique, auto-adaptatif et peut donc être ré-exécuté sur le nouveau portefeuille à partir des résultats obtenus sur l'ancien.

La modélisation considérée présente certaines limites. Tout d'abord, l'analyse ne prend pas en compte la charge de sinistres car cette variable n'était pas disponible. Or, une poche de clients qui résilient peu est peut être aussi une poche de clients très sinistrés donc celle-ci n'est pas forcément intéressante pour l'assureur. De plus, les nombreux retraitements effectués et notamment la création de *dummy variables* introduisent un biais qui peuvent donner plus d'importance à certaines variables, notamment les variables ayant beaucoup de modalités. Les résultats obtenus sont donc dépendants des retraitements effectués.

Pour obtenir une nouvelle validation de la modélisation, il serait intéressant de tester l'approche sur un autre portefeuille d'assurance automobile pour comparer les résultats obtenus. De plus, la prise en compte de la sinistralité peut apporter de nouvelles pistes et permettre à terme l'élaboration d'un modèle de la valeur-client basé en partie sur l'approche du mémoire.

Enfin, pour aller plus loin, il aurait été intéressant de décloisonner les informations disponibles, d'une part à la maille de la poche client et d'autre part, au sein du portefeuille tout entier. En effet, quantifier la probabilité d'appartenance aux différentes poches client plutôt que d'avoir l'information sur l'appartenance à une poche donnée peut apporter de l'information supplémentaire sur la résiliation.

Executive summary

Keywords : *Predictive Analytics, self-organizing maps, segmentation, lift, churn, loi Hamon Consumer Bill, car insurance, random forests, data science*

The Predictive Analytics is the use of various statistical methods such as machine learning, data mining or even game theory in order to make predictive hypotheses on future events from analyzing the present and the past. These models aim at finding data patterns to identify risks and opportunities. The thesis will use the Predictive Analytics background.

The goal of the thesis is to allow a better knowledge of churn risk in non-life insurance. The French loi Hamon Consumer Bill, whose entry into force was on January 1st 2015, allows French car and home insurance owners to cancel their policy later than one year after the underwriting without charge or penalty. This bill adds a pressure onto non-life insurers that are already evolving in a competitive and reduced margins market. It forces insurers to better select their customers, particularly according to their churn behavior. The goal is to keep low churners and get rid of high churners. This allows the insurer to have an amortization of acquisition costs on a longer period and therefore increases the profitability of the company.

The actuarial study is based on a car insurance portfolio provided by a French insurer. The improvement of the churn knowledge in the portfolio is divided into three steps :

— segmentation of the heterogeneous portfolio into homogeneous clusters and analysis of the churn rates

— Choice of a cluster and finding of small customer groups with a churn behavior different from the average churn rate of the cluster

— Correlation of the churn action with predictor variables into the small customer groups

The main interest of the method is to identity small customer groups with a counter-intuitive churn behavior. For example, identifying among the young drivers, that are apparently high churners, a small customer group that has a low churn rate is value-adding for the insurer. Indeed, it can target its marketing actions linked to the young drivers to these particular types of customers. Moreover, the retention actions or tariffs reduction can preferentially target these customers. The acquired knowledge is used to increase the insurer's profitability.

The thesis begins with the definition of the non-life insurance context, describing the main insurance coverages and the key figures of car and home insurance. The loi Hamon Consumer Bill and the issues being raised because of it are then presented. The Predictive Analytics is then defined and linked to the Big Data revolution. What are the fields of application of the Predictive Analytics in order to create value ? The thesis will then focus on the underlying phenomenon motivating the data analysis. The Big Data is just Predictive Analytics extended to huge data volumes, with a variety of data sources, and velocity in the processing of the information. This is the 3 V of Gartner's Big Data definition. The regulatory framework of data privacy is then discussed.

The actuarial study really begins with acquiring a better understanding of the provided database. The car insurance portfolio is formed by 2 millions of policies. The thesis explains the reprocessing of data in order to achieve the goal. The missing values in the data are imputed with machine learning algorithms and it allows a better completion of missing values than the usual approach that imputes the mean or the median.

Moreover, the missing values imputation using these techniques and using Google Maps data allows to keep some variables or data rows. It is possible to use, if the bias is not too important, the reprocessed variables.

The portfolio is then segmented in order to get homogeneous clusters. The self-organizing map algorithm has been used. It is a method of the artificial neural network class. This technique is modelling the way the visual field of the eye and the posterior cortex in the brain are jointly arranged. Two close retinal captors in the eye are associated with close neural zones in the posterior cortex. The self-organizing maps summarize input data in an easily understandable neural grid, so that two input vectors are linked to two neighbor neurons in the self-organizing map. This approach allows to obtain easily interpretable clusters, in opposition to methods such as K-means.

The 12 obtained clusters are young drivers or even multiple policies customers. The cancellation rates in the twelve clusters are then analyzed and the study continues looking for small customer groups that are interesting to focus on for the partner company. The model aims at finding small customer groups with a churn behavior different from the average churn rate of the related cluster, so that these profiles previously not interesting for the insurer are identified as interesting to prospect.

The chosen cluster for the study is the one of old customers that have a long policy duration and no tariff reduction. It has a churn rate and a policy duration higher than the average. Among this cluster, a clustering with a self-organizing map has been used in order to obtain 10 small customers groups with some having a remarkable profile compared to the average churn rate of the cluster.

After this clustering, two small customer groups has been studied, one composed of low churners and the other made of high churners when compared to the whole cluster.

The main interest of the modelling is that it gives a clear overview of the portfolio, thanks to the self-organizing map algorithm. Moreover, it makes the clusters and small customer groups easily interpretable. The implemented model is very robust compared to other clustering techniques. The obtained clusters are therefore really stable. A small modification of the input data will not change the resulting clusters. Moreover, this approach allows an individual understanding of the churn behavior.

The approach, based on the self-organizing maps, has the advantage of being easily updated in the case the data structure slightly changes. Indeed, the algorithm can be recomputed starting from the results obtained with the data before modification.

But the model has some limits. The first one is that the claim severity is not taken into account in the study because it was not available in the input data. Because low churners can also be risky customers in terms of severity, taking this variable into account can be interesting for the insurer. Moreover, the preprocessing of the data introduced an unavoidable bias, mostly because of the created dummy variables. The results are in consequence dependent of the variables reprocessing.

To go further into the model, it would have been interesting to test the approach on an other car insurance portfolio and to compare the results. Moreover, taking into account the claim severity would allow to build a customer lifetime value model.

Finally, to go further, it would have been a good idea to quantify the probability for a customer to be in the portfolio obtained small groups. Indeed, the knowledge of this probability would have brought additional information because the churn rate would have been more individual than with our approach.

Remerciements

Je remercie Jean-Guillaume Zanotti et Charlotte Choquet, respectivement directeur et manager dans l'équipe Actuariat Assurance du cabinet Deloitte Conseil, pour leur suivi régulier, leurs conseils et l'intérêt porté à mes travaux.

Je tiens également à remercier Claude Chassain, associée du cabinet Deloitte Conseil, pour m'avoir accueilli au sein de son département et m'avoir permis de réaliser ces travaux.
Mes remerciements vont également à l'entreprise partenaire ayant fourni les données grâce auxquelles j'ai pu effectuer mes travaux et plus particulièrement aux collaborateurs du service concerné pour le temps qu'ils m'ont consacré.

Je tiens à remercier Lionel Gabet, directeur de l'option Mathématiques Appliquées à l'Ecole Centrale Paris, Valérie Ferrebœuf, responsable de la filière Stratégie Finance à l'Ecole Centrale Paris, ainsi que Marc Hoffmann, responsable du Master Actuariat de l'Université Paris Dauphine. Les enseignements reçus à l'Ecole Centrale Paris et à l'Université Paris Dauphine m'ont passionné et me permettent d'arriver sereinement dans le monde professionnel.

Merci à Laurence Lévêque et Vincent Rivoirard, mes tuteurs à l'Ecole Centrale Paris et à l'Université Paris Dauphine, pour m'avoir accompagné tout au long de mon stage.

Enfin, je tiens à remercier ma famille, mes parents Laurent et Hélène et ma sœur Marion, pour le soutien sans faille que vous m'avez apporté pendant toutes ces années et sans lequel rien n'aurait été possible.

Table des matières

Introduction générale

Le Big Data est aujourd'hui une expression employée dans tous les secteurs d'activité. Cependant, peu d'emplois de ce terme désignent réellement le Big Data. Ce dernier a une définition bien précise qui passe par les 3 V du Big Data : le Volume (de données), la Vitesse (d'exécution) et la Variété (des données utilisées). La révolution en cours, souvent désignée comme celle du Big Data, est en fait la révolution des données. Le coût de stockage d'un gigaoctet de données a été divisé par 10000 entre 1995 et 2015 avec pour conséquence la multiplication des données disponibles.

Cette révolution des données a accompagné le développement d'Internet. De nouveaux besoins en termes d'analyse de données ont émergé, en particulier grâce aux moteurs de recherche (par exemple Google) qui cherchent à valoriser les données dont ils disposent. La recherche académique a permis le développement de nouvelles méthodes statistiques, dites de l'apprentissage automatique, qui ont la capacité d'analyser de gros volumes de données pour en déduire des comportements ou règles sans à priori sur les données sous-jacentes. Ces méthodes ont permis l'exploration de la masse de données nouvellement à disposition sans utiliser les outils statistiques classiques (tests statistiques, régression linéaire multiple, ...) non adaptés aux besoins du Web.

Le secteur de l'assurance dont la rentabilité dépend étroitement de la capacité à gérer les risques souscrits à partir des données collectées est amené à être de plus en plus concerné par ce bouleversement. Les outils de l'apprentissage automatique, parfois appelé l'Analytics, sont appelés à faire entrer l'analyse des données en assurance dans une nouvelle ère. Or, les algorithmes comme les réseaux de neurones (1957 pour le perceptron) ou les forêts aléatoires (2001) sont traditionnellement encore très peu utilisés par les actuaires. Cela se justifiait par l'aspect encore exploratoire de ces techniques mais dès lors que la puissance de calcul n'est plus un frein à leur implémentation, un nouveau champ s'ouvre aux sciences actuarielles.

L'objectif de ce mémoire est de faire le lien entre cette révolution et l'actuariat, à travers à un sujet d'actualité : la loi Hamon. En permettant aux assurés détenant depuis plus d'un an leurs contrats auto ou MRH de les résilier sans frais ni pénalités, la loi Hamon rend plus délicat encore le maintien des faibles marges des assureurs non-vie dans un domaine déjà fortement concurrentiel. Dans ce contexte, mieux comprendre les catégories d'assurés ayant une propension à résilier éloignée du comportement moyen devient donc crucial pour les assureurs car cela peut permettre, par exemple, de mieux cibler des actions marketing et de les différencier selon les populations. En traitant ces problématiques, l'Analytics peut constituer un fort levier de développement commercial rentable

Le mémoire commence avec la description du contexte motivant l'étude de la résiliation non-vie et l'approche considérée qui utilise les techniques de l'Analytics. Cette étude de la résiliation sera basée sur le portefeuille de contrats automobiles d'un grand assureur français avec lequel un partenariat a été établi dans le cadre du mémoire. Celui-ci se poursuivra avec l'analyse préliminaire des données fournies, la segmentation du portefeuille avec des méthodes innovantes, la découverte de poches clients au sein de chacun des segments du portefeuille et enfin la compréhension des facteurs de résiliation au sein des poches clients obtenues.

Chapitre 1

Contexte

Sommaire

Les raisons motivant l'étude de la résiliation en assurance non-vie sont nombreuses. Il est donc essentiel de définir le contexte de l'assurance non-vie et de fixer un cadre sur ce que signifie réellement le *Predictive Analytics* afin de commencer sereinement l'étude à proprement parler.

1.1 Le marché de l'assurance de biens et de responsabilité

L'assurance de biens et de responsabilité sert à protéger les particuliers et les entreprises face aux aléas en indemnisant les victimes de sinistres. Le marché de l'assurance de biens et de responsabilité a une croissance modérée. Le chiffre d'affaires a été multiplié par 2 entre 1984 et 2014. C'est un marché très concurrentiel (voir SCHAAL O. (2014)).

Le marché de l'assurance de biens et des responsabilité a pour caractéristiques :

— les marges réduites dues à un marché très concurrentiel,

— le caractère cyclique des résultats,

— l'impact de la gestion financière sur le résultat (car la gestion financière est censée couvrir les frais de gestion).

Les différentes catégories d'assurance des biens et des responsabilités ont la répartition suivante, en termes de primes :

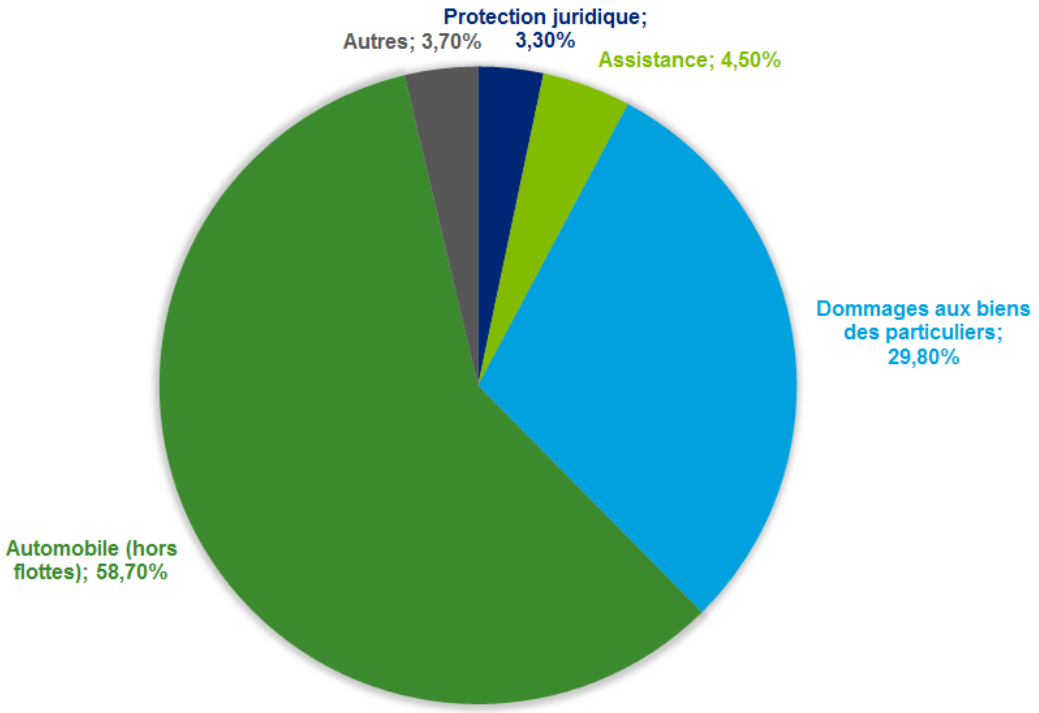


FIGURE 1.1 – Répartition des différentes catégories d'assurance de biens et de responsabilité en 2013 (source : FFSA)

Seules les caractéristiques de l'assurance automobile seront détaillées car les autres types d'assurance (notamment l'assurance habitation) ne sont pas l'objet du mémoire.

1.1.1 L'assurance automobile

L'assurance automobile est une des branches de l'assurance des biens et de responsabilité. Cette partie en expose les principales caractéristiques.

1.1.1.1 Les garanties en assurance automobile

Responsabilité civile (obligatoire) Elle permet l'indemnisation des dommages causés au tiers [1] par le gardien, le conducteur du véhicule ou un passager :

— Corporels : blessures ou décès subis par un piéton, un passager, ou un occupant d'un autre véhicule ...

— Matériels : dégâts causés aux autres voitures, deux-roues, immeubles, ...

Sont ainsi couverts au titre de cette garantie RC [2] les conducteurs autorisés ou non autorisés ; cependant, après avoir indemnisé les victimes, l'assureur peut disposer d'un recours à l'encontre des conducteurs non autorisés.

Exclusions La garantie ne s'applique pas dans les cas suivants :

— les dommages résultant d'un fait volontaire

— les accidents survenus au cours d'épreuves, courses, ...

— la non-validité du permis de conduire

Ces deux dernières exceptions ne s'appliquent pas aux tiers. L'assureur doit indemniser les dommages subis par les victimes et peut demander au conducteur responsable de le rembourser.

Garanties complémentaires (facultatives) Il ne s'agit plus de garantir une dette que le responsable a contractée vis-à-vis d'un tiers mais de protéger un patrimoine (les dommages subis par le véhicule par exemple) ou une personne.

— **Dommages tous accidents** : l'assureur garantit tous les dommages subis par le véhicule lorsqu'ils résultent d'une collision avec un corps fixe ou mobile ou d'un versement sans collision.

— **Dommages collision** : l'assureur garantit tous les dommages subis par le véhicule lorsqu'il entre en collision avec un piéton, un autre véhicule ou un animal dont le propriétaire est identifié.

— **Incendie** : l'assureur garantit les dommages résultant d'un incendie qu'il ait pris ou non pris naissance dans le véhicule (sont exclus les dommages résultant de brûlures causées par les fumeurs).

1. Les tiers sont toutes personnes autres que le conducteur
2. Responsabilité Civile.

— **Vol** : si le véhicule est retrouvé dans le mois suivant le vol, l'assureur prend en charge les frais de remise en état dans la limite de sa valeur avant sinistre. Sinon une indemnité contractuelle est due. Les assureurs proposent des conditions tarifaires adaptées selon les mesures de protection prises (indice SRA [3]).

— **Bris de glace** : l'assureur garantit les dommages liés aux pare-brises, aux glaces latérales ...

— **Assurance du conducteur** : Cette assurance couvre le conducteur qui n'est pas inclus dans la garantie RC.

— **Catastrophes naturelles** : l'assureur garantit les dégâts dus aux catastrophes naturelles (inondation, avalanche, tremblement de terre ...) sous réserve de parution au Journal officiel de l'arrêté interministériel constatant l'état de catastrophe naturelle. Une franchise de 380 euros est applicable.

— **Protection juridique** : Couvre les frais de défense de l'assuré devant les tribunaux.

— **Assistance** : cette assurance couvre le versement d'une prestation dans le cadre de l'assistance aux véhicules et aux personnes

1.1.1.2 La sinistralité en assurance automobile

En 2013, le chiffre d'affaires de l'assurance automobile est de 19,7 milliards d'euros pour un ratio combiné [4] de 103% avant réassurance et 104% après. La charge de sinistres par garantie en assurance automobile se répartit comme telle :

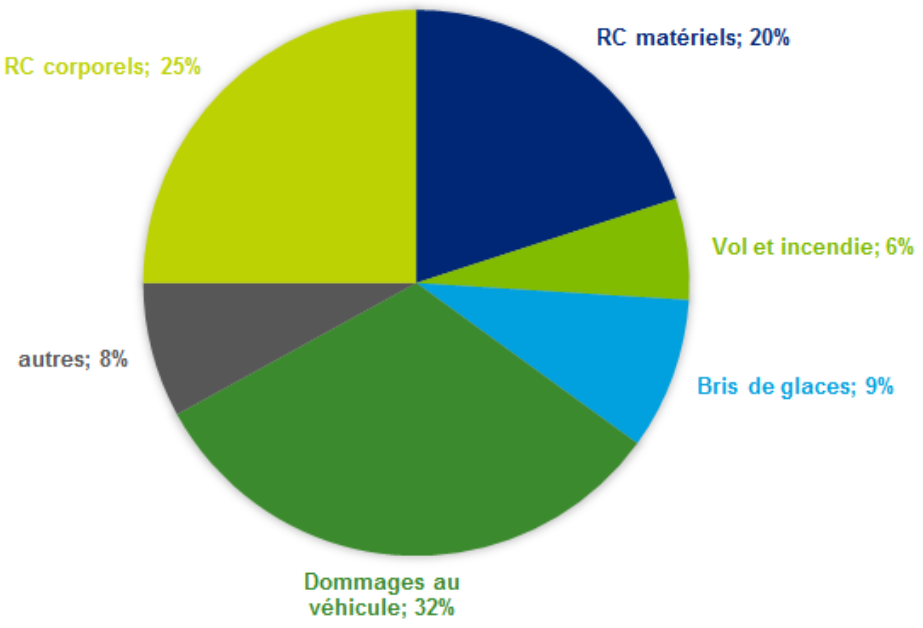


FIGURE 1.2 – Répartition de la charge de sinistres automobile par garantie en 2011 (sources : FFSA, GEMA)

3. Sécurité et Réparation Automobiles

4. Le ratio combiné d'un assureur est égal à $\frac{S+F}{P}$ avec S le coût des sinistres, P les primes encaissées et F les frais. Un ratio combiné supérieur à 100% ne permet pas à une compagnie d'assurance de réaliser des profits sauf si les produits financiers compensent l'insuffisance des primes par rapport aux coûts.

1.2 La loi Hamon

La loi Hamon (voir LEGIFRANCE (2014)), ou Loi n° 2014-344 du 17 mars 2014 relative à la consommation, instaure de nombreux changements au sein du Code de la Consommation.

Elle apporte notamment, avec son article 61, des nouveaux droits en matière de résiliation pour certains contrats d'assurance mais définit aussi d'autres mesures dont la volonté générale est d'augmenter le pouvoir d'achat des français.

1.2.1 Article 61 de la loi Hamon

L'article 61 de la loi Hamon a pour champ d'application :

1. Contrats d'assurance couvrant les personnes physiques en dehors de leurs activités professionnelles
2. Contrats d'assurance relevant des branches définies par décret en Conseil d'État.

Cet article autorise les assurés concernés à résilier, « à l'expiration d'un délai d'un an à compter de la première souscription, sans frais ni pénalités les contrats et adhésions tacitement reconductibles ».

1.2.1.1 Conditions

« La résiliation prend effet un mois après que l'assureur en a reçu la notification par l'assuré, par lettre ou tout autre support durable. »

« Le droit de résiliation doit être mentionné dans chaque contrat d'assurance, il doit également être rappelé avec chaque avis d'échéance de prime ou de cotisation. »

« Lorsque le contrat est résilié dans les conditions prévues, l'assuré n'est tenu qu'au paiement de la partie de prime ou de cotisation correspondant à la période pendant laquelle le risque est couvert. »

« Pour l'assurance de responsabilité civile automobile définie à l'article L. 211-1 et pour l'assurance mentionnée au g de l'article 7 de la loi n° 89-362 du 6 juillet 1989 tendant à améliorer les rapports locatifs et portant modification de la loi n° 86-1290 du 23 décembre 1986, le nouvel assureur effectue pour le compte de l'assuré souhaitant le rejoindre les formalités nécessaires à l'exercice du droit de résiliation dans les conditions prévues au premier alinéa du présent article. Il s'assure en particulier de la permanence de la couverture de l'assuré durant la procédure. »

La loi s'applique à tous les contrats conclus ou tacitement reconduits après la publication du décret, le 18 mars 2014. La loi est entrée en vigueur le 1er janvier 2015.

1.2.2 Objectif de la loi Hamon

La loi Hamon couvre l'assurance auto, l'assurance multirisques habitation, et les assurances dites affinitaires car constituant le complément d'un bien ou d'un service. D'après le gouvernement, cette mesure se veut déterminante (voir REPUBLIQUE FRANCAISE (2014)) pour aider les consommateurs à trouver le tarif qui leur correspond le mieux pour les contrats d'assurance qui constituent 5% de leur budget. Cela devrait permettre une hausse de leur pouvoir d'achat. Cette mesure devrait aussi créer aussi une plus grande fluidité du marché de l'assurance et garantir aux consommateurs de pouvoir faire mieux jouer la concurrence afin de bénéficier des meilleures offres en terme de prix et services.

1.2.3 Impact sur les compagnies d'assurance non-vie

Cette partie expose les problématiques qui ont motivé la mise en place de la loi Hamon mais aussi les critiques pouvant être faites de celle-ci.

1.2.3.1 Problématiques

La loi Hamon incite les assurés à résilier davantage. En 2010, le taux de résiliation en assurance auto était de 15,7% tandis que celui de l'assurance MRH était de 11.8%. Les facteurs venant compliquer l'analyse de l'impact de la loi Hamon sur le marché de l'assurance sont :

— La contexte de crise pourrait avoir une influence sur le taux de résiliation, les assurés en difficulté préférant avoir une prime moins chère

— Les comparateurs d'assurance comme Assurland ou Hyperassur pourraient inciter les assurés à résilier leur contrat et ainsi rendre le marché plus fluide.

— La majorité des nouveaux clients potentiels est actuellement les primo-conducteurs et les résiliés. Cela pourrait changer avec la loi Hamon.

1.2.3.2 Conserver ses clients : l'enjeu principal de la loi Hamon

FIGURE 1.3 – Enjeu de la résiliation non-vie : amortissement des coûts d'acquisition

Les coûts d'acquisition d'un client étant amortis sur la durée de vie en portefeuille de celui-ci, il sera d'autant plus rentable qu'il résiliera tardivement. L'enjeu pour l'assureur est donc de chercher à conserver ses clients le plus longtemps possible, si tenté qu'ils soient rentables, sachant qu'il faut définir la notion de rentabilité.

Une stratégie commerciale adaptée à la loi Hamon peut se composer de deux axes :

Renouvellement L'assureur peut augmenter ses interactions avec les clients tout en en maîtrisant les coûts (applications numériques, réseaux sociaux, agences, etc...). De plus, il peut augmenter ses actions de fidélisation pendant les périodes à haut risque (saisonnalité des résiliations, date du premier renouvellement). Enfin, l'assureur, pour diminuer la probabilité de résiliation de ses contrats, peut proposer des conditions avantageuses aux assurés qui choisissent de souscrire plusieurs contrats chez lui.

Conquête L'assureur peut identifier les segments de clients rentables parmi ceux qu'il pourrait prospecter.

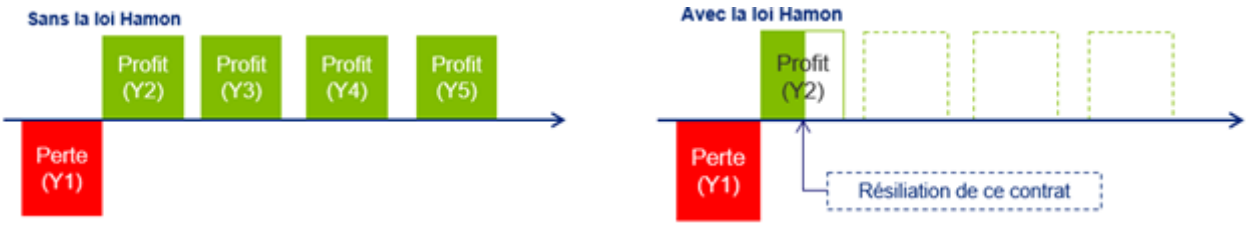


FIGURE 1.4 – Impact de la loi Hamon

1.2.3.3 Critiques

La principale critique envers la loi Hamon est que la mise en place de la résiliation infra-annuelle pourrait ne pas avoir l'effet escompté, à savoir fluidifier le marché et permettre une baisse des primes d'assurance. En effet, au Royaume-Uni où le taux de turn-over est très élevé (à environ 50% par an), les britanniques paient les primes les plus élevées d'Europe tout en étant les meilleurs conducteurs d'Europe. Ce paradoxe s'explique par le fait déjà expliqué qu'une compagnie d'assurance amortit ses coûts d'acquisition sur la durée de vie d'un client en portefeuille.

La FFSA [5] estime dans une étude qu'un point de turn-over correspond à 0,4 points de frais de gestion supplémentaires. Si le taux de résiliation annuel passe de 10% à 20%, les frais de gestions pourraient donc augmenter de 4 points. Cela est contraire au but initial de la loi Hamon.

De plus, les assureurs et assurés craignent que la loi Hamon entraîne certains autres problèmes :

- Perte de la mutualisation des risques, principe même de l'assurance, avec une distinction de plus en plus forte des bons et des mauvais risques.
- Remise en cause du calcul technique des primes si l'on peut résilier à tout moment
- Inflation probable des fausses attestations
- Risque de résiliation abusive de la part d'acteurs peu scrupuleux cherchant à capter des contrats
- Flou juridique entre l'application de la résiliation en loi Châtel ou en loi Hamon

5. Fédération Française des Sociétés d'Assurance

1.3 Solvabilité 2 : Risque de cessation en non-vie

L'étude de la résiliation en non-vie permet d'optimiser le capital économique dans le cadre de la Directive 2009/138/EC (Solvabilité 2). On détaille ici les méthodes de calcul du capital liées à la résiliation non-vie en formule standard.

Le 10 octobre 2014, la Commission Européenne a adopté les actes délégués de la Commission de Régulation venant en complément de Solvabilité 2. L'article 118 détaille la méthode de calcul du sous-module de cessation en non-vie : NL_{lapse}.

Les fonds propres de base (*basic own funds* ou BOF) sont la différence de l'actif avec le passif, plus les passifs subordonnés.

L'exigence en capital pour couvrir le risque de résiliation non-vie doit être égal à la perte en fonds propres de base résultant de la combinaison de deux chocs :

$$NL_{\text{lapse}} = \Delta BOF|(\text{choc résiliation}_1, \text{choc résiliation}_2)$$

avec

- NL_{lapse} : Exigence de capital pour couvrir le risque de résiliation non-vie
- ΔBOF : Variation de la valeur des fonds propres de base (à l'exception des variations dans la marge de risque des provisions techniques)
- choc résiliation$_1$: la cessation de 40% des contrats d'assurance dans le cas desquels cette cessation a pour effet d'entraîner une augmentation des provisions techniques sans la marge de risque.
- choc résiliation$_2$: lorsque des contrats de réassurance couvrent des contrats d'assurance ou de réassurance qui seront émis à l'avenir, la baisse de 40% du nombre de ces futurs contrats d'assurance ou de réassurance utilisé dans le calcul des provisions techniques.

choc résiliation$_1$ et choc résiliation$_2$ s'appliquent uniformément à l'ensemble des contrats d'assurance et de réassurance concernés. En ce qui concerne les contrats de réassurance, choc résiliation$_1$ s'applique aux contrats d'assurance sous-jacents.

Pour déterminer la perte de fonds propres de base de l'entreprise d'assurance ou de réassurance dans le contexte de l'évènement choc résiliation$_1$, l'entreprise se fonde sur le type de cessation qui affecte le plus négativement ses fonds propres de base selon un calcul contrat par contrat.

1.4 Big Data et Predictive Analytics

Après définition du contexte de l'assurance non-vie, le mémoire s'intéresse à la définition du *Predictive Analytics* et des applications possibles de celui-ci.

Big Data est un terme générique qui désigne les jeux de données de taille très importante (au moins plusieurs Gio[6]). La taille importante de ces jeux de données fait que les méthodes habituelles de traitement deviennent obsolètes car inadaptées. Les enjeux des Big Data sont l'analyse, la capture, la recherche, le partage, le transfert et le caractère privé des données.

Communément, Big Data désigne toute activité liée à l'exploitation de ces données afin d'en extraire des informations utiles. Or, *Big Data* est un terme réservé à des activités aux caractéristiques bien précises. Pour toute activité qui consiste à analyser des données afin d'en extraire de la valeur, on préfèrera le terme Predictive Analytics (ou analyse prédictive). Le *Predictive Analytics* est un grand enjeu du secteur des assurances. C'est moins vrai pour le Big Data qui est plus lié à des domaines comme le Web (chez Google, Amazon ou Facebook par exemple) ou la recherche scientifique[7].

1.4.1 Big Data

FIGURE 1.5 – Le Big Data (source : Analytics Seo)

1.4.1.1 Les 3V du Big Data

Volume Le volume est une des caractéristiques du Big Data. C'est une conséquence directe de la croissance de la capacité mondiale de stockage de données. Avec l'essor de

6. Gio, symbole d'unité du gibioctet valant 1 073 741 824 = 2^{30} octets
7. Le LHC, ou *Large Hadron Collider*, peut produire jusqu'à 5×10^{20} octets de données expérimentales par jour

la vente en ligne, certains sites peuvent avoir un catalogue dépassant le million de produits (126 millions de références proposées sur Amazon.fr en 2015). Cette explosion du volume de données disponibles touche tous les secteurs d'activité et nécessite des nouvelles technologies de traitement. C'est le cas de MapReduce, technologie développée par Google afin de réaliser des opérations sur des gros volumes de données. MapReduce permet la manipulation et le traitement d'un nombre important de données au sein d'un cluster de nœuds. Il s'agit d'une technologie basée sur le calcul parallèle. Le framework Hadoop, basé sur MapReduce, est utilisé par Yahoo ! ou Facebook et permet de gérer le volume important généré par ces sites Internet. Facebook a créé, par exemple, 10 Tio de données chaque jour en 2013.

Vitesse La masse de données disponible peut être traitée et des algorithmes existent pour extraire l'information utile de ces données comme, par exemple, les outils du Machine Learning. Cependant, il est parfois nécessaire de traiter ces données en un temps très court afin de donner une information rapidement. On peut prendre l'exemple de Criteo, entreprise française cotée au NASDAQ spécialisée dans le ciblage publicitaire sur Internet, qui doit pouvoir afficher une publicité sur le site de ses clients en fonction des caractéristiques du visiteur. Cette publicité doit être la plus pertinente possible afin d'augmenter la probabilité de clic, qui est l'évènement rémunérant Criteo. Les algorithmes d'apprentissage automatique déterminant la publicité la plus adaptée au client doivent s'exécuter en moins de 6 millisecondes. On comprend donc que la vitesse soit une caractéristique du Big Data. Cette problématique peut avoir un écho en assurance avec, par exemple, la souscription prédictive (*predictive underwriting*). Les modèles de tarification doivent être implémentés pour pouvoir proposer un niveau de prime le plus rapidement possible et ainsi alimenter un comparateur d'assurances. Comme le souligne Jérome Cornillet, responsable *solutions business analytics* chez SAS, dans un article de l'Argus de l'assurance (voir CHEVRIER C. (2013)), « *Là où il fallait cinq heures pour un traitement, trois minutes suffisent. Plus de cent itérations sont possibles contre sept auparavant. Cela permet d'améliorer la qualité des modèles et de travailler sur des bases de données globales et non plus des échantillons* ». Les assureurs doivent tirer parti de ces technologies afin de se forger un avantage concurrentiel, même s'il est temporaire. Ils doivent mieux gérer les risques et axer leur stratégie sur la prévention pour réduire leurs risques. Ce doit notamment être le cas en assurance santé.

Variété De plus, un autre des enjeux soulevé par le Big Data est la variété des sources de données. Les nouvelles sources de données peuvent provenir, par exemple, des réseaux sociaux, des terminaux mobiles, ou des parcours clients enregistrés suite à des visites sur Internet. Contrairement aux données habituellement utilisées, ces nouvelles sources de données présentent la caractéristique d'être peu ou pas structurées. La vraie difficulté, quand il s'agit de la variété de ces données, est de savoir quoi en faire. Il faut tout d'abord s'assurer de la véracité des données récupérées. En effet, les données issues des réseaux sociaux peuvent être peu fiables. De plus, ces données sont souvent parcimonieuses (beaucoup de valeurs non renseignées) et les algorithmes de traitement doivent pouvoir s'y adapter. Il faut ensuite savoir comment utiliser, notamment en assurance, ces nouvelles sources de données qui présentent un potentiel exploitable. Cela pose aussi des problématiques de compétences à développer chez les assureurs si les activités liées au Big Data sont internalisées. On peut, par exemple, noter l'émergence de formations professionnelles complémentaires comme celle d'actuaire data-scientist. Celle-ci est dispensée par l'Institut du Risk Management, rattaché à l'Institut des Ac-

tuaires, et permet aux professionnels en gestion des risques et de la fonction actuarielle de développer des compétences en data-sciences.

1.4.2 Predictive Analytics, ou analyse prédictive

Le *Predictive Analytics* est l'utilisation de différentes méthodes statistiques comme l'apprentissage automatique (machine learning), l'analyse exploratoire de données (data mining) ou la théorie des jeux afin d'analyser le présent et le passé pour faire des hypothèses prédictives sur des évènements futurs. Ces modèles prédictifs cherchent à établir des schémas dans les données (patterns) pour identifier les risques et les opportunités. Le Predictive Analytics est utilisé en actuariat, dans les services financiers, l'assurance, les télécommunications, le commerce de détail, le tourisme, la santé, l'industrie pharmaceutique et de nombreux autres domaines.

Une prédiction dépend de nombreux facteurs : les caractéristiques de chaque patient, chaque emprunteur, chaque assuré. L'enjeu est de mettre ensemble toutes ces sources de données afin de faire une prédiction. Le machine learning utilise les capacités informatiques d'un ordinateur pour exploiter ces données. Un algorithme de machine learning fonctionne comme ceci :

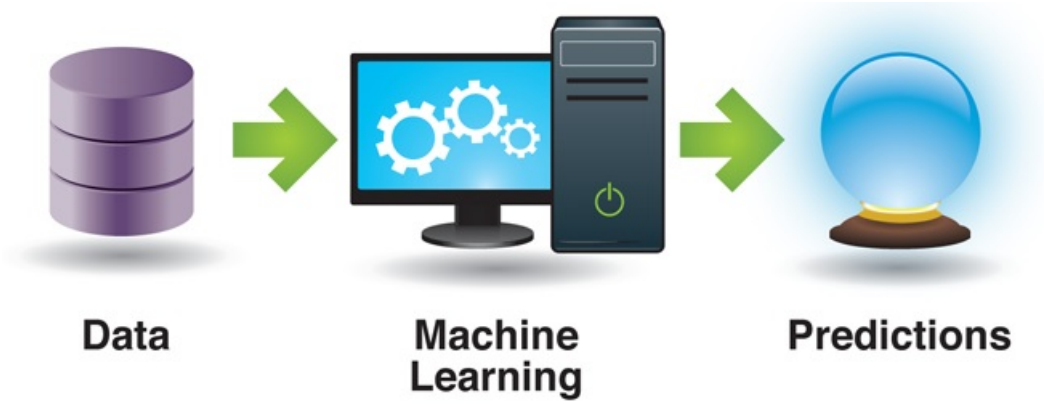


FIGURE 1.6 – Modèle prédictif (source : Eric Siegel)

Le machine learning peut donner des informations sans hypothèse à priori sur les données comme (voir SIEGEL E. (2013)) :

— partir à la retraite plus tôt réduit l'espérance de vie
— la criminalité augmente après un évènement sportif
— les végétariens manquent moins leurs vols à l'aéroport

Ces conclusions sont dites *data-driven*. C'est l'algorithme qui a trouvé tout seul le *pattern* entre le fait d'être végétarien et celui de manquer moins ses vols, par exemple. Celui qui a mis en place le modèle n'a pas étudié à priori le lien entre ces deux variables.

Le *Predictive Analytics* est à distinguer de la prévision. La prévision est macroscopique tandis que la prédiction est individuelle. Dans le cas de la modélisation de la résiliation, on peut distinguer la prévision du taux de résiliation au cours du temps et la prédiction de la probabilité de résiliation pour chaque assuré. Les approches pour traiter ces deux problèmes peuvent être très différentes en termes de techniques statistiques utilisées.

1.4.2.1 Les types de modèle

Modèles descriptifs Un modèle descriptif sert à expliquer les variables qui vont influer sur un phénomène. Les indications données par un tel modèle peuvent se révéler très utiles car elles peuvent donner de bonnes intuitions sur l'orientation stratégique, en termes de marketing, par exemple. Les données à disposition des entreprises étant souvent à grandes dimensions, il peut être très difficile de piloter son activité en les exploitant. En effet, la grande dimension empêche une bonne visualisation des données et découvrir des *patterns* entre les variables peut donc être très difficile. Le *Predictive Analytics* dans le cadre de modèles descriptifs se rapproche ici du *Business Analytics* qui, lui, vise à permettre une visualisation en temps réel des KPI [8] de l'entreprise pour aider à la décision.

Modèles prédictifs Le *Predictive Analytics* a plusieurs buts. Tout d'abord, il peut s'agir d'élaborer un modèle prédictif. Dans ce cas, le but n'est pas d'expliquer les relations entre les données mais de prédire une variable d'intérêt à l'aide de données existantes. Ce type de modèle a pour but la précision dans l'estimation de la variable d'intérêt. Il est possible, par exemple, de chercher à prédire la probabilité de survenue d'un cambriolage pour un assuré, en temps réel. Le but d'un tel modèle ne serait pas d'expliquer pourquoi le cambriolage survient. Bien sûr, construire un tel modèle prédictif nécessite de choisir des variables liées à la survenue d'un cambriolage et donc requiert d'établir au préalable d'un modèle afin de décrire les données dont on dispose.

Modèles de décision Ces modèles se basent sur les données ainsi que sur le résultat des modèles prédictifs pour aboutir à une décision. Par exemple, il est possible d'utiliser un modèle de décision pour savoir à quelle personne proposer quel produit d'assurance parmi un fichier de prospects.

1.4.2.2 Les 5 effets de la prédiction

Ces effets sont décrits par Eric Siegel, ex-professeur de Machine Learning à Columbia University, dans son livre : *Predictive Analytics : The Power to Predict Who Will Click, Buy, Lie or Die* (voir SIEGEL E. (2013)). Ils sont très généraux et les principes de base qui motivent l'exploration des données.

L'effet "Prediction" Une prédiction, même très approximative, peut créer de la valeur. Il n'est pas forcément nécessaire d'avoir une estimation précise de la probabilité de résiliation individuelle. Par contre, pouvoir identifier une population significative qui va résilier 3 fois moins que la population générale peut être très utile. Identifier une telle population et la prospecter en priorité peut permettre d'augmenter la durée de vie moyenne d'un contrat en portefeuille et ainsi amortir les coûts d'acquisition de la police

8. Key Performance Indicators

sur une plus grande période. De plus, cet effet indique qu'il est possible d'extraire de la valeur de données même si leur qualité n'est pas optimale.

L'effet "Data" Les données ont toujours quelque chose à raconter.

L'effet "Induction" C'est le procédé par lequel un algorithme de machine learning traite les données afin de produire un modèle prédictif.

L'effet "Ensemble" Mis ensemble, chacun des modèles prédictifs compense les limites des autres modèles. L'ensemble qui en résulte est plus performant que chacune de ses composantes prises séparément. Cela se nomme *meta-learning*.

L'effet "Persuasion" L'impact d'une action (marketing par exemple) peut être quantifié. Par exemple, on peut valider statistique l'effet d'un médicament contre un placebo. La modélisation « Uplift » est un moyen de prédire, avec un certain taux d'erreur, l'impact d'une action sur le comportement de quelqu'un. Ce type de modélisation peut être très complexe. Dans le cas de la résiliation, une fois que le modèle pour prédire qui va résilier a été créé, il serait possible de développer un modèle uplift pour savoir qui appeler parmi ces personnes. En effet, savoir qui il est utile d'appeler est créateur de valeur car il n'est pas utile d'appeler ceux qui vont partir quand même malgré un appel et une éventuelle remise. Cette modélisation est difficile à mettre en œuvre.

1.4.3 Applications du Predictive Analytics

Le Predictive Analytics est utile dans de nombreux secteurs d'activité. Il peut permettre d'améliorer l'efficacité opérationnelle ou même la prise de décisions. L'apport du Predictive Analytics est néanmoins limité par des contraintes légales ou d'éthique. Cette partie expose les différents enjeux de l'analyse prédictive.

1.4.3.1 Quelles sont les applications possibles du Predictive Analytics ?

Ciblage marketing Il est possible de prédire qui, parmi les clients potentiels, va répondre s'ils sont contactés. Cela permet de cibler en priorité les personnes répondant plus que la moyenne.

Ciblage publicitaire sur Internet La prédiction de la publicité sur laquelle un client en particulier va avoir plus de chances de cliquer est effectuée.

Détection de grossesse Il est prédit quelles sont les clientes potentielles qui vont avoir un enfant dans les mois à venir afin de leur proposer des offres spécifiques.

Prédiction de la délinquance Les endroits dans lesquels risquent d'avoir lieu des crimes futurs sont identifiés par une prédiction afin d'y envoyer une patrouille de police en prévention.

Détection de fraudes Il est prédit lesquels sont frauduleux parmi les sinistres déclarés, les transactions bancaires et autres. Cela permet d'améliorer l'efficacité opérationnelle.

Rétention des employés Un algorithme réalise une prédiction de quels employés ont une probabilité de départ élevée. Les managers utilisent les prédictions à propos de ceux qu'ils supervisent afin de prendre des mesures éventuelles pour tenter de les garder.

Prédiction de la récidive Une prédiction indique quels délinquants risquent d'enfreindre à nouveau la loi afin d'adapter la peine de prison. Cela pose des problèmes d'éthique et notamment des problèmes soulevés par les faux positifs.

Détection de sentiments dans des posts de blog Une prédiction indique quels posts expriment de l'anxiété afin de construire un indicateur de sentiments pour la population globale. Il est, par exemple, possible d'avoir l'intuition que des personnes anxieuses vont moins résilier leurs garanties d'assurance ou souscrire à plus de garanties afin de se couvrir.

Rétention des clients (Churn) Les modèles de *churn* sont très utilisés dans le secteur des télécommunications pour savoir qui de leurs clients risque de résilier son contrat téléphonique. Le *churn modelling* peut aussi s'appliquer dans d'autres domaines. Une fois l'information sur le risque de résiliation quantifiée, il est possible de prendre des mesures pour tenter de conserver le client.

Prédiction du remboursement anticipé d'un prêt Il est possible de prédire qui va rembourser son prêt de manière anticipée afin de tenter de le retenir.

Marketing ciblé grâce à un modèle Uplift Cette modélisation répond à la question : Parmi les clients qui risquent de partir, lesquels sont ceux que je pourrai convaincre de rester ?

Souscription Une prédiction du risque associé à un fichier de prospects est effectuée. Cela permet de connaître la prime de risque associée à un client donné. Le *predictive underwriting*[9] permet aussi d'améliorer l'efficacité opérationnelle, par exemple en réduisant le temps destiné à remplir le formulaire de souscription.

1.4.3.2 Exemples concrets de réalisation

Voici quelques exemples d'utilisation du Predictive Analytics en assurance. Ces exemples sont rapportés par Eric Siegel dans *Predictive Analytics : The Power to Predict Who Will Click, Buy, Lie or Die* (voir SIEGEL E. (2013)).

Coût des dommages corporels après un accident de voiture *Allstate* : En 2012, et en se basant uniquement sur les caractéristiques de véhicule assuré, la compagnie a pu prédire le coût ultime des dommages corporels pour l'assureur. Cela a permis de dégager 40 millions de dollars par an.

9. En français, souscription prédictive

Coût des accidents du travail *Accident Fund Insurance* : En utilisant les maladies déclarées (diabète, obésité, ...) des employés ayant souscrit, la compagnie a pu savoir quelles maladies déclarées permettaient de prévoir les accidents les plus coûteux. Les employés présentant les maladies provoquant les sinistres les plus chers ont ensuite été ciblés par des actions de prévention.

Charge des sinistres *Un leader mondial de l'assurance* : Mise en place d'un modèle prédictif ayant permis de diminuer le ratio S/P de 50 points de base, ayant eu pour conséquence une économie de 50 millions de dollars.

Décès *Compagnie d'assurance santé américaine* : Prédiction de la probabilité de décès d'une personne âgée dans les 18 mois à venir afin de lui proposer des produits de fin de vie.

Remboursement anticipé de prêts immobiliers *Chase* : La compagnie a généré des centaines de millions de dollars en prédisant qui, parmi les personnes à qui elle a accordé un prêt, va refinancer son prêt et ainsi transférer les intérêts futurs à une banque concurrente.

Risque de défaut sur les prêts *Citigroup* : A développé des modèles par zone géographique pour prédire le risque de défaut d'un futur débiteur et l'a implémenté opérationnellement en agence.

Non-paiement *Institution financière* : A économisé 2.1 millions de dollars en offrant des facilités de paiement aux clients qui n'auraient pas payé sans, et en n'en offrant pas à ceux qui auraient payé de toute façon.

Cours d'une action *London Stock Exchange* : 40% des ordres de trading sur le London Stock Exchange sont effectués automatiquement par des algorithmes.

1.4.4 Les nouvelles données en assurance

L'essor des NTIC [10] a permis le développement de nouvelles sources de données. Les assureurs ont maintenant à leur disposition des données externes : données des réseaux sociaux, navigation sur Internet, capteurs bio-métriques, données liées les transactions bancaires, coordonnées GPS ...

1.4.4.1 Les nouvelles sources de données

Les nouvelles sources de données possibles sont listées ci-après. Leur utilisation peut être ou non possible en raison des contraintes liées à l'éthique et à la Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés. Voir section 1.4.4.3 p. 20.

10. Nouvelles Technologies de l'Information et de la Communication

Données géospatiales Les données géographiques peuvent être utilisées comme variables explicatives dans un modèle actuariel. Ces données peuvent être des types suivants :

— Données de conduite : Plusieurs compagnies d'assurance ont lancé des offres « Pay As You Drive » en assurance auto. Ces offres se basent sur un capteur connecté fixé dans la voiture et qui enregistre les déplacements de l'assuré (distance parcourue). L'assuré paye alors en fonction de la distance parcourue et de son tarif au kilomètre. Une offre plus évoluée, le « Pay How You Drive » utilise la vitesse, l'accélération, le comportement en agglomération, ...[11] pour calculer le risque et adapter le tarif de l'assuré chaque mois en fonction de sa conduite.

— Données issues des NTIC : Les smartphones et autres périphériques enregistrent les déplacements de leur possesseur.

— Données de cartographie : Il est possible d'utiliser les données de cartographie dans un modèle actuariel. Ces données sont par exemple les coordonnées GPS des commissariats qui peuvent être utilisées pour expliquer la survenue d'un cambriolage. Il est aussi possible d'utiliser des données cartographiques (fleuves, lacs, ponts, ...) en réassurance pour la modélisation des catastrophes naturelles. Ces informations sont, par exemple, en accès libre sur le portail OpenStreetMap.org.

FIGURE 1.7 – Portail OpenStreetMap.org (Source : Openstreetmap.org)

Données bio-métriques Les capteurs connectés permettent de surveiller les indicateurs de santé de l'assuré : le poids, la fréquence cardiaque, la qualité du sommeil ou des paramètres non propres à l'individu comme la qualité de l'air. Ces paramètres peuvent être utilisés pour mener des campagnes de prévention afin de réduire le nombre de prestations en assurance santé. L'assuré doit comprendre le bénéfice qu'il peut tirer de l'utilisation de ces objets connectés et le mesurer à l'inconvénient d'une intrusion dans la vie privée. Bien entendu, des obstacles juridiques compliquent l'utilisation de ces données.

11. Ces données sont appelées « données télématiques ».

Données issues des réseaux sociaux L'utilité des réseaux sociaux en assurance passe notamment par la détection de fraudes. L'assureur pourrait utiliser les informations qu'il trouve sur les réseaux sociaux afin d'établir une contradiction entre la déclaration d'un sinistre et l'activité de l'assuré sur Facebook ou Twitter. Par exemple, un assuré santé pourrait mentir et déclarer un arrêt de travail tout en étant identifié sur Facebook dans une activité incompatible avec la déclaration du sinistre (voir D'CAMERA (2011)).
De plus, il serait possible d'obtenir l'information sur l'interconnexion entre les assurés en utilisant les réseaux sociaux. Si on sait que plusieurs assurés sont interconnectés, une famille par exemple, alors on mesurera mieux le risque d'augmenter le tarif de l'enfant ou de le résilier. Une référence peut être faite au mémoire d'actuariat suivant qui traite de l'interconnexion entre assurés et de la création d'une offre associée (voir CHOQUET C. (2011)).
Ces données sont aussi utiles pour identifier les évènements de la vie d'un assuré et lui proposer des produits adéquats. En utilisant l'information issue, par exemple, de Facebook sur les *likes*, il serait possible d'améliorer la connaissance sur les types de profils d'assuré afin de prendre des actions marketing comme proposer de la vente croisée (comme le ferait Amazon avec son « Cet article pourrait aussi vous plaire » mais avec des produits d'assurance) ou créer des produits pour des types de profils non soupçonnés jusqu'alors.

Données comportementales sur Internet Le comportement de navigation de l'assuré permettrait d'en déduire des conclusions en termes de profils (voir COSTES Y. (2000)). La plupart des assureurs dispose d'un site internet sur lequel l'assuré peut gérer son contrat. Par exemple, imaginons une interface de gestion pour un contrat d'assurance vie. L'information sur les fréquences de connexions de l'assuré au portail pourrait avertir l'assuré à propos d'un rachat imminent du contrat. Des mesures pourraient alors être prises pour empêcher ce rachat. De plus, si une compagnie d'assurance pouvait faire le lien entre ses clients et les visiteurs d'un site de comparateur d'assurance, alors elle pourrait connaître la volonté de résilier d'un assuré. Cependant, à l'heure actuelle (2015), la CNIL [12] définit strictement le cadre légal des conditions de collecte et de conservation des adresses IP.

Données socio-économiques Les facteurs socio-économiques peuvent avoir une influence sur la sinistralité (voir ENTORF H. and SPENGLER H. (2000)). Ces données peuvent être le revenu moyen, le taux de chômage, le pourcentage de cadres, etc . . . à la maille d'une zone IRIS [13]. Il est alors possible de corréler ces informations avec l'information sur la sinistralité, la résiliation ou autres pour en tirer des conclusions.Il est possible de trouver ce type de données sur le site de l'INSEE [14].

Autres données Plusieurs autres types de données pourraient être utiles en assurance. Les données conjoncturelles comme la météo ou des évènements récents dans un périmètre donné (plusieurs voitures dégradées récemment dans une ville donnée) pourraient notamment être utilisées pour faire de la prévention. L'assureur pourrait créer de la valeur en se basant sur l'adage : « Mieux vaut prévenir que guérir. ». Plusieurs

12. Commission nationale de l'informatique et des libertés
13. Îlots Regroupés pour l'Information Statistique : zone de plusieurs milliers d'habitants sur laquelle est calculée plusieurs indicateurs statistiques.
14. L'Institut national de la statistique et des études économiques

assureurs ont lancé des projets visant à créer des applications smartphone afin de faire de la prévention.

1.4.4.2 Le mouvement OpenData

Le mouvement Open Data vise à promouvoir la mise à disposition de données sur Internet. Ces données doivent servir l'intérêt général. La République Française, via son site Data.gouv.fr, suit ce mouvement. Elle propose des jeux de données sur les thématiques suivantes :

- Agriculture et alimentation
- Culture
- Economie et Emploi
- Education et Recherche
- International et Europe
- Logement, Développement Durable et Energie
- Santé et Social
- Société
- Territoires et Transports

Ces données sont utilisables notamment dans un modèle explicatif.

FIGURE 1.8 – Mouvement Open Data en France (Source : data.gouv.fr)

1.4.4.3 Respect de la vie privée et éthique

La Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés, ou loi informatique et libertés règlemente les droits essentiels des particuliers vis-à-vis des données qu'il pourrait exister sur eux. Le Parlement a instauré en parallèle de cette loi la Commission nationale de l'informatique et des libertés (CNIL) comme organe de contrôle de la bonne application de la loi informatique et libertés. Celle-ci veille à ce que l'informatique n'enfreigne pas les droits de l'homme, la vie privée ainsi que les libertés individuelles.

L'utilisation des données par l'assureur doit respecter les principes suivants (voir FROIDEFOND E.A. (2014)) :

- **Finalité du traitement** : pas de fichiers administratifs utilisés à des fins de prospection commerciale
- **Pertinence des données** : interdiction de collecter les données sensibles qui font apparaître, directement ou indirectement, les origines raciales ou ethniques, les opinions politiques, philosophiques ou religieuses ou l'appartenance syndicale ainsi que les données relatives à la santé ou à la vie sexuelle. Sauf exceptions (consentement, intérêt public...).

— **Conservation limitée des données** : les données ne peuvent être conservées dans les fichiers au-delà de la durée nécessaire à la finalité poursuivie qu'à des fins statistiques.

— **Obligation de sécurité** : Respect de l'intégrité et de la confidentialité des données : empêcher que les données soient déformées, endommagées ou que des tiers non autorisés y aient accès.

— **Respect des droits des personnes** : les personnes doivent être informées, lors du recueil, de l'enregistrement ou de la première communication des données. Le transfert de données hors Union Européenne est encadré.

1.4.5 Qualité des données

Les données doivent être d'une qualité suffisante pour que leur étude soit pertinente et pour pouvoir être utilisées dans certains calculs assurantiels. La Directive Solvabilité 2 réglemente avec son article 82 « *Qualité des données et application d'approximations, y compris par approches au cas par cas, pour les provisions techniques* » le niveau de qualité que les données doivent avoir pour le calcul des provisions techniques. Ceci est détaillé dans le texte « *CEIOPS' Advice for Level 2 Implementing Measures on Solvency II : Technical Provisions - Article 86 f Standard for Data Quality* » publié en 2009 par le CEIOPS [15] qui, depuis, est devenu l'EIOPA [16].

Les exigences sur la qualité des données ne s'appliquent pas seulement à l'évaluation des provisions techniques mais aussi à l'utilisation des paramètres propres à l'entreprise (*Undertaking Specific Parameters*) dans la formule standard ainsi qu'aux normes de qualité statistique.

1.4.5.1 Les critères de qualité des données

Des données sont de qualité si elles sont pertinentes, exhaustives et exactes.

Pertinence Les données sont pertinentes si :

— elles sont adaptées à l'utilisation qui en est fait,

— elles sont représentatives des facteurs de risque auxquels l'entité est exposée,

— elles sont représentatives du portefeuille valorisé,

— elles sont adaptées à une utilisation pour estimer des flux de trésorerie futurs.

Exhaustivité Les données sont exhaustives si :

— elles couvrent les principaux groupes de risques homogènes,

— elles ont une granularité suffisante pour permettre une identification des tendances et une bonne compréhension du comportement des risques sous-jacents,

— elles doivent permettre l'utilisation des méthodes de provisionnement classiques,

— un historique suffisamment conséquent est disponible.

15. Committee of European Insurance and Occupational Pension Supervisors
16. European Insurance and Occupational Pensions Authority

Exactitude Les données sont exactes si :

— il n'y a pas d'erreurs matérielles [17],

— elles sont enregistrées régulièrement et sans changement de méthode au cours du temps,

— l'entité peut démontrer qu'elle considère les données comme bonnes à utiliser (en prouvant qu'elle les utilise de manière opérationnelle et dans les processus de décision).

1.4.5.2 Processus de gestion de la qualité des données

Maintenir une qualité de données suffisante est un processus continu constitué des étapes ci-après :

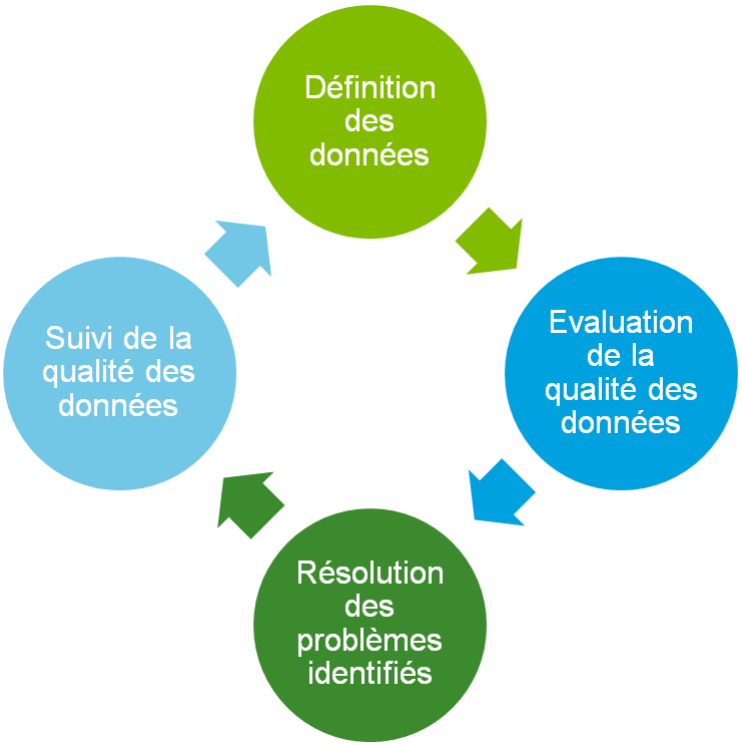


FIGURE 1.9 – Processus de gestion de la qualité des données

1.4.5.3 Qualité des sources de données externes

Dans le cadre du provisionnement, l'utilisation de données externes produites par des tiers ou des données de marché est réglementée. Leur qualité doit vérifier certains critères. Les données doivent être pertinentes, exhaustives et exactes. Dans le cas de données externes, ces critères signifient :

— **Pertinence et exhaustivité** : Les vérifications de pertinence et d'exhaustivité doivent se faire à la maille du portefeuille et à un niveau plus granulaire quand

17. C'est-à-dire des erreurs qui ne portent pas à conséquence sur le résultat final ; portant à conséquence étant à quantifier.

c'est pertinent. Il doit pouvoir être prouvé que l'utilisation de ces données externes améliore la pertinence et l'exhaustivité nécessaire à l'étude réalisée.

— **Exactitude** : Les parties prenantes doivent vérifier la fiabilité des sources de données externes. Il faut pouvoir prouver au régulateur que les données externes seront toujours disponibles. Dans le cas contraire, la compagnie s'exposerait au risque de devoir changer sa méthode de calcul si les données externes venaient à ne plus être publiées.

1.5 Les techniques pour le Predictive Analytics

Le but de cette partie n'est pas de détailler les algorithmes d'Analytics. Le lecteur se rapportera à KRIEGER J. (2014) pour plus de détails.

L'apprentissage automatique, ou *machine learning*, permet d'analyser des masses de données plus ou moins importantes afin de prédire pour un nouveau vecteur de données son appartenance à un sous-ensemble de la base globale ou alors pour estimer une valeur cible associée à ce vecteur (probabilité, valeur réelle). Il existe plusieurs types d'algorithmes d'apprentissage automatique.

Le choix de l'algorithme adapté au problème à traiter se généralement en testant plusieurs algorithmes et en choisissant celui qui a les meilleures performances sur la base de données disponible. Il n'existe pas de notion de meilleur algorithme et cela dépend des données.

Les 6 étapes essentielles Tout problème de data-mining ou machine learning se divise en 6 étapes indispensables au succès de l'étude :

1. Définir le problème
2. S'approprier les données
3. Retraiter et faire des statistiques de base sur les données
4. Implémenter le modèle
5. Évaluer le modèle et ses limites
6. Tirer des conclusions de l'étude

1.5.1 Types d'algorithmes

Tout d'abord, on distingue les méthodes de **classification**. Celles-ci regroupent les vecteurs en entrée dans différentes classes connues à l'avance. Un algorithme de ce type peut répondre à la question « Est-ce que cet individu va résilier son contrat cette année ? ». On peut citer les Support Vector Machines (SVM) comme exemple d'algorithmes de classification.

Les méthodes de **régression** utilisent une base d'apprentissage de taille n $(X_i, Y_i)_{i\in[\![1,n]\!]}$ pour prédire la valeur de la variable cible Y_m d'un nouveau vecteur de données (X_m, Y_m) avec $m \in \mathbb{N}\backslash[\![1,n]\!]$. On cherche généralement à estimer Y_m par $\hat{Y}_m = h(X_m)$ avec h une fonction telle que la quantité $\frac{1}{n}\sum_{k=1}^{n}\psi(Y_k, h(X_k))$ soit minimale[18]. Les réseaux de neurones sont un type d'algorithme de régression.

Le **clustering** est une technique qui consiste à détecter des groupes homogènes dans une base d'apprentissage de taille n $(X_i)_{i\in[\![1,n]\!]}$. Les groupes formés sont appelés clusters. Leur nombre n'est pas connu à priori et il existe des critères pour en déterminer le nombre optimal. L'algorithme K-means en est l'exemple le plus connu.

18. Avec $\psi(\cdot,\cdot)$ une fonction de perte. On a souvent $\begin{array}{lll}\psi : \mathbb{R}^2 & \longrightarrow \mathbb{R} \\ (u,v) & \longmapsto (u-v)^2\end{array}$

Les méthodes de **réduction de la dimension** sont des techniques qui permettent projeter les données dans un espace de dimension plus petite afin d'en faciliter la compréhension. Citons par exemple l'analyse en composantes principales (ACP).

Enfin, les techniques d'**estimation de densité** permettent de connaître la distribution des variables.

1.5.2 Méthodes supervisées et non supervisées

Il existe deux types de techniques.

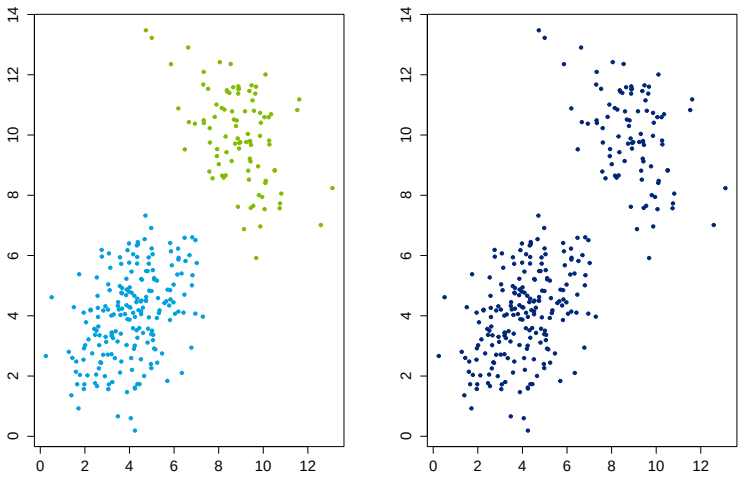


FIGURE 1.10 – Données adaptées à une classification supervisée (à gauche) et non supervisée (à droite)

L'objectif des techniques supervisées est de définir des règles permettant de classer des objets dans des classes à partir de variables qualitatives ou quantitatives associées à ces objets. Ces méthodes reposent donc sur l'existence d'une variable d'intérêt. On parle de classification lorsque la variable d'intérêt est qualitative et de régression lorsque la variable d'intérêt est une valeur continue. L'algorithme se base sur des données permettant d'édicter des règles afin de prédire la valeur de la variable d'intérêt pour des nouveaux vecteurs de données.

Les techniques non supervisées ne reposent pas sur l'existence d'une variable d'intérêt. Ces techniques servent à diviser les données en sous-groupes homogènes, appelés clusters, de telle sorte que deux vecteurs très éloignés l'un de l'autre [19] soient groupés dans des clusters distincts.

19. Au sens d'une distance à définir.

Chapitre 2

Analyse préliminaire des données fournies

Sommaire

2.1 Présentation des données

Les données ont été fournies par un partenaire dans le cadre d'une collaboration sur le Big Data. Par souci de confidentialité, les ordres de grandeur, noms des produits et autres indications pouvant permettre d'identifier le Partenaire ont volontairement été modifiés sans entraîner de changement dans les conclusions de l'étude.

2.1.1 La base de données

Le cadre de l'étude est l'assurance automobile. Le périmètre d'étude est les véhicules à 4 roues et de moins de 3,5 tonnes. Les contrats sont uniquement ceux souscrits par des personnes physiques.

Les données se présentent sous la forme de 4 bases, préparées par le partenaire :

1. Une base image au 31/12/2009 : Elle contient tous les contrats en cours au 1er janvier 2010.
2. Une base image au 31/12/2014 :
   - les mêmes informations sur les contrats et assurés, actualisées.
   - le nombre de sinistres observés entre le 01/01/2010 et le 31/12/2014 vu au 31/12/2014.
   - le nombre de contacts entre l'assureur et l'assuré, peu importe de qui vient l'initiative, entre le 01/01/2012 et le 31/12/2014.
3. Une base de sinistres entre le 01/01/2010 et le 31/12/2014 vue au 31/12/2014 contenant pour chaque sinistre la date du sinistre, le descriptif, la responsabilité de l'assuré et un indicateur clos sans paiement.
4. Une base de contacts réalisés entre le 01/01/2012 et le 31/12/2014 contenant pour chaque contact la date, la motivation du contact, le type de contact et le détail des informations sur le contact.

Les bases image représentent avant tout retraitement environ 2 300 000 contrats, la base de sinistres représente 1 900 000 sinistres sur les cinq années d'observation et la base de contacts représente environ 10 000 000 de contacts.

Les bases de **contrats** au 01/01/2010 et au 31/12/2014 contiennent les types de variables suivantes :

- les types de garanties souscrites,
- les dates d'effet et de résiliation éventuelle[1] des contrats,
- les caractéristiques de l'assuré (sexe, adresse, date d'obtention du permis, $\cdots$),
- les caractéristiques du véhicule,
- des informations sur les conditions tarifaires (prime, bonus/malus, réduction éventuelle, $\cdots$)
- des informations sur les autres contrats souscrits chez le partenaire,
- les nombres de sinistres et de contacts.

1. Résiliation éventuelle des contrats actifs au 31/12/2009 vue au 31/12/2014.

La base **sinistres** contient entre autres la date du sinistre, une description du sinistre (collision, remplacement pare-brise, ...), des informations sur la responsabilité éventuelle de l'assuré.

La base de **contacts** donne des informations dur la date, la raison et le type de contact ainsi que sur l'intérêt pour une souscription et la présence éventuelle de devis.

2.1.2 Nettoyage des données

Une étude statistique nécessite obligatoirement des pré-traitements sur les données afin de ne pas être perturbé par les points suivants :

— des incohérences dans les données (comme un assuré qui a obtenu son permis avant la naissance),
— des valeurs aberrantes (comme des sinistres de charge négative),
— des chaînes de caractère mal mises en forme (modèle de voiture, etc...),
— des valeurs manquantes,
— etc...

En fonction des problèmes rencontrés, plusieurs solutions se présentent. Si une variable possède trop de valeurs manquantes, il peut être préférable de la supprimer. Si des lignes présentent des incohérences, il est de même possible de les supprimer si celles-ci ne sont pas trop nombreuses. Dans les autres cas, si supprimer la ligne ou supprimer la variable n'est pas souhaitable, il est nécessaire d'effectuer des retraitements sur les données.

2.1.2.1 Pré-traitements génériques

Avant tout autre retraitement, il est nécessaire de supprimer plusieurs lignes de cette base.

Les lignes pour lesquelles la marque de la voiture est rencontrée une seule fois dans la base sont considérées comme incohérentes. C'est le cas par exemple de la marque « PEUEOT » qui est bien évidemment « PEUGEOT » avec une faute de frappe. Il serait possible de corriger directement ces champs à la main ou en utilisant la marque qui minimise la distance entre « PEUEOT » et les marques disponibles dans la base de données[2]. Vu que les marques rencontrées une seule fois dans la base de données ne représentent qu'une centaine de lignes, on supprime celles-ci sans se poser plus de question.

De même, les lignes présentant une incohérence entre date de permis, date de naissance et date de premier effet du contrat sont supprimées. C'est par exemple le cas des lignes pour lesquelles la date de permis est antérieure à la date de naissance.

D'autres lignes sont aussi supprimées, notamment celles ayant un nombre de champs vides trop important. Après toutes ces corrections, environ 0,1% des lignes ont été supprimées.

2. Avec une distance à définir.

Une fois ces pré-traitements effectués, il est nécessaire d'effectuer des corrections pour les champs ou lignes qui ne peuvent être supprimés sans introduire un biais trop important (perte d'information ou suppression d'un trop grand nombre de lignes).

2.1.2.2 Traitement des valeurs manquantes : Date d'obtention du permis

Au préalable, 711 lignes pour lesquelles la date de naissance est manquante sont supprimées.

La base contient environ 8000 lignes pour lesquelles la date de permis n'est pas renseignée. Il n'est pas envisageable de supprimer les lignes associées car cela induirait un biais trop important dans l'étude. Il n'est aussi pas possible de supprimer la variable « Date d'obtention du permis » car il est naturel de penser que la durée depuis l'obtention du permis peut avoir une influence sur le comportement de résiliation. Il faut donc trouver une solution pour retraitement le champ « Date d'obtention du permis ».

Une premier solution est de remplacement les valeurs manquantes par :

$$\text{Date de naissance de l'assuré} + \text{Moyenne de l'âge d'obtention du permis}$$

C'est le choix du remplacement de la valeur manquante par la valeur moyenne. Cependant, il est possible de raffiner cette correction des valeurs manquantes en utilisant les méthodes de l'apprentissage automatique.

L'idée est d'utiliser des arbres CART de régression (voir 4.5 pour la théorie) afin de prédire l'âge d'obtention du permis en fonction des caractéristiques du contrat. Cette méthode est plus fine que de remplacer directement par un âge moyen d'obtention du permis.

Les variables disponibles dans la base de données qui peuvent à priori permettre de prédire l'âge d'obtention du permis sont :

— la date d'adhésion de l'assuré,
— la date de premier effet [3] du contrat,
— le sexe de l'assuré.
— la date de naissance.

On calcule alors le temps en années entre les différentes dates et la date de naissance pour obtenir les variables **AgeObtentionPermis**, **AgeAdhesion** et **AgeDebutContrat**.

3. Date où le contrat auto a commencé, même s'il y a eu des modifications du contrat depuis.

La variable cible de l'arbre de régression est alors **AgeObtentionPermis** et les variables prédictives sont :

— **AgeAdhesion**
— **AgeDebutContrat**
— **Sexe**

La base utilisée pour l'arbre constitue l'ensemble des lignes pour lesquelles la date d'obtention du permis est renseignée.

On découpe alors cette base en 2 parties afin de tester les performances de l'arbre :

— la base d'apprentissage (66% de la base)
— la base de test (34% de la base)

Comme son nom l'indique, la base d'apprentissage sert de données sur lesquelles l'arbre va apprendre. Il convient alors de tester les performances de l'arbre ainsi obtenu en calculant l'erreur quadratique moyenne sur la base de test. Celle-ci est est définie comme :

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

avec y_i l'âge d'obtention du permis associée au contrat i sur la base de test, $\hat{y}_i$ l'âge d'obtention prédit pour le même contrat, et n le nombre de contrats dans la base de test.

L'erreur quadratique moyenne est de l'ordre de 4.95 années pour un split minimum des feuilles de l'arbre de 300000 contrats et de 4.75 années pour un split minimum de 20000 contrats. On peut voir que l'erreur quadratique moyenne diminue, tant sur la base de test que sur la base d'apprentissage, quand le split minimum diminue. Cela est intuitif car un arbre avec plus de feuilles sera à priori plus précis. Comme le but ici n'est pas d'avoir une très grande précision sur l'âge d'obtention du permis, mais plutôt de proposer une méthode d'imputation des valeurs manquantes plus précise qu'un simple remplacement par l'âge moyen d'obtention du permis, on ne poussera pas le split minimum jusqu'à obtenir le phénomène de sur-apprentissage. Ce dernier se matérialiserait par une très faible erreur sur la base d'apprentissage et une erreur importante sur la base de test, prouvant que la capacité du modèle à généraliser les prédictions sur des nouvelles données n'est pas bonne.

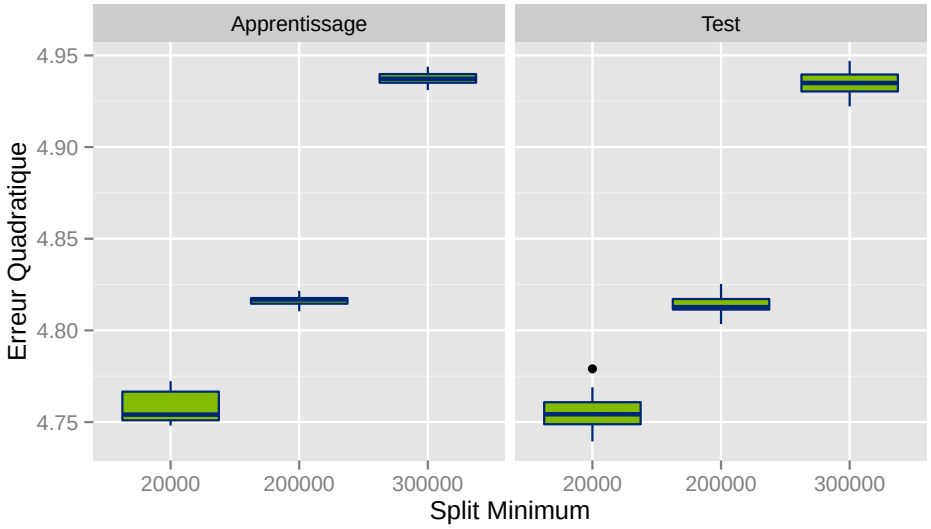


FIGURE 2.1 – Performances de l'arbre sur les bases d'apprentissage et de test (sur 10 modèles)

Un split minimum de 20000 contrats par feuille est donc choisi. L'arbre en résultant est tracé en figure 2.2. Pour le lire, par exemple, si l'âge d'adhésion est inférieur à 34 ans, puis à 26 ans, la moyenne de l'âge d'obtention du permis est de 19 ans et la feuille contient environ 37% des données.

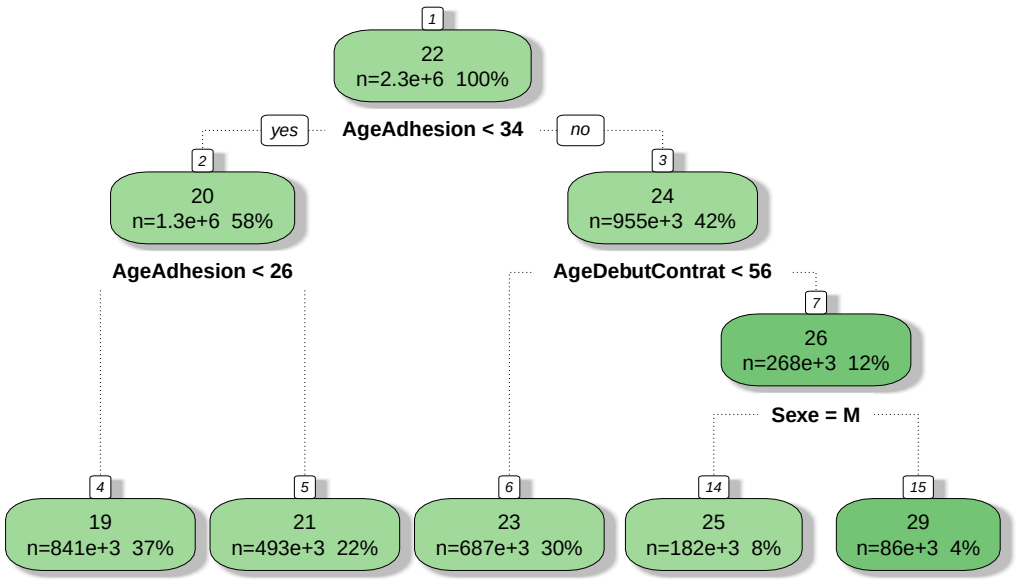


FIGURE 2.2 – Arbre obtenu pour un split minimum de 20000 contrats par feuille

2.1.2.3 Traitement des valeurs manquantes : Puissance DIN

Un autre champ intéressant est la puissance DIN de la voiture. Elle représente la puissance du moteur de la voiture dans l'unité *cheval DIN* avec 1 cheval DIN = 0.736

kW. La base de données comportent aussi le champ **Puissance fiscale** qui est un indicateur normalisé en France.

Depuis juillet 1998, cette dernière se définit, en notant C la consommation en CO_2 par kilomètre (g/km) et P la puissance maximale du moteur en kilowatts, par :

$$P_f = \frac{C}{45} + \left(\frac{P}{40}\right)^{1.6}$$

Dans la base de données, toutes les puissances fiscales sont renseignées mais il manque la puissance DIN pour environ 25000 contrats. Or, la puissance DIN semble plus importante pour l'étude que la puissance fiscale qui n'est qu'un indicateur normalisé sans sens physique. Comme un lien fort existe entre ces deux variables, l'idée est d'utiliser la puissance fiscale pour compléter les valeurs manquantes pour la puissance DIN.

Pour cela, une première idée est de découper les valeurs de la puissance DIN en intervalles et de calculer la consommation en dioxyde de carbone moyenne pour chaque intervalle en utilisant la formule précédente.

Pour affiner cette approche, on choisit d'utiliser une méthode à noyau de Nadaraya-Watson.

Connaissant des données $(x_i, y_i)_{i\in[\![1,n]\!]}$, l'estimateur de Nadaraya-Watson de y appartenant au couple (x, y) se définit par :

$$\hat{f}_n(x) = \sum_{i=1}^{n} W_{h,i} y_i$$

avec $W_{h,i}$ défini en fonction du noyau $K(\cdot)$ et h la largeur de bande :

$$W_{h,i} = \frac{K\left(\frac{x-x_i}{h}\right)}{\sum\limits_j K\left(\frac{x-x_j}{h}\right)}$$

Pour le problème, le noyau gaussien est choisi :

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$$

L'objectif est alors, connaissant les données $(P_{\text{fiscale,i}}, P_{\text{DIN,i}})$ pour lesquelles les deux champs sont renseignés, de trouver une fonction $\hat{f}$ permettant de prédire, si la valeur est manquante, P_{DIN} en fonction de P_{fiscale}.

L'estimateur s'écrit :

$$P_{\text{DIN}} = \hat{f}_n(P_{\text{fiscale}}) = \frac{\sum\limits_{i=1}^{n} \exp\left(-\frac{\left(P_{\text{fiscale}} - P_{\text{fiscale,i}}\right)^2}{h^2}\right) P_{\text{DIN,i}}}{\sum\limits_{i=1}^{n} \exp\left(-\frac{\left(P_{\text{fiscale}} - P_{\text{fiscale,i}}\right)^2}{h^2}\right)}$$

Le choix de la largeur de bande est le principal enjeu des estimateurs à noyau. Une des possibilités pour obtenir un h satisfaisant est de découper la base en bases d'apprentissage et de test et choisir h qui minimise l'erreur quadratique moyenne sur la base de test. Par soucis de simplicité, on prend pour largeur de bande une valeur pour laquelle la courbe de lissage semble visuellement bien suivre les points de données.

La largeur de bande choisie est de $h = 10$ et l'estimateur de Nadaraya-Watson obtenu est en figure 2.3.

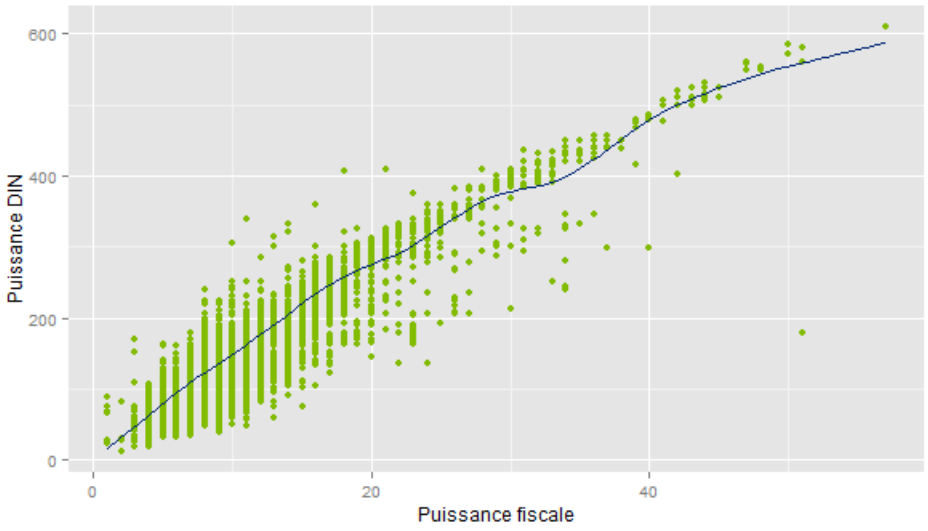


FIGURE 2.3 – Estimation par noyau gaussien de la puissance DIN ($h = 10$)

2.1.2.4 Imputation des types d'habitation par géolocalisation

Pour les assurés n'ayant pas par ailleurs de contrat MRH, le champ **Type d'habitation** (Maison, appartement ou autre) est vide. On propose une solution pour remplacer les valeurs manquantes par la valeur probable pour le type d'habitation.

L'intuition est que les maisons ont plus de chance d'être entourées par d'autres maisons que par des appartements. En effet, par exemple à Paris, on considère un assuré pour lequel le champ en question est manquant. En regardant le type d'habitation majoritaire parmi les assurés voisins pour lesquels le champ est disponible, il est possible d'en déduire que l'assuré habite dans un appartement.

Pour formaliser cette méthode, il est nécessaire de connaître les coordonnées GPS du lieu de résidence de chaque assuré pour pouvoir calculer des distances entre assurés.

Pour cela, on utilise le paquet **ggmap** de **R** qui permet d'effectuer automatiquement des demandes de coordonnées GPS associées à une adresse en effectuant une requête à Google Maps. Google limitant à 2500 requêtes par jour la fonction de géocoding présente dans ce paquet, il n'est pas envisageable d'obtenir les coordonnées associées à chaque adresse. De plus, on ne dispose pas de l'adresse de l'assuré dans la base mais seulement le code postal de son lieu de résidence.

On obtient via Google Maps les coordonnées GPS associées à chacun des codes postaux présents dans la base. Cela représente environ 7000 codes postaux uniques. Pour géocoder par exemple le 75014, on effectue la requête *75014 France* à Google Maps qui renvoie la longitude et la latitude associée au centroïde [4] du 14ème arrondissement de Paris.

Une fois les coordonnées GPS des codes postaux obtenues, il est nécessaire de corriger les erreurs de géocoding en supprimant les coordonnées qui ne sont pas dans les limites de la France Métropolitaine. Une fois cela effectué, la base dispose des champs **longitude** et **latitude**.

Enfin, pour pouvoir calculer des distances entre assurés, il faut projeter les coordonnées sphériques de la latitude et longitude en un système de coordonnées cartésiennes. Cela peut déformer les distances par rapport à une formule exacte de la distance entre 2 points dans un système de coordonnées sphériques. On utilise la projection Lambert II étendue dont l'explication sort du contexte ici. Cela transforme longitude et latitude en coordonnées X et Y en mètres.

L'approche des plus proches voisins, parfois appelées approche k-NN (*k-Nearest Neighbors*), est un type d'apprentissage supervisé qui consiste à prédire une variable cible en fonction de la valeur de ses voisins. Dans le cas de notre problème, la distance entre deux points $\mathrm{P}_1 = (x_1, y_1)$ et $\mathrm{P}_2 = (x_2, y_2)$ est simplement la distance euclidienne :

$$d(\mathrm{P}_1, \mathrm{P}_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Pour connaître le type d'habitation de l'assuré, on regarde, parmi les k plus proches voisins de l'assuré, quelle est le type d'habitation majoritaire. Le type majoritaire définit la prédiction.

Pour cela, on constitue une base constituée d'une ligne par assuré ainsi que le type d'habitation et les coordonnées associés à l'assuré.

Les modalités sont distribuées comme ceci dans la base, parmi les assurés pour lequel le champ est disponible :

On choisit de négliger la modalité « Mobile Home » car elle serait de toute façon impossible à prévoir en utilisant l'approche des k plus proches voisins. La base sur laquelle le modèle va être entraîné et testé est donc celle constituée de tous les contrats pour lesquels la variable **Type d'habitation** est renseignée, et pour lesquelles la modalité est soit *Maison* soit *Appartement*, ce qui représente plus de 99,9% de la base initiale.

Pour le modèle, la base d'apprentissage constitue 75% de la base ainsi créée et le base de test regroupe les 25% d'observations restantes.

4. Le centre de gravité associé au polygone délimitant les limites du 14ème arrondissement.

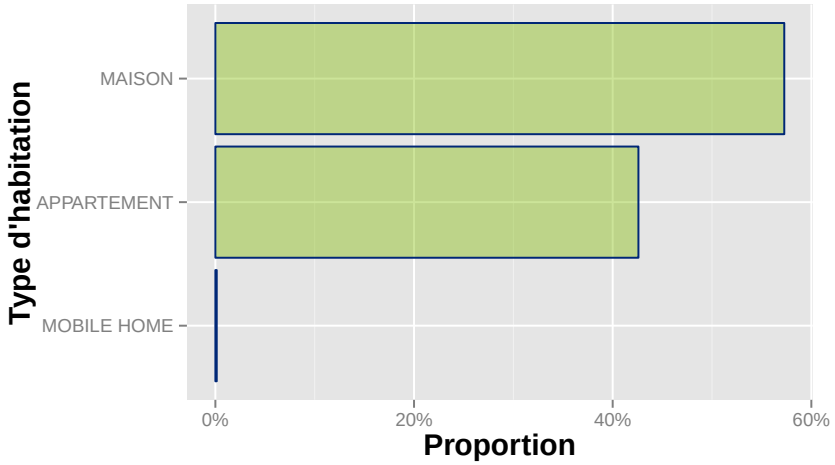


FIGURE 2.4 – Proportion des différents types d'habitation

L'objectif est de déterminer la valeur de k qui va minimiser l'erreur de prédiction sur la base de test. Un des problèmes soulevés est que la précision du géocoding n'est qu'au code postal près. Plusieurs assurés étant géocodés en un même point, définir lesquels sont les plus proches voisins peut être problématique si plus de k assurés sont présents en un code postal donné. Pour régler ce problème, il convient juste de choisir arbitrairement lesquels sont les k plus proches voisins avec un k maximisant les performances de prédiction sur la base de test.

Pour éviter d'avoir un vote à égalité parmi les modalités de **Type d'habitation** des k plus proches voisins, il convient de choisir des valeurs de k impaires. De plus, il est nécessaire de ne garder qu'une seule ligne par assuré car un biais existerait en gardant les plusieurs contrats d'un même assuré. En effet, pour un contrat donné, son plus proche voisin pourrait être un autre contrat correspondant au même assuré et cela reviendrait à prédire une valeur grâce à elle-même car le type d'habitation ne change pas à priori entre les différents contrats d'un même assuré.

Les erreurs de classification pour les valeurs de k de 1, 5, 11, 51, 75 et 101 sont affichées en figure 2.5. L'erreur de classification décroit quand k augmente et semble se stabiliser aux alentours de 25%. Elle semble toujours légèrement supérieure pour l'échantillon de test par rapport à la base d'apprentissage à k donné.

Il est nécessaire de savoir si 25% d'erreur est un résultat satisfaisant ou non. En effet, l'algorithme doit battre le hasard pour être utile.

Soit $\mathbf{U} = (u_1, \cdots, u_n) \in \{0,1\}^n$ et $p = \frac{1}{n}\sum_{i=1}^{n} u_i$. $\mathbf{U}$ représente les valeurs du type d'habitation pour les différents assurés.

On considère le vecteur aléatoire $\mathbf{X} = (X_1, \cdots, X_n)$ de composantes indépendantes deux à deux et suivant une loi de Bernoulli $\mathcal{B}(p)$.

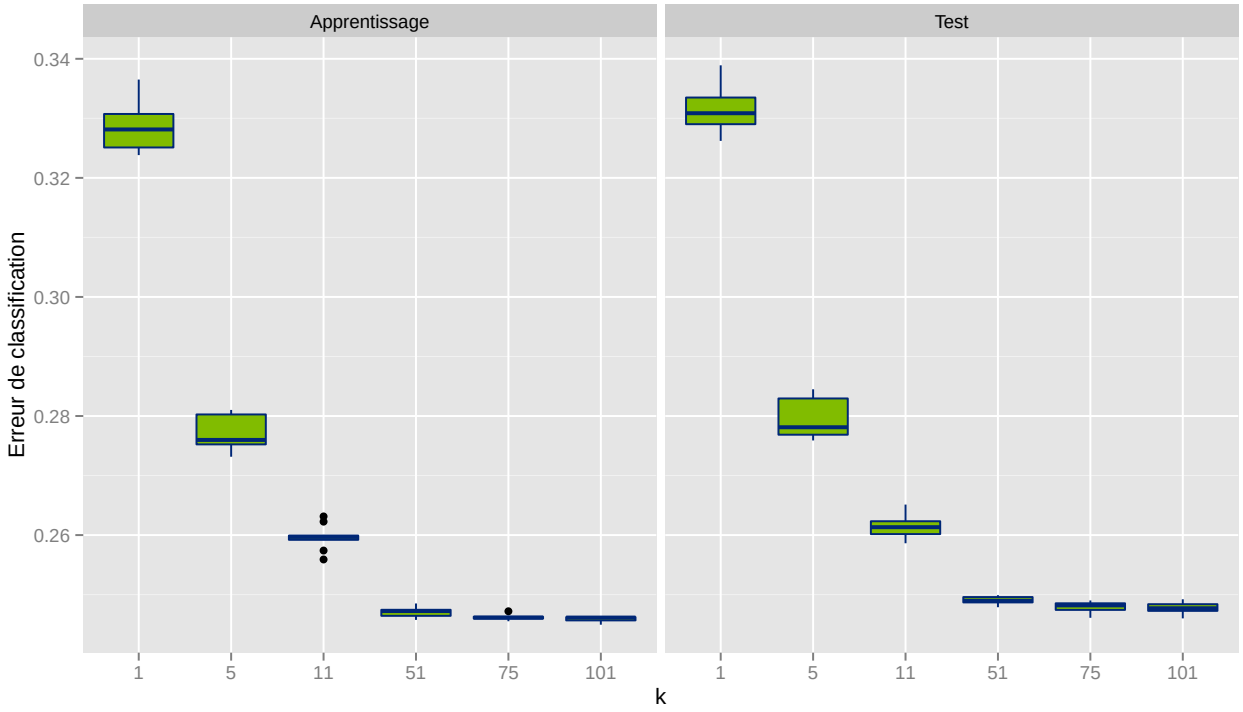


FIGURE 2.5 – Erreur de classification en fonction de k

Soit alors la variable aléatoire $\mathbf{T}(\mathbf{X})$ définie par :

$$\mathbf{T}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(X_i = u_i)$$

$1 - \mathbf{T}(\mathbf{X})$ représente l'erreur de classification, ayant prédit les valeurs de $\mathbf{U}$ aléatoirement avec la variable $\mathbf{X}$.

Pour décider que l'algorithme des k-plus proches voisins a battu le hasard avec une confiance de 95%, il faut [5] que l'erreur de classification de l'algorithme, 25%, ne soit pas dans l'intervalle de confiance de l'estimateur de l'espérance de $1 - \mathbf{T}(\mathbf{X})$ à 95%.

Soit m réalisations indépendantes de $\mathbf{X}$.

L'estimateur de l'espérance de $\mathbf{C} = 1 - \mathbf{T}(\mathbf{X})$ sur ces m réalisations s'écrit :

$$\bar{\mathbf{C}} = \frac{1}{m}\sum_{j=1}^{m}\mathbf{C}_j = 1 - \frac{1}{m}\sum_{j=1}^{m}\mathbf{T}(\mathbf{X})_j$$

Son espérance vaut :

$$\mathbb{E}\left[\bar{\mathbf{C}}\right] = 1 - \left[p^2 + (1-p)^2\right] = 2p(1-p)$$

5. En réalité, 25% d'erreur a été trouvé en utilisant 10 échantillons de test et n'est donc qu'un estimateur de l'erreur de classification. Cependant, on fait l'approximation que cet estimateur est l'espérance de l'erreur de classification de l'algorithme.

et sa variance :

$$\mathbb{V}\left[\bar{\mathbf{C}}\right] = \frac{1}{m}\mathbb{V}[\mathbf{T}(\mathbf{X})] = \frac{2p(1-p)(2p^2-2p+1)}{m \times n}$$

Pour estimer l'intervalle de confiance à 95% de $\bar{\mathbf{C}}$, la technique du bootstrap est utilisée. Avec m réalisations de $\mathbf{C}$, on procède par ré-échantillonnage avec remise et la moyenne sur chaque échantillon est calculée. L'intervalle de confiance à 95% est délimité par les quantiles à 2,5% et à 97,5% obtenus.

Pour $m = 20000$, $n = 1036266$ et un nombre de ré-échantillonnage de 2000, on obtient un quantile à 2,5% de 0.4891837 et un quantile à 97,5% de 0.4891963. La valeur de 0.25 est bien en dehors de l'intervalle, ce qui valide bien que l'algorithme mis en place est plus performant que le hasard pour l'imputation des types d'habitation manquants.

L'erreur globale restant cependant élevée, on exclut de la base de données le type d'habitation.

2.1.3 Obtention d'une base utilisable

La phase de retraitement des données est critique. Si les données sont mal retraitées, les algorithmes de Machine Learning ne pourront pas fonctionner convenablement.

2.1.3.1 Regroupement des modalités rares des variables qualitatives

Le principal retraitement des données concerne les variables qualitatives. Il est nécessaire de réduire le nombre de modalités en reclassant les modalités trop rares dans une modalité « Autre ». Cette étape est essentielle et nécessite de fixer un seuil de représentativité en dessous duquel une modalité doit être reclassée dans « Autre ».

Laisser trop de modalités revient à laisser du bruit dans les données et peut perturber notamment les arbres de classification et régression.

La variable **Classe Socio-Professionnelle** a 23 modalités. Il n'est pas envisageable de les conserver toutes et notamment les modalités très rares. Les classes similaires sont regroupées, par exemple Retraité et retraité du secteur public. On obtient grâce à cela 9 regroupements de modalités.

Cette opération est difficilement automatisable et constitue l'un des retraitements qui doit être fait à la main.

2.1.3.2 Fusion des bases de données fournies

Le Partenaire a fourni 3 bases pour l'étude :

— la base contenant les informations sur les contrats actifs au 1er janvier 2010 (vue au 1er janvier 2010 et au 31 décembre 2014),

— la base contenant les sinistres associés à ces contrats entre ces deux dates,

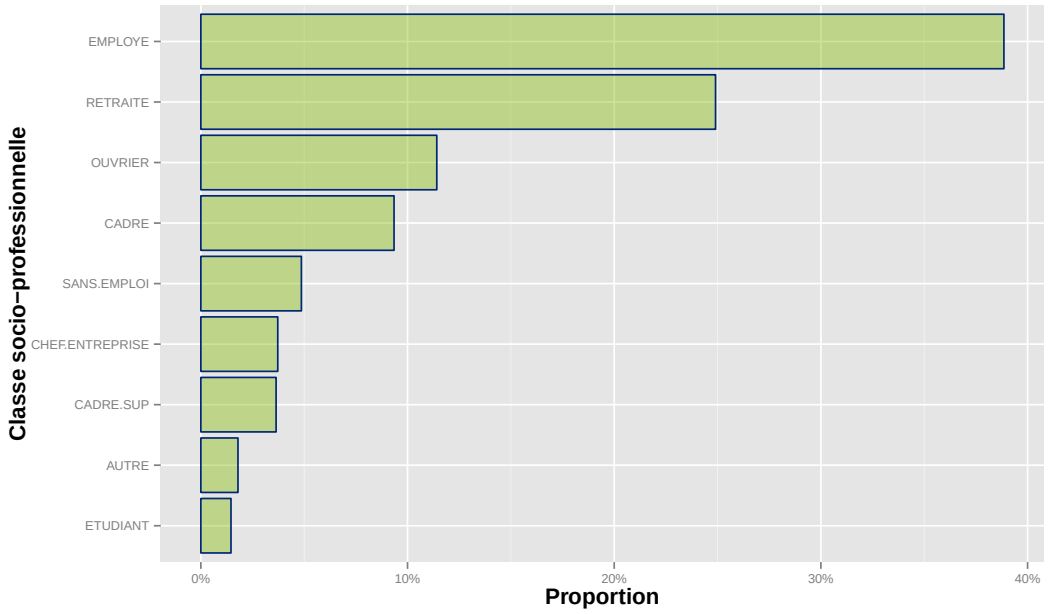


FIGURE 2.6 – Classe socio-professionnelle après retraitement

— la base contenant les contacts associés à ces contrats entre le 1er janvier 2012 et le 31 décembre 2014.

Il est nécessaire de pouvoir relier ces trois bases afin d'en extraire des informations utiles à la compréhension des données.

Pour connecter ces bases entre elles, il est indispensable de définir un identifiant unique permettant de relier une ligne donnée d'une base avec les lignes correspondantes au même contrat dans les autres bases.

Dans la base contrats, chaque ligne est représentée par un numéro d'assuré et un numéro de contrat. Un assuré peut posséder plusieurs contrats. Le couple **Numéro d'assuré - Numéro de contrat** identifie de manière unique un contrat. Dans cette base, ce couple représente la clé primaire d'identification.

Dans la base sinistres, le couple Numéro d'assuré - Numéro de contrat permet de faire le lien avec les données de la base contrats. Une ligne de la base contrats peut cependant être associée à plusieurs lignes de la base sinistres car un contrat peut avoir plusieurs sinistres.

Pour chaque couple Numéro d'assuré - Numéro de contrat, il est donc possible de calculer le nombre de sinistres et le type de sinistre le plus courant pour une période d'observation donnée. Afin de nettoyer les données, certaines modalités rares de type de sinistre sont regroupées et la modalité *Pas de Sinistre* est rajoutée pour éviter d'avoir une valeur manquante pour les contrats non sinistrés. La figure 2.7 montre la répartition du type le plus courant de sinistre par contrat sur la période 2010-2014.

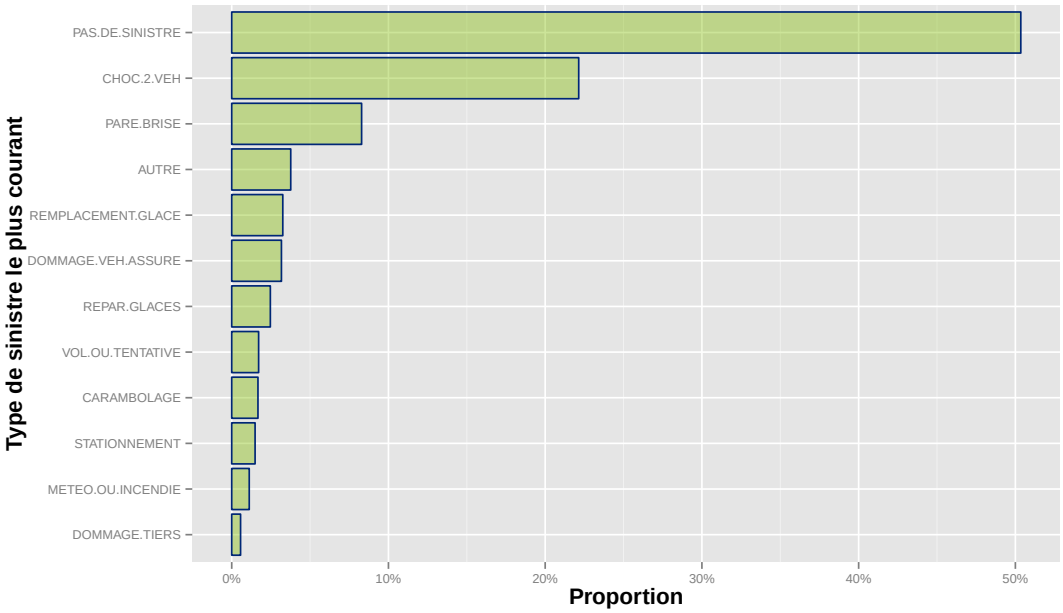


FIGURE 2.7 – Type de sinistre le plus courant par couple Numéro d'assuré - Numéro de contrat

La base contacts diffère des deux autres car un contact est seulement identifié par le numéro d'assuré qui lui est associé. En effet, un contact comme un appel téléphonique ne concerne pas un contrat donné car l'assuré peut en posséder plusieurs. Il n'est donc pas possible de relier directement cette base aux autres.

Le choix est fait est d'associer à chaque couple Numéro d'assuré - Numéro de contrat le nombre de contacts et le type de contact le plus courant correspondant au même numéro d'assuré dans la base contacts, pour une période donnée.

La méthode permettant de relier ces bases ayant été expliquée, il est possible de commencer l'étude statistique.

2.1.4 Description des données

Avant toute étude statistique, il est nécessaire de décrire la base de données pour comprendre la distribution des variables. Cela permet d'obtenir les intuitions indispensables à la suite de l'étude.

L'étude statistique préliminaire est réalisée sur la base des contrats retraitée et enrichie avec les données des bases sinistres et contacts. Les variables disponibles dans la base après retraitements sont affichées dans le tableau 2.8.

Certaines variables (**NbSinistres**, **TypeSinistre**, **NbContacts**, **TypeContact** et **TauxSinResponsable**) doivent être calculées sur une période donnée, par exemple l'année 2010.

Variable	Description
Formule	Formule du contrat
Option	Options de la formule
MarqueVehicule	Marque du véhicule (Exemple : Renault)
ModèleVehicule	Modèle du véhicule (Exemple : Twingo, ...)
AppellationCommerciale	Appellation commerciale du véhicule (Exemple : 1.5L TDI)
Moteur	Type de moteur (Diesel, essence, ...)
PuissanceVehicule	Puissance DIN du véhicule
BonusMalus	Bonus-Malus en pourcent
MajorConducteurNovice	Majoration du tarif car conducteur novice (Oui/Non)
ReducFamille	Réduction de tarif pour les familles (Oui/Non)
UsageVehicule	Type d'usage (Privé, Privé et travail, Travail uniquement)
ReducTarif	Réduction de tarif (Oui/Non)
NbConducteurs	Nombre de conducteurs associés au véhicule
NbConduiteAcc	Nombre de conducteurs en conduite accompagnée
Sexe	Sexe
AgeAssure	Âge de l'assuré au 1er janvier 2010
SitFamiliale	Situation familiale (marié ou autre)
CSP	Classe socio-professionnelle
LienAutreAssure	Contrat lié à un autre contrat de la base (Oui/Non)
NbSinistres	Nombre de sinistres
TypeSinistre	Type de sinistre majoritaire (Choc véhicule, météo, ...)
TauxSinResponsable	Taux de sinistres responsables
NbContacts	Nombre de contacts entre début 2012 et fin 2014
TypeContact	Type de contact majoritaire
NbContratAuto	Nombre de contrats auto
NbContratMRH	Nombre de contrats MRH
NbContratAutreIARD	Nombre de contrats IARD hors auto et MRH
AgeObtentionPermis	Âge d'obtention du permis de conduire
AgeAdhesion	Âge auquel l'assuré a souscrit son premier contrat
AgeVoiture	Durée entre la date de 1ère circulation et début 2010
AgeDebutContrat	Âge auquel l'assuré a souscrit son premier contrat auto
DurationContrat	Durée entre la date d'effet du contrat et début 2010
X	Coordonnée GPS en projection Lambert II étendue
Y	Coordonnée GPS en projection Lambert II étendue
Depart	Résiliation du contrat auto entre 2010 et 2014 (Oui/Non)
DureeVieContrat	Pour les contrats résiliés, durée de vie totale du contrat
DureeVieClient	Pour les clients quittant le Partenaire, durée d'adhésion totale
DureeVieContratApres2010	Durée en portefeuille après début 2010
DateResilContrat	Date de résiliation du contrat
MotifResilContrat	Motif de résiliation du contrat
DateResilAssure	Date de départ de l'assuré de la compagnie d'assurance
MotifResilAssure	Motif de départ de l'assuré

FIGURE 2.8 – Variables de la base

2.1.4.1 Statistiques descriptives de la base de données

Le choix est fait, par souci de confidentialité, de ne pas décrire la base de données en tant que telle (répartition des âges des assurés, etc...) en détail. Les statistiques descriptives se résument à l'étude des différentes variables sur la résiliation.

La distribution de la durée de vie des contrats (en ne considérant que ceux résiliés pendant la période d'observation) est affichée en figure 2.9. La moyenne des durées de vie est d'environ 11 ans pour une médiane d'environ 8 ans et un écart-type de 9 ans.

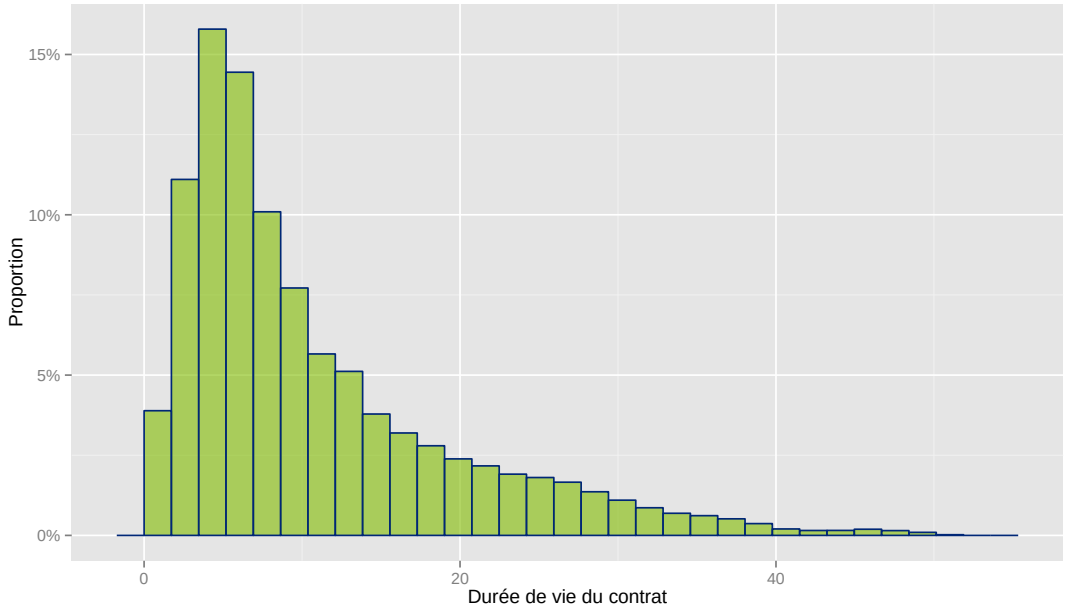


FIGURE 2.9 – Distribution de la durée de vie d'un contrat d'assurance

Les contrats d'assurance auto sont définis par une formule et des options associées. La figure 2.10 détaille la répartition des options souscrites toutes formules confondues. L'option 2 est l'option de base choisie par les assurés et l'option 3 est une option qui n'existe plus mais qui est encore détenue par quelques assurés. L'option choisie est liée à l'aversion au risque de l'assuré et donc à son comportement. On peut donc penser que l'option choisie a une influence sur la propension à résilier.

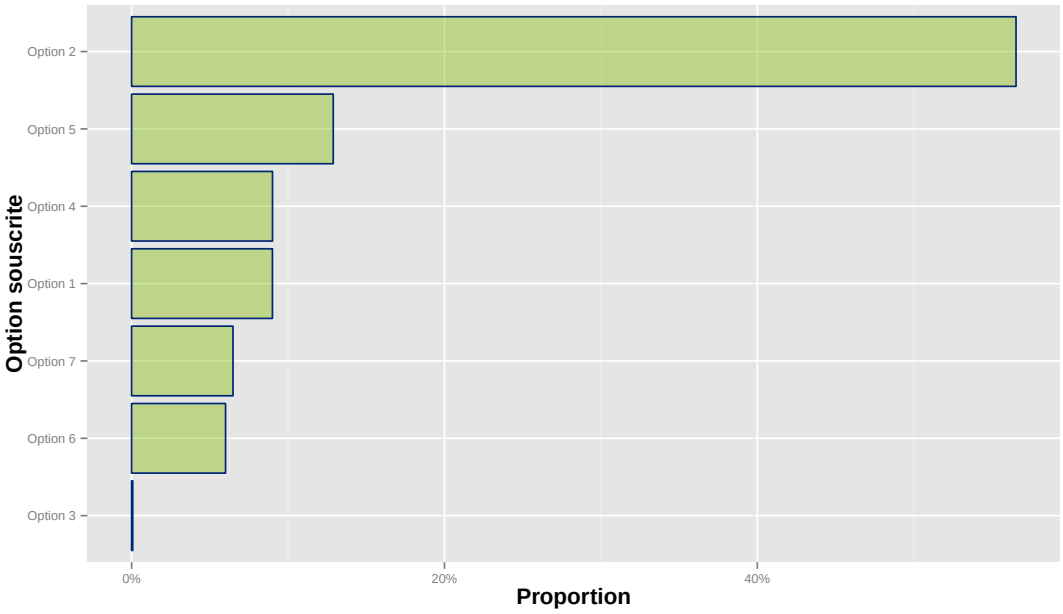


FIGURE 2.10 – Répartition des options souscrites

La durée de vie des contrats en fonction de l'option choisie ne montre cependant pas de différences majeures entre les options.

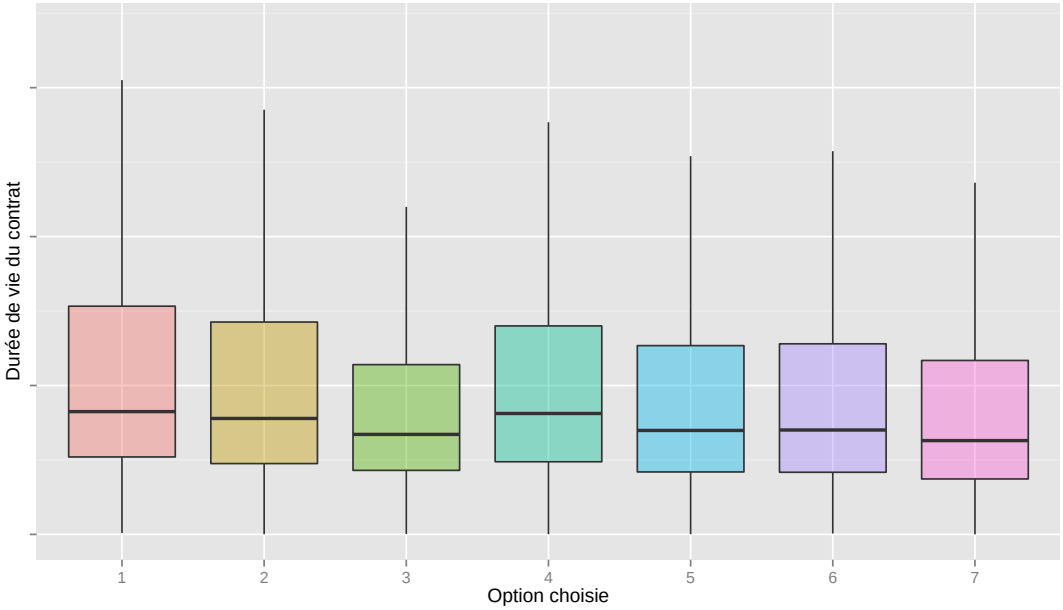


FIGURE 2.11 – Durée de vie du contrat en fonction de l'option choisie (échelle anonymisée)

La répartition des contrats en fonction de s'il y a réduction tarifaire ou non est affichée en figure 2.12.

Il y a 5,2% de contrats avec une réduction tarifaire appliquée et 94,8% sans réduction tarifaire.

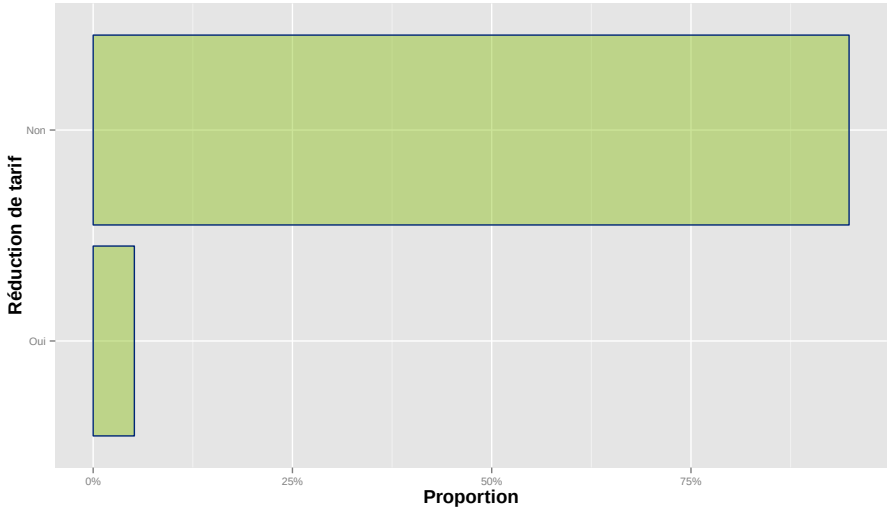


FIGURE 2.12 – Réduction tarifaire par contrat

Ensuite, la distribution des durées des contrats de la base résiliés entre 2010 et 2014 en fonction de la réduction tarifaire est affichée en figure 2.13.

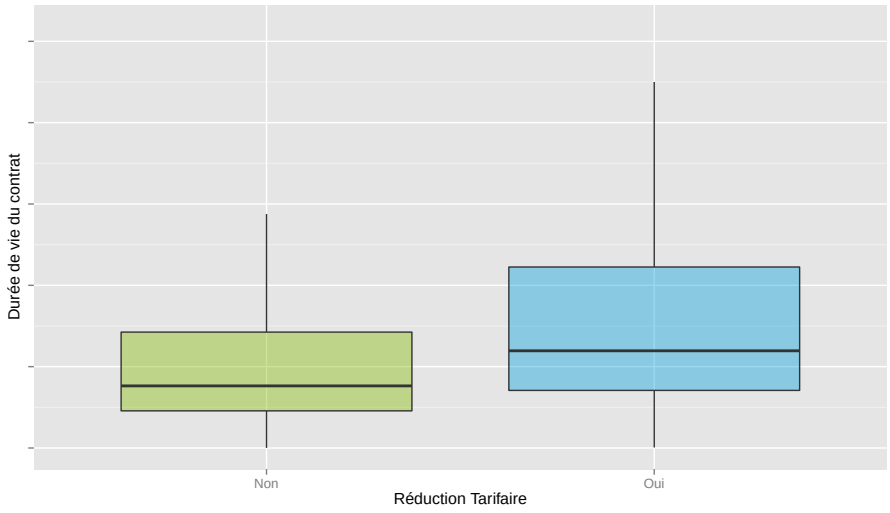


FIGURE 2.13 – Réduction tarifaire par contrat (échelle des durées anonymisée)

On semble voir, malgré quelques valeurs aberrantes, que la durée de vie en portefeuille des contrats avec réduction tarifaire est supérieure à celles sans réduction tarifaire.

L'idée est d'utiliser un test statistique pour tester l'égalité des moyennes. Comme les variances ne sont pas égales, le t-test de Welch est utilisé.

En notant μ_O (resp. μ_N) la moyenne des durées de vie s'il y a réduction tarifaire (resp. sans réduction tarifaire), σ_O (resp. σ_N) l'écart-type des durées de vie avec réduction tarifaire et n_O (resp. n_N l'effectif avec réduction tarifaire (resp. sans), on a la t-statistique de Welch telle que :

$$t = \frac{\mu_O - \mu_N}{\sqrt{\frac{\sigma_O^2}{n_O} + \frac{\sigma_N^2}{n_N}}}$$

Le nombre de degrés de liberté ν est approché par l'équation de Welch-Satterthwaite :

$$\nu \approx \frac{\left(\frac{\sigma_O^2}{n_O} + \frac{\sigma_N^2}{n_N}\right)^2}{\frac{\sigma_O^4}{n_O^3 \cdot \nu_O} + \frac{\sigma_N^4}{n_N^3 \cdot \nu_N}}$$

avec $\nu_i = n_i - 1$ pour $i = \{O, N\}$ les degrés de liberté associés au calcul de la variance dans chaque échantillon.

Une fois la t-statistique et le degré de liberté ν calculé, le test unilatéral de Student suivant est appliqué : $H_0 : \mu_O > \mu_N$.

Pour $\mu_O = 15.58$, $\mu_N = 10.71$, $\sigma_O = 10.63$, $\sigma_N = 8.78$, $n_O = 11238$ et $n_N = 700313$, on obtient une p-value de 0 [6], ce qui signifie que l'on accepte l'hypothèse nulle H_0 avec une confiance d'environ 100%.

Le fait que les assurés avec une réduction tarifaire résilient moins que ceux sans est donc validé statistiquement.

On s'intéresse maintenant à l'impact du bonus-malus sur la résiliation. Les deux groupes considérés pour l'étude sont ceux pour lesquels le bonus-malus est égal à sa valeur minimale [7] contre le cas où la valeur est supérieure à celle-ci. Cela donne en proportions :

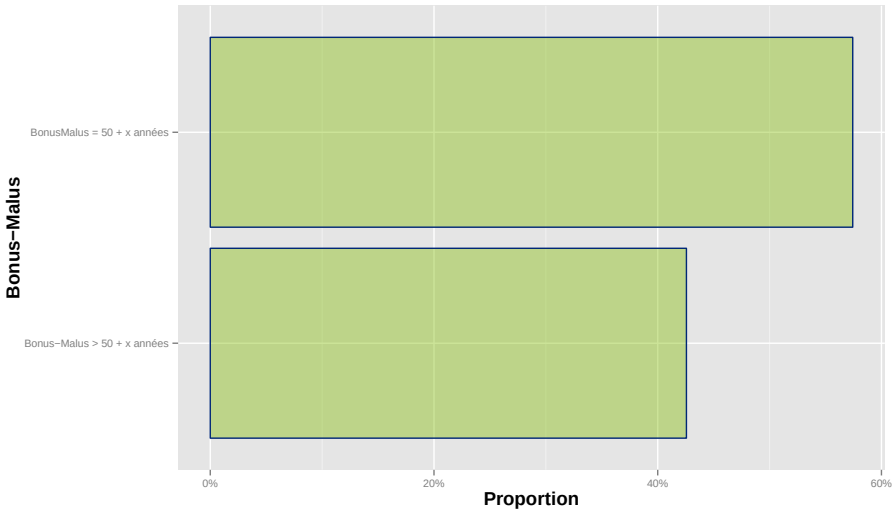


FIGURE 2.14 – Répartition des bonus-malus

6. Numériquement, cette valeur vaut 0 mais n'est pas nulle en réalité. La valeur est seulement en dessous du seuil de précision des variables de type **double** sur **R**.

7. La valeur minimale du coefficient de bonus-malus vaut 50% + x années, c'est-à-dire 50% plus x années sans sinistre. La valeur exacte n'est pas donnée pour des raisons de confidentialité.

En figure 2.15, la durée de vie associée à ces groupes est affichée. Le même test statistique que précédemment montre que les contrats au bonus-malus minimal résilient moins que ceux au bonus-malus plus élevé. Comme un coefficient de bonus-malus minimal donne une indication sur le nombre d'années sans sinistre, on peut donc voir que les individus n'ayant pas eu de sinistre depuis longtemps ont une durée de vie en portefeuille plus longue. Un biais existe car si un individu n'a pas eu de sinistre depuis un certain temps, cela veut déjà dire qu'il est resté longtemps assuré et donc il est normal que sa durée de vie moyenne soit élevée.

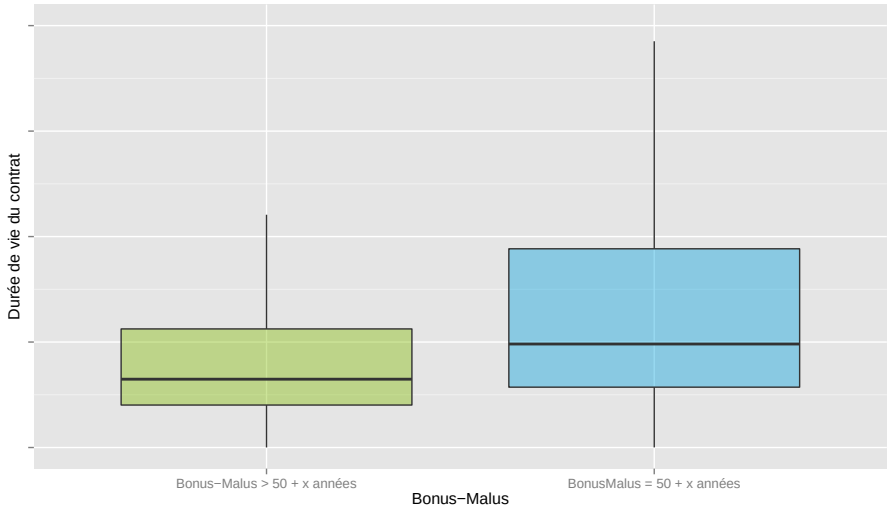


FIGURE 2.15 – Répartition des durées de vie par groupe de bonus-malus (échelle des durées anonymisée)

2.2 Conclusion de l'analyse préliminaire des données

Cette partie a permis, après raffinement des données brutes à l'aide de méthodes innovantes, d'obtenir une base de données relativement propre pour la suite de l'étude.

Les problématiques rencontrées dans cette partie ont été multiples. Tout d'abord, il a été nécessaire de transformer les données brutes en données exploitables. Cela a nécessité de nombreux retraitements. Ceux-ci dépendent de l'étude et justifient qu'une étude réalisée sur les même données puisse avoir des résultats différents. A un moment donné, lors des retraitements, il est notamment nécessaire de supprimer volontairement de l'information afin d'obtenir des données exploitables comme cela a été le cas avec la fusion des bases fournies.

L'apport des techniques de l'apprentissage automatique a permis d'effectuer un retraitement intelligent des valeurs manquantes dans les données.

Le traitement des valeurs manquantes a aussi été permis par l'apport des données externes. Utiliser les coordonnées GPS associées aux différents contrats grâce à Google Maps a permis de compléter les types d'habitation manquants.

Les données sont maintenant prêtes pour effectuer la segmentation du portefeuille, étape qui va permettre d'obtenir des clusters homogènes d'assurer afin de mieux comprendre les données. Cette segmentation va utiliser des méthodes innovantes issues des réseaux de neurones artificiels : les cartes auto-adaptatives.

Chapitre 3

Segmentation du portefeuille

Sommaire

Segmenter le portefeuille consiste à identifier des sous-groupes homogènes au sein des données. La segmentation permet d'améliorer la connaissance du portefeuille en la résumant à un nombre limité de profils moyens. Le comportement de résiliation au sein des différents clusters peut être très éloigné du comportement moyen et identifier des populations résiliant beaucoup plus ou beaucoup moins que la moyenne est très intéressant. Le but de cette partie est de montrer l'intérêt des cartes auto-adaptatives pour l'analyse de données en actuariat, et en particulier pour l'étude de la résiliation.

3.1 Cartes auto-adaptatives (Self-Organizing Maps)

Une carte auto-adaptative (ou en anglais Self-Organizing Map, SOM), parfois aussi appelée réseau de Kohonen, est un type de réseaux de neurones non supervisés qui produit une vue en basse dimension (généralement 2) de l'espace d'entrée (la base d'apprentissage). Ces cartes ont été décrites pour la première fois dans un article de Teuvo Kohonen : *Self-organized formation of topologically correct feature maps*, 1982 (voir KOHONEN T. (1982)). Il existe plusieurs types de SOM mais le mémoire s'intéressera uniquement aux cartes auto-adaptatives de Kohonen. Cet algorithme seul ne permet pas de segmenter des données mais uniquement de les décrire aisément.

3.1.1 Idée sous-jacente

Le principe des SOM [1] est très simple à comprendre. Il est exposé ici de manière très grossière. Considérons une foule sur un terrain de football. Chaque personne a des attributs (son âge, sa taille, son sexe, etc...). Pendant que l'algorithme SOM tourne, chaque personne se déplace jusqu'à ce que les personnes qui ont des attributs proches soient à côté les unes des autres. Regardons alors le stade vu d'en haut (la carte auto-adaptative) et demandons à chaque personne de tenir une pancarte avec son âge, par exemple. Cela fait une photographie de la répartition des âges sur la carte. Faisons de même avec le salaire, le sexe et cela permet d'obtenir plusieurs cartes. Cela permet d'identifier graphiquement les liens entre les différentes variables.

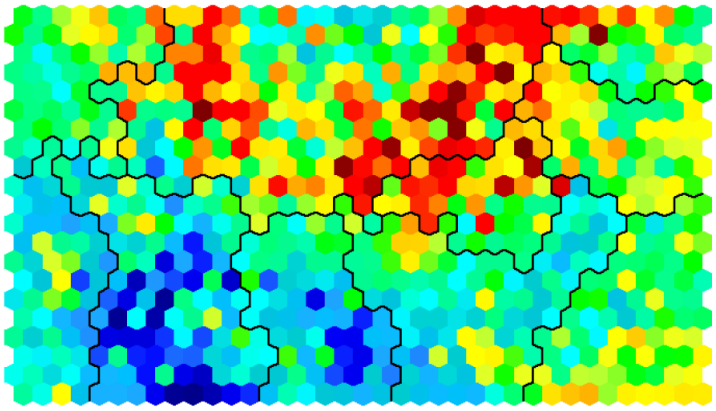


FIGURE 3.1 – Distribution d'une variable donnée sur la carte auto-adaptative [2]

1. Self-Organizing Maps
2. Source : Deloitte Analytics

3.1.2 Théorie des cartes auto-adaptatives

Le SOM est un réseau de neurones non supervisé. C'est une technique de réduction de la dimension. L'algorithme est un modèle mathématique abstrait de la façon dont fonctionnent conjointement les capteurs de la rétine et le cortex cérébral (voir Figure 3.2). Dans le cerveau, les aires visuelles sont organisées de telle façon que deux neurones physiquement proches dans le cortex visuel traitent des entrées physiquement proches dans la rétine. C'est l'organisation rétinotopique. Elle n'est pas déterminée génétiquement mais résulte d'un apprentissage.

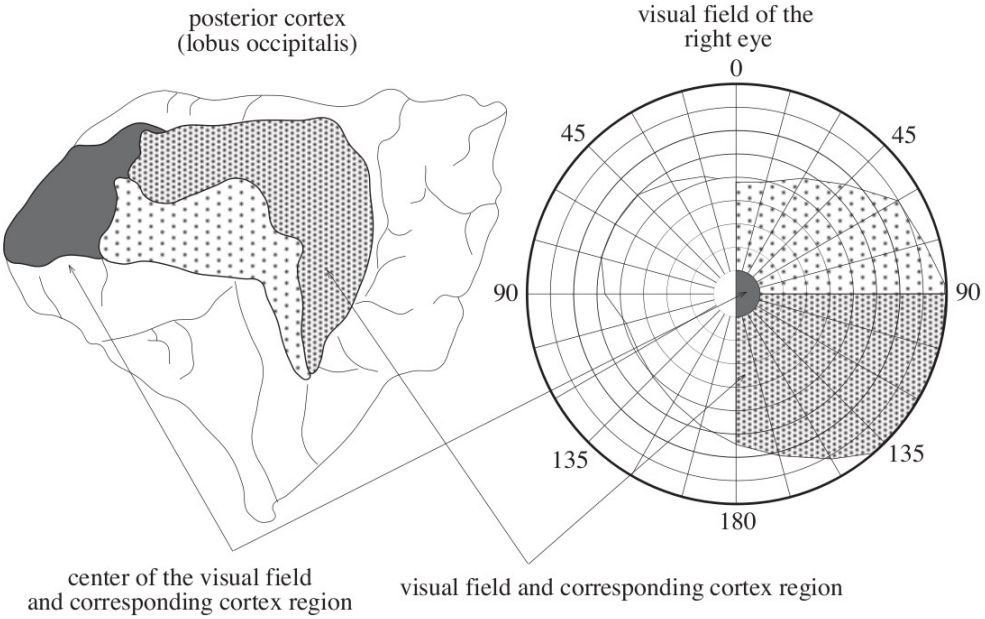


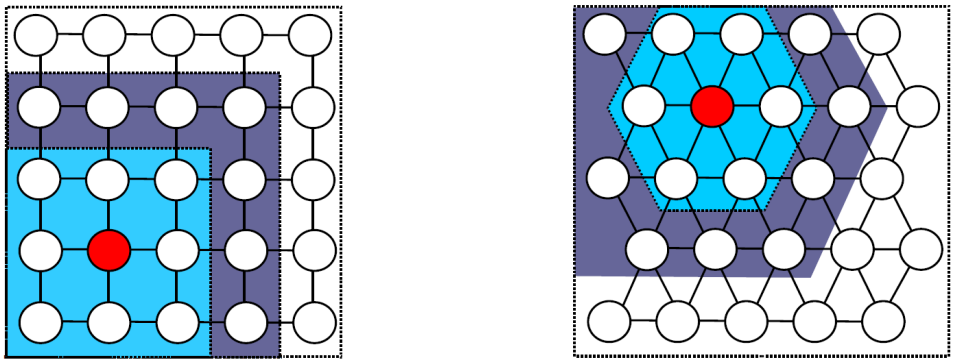
FIGURE 3.2 – Origine biologique des cartes auto-adaptatives (Source : Nicolas P. Rougier - Université de Bordeaux)

Soit n de vecteurs de l'espace d'entrée et p la dimension de ces vecteurs.

On note $\left\{\mathrm{x}^{(k)}\right\}_{k\in[\![1,n]\!]} = \left\{\left(x_1^{(k)},\ldots,x_p^{(k)}\right)\right\}_{k\in[\![1,n]\!]}$ les données en entrée (la base d'apprentissage).

Une carte auto-adaptative est composée de M neurones tous reliés les uns les autres. Cette carte a une topologie, c'est-à-dire que l'on définit une distance entre les neurones (souvent la distance euclidienne). De plus, on associe à chaque neurone i son représentant dans l'espace d'entrée $\mathrm{w}_i \in \mathbb{R}^p$. w_i est appelé **prototype** du neurone i ou vecteur de poids (voir Figure 3.5).

Les grilles neuronales sont souvent en deux dimensions et la forme des neurones est généralement hexagonale ou carrée de telle sorte qu'un neurone ait respectivement 6 ou 4 neurones immédiatement voisins.



(a) Grille neuronale carrée (b) Grille neuronale hexagonale

FIGURE 3.3 – Deux types de grilles neuronales 2-D (Source : Cartes auto-adaptatives pour l'analyse de données, Vincent Lemaire)

L'algorithme réalise une quantification vectorielle (cf. annexe 4.5) de l'espace de données. Les prototypes (ou vecteurs de poids) sont tous reliés entre eux et forment un réseau. Le but de l'algorithme est que la topologie du réseau s'adapte le mieux possible à la topologie de l'espace des données (voir Figure 3.4). Chaque vecteur x des données d'entrée est rattaché au neurone qui a son prototype le plus proche de x.

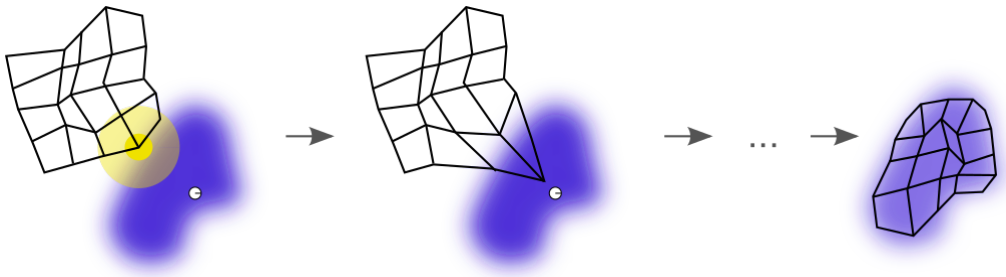


FIGURE 3.4 – Adaptation du réseau de prototypes (la grille noire) à la distribution des données (le nuage bleu). Le réseau s'adapte de plus en plus aux données au fil des itérations (Source : Wikipedia.com - Self-Organizing Maps).

Algorithme L'algorithme du SOM se base sur 4 étapes (voir YIN H. (2008)) :

1. **Initialisation** : Chaque vecteur de poids est initialisé à une valeur aléatoire de telle sorte que les vecteurs initialisés recouvrent une partie importante de l'espace des données.
2. **Compétition** : A chaque itération t est choisi aléatoirement un vecteur de données $\mathrm{x}(t)$ à la grille de prototypes $\{\mathrm{w}_i\}_{i\in[\![1,M]\!]}$. Le prototype le plus proche de $\mathrm{x}(t)$ au sens de la distance considérée (par exemple, la distance euclidienne), $\mathrm{w}_\nu(t)$ avec $\nu \in [\![1, M]\!]$, est nommé vainqueur (Best Matching Unit) et l'algorithme passe à l'étape d'après.
3. **Coopération** : L'algorithme évalue le voisinage du neurone gagnant dans l'espace neuronal. Ce voisinage représente les neurones excités par x. On définit une fonction de voisinage telle que si un neurone appartient au voisinage du neurone

vainqueur, la fonction vaut 1 et 0 sinon. En pratique, la voisinage représente toute la grille et la fonction de voisinage a une forme gaussienne [3].

4. **Adaptation** : L'algorithme met à jour la valeur du prototype associée au neurone vainqueur ainsi que celle de ses voisins pour se rapprocher de l'entrée présentée. La grille de prototypes est alors déformée. On a $\mathrm{w}_k(t+1) = \mathrm{w}_k(t) + \Delta\mathrm{w}_k(t)$ avec $\Delta\mathrm{w}_k(t) = \alpha(t)\eta(\nu,k,t)[\mathrm{x} - \mathrm{w}_\nu(t)]$ avec $\alpha(t)$ une fonction de t appelée taux d'apprentissage (*learning rate*). $\alpha(\cdot)$ est souvent sous la forme $\alpha(t) = \alpha_0 \exp\left(-\frac{t}{\lambda}\right)$ avec λ le nombre d'itérations désiré.
5. Tant que la carte n'a pas convergé (c'est-à-dire que la grille de prototypes recouvre mal l'espace des données), retour à l'étape 2.

En pratique, le critère d'arrêt est souvent le nombre de fois que l'on présente la base de données complète à la grille de prototypes. La valeur par défaut du paquet **R** utilisé par la suite est de 100 présentations de la base complète.

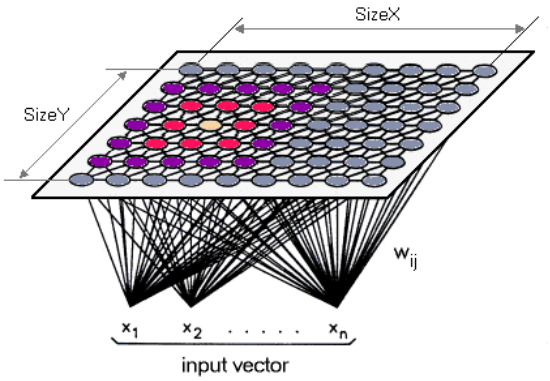


FIGURE 3.5 – Attribution du **Best Matching Unit** associé à une entrée donnée et définition de son voisinage (Source : pitt.edu)

Il existe plusieurs variantes mais toutes reposent sur la même idée de compétition entre les neurones. Afin de modéliser en plus une notion appelée *inhibition latérale*, on utilise une forme modifiée pour $\Delta\mathrm{w}_k(t)$. L'inhibition latérale réduit l'activité des voisins des neurones excités. Dans l'œil, cela permet d'augmenter le contraste et améliore ainsi la perception sensorielle. Un capteur rétinien excité réduira donc l'activité de ses voisins de telle sorte que la partie de l'image perçue par ce capteur sera moins floue. L'adaptation devient alors, pour $\beta(\cdot)$ une fonction donnée :

$$\Delta\mathrm{w}_k(t) = \alpha(t)\eta(\nu,k,t)[\mathrm{x} - \mathrm{w}_\nu(t)] + \beta(t)[\mathrm{w}_\nu(t) - \mathrm{w}_k(t)]$$

3.1.3 Performances d'un Self-Organizing Map

Intuitivement, on voit que l'algorithme a tourné avec succès si la topologie de la grille de prototypes s'est adaptée à la distribution de l'espace d'entrée.

3. Soit r_ν la position sur la grille neuronale du neurone vainqueur et r_k la position d'un neurone donné indexé par k. Alors la fonction de voisinage à l'itération t a la forme $\eta(\nu,k,t) = \exp\left(\frac{\|\mathrm{r}_\nu - \mathrm{r}_k\|^2}{2\sigma(t)^2}\right)$ avec $\sigma(\cdot)$ une fonction décroissante de t qui représente l'ordre de grandeur de la valeur du voisinage d'un neurone.

L'aptitude d'un SOM à remplir sa fonction est déterminée par plusieurs paramètres qui sont réglés empiriquement. Il est courant parmi les algorithmes de machine learning de devoir adapter des paramètres empiriquement. Certains algorithmes ont d'ailleurs des bonnes performances qui ne sont pas justifiées mathématiquement. C'est par exemple le cas des forêts aléatoires.

— **le taux d'apprentissage** : Il détermine en partie les performances du SOM. Kohonen recommande (voir KOHONEN T. (1998)) d'utiliser une fonction de la forme $t \mapsto \frac{A}{t+B}$ avec A et B des constantes. Généralement, une forme de décroissance exponentielle est utilisée. Une forme décroissante permet un apprentissage important dans la phase d'adaptation (lorsque la grille s'adapte aux données) et un faible apprentissage en phase de convergence.

— **la fonction de voisinage** : Elle joue aussi sur les performances. Kohonen précise que cette fonction doit avoir une grande valeur dans les premières itérations, de telle sorte qu'un voisinage recouvre au début au moins 50% de l'espace d'arrivée (les neurones). Cela permet, selon lui, d'éviter l'effet non souhaité d'une convergence vers des configurations méta-stables. En pratique, il faudra chercher à recouvrir $^2/_3$ de l'espace d'arrivée dès la première itération.

— **le nombre d'itérations** : Selon Kohonen, il y a généralement convergence pour un nombre d'itérations de l'ordre de 1000 plus 500 fois le nombre de neurones mais cela dépend des données.

— **le nombre de dimensions de l'espace d'arrivée** : Généralement, cet espace est en 2 dimensions mais le concept de SOM est généralisable à un plus grand nombre de dimensions. Il est aussi possible d'avoir un espace d'arrivée en forme de tore pour les neurones aux extrémités des cartes soient reliés entre eux.

— **la configuration spatiale des neurones** : Elle peut être hexagonale ou rectangulaire dans la majorité des cas. Une configuration hexagonale permet d'avoir 6 voisins immédiats.

— **le nombre de neurones** : La pratique [4] détermine empiriquement de régler le nombre de neurones à $5 \times n^{0.54321}$ où n est le nombre de vecteurs de données.

L'évolution de la convergence de l'algorithme grâce à la mesure **Mean Distance to Closest Unit** [5] qui est la moyenne des distances des données à l'ensemble des vecteurs de poids (voir Manuel du package **R** kohonen, WEHRENS R. and BUYDENS L. (2015)) est affichée en figure 3.6. Il est compréhensible qu'une telle mesure soit minimale quand la grille des prototypes est bien adaptée aux données car il existera toujours un vecteur de poids proche d'un vecteur de données. On note $\Lambda(t)$ cette mesure, avec t l'itération en cours [6] :

$$\Lambda(t) = \frac{1}{n} \sum_{i=1}^{n} \min_{k \in [\![1, M[\![} \mathrm{d}\left(\mathrm{x}^{(i)}, \mathrm{w}_k\right)$$

où $\mathrm{d}(\cdot, \cdot)$ est une distance dans l'espace d'entrée (généralement la distance euclidienne).

4. Voir http ://www.cis.hut.fi/somtoolbox/package/docs2/som_make.html
5. Aussi appelée **Average Quantization Error** ou erreur moyenne de quantification.
6. Dans ce cas, une itération signifie une présentation de la base complète à la grille de prototype.

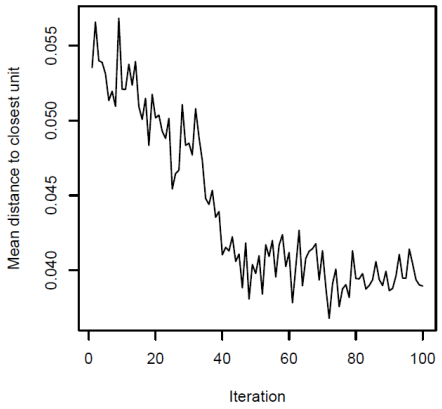


FIGURE 3.6 – Convergence en terme de **Mean Distance to Closest Unit** pour un exemple d'apprentissage

La **Mean Distance to Closest Unit** est une mesure de performance relative. Elle permet de déterminer le nombre d'itérations de l'algorithme nécessaires pour obtenir une convergence. Augmenter le nombre d'itérations plus loin que cette limite ne permettra pas d'améliorer $\Lambda(t)$. Mais l'algorithme SOM peut converger vers un état méta-stable[7]. Il est donc nécessaire de se munir d'une autre mesure de performance pour estimer la qualité d'un SOM.

L'erreur topographique (ou *topographic error*) est la plus simple des mesures de préservation de la topologie[8] (voir PÖLZLBAUER G. (2004)). Soit u la fonction telle que $u(\mathrm{x}) = 0$ si le prototype vainqueur associé à x et le second prototype le plus proche de x dans l'espace d'entrée sont associés à des neurones voisins dans la carte neuronale. On a $u(\mathrm{x}) = 1$ sinon. On définit alors l'erreur topographique $\tau(t)$ par :

$$\tau(t) = \frac{1}{n} \sum_{i=1}^{n} u(\mathrm{x}_i)$$

$\tau(t)$ est compris entre 0 et 1 et mesure le fait que des voisins dans l'espace d'entrée sont aussi des voisins dans la carte neuronale. La qualité du SOM est d'autant plus grande que $\tau(t)$ est proche de 0.

3.1.4 Clustering d'une Self-Organizing Map

Outre le fait de permettre d'analyser graphiquement les relations entre les variables, une carte auto-adaptative permet aussi de réaliser un clustering des données en entrée. Cela permet de réduire les données à plusieurs groupes homogènes. Il est pour cela nécessaire d'appliquer un algorithme de clustering en plus de l'algorithme SOM. Le mémoire détaillera la classification ascendante hiérarchique mais l'algorithme K-means qui a donné de meilleures performances lors des tests préliminaires sera utilisé (cf. annexe 4.5). Une segmentation grâce aux SOM permet de s'affranchir des valeurs aberrantes dans les données, ce qui n'est pas le cas avec un algorithme K-means.

7. Par exemple si les données forment 2 nuages de points et que l'algorithme converge en n'en recouvrant qu'un seul.

8. Par préservation de la topologie, on entend que la structure de l'espace d'entrée se retrouve dans la carte des neurones.

Avant de réaliser le clustering par classification ascendant hiérarchique, il est nécessaire de définir plusieurs notions.

U-matrix La U-matrix (ou U-matrice) est une carte des distances entre les différents neurones. La distance moyenne à ses voisins est calculée pour chaque neurone de la carte et est affichée, généralement grâce à des niveaux de couleur en fonction de la valeur. Pour un neurone donné i, la mesure de distance associée (U-height dans la littérature, voir ULTSCH A. (2003)) est, en notant $\mathcal{V}(i)$ la voisinage du neurone i :

$$\mathrm{U}(i) = \frac{1}{|\mathcal{V}(i)|} \sum_{k \in \mathcal{V}(i)} \mathrm{d}\left(\mathrm{w}_i, \mathrm{w}_k\right)$$

Le lecteur trouvera en figure 3.7 la représentation d'une U-matrice. Les neurones gris ne sont associés à aucun vecteur de l'espace d'entrée.

(a) Nombre de vecteurs d'entrée par neurone (b) U-matrix

FIGURE 3.7 – Aspect des représentations utiles au clustering

En regardant la U-matrice, il est possible de voir que la zone en bas à gauche est éloignée des autres neurones. C'est à priori un cluster.

Ayant une notion de distance entre les neurones définie comme la distance entre les vecteurs de poids associés, il est possible d'appliquer une classification ascendante hiérarchique.

Classification ascendante hiérarchique La classification ascendante hiérarchique est une technique de classification non supervisée. Elle vise à créer une hiérarchie à partir d'un ensemble de vecteurs $\mathcal{X} = \left\{\mathrm{x}^{(k)}\right\}_{k \in [\![1,n]\!]}$ avec $n \in \mathbb{N}^*$.

Pour cela, l'algorithme cherche à regrouper les vecteurs les plus proches dans des sous-groupes puis à rassembler ces sous-groupes dans des groupes plus importants jusqu'à obtenir un groupe contenant tous les éléments.

On appelle hiérarchie de $\mathcal{X}$ un ensemble de parties telles que :

1. La partie vide en fait partie
2. Les parties réduites à un seul élément en font partie.
3. L'ensemble $\mathcal{X}$ en fait partie.
4. Si $\mathcal{A}$ et $\mathcal{B}$ en font partie, alors soit $\mathcal{A}$ et $\mathcal{B}$ sont disjointes, soit $\mathcal{A}$ contient $\mathcal{B}$, soit $\mathcal{B}$ contient $\mathcal{A}$.

Par exemple, si $\mathcal{X} = \{a, b, c, d, e\}$ alors un exemple de hiérarchie de $\mathcal{X}$ est :

$$\{\{a\}, \{b\}, \{c\}, \{d\}, \{e\}, \{a, b\}, \{a, b, c\}, \{d, e\}, \{a, b, c, d, e\}\}$$

La visualisation de l'arbre associé à l'exemple de hiérarchie précédent est disponible en figure 3.8. Cet arbre est appelé dendrogramme.

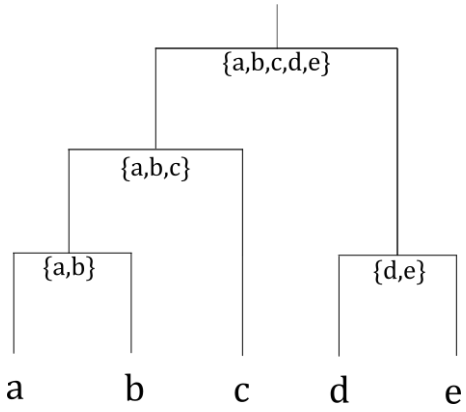


FIGURE 3.8 – Exemple de hiérarchie de $\mathcal{X} = \{a, b, c, d, e\}$

Pour appliquer la classification ascendante hiérarchique à $\mathcal{X}$, classe d'éléments de $\mathbb{R}^p$ [9], il faut d'abord définir deux distances :

Distance entre deux éléments Une distance entre les éléments composants la classe. Soient $\mathrm{x}^{(i)} = \left(x_1^{(i)}, \cdots, x_p^{(i)}\right)$ et $\mathrm{x}^{(j)}$ deux éléments d'une classe $\mathcal{P} \subset \mathcal{X}$. On définit généralement $\mathrm{d}(\cdot, \cdot)$ comme étant la distance euclidienne entre $\mathrm{x}^{(i)}$ et $\mathrm{x}^{(j)}$.

$$\mathrm{d}(\mathrm{x}^{(i)}, \mathrm{x}^{(j)}) = \sqrt{\sum_{k=1}^{p} \left(x_k^{(i)} - x_k^{(j)}\right)^2}$$

Distance inter-classes Une distance entre deux classes est nécessaire. Il existe plusieurs définitions possibles pour la distance inter-classes. Les 3 plus courantes ne sont pas des distances au sens strict (car elles n'en vérifient pas tous les axiomes) mais des pseudo-distances. Le lecteur remarquera que la distance d'un élément à une classe n'est autre que la distance du singleton contenant l'élément à la classe en question.

En notant $\mathcal{U}$ et $\mathcal{V}$ des classes contenues dans $\mathcal{X}$, on a :

1. **Single linkage** : $\mathrm{d}_c(\mathcal{U}, \mathcal{V}) = \inf_{\mathrm{x} \in \mathcal{U}, \mathrm{y} \in \mathcal{V}} \mathrm{d}(\mathrm{x}, \mathrm{y})$
2. **Complete linkage** : $\mathrm{d}_c(\mathcal{U}, \mathcal{V}) = \sup_{\mathrm{x} \in \mathcal{U}, \mathrm{y} \in \mathcal{V}} \mathrm{d}(\mathrm{x}, \mathrm{y})$
3. **Average linkage** : $\mathrm{d}_c(\mathcal{U}, \mathcal{V}) = \frac{1}{|\mathcal{U}| \cdot |\mathcal{V}|} \sum_{\mathrm{x} \in \mathcal{U}} \sum_{\mathrm{y} \in \mathcal{V}} \mathrm{d}(\mathrm{x}, \mathrm{y})$

Algorithme Détaillons l'algorithme de classification ascendante hiérarchique :

9. Avec p un entier non nul.

Étape initiale On forme l'ensemble $\pi(0) = \left\{\{\mathrm{x}^{(1)}\}, \cdots, \{\mathrm{x}^{(n)}\}\right\} = \{\mathrm{A}_1, \cdots, \mathrm{A}_n\}$.

Agglomération : Connaissant à l'étape t les $n(t)$ clusters $\pi(t)$ avec $n(1) = n$ et $n(t+1) = n(t) - 1$, on cherche (k^*, l^*) tels que :

$$(k^*, l^*) = \mathrm{argmin}_{k \neq l}\, \mathrm{d}_c\left(A_k, A_l\right)$$

et on fusionne en $\mathrm{A}^* = \mathrm{A}_{k^*} \cup \mathrm{A}_{l^*}$. Cela forme $\pi(t+1) = [\{\pi(t) \backslash \mathrm{A}_{k^*}\} \backslash \mathrm{A}_{l^*}] \cup \mathrm{A}^*$.

Condition d'arrêt Tant que $\pi(t)$ n'est pas égal à $\mathcal{X}$, $t = t+1$ et retour à l'étape d'agglomération.

Nombre de clusters optimal L'algorithme de classification ascendante hiérarchique crée par défaut autant de clusters que d'éléments de $\mathcal{X}$. Il est nécessaire de réaliser un élagage du dendrogramme pour diminuer le nombre d'étages de l'arbre. Ceci permet de regrouper certaines feuilles de l'arbre afin d'avoir un dendrogramme plus simple à comprendre et d'obtenir un nombre limité de clusters.

Il existe plusieurs méthodes pour estimer le nombre optimal de clusters pour une classification ascendante hiérarchique. Une de ces méthodes est exposée ci-après.

Définissons alors les notions d'inertie intra et inter-classes. On a, si la partition π de $\mathcal{X}$ a N parties $\mathcal{E}_i$ de centres respectifs [10] c_i :

$$\mathrm{I}_{\mathrm{intra}} = \sum_{i=1}^{N} \sum_{\mathrm{x} \in \mathcal{E}_i} \mathrm{d}\left(\mathrm{x}, c_i\right)^2$$

$$\mathrm{I}_{\mathrm{inter}} = \sum_{i=1}^{N} |\mathcal{E}_i| \cdot \mathrm{d}\left(c_i, C\right)^2$$

avec C le vecteur moyenne de $\mathcal{X}$.

$\mathrm{I} = \mathrm{I}_{\mathrm{intra}} + \mathrm{I}_{\mathrm{inter}}$ est l'inertie totale du nuage de points.

On note $R^2 = \frac{\mathrm{I}_{\mathrm{inter}}}{\mathrm{I}}$ la proportion de la variance expliquée par les classes.

Le nombre de clusters optimal est celui pour lequel la courbe de R^2 en fonction du nombre de clusters forme un coude.

10. Cette notion est à définir clairement. On peut considérer que le centre de $\mathcal{E}_i$ est le vecteur égal à la moyenne des éléments de $\mathcal{E}_i$.

Application aux Self-Organizing Maps La classification ascendante hiérarchique (CAH) permet de créer des clusters de neurones sur la carte auto-adaptative (voir VESANTO J. and ALHONIEMI E. (2000)). Pour cela, on applique la classification ascendante hiérarchique aux vecteurs de poids $\{w_i\}_{i\in[\![1,M]\!]}$ qui représentent les neurones. Des classes obtenues parmi les vecteurs de poids sont déduites les classes pour les neurones. Après une phase d'élagage des clusters, on obtient ce type de cartes (voir figure 3.9) qui résume les données à différentes classes. Les cartes de niveau obtenues dans la première phase de mise en place des SOMs permettent d'expliquer à quels types de profils correspondent ces classes.

Il est possible de voir l'algorithme SOM comme une méthode de clustering car elle associe les vecteurs de l'espace d'entrée à un faible nombre de neurones. Les proto-clusters obtenus sont alors regroupés à l'aide d'une classification ascendante hiérarchique. Cela permet de réduire le temps de calcul afin d'obtenir les clusters finaux par rapport à une classification ascendante hiérarchique appliquée directement aux vecteurs de l'espace d'entrée (voir VESANTO J. and ALHONIEMI E. (2000)).

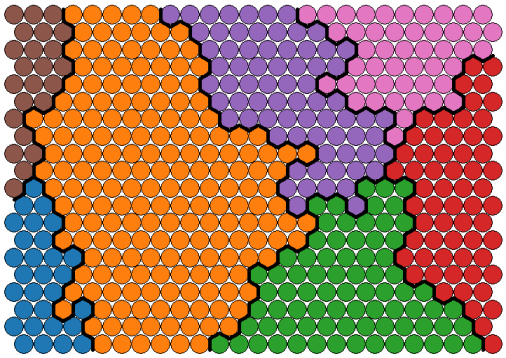


FIGURE 3.9 – Clusters sur une SOM (Source : Deloitte Analytics)

3.1.5 Avantages et limites des cartes auto-adaptatives

Avantages Les cartes auto-adaptatives présentent plusieurs avantages intéressants :

- Tout d'abord, ils permettent de visualiser des données de grande dimension dans un plan en 2 dimensions.
- Leur principe de fonctionnement est plutôt simple. Après implémentation, il est plutôt aisé d'exploiter ces cartes et d'en extraire des informations.
- Les SOM sont un algorithme non linéaire contrairement à l'Analyse en Composantes Principales qui suppose que les données forment un nuage de points gaussien.
- Les SOM permettent aussi de prédire dans quel cluster se trouvera un nouveau vecteur de données.
- C'est un algorithme très robuste. Bien que dépendant de nombreux paramètres (le taux d'apprentissage et la fonction de voisinage notamment), un réglage approximatif des paramètres aboutit souvent à un résultat très satisfaisant.

Limites Les SOM présentent plusieurs limites :

— L'algorithme peut demander un temps de calcul important.

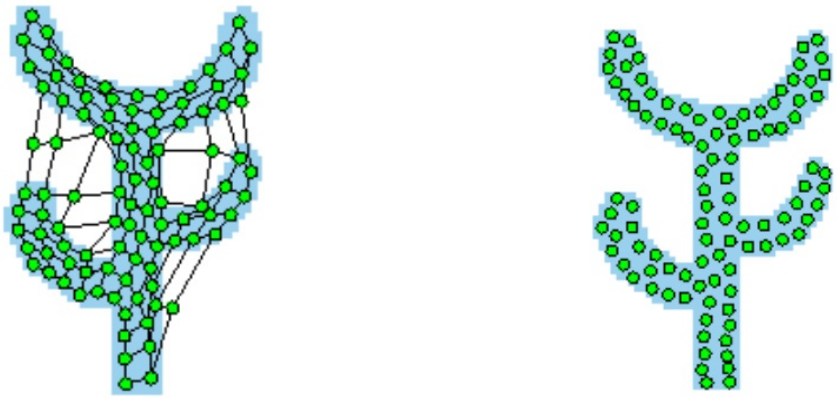
— Les données en entrée doivent être numériques et ne pas présenter de valeur manquante car l'algorithme ne sait pas les gérer.

— Il y aura autant de cartes représentant la distribution des variables que de variables. L'exploitation des cartes pour trouver des informations peut donc être longue.

— Il peut parfois être difficile de représenter des données ayant une grande dimension dans un plan en 2 dimensions à cause de la perte d'informations induite.

— Le résultat de l'algorithme peut être très différent si on l'exécute 2 fois à la suite car il dépend du hasard (tirage aléatoire des vecteurs d'entrée). De plus, à cause de ce hasard, les prototypes peuvent ne pas se dérouler correctement dans l'espace d'entrée.

Pour des formes de nuage de points particulières, il est parfois pratique de vouloir casser des liaisons entre des neurones. En effet, prenons par exemple le nuage de points suivant [11] :



(a) Self Organizing Map

(b) Neural Gas

FIGURE 3.10 – Différence entre les résultats des SOM et Neural Gas

Cet exemple illustre que la structure rigide d'un SOM n'est pas adaptée à tous les types de données. Afin d'obtenir une quantification vectorielle plus fidèle à la structure des données, il est souhaitable de casser les liaisons entre les prototypes se trouvant sur 2 branches du « cactus ». L'algorithme « *Gaz neuronal* » [12] et son dérivé le « *Growing Neural Gas* » permettent cela et sont donc une extension plus robuste des cartes auto-adaptatives.

11. Voir http ://www.demogng.de/JavaPaper/node22.html
12. En anglais, *Neural Gas*

3.2 Un exemple simple pour mieux appréhender les SOMs

L'objectif est de prouver l'intérêt des Self-Organizing Maps pour l'analyse des données, tant au niveau de la visualisation des variables que pour la segmentation.

Les cartes auto-organisatrices, pour le type d'algorithme décrit dans le mémoire, sont uniquement capables de traiter des variables numériques mais pas des catégories. Cependant, certaines variables de la base sont catégoriques. Pour transformer une variable catégorique en variable numérique, il existe une astuce : les *dummy variables*.

Soit une variable $\mathbf{X}$ catégorique, à P modalités $\{1, \cdots, P\}$. Pour transformer $\mathbf{X}$ en variable numérique, l'astuce revient alors à remplacer $\mathbf{X}$ par $P-1$ variables $\mathbf{X}m^1, \cdots, \mathbf{X}m^{P-1}$ telles que :

$$\forall i \in \{1, \cdots, P\}, \mathbf{X}m^i = \mathbb{I}(\mathbf{X} = i)$$

De plus, si toutes ces variables sont nulles, c'est que $\mathbf{X}$ est égal à P. Connaître les *dummy variables* est donc équivalent à la connaissance de la variable $\mathbf{X}$.

Pour utiliser les SOMs, il est donc nécessaire de transformer en *dummy variables* les variables catégoriques.

Cela pose un problème pour la variable **Marque du véhicule** car le portefeuille compte un grand nombre de marques de voiture, allant des plus courantes et représentées par la majorité des contrats dans la base aux marques rares représentées par une très faible proportion de voitures. Il est exclu de créer autant de *dummy variables* en lieu et place de la marque de voiture et il est donc nécessaire de créer des groupes de marques de voiture.

3.2.1 Calibration des paramètres de la carte auto-adaptative

Une des difficultés inhérentes à la prise en main des cartes auto-adaptatives est la compréhension de la façon dont les différents paramètres doivent être réglés de façon à obtenir un résultat exploitable. Comme précisé, un réglage approximatif des paramètres peut aboutir à des résultats très satisfaisants.

Cependant, il est indispensable de comprendre la façon dont une carte auto-adaptative doit être calibrée.

La fonction de voisinage est réglée de façon à ce que le voisinage d'un neurone contienne au moins $^2/_3$ des neurones de la grille à la première itération. Ensuite, elle décroit linéairement jusqu'à 1 pendant le premier tiers des itérations. Pendant le reste des itérations, seul le neurone gagnant est donc ensuite mis à jour.

Le paramètre le plus important des SOMs est le **taux d'apprentissage**. Ce dernier est choisi pour décroître linéairement d'une valeur α à une valeur β, telle que la carte converge. Par défaut dans le paquet **R** utilisé, α vaut 0,05 et β 0,01.

Testons tout d'abord l'impact du réglage du couple (α, β) sur des données *jouets*.

Soit le nuage de points en deux dimensions suivant :

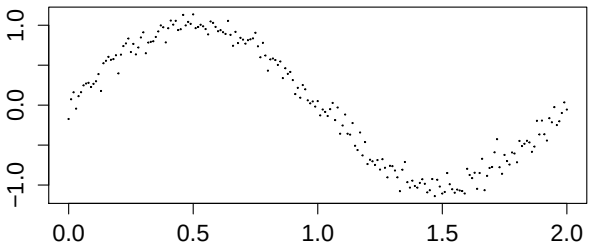


FIGURE 3.11 – Nuage de points en deux dimensions

Un SOM avec $(\alpha, \beta) = (0.05, 0.01)$, 500 présentations complètes des données à l'algorithme, et de taille 10×1 neurones, donc une grille à une dimension, donne le résultat suivant. Ce sont les vecteurs prototypes qui sont représentés en bleu dans la figure 3.12

Le SOM réalise donc, sur cet exemple simple sa tâche consistant à reproduire la « forme » des données.

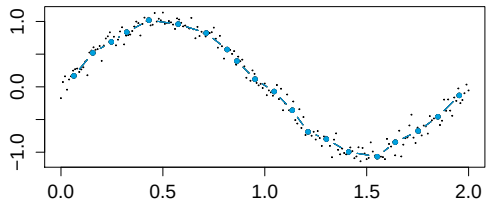
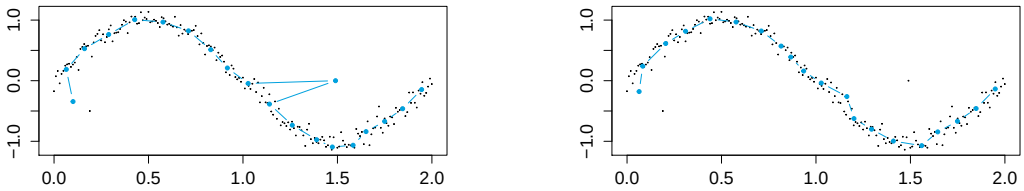


FIGURE 3.12 – Vecteurs prototypes d'une grille SOM à une dimension recouvrant les données

Voyons en figure 3.13 l'effet des paramètres du taux d'apprentissage quand des valeurs aberrantes sont ajoutées dans les données.



(a) $(\alpha, \beta) = (0.1, 0.01)$ (b) $(\alpha, \beta) = (0.01, 0.001)$

FIGURE 3.13 – Résultats des SOMs en fonction de (α, β)

Un taux d'apprentissage trop grand a tendance à déplacer la grille de prototypes vers les valeurs aberrantes. Or, le SOM est justement censé ne pas en tenir compte. On voit donc qu'il est nécessaire de bien régler le taux d'apprentissage pour éviter ce phénomène.

Ces exemples sur des cas élémentaires permettent d'appréhender les limites des SOMs ainsi que l'influence des différents paramètres. La compréhension des paramètres influant sur un algorithme de Machine Learning peut, de la même façon, être acquise en testant l'algorithme sur des données très simples.

3.2.2 Obtention de clusters pour les marques de voiture

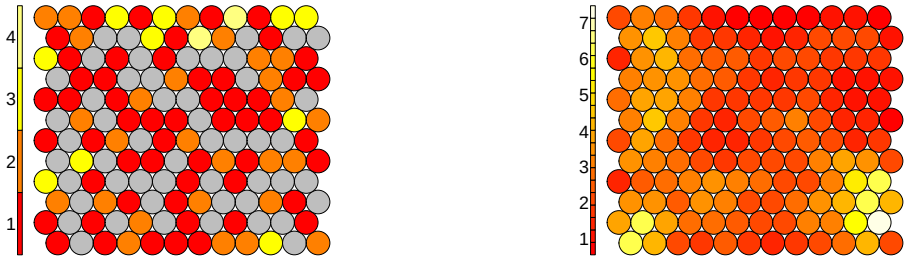
Afin de réduire le nombre de *dummy variables* créées, la variable **Marque du véhicule** sera résumée à l'aide d'une carte auto-adaptative.

Pour cela, la base de données des contrats est agrégée par marque et les variables suivantes sont utilisées :

— Puissance DIN médiane par marque,
— Effectif en nombre de voitures en portefeuille par marque,
— Proportion d'hommes parmi les possesseurs de la marque dans le portefeuille.

Les effectifs et les puissances calculées possèdent des densités très étendues. Par conséquent, le rendu des couleurs sur la carte sera non satisfaisant. On applique donc des transformations d'échelle à ces deux variables. Un exemple de transformation d'échelle est de prendre le logarithme de l'effectif comme variable à la place de l'effectif.

Après plusieurs tests, on choisit $(\alpha, \beta) = (0.05, 0.01)$ et 400 itérations.



(a) Nombre de vecteurs d'entrée par neurone (b) U-matrix

FIGURE 3.14 – Qualité de l'exécution de la carte auto-adaptative

Dans la figure 3.14 (b), on détecte un cluster en bas à droite. La carte de nombre de vecteurs par neurones montre des zones non associées à des neurones, ce qui est problématique et peut être dû à une carte trop grande pour le nombre de vecteurs dans l'espace d'entrée. Cependant, pour plus de lisibilité, on préfère surestimer que sous-estimer la taille optimale de la grille.

L'erreur topographique est de 0.36984 et l'erreur de quantification semble atteindre un plateau indiquant que la carte a convergé.

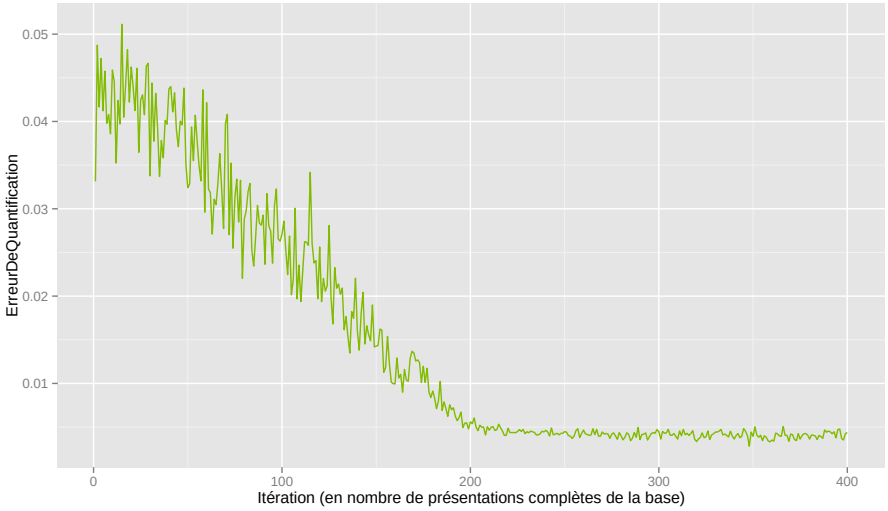


FIGURE 3.15 – Erreur de quantification

Un clustering K-means est ensuite appliqué aux vecteurs prototypes et la proportion de variance expliquée par les classes en fonction du nombre de clusters est affichée en figure 3.16. Le coude semble survenir pour des valeurs de 5 ou 6 clusters. On choisit un nombre de 5 clusters même si le nombre optimal semble être de 6 ou 7 d'après la figure 3.16 car ces clusters vont être transformés en *dummy variables* et on préfère en limiter le nombre.

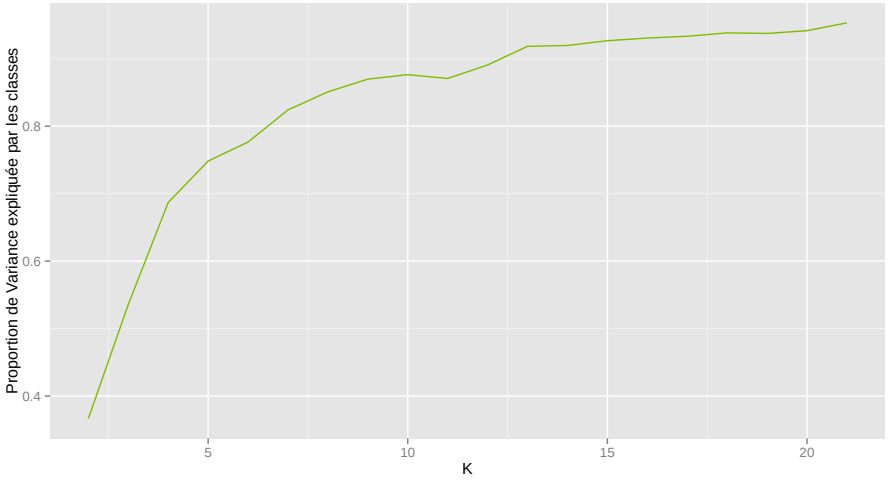
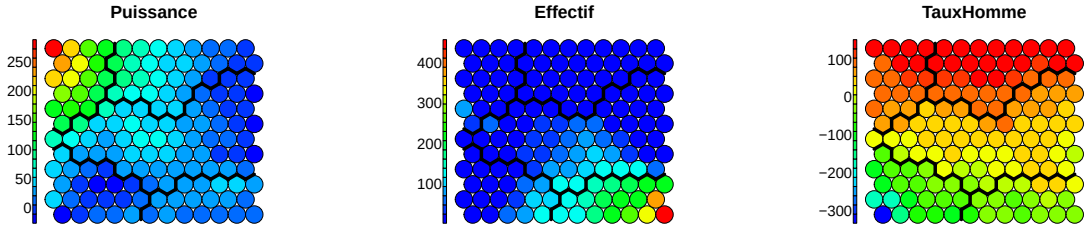


FIGURE 3.16 – Détermination du nombre optimal de clusters

Cela donne les résultats suivants :



(a) Puissance (b) Effectif (c) Proportion d'hommes

FIGURE 3.17 – Distribution des variables sur la carte auto-adaptative

Les frontières des clusters sont affichées en noir sur les figures de distribution des variables. On voit qu'il existe un cluster constitué de voitures très courantes en bas à droite. Les clusters en haut à droite et celui en haut à gauche semblent avoir une proportion d'hommes supérieure à la moyenne. De plus, le cluster en haut à gauche semble être constitué de voitures très puissantes. Il y aurait donc une corrélation entre puissance de la voiture et sexe de l'assuré qui peut être vérifiée par un test classique de corrélation. Enfin, on constate que la proportion d'hommes, sur laquelle aucune transformation d'échelle n'a été effectué, a une distribution de variables entre -300% et +100%, ce qui n'a pas de sens pour les valeurs négatives. Cela signifie que la carte s'est déroulée en dehors des limites des données. Cependant, il est possible d'interpréter une valeur négative comme une proportion d'hommes faible.

Enfin, le résultat du clustering de la carte auto-adaptative permet de tracer en figure 3.18 les 5 clusters de marques de voiture.

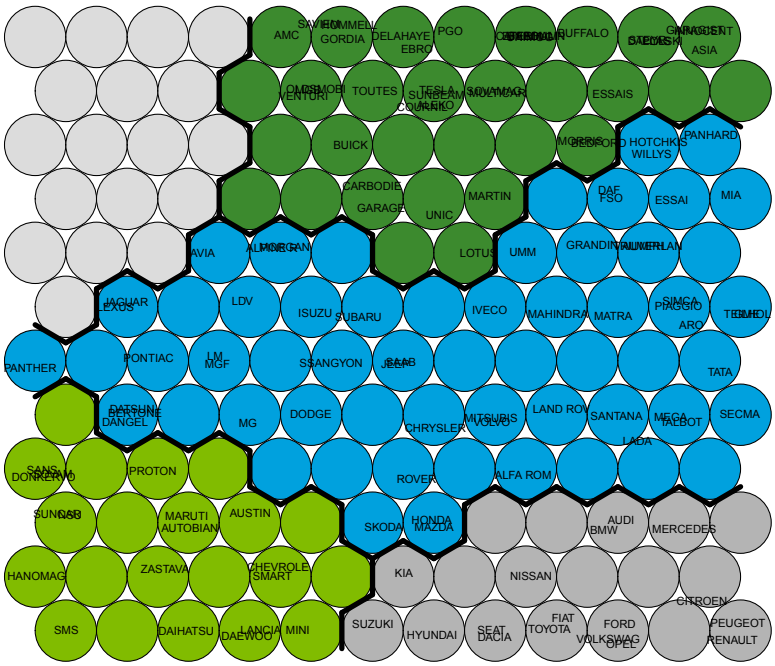


FIGURE 3.18 – Clusters de marques de voiture (cluster gris clair anonymisé)

Le cluster gris en bas à droite représente les voitures très courantes comme Peugeot et Renault. De plus, ces voitures ont une puissance médiane assez faible. Les voitures du cluster vert en bas à gauche sont les voitures un peu moins courantes comme les Mini ou les Lancia. Les clusters bleu et vert foncé sont les voitures très rares. Enfin, le cluster gris en haut à gauche représente les voitures de luxe comme Ferrari et ont une proportion d'hommes plus élevée que la moyenne. Pour des raisons de confidentialité, les marques du cluster gris clair ne sont pas affichées mais ce dernier était facilement explicable. La présence de certaines marques est difficile à justifier comme les voitures Alfa Romeo dans le cluster des marques rares. Cependant, les résultats sont plutôt interprétables malgré des difficultés d'interprétation ponctuelles.

3.3 Application aux données du portefeuille

L'objectif de cette section est d'appliquer les cartes auto-adaptatives aux données du portefeuille. L'apport de l'utilisation des SOMs est de réaliser une segmentation du portefeuille mais aussi de décrire l'interaction entre les différentes variables.

L'application des cartes auto-adaptatives à des données réelles pose plusieurs problèmes.

Tout d'abord, les données de la base présentent des variables catégoriques qu'il faut transformer en *dummy variables*.

De plus, un des intérêts des cartes auto-adaptatives étant la visualisation de la distribution des variables, il est nécessaire de réaliser des retraitements pour les variables comportant des valeurs très éloignées de la moyenne afin de ne pas avoir qu'une seule couleur sur la carte. En effet, si une variable positive a 99% de ses données entre 0 et 1 et 1% supérieur à 10, la carte de cette variable sera presque uniformément bleue (avec les conventions de couleur adoptées).

Enfin, le choix des différents paramètres de la carte auto-adaptative n'est pas aisé. En effet, avec environ 2 000 000 de vecteurs de données, le temps d'exécution peut être extrêmement long et il n'est pas possible de jouer avec les paramètres afin de trouver ceux qui vont donner les meilleurs résultats.

3.3.1 Optimisation de la carte auto-adaptative

Un retraitement des données s'impose avant d'exécuter l'algorithme SOM. Les données réelles, notamment en assurance, comportent des valeurs très éloignées de la moyenne (par exemple, les charges de sinistres). Afin d'améliorer la qualité du SOM, un retraitement s'impose.

3.3.1.1 Préparation des données

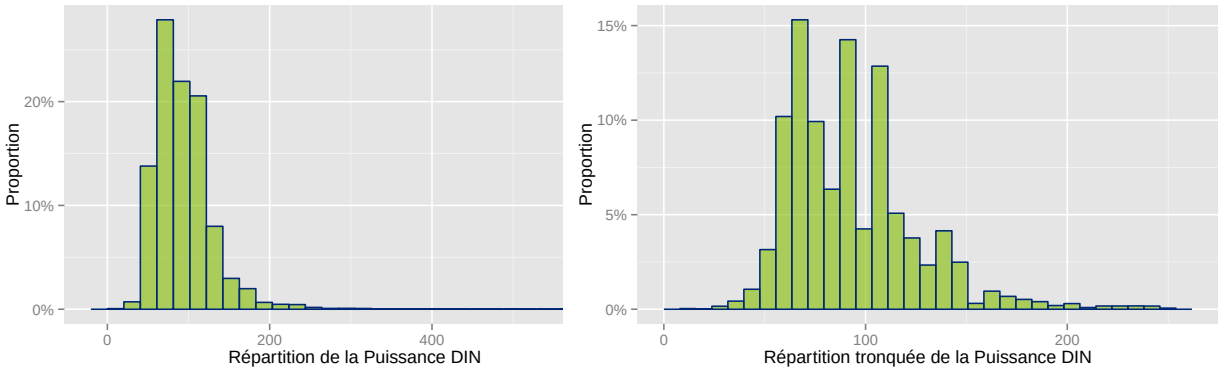
Il est nécessaire de retraiter au préalable les variables dont la représentation pourrait ne pas être adaptée à la visualisation. C'est le cas de la puissance DIN du véhicule. Cette dernière a une moyenne de 94,26 chevaux DIN pour une médiane de 90 chevaux DIN et un écart-type de 35 chevaux DIN. Cependant, cette variable a un quantile à 99% de 340 chevaux DIN, ce qui aura tendance à polluer la visualisation de la carte auto-adaptative pour cette variable. Deux solutions sont possibles pour corriger ce problème :

— Tronquer la distribution en supprimant les lignes trop éloignées de la moyenne en supprimant les lignes pour lesquelles la puissance DIN est supérieure à un quantile donné, par exemple celui à 99%.
— Appliquer une transformation d'échelle à la variable afin de conserver toutes les lignes.

Ces deux possibilités introduisent des biais. La première nécessite de supprimer des lignes. Dans le cas précédent, cela revient à supprimer les lignes associées à des voitures très puissantes. Or analyser ce segment peut potentiellement être intéressant et représente plusieurs dizaines de milliers de contrats. De plus, la suppression de ces lignes peut introduire un biais dans l'analyse.

L'autre solution est d'appliquer une transformation d'échelle à la variable en question comme, par exemple, remplacer la variable par son logarithme. Cette option fait cependant perdre de son sens à la visualisation des données.

La distribution de cette variable et de celle-ci après troncature au seuil 250 est affichée en figure 3.19.



(a) Puissance DIN (b) Puissance DIN tronquée

FIGURE 3.19 – Distribution des variables

Cette troncature fait perdre environ 3000 lignes à la base mais améliore la lisibité des cartes produites.

Le bonus-malus présente une distribution similaire avec des valeurs allant jusqu'à 350. Cette variable sera tronquée à 130.

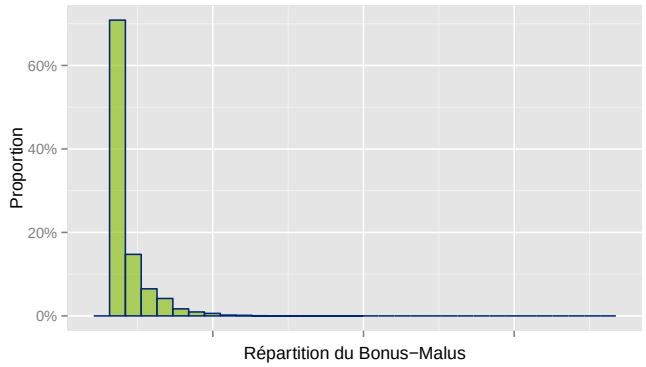


FIGURE 3.20 – Répartition du bonus-malus (échelle des abscisses anonymisée)

Un retraitement est aussi effectué sur la variable **NbConduiteAcc** qui donne le nombre de conducteurs en conduite accompagnée associés au contrat. Sachant que 97% des contrats n'ont pas de conducteur en conduite accompagnée et un peu moins de 3% en ont un seul, on remplace par 1 les valeurs de NbConduiteAcc supérieures à 1.

Les variables **NbSinistres** et **NbContacts** sont exclues du SOM car l'objectif est de réaliser un clustering à priori du portefeuille [13]. Comme ces deux variables sont observées après le 1er janvier 2010, le clustering ne l'aurait pas été.

Le même problème se pose pour les variables **TypeSinistre** et **TypeContact** qui sont aussi exclues du SOM.

La variable **Formule** n'est pas considérée car une des formules représente la grande majorité des contrats donc la variable n'apportera pas d'information.

Les variables géographiques **X** et **Y** ne sont pas conservées dans le clustering pour ne pas séparer deux contrats similaires en tous points mais géographiquement éloignés.

3.3.1.2 Optimisation de la taille de la carte

Le choix du nombre de neurones n'est pas anodin. Il détermine la qualité de l'exécution de la carte auto-adaptative mais aussi le temps d'exécution de l'algorithme.

Plus le nombre de neurones est grand, plus la résolution de la carte sera importante et plus l'algorithme sera à même de fournir des analyses détaillées. Un critère pouvant orienter le choix est le nombre de vecteurs de l'espace d'entrée par neurone dans la carte. Pour 10 000 de vecteurs de l'espace d'entrée et une carte de 10×10, il y aura en moyenne $\frac{10000}{10\times10} = 100$ vecteurs par neurone. Plus le nombre de vecteurs par neurone est important, moins la perte d'information associée à la réduction de la dimension est importante. Plusieurs heuristiques existent pour le nombre optimal de neurones. En notant n le nombre de vecteurs de l'espace d'entrée, les heuristiques sont :

— $5\sqrt{n}$,
— $5 \times n^{0.54321}$,
— 10 fois le nombre de clusters que l'on prédit avant exécution.

Aucune de ces valeurs n'est la formule exacte (ce qui d'ailleurs n'existe pas) et le choix dépendra des données.

Le temps d'exécution étant aussi un facteur déterminant dans la calibration de la taille de la carte, il est aussi nécessaire de réaliser une analyse comparative du temps d'exécution pour une itération en fonction de la dimension de la carte (égale à 50 pour une carte de 50×50 par exemple).

13. C'est-à-dire un clustering en fonction de variables que l'on connaît en début d'année si on observe le contrat du début à la fin d'une année.

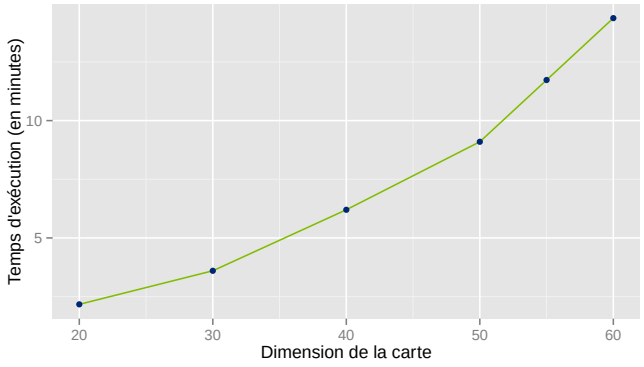


FIGURE 3.21 – Evolution du temps d'exécution en fonction de la dimension de la carte

Le figure 3.21 montre que le temps d'exécution [14] pour une itération croît vite en fonction de la dimension de la carte, passant de 2,17 minutes pour une dimension de 20 à 14,37 minutes pour une dimension de 60. Le nombre d'itérations avant d'aboutir à la convergence étant de l'ordre de 100 itérations, le temps d'exécution total peut être très long. En pratique, il faut vérifier la convergence après exécution de l'algorithme, ce qui peut pousser à surestimer le nombre d'itérations nécessaires et donc à augmenter le temps d'exécution.

14. Les calculs ont été réalisés sur un ordinateur de processeur Intel (R) Core (TM) i5-2520M CPU 2.50 GHz et ayant 4 Gio DDR3 de mémoire vive.

3.3.2 Algorithme SOM

Pour les données du portefeuille, qui représentent plus de 2 000 000 vecteurs dans l'espace d'entrée, le choix est fait d'avoir une carte de 50×50 neurones. Cela permet d'avoir environ 800 vecteurs par neurones. On choisit d'effectuer 50 itérations après test sur un nombre d'itérations plus grand au préalable. On choisit les paramètres par défaut pour le taux d'apprentissage : $(\alpha, \beta) = (0.05, 0.01)$. La fonction de voisinage est choisie pour recouvrir $^2/_3$ de l'espace des neurones à la première présentation des données et décroître linéairement jusqu'à 1 pendant le premier tiers des itérations. La durée d'exécution de l'algorithme a été d'environ 7 heures.

Avant exécution de l'algorithme, les 51 variables utilisées sont normalisées (à l'exception des *dummy variables*). En effet, des distances sont calculées dans l'espace d'entrée et ne pas normaliser les variables aurait donné plus d'importance à certaines variables par rapport aux autres.

L'évolution de l'erreur de quantification est affichée en figure 3.22. Il semble y avoir convergence pour un nombre d'itérations de 25 mais cela peut être dû à la fonction de voisinage. Cela ne signifie cependant pas forcément que le résultat sera correct. En effet, l'erreur topographique vaut 0.8382091. Cela aura tendance à produire un clustering de mauvaise qualité.

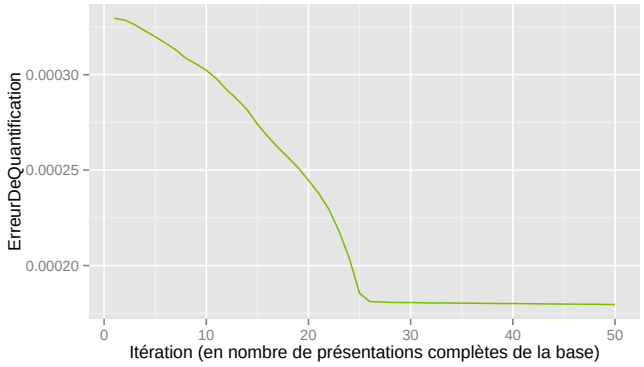
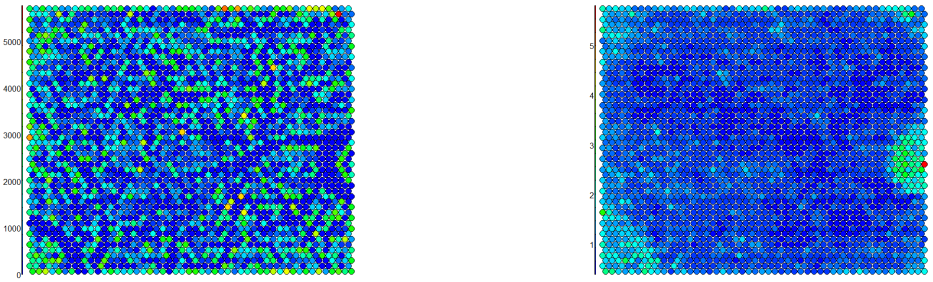


FIGURE 3.22 – Evolution de l'erreur de quantification

Pour améliorer le clustering, qui est l'objectif de cette partie, le choix est fait de poursuivre avec 3 itérations en utilisant comme vecteurs de poids à l'initialisation de l'algorithme les vecteurs de poids en sortie de la carte précédente. De plus, comme la carte a déjà convergé, on choisit un très faible taux d'apprentissage : $(\alpha, \beta) = (0.01, 0.001)$. La nouvelle erreur topographique est de 0.6464172, ce qui constitue une amélioration par rapport à la carte précédente. La qualité de la carte n'est cependant pas aussi bonne que pour le clustering des marques de voiture (0.36984). Cette moindre qualité est en partie due au fait que le nombre de vecteurs de l'espace d'entrée par neurone est élevé pour cette carte. En effet, la heuristique $5 \times n^{0.54321}$ avait été utilisée mais un tel nombre de neurones n'est pas envisageable pour plus de 2 millions de vecteurs dans l'espace d'entrée. En effet, le temps de calcul à nombre d'itérations identique aurait été trop important et la mémoire vive disponible n'aurait pas été suffisante.

La figure 3.23 montre, avec les conventions de couleur un dégradé de bleu à rouge en allant de la plus petite à la plus grande valeur, la qualité de l'exécution du SOM. La répartition du nombre de vecteurs par neurone ne semble pas équi-répartie tandis que la U-matrice semble montrer un cluster au milieu à droite de la carte.



(a) Nombre de vecteurs d'entrée par neurone (b) U-matrix

FIGURE 3.23 – Qualité de l'exécution de la carte auto-adaptative

Le clustering K-means est utilisé. Pour déterminer le nombre optimal de clusters, la heuristique de la proportion de variance expliquée par les classes est utilisée. La figure 3.24 montre, à l'aide du critère du coude, que le nombre optimal de clusters est aux alentours de 12. L'algorithme K-means étant NP-difficile, une approximation de l'algorithme est utilisée. Celle-ci rajoute de l'aléatoire et il est donc nécessaire de fixer la *seed* du générateur aléatoire de **R** afin de fixer un clustering une fois pour toutes. Cette part aléatoire fait que le clustering n'est qu'un clustering parmi d'autres et reste subjectif.

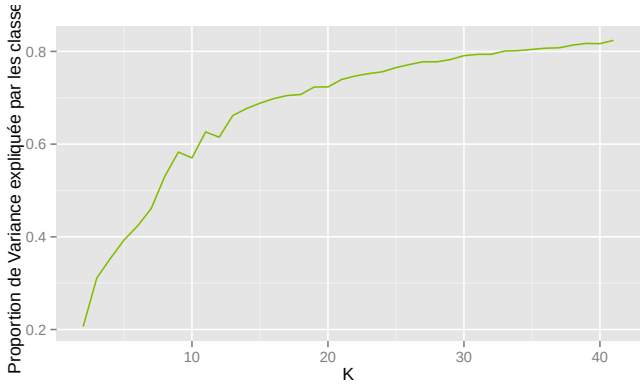


FIGURE 3.24 – Recherche du nombre optimal de clusters

Le rendu du clustering est affiché en figure 3.25. Le clustering semble visuellement plutôt satisfaisant malgré quelques points orphelins au milieu de certains clusters. Les tests préliminaires avec un nombre d'itérations moindre ou sans la ré-exécution du SOM sur 3 itérations donnaient des clusterings beaucoup moins réguliers. Le point notable est que certains clusters, les 1, 7 et 10, ne sont pas contigus. Cela est tout à fait normal car le clustering est réalisé sur les vecteurs de poids qui se sont adaptés aux vecteurs de l'espace d'entrée, peu importe la forme du nuage de points.

Un des apports des cartes auto-adaptatives est de visualiser directement le clustering. En effet, plusieurs exécutions de l'algorithme K-means ont été testées et la visualisation directe des clusters a permis de choisir la partition de l'espace qui semblait la plus cohérente vue la distribution des variables. Cette visualisation directe n'est pas possible en effectuant directement un algorithme de clustering classique comme les K-means ou la classification ascendante hiérarchique car le clustering est réalisé dans un espace à grandes dimensions ($n > 3$). Hormis la visualisation des données dans un plan constitué des 2 directions principales du nuage de points (obtenues avec une analyse en composantes principales), il n'est pas possible de voir à quoi ressemble le clustering sauf si les données ont moins de 3 dimensions. Cet avantage est compensé par le fait qu'il est difficile de savoir si l'exécution de la carte auto-adaptative a bien fonctionné, hormis avec les 2 indicateurs de qualité présentés.

De plus, il n'existe pas de théorie donnant le nombre optimal de clusters pour un algorithme K-means [15]. La part d'inertie expliquée par les classes n'est en effet qu'une heuristique parmi d'autres. La possibilité de visualiser directement les clusters a aussi donné la possibilité de tester des partitions à 11 ou 13 clusters afin de voir si ces valeurs, proches du nombre de 12 choisi, étaient pertinentes.

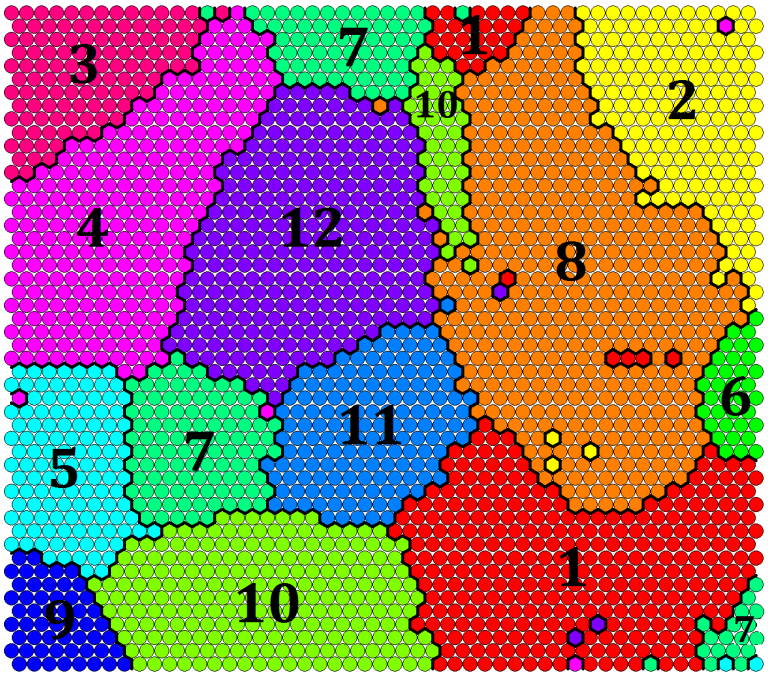


FIGURE 3.25 – Clustering de la carte auto-adaptative

Il est nécessaire d'expliquer la signification de chacun de ces clusters. La force des cartes auto-adaptatives est de donner une visualisation des clusters et des distributions des variables au sein de ceux-ci. Contrairement aux algorithmes de clustering plus classiques comme la classification ascendante hiérarchique ou les K-means appliqués directement aux vecteurs de la base, pour lesquels l'interprétation d'un cluster nécessite une analyse statistique complémentaire, les SOM rendent l'exploitation du clustering aisée.

15. L'algorithme K-means nécessite en effet de fournir en paramètre le nombre de clusters.

Les différents clusters sont explicables en mettant en regard les différentes cartes. Par exemple, le cluster n° 2 représente les jeunes, ayant une majoration conducteur novice et un bonus malus élevé. En effet, les cartes du bonus-malus, de la majoration conducteur novice (la valeur vaut 0 si une majoration conducteur novice est appliquée) et de l'âge des assurés montrent des particularités pour ce cluster en particulier.

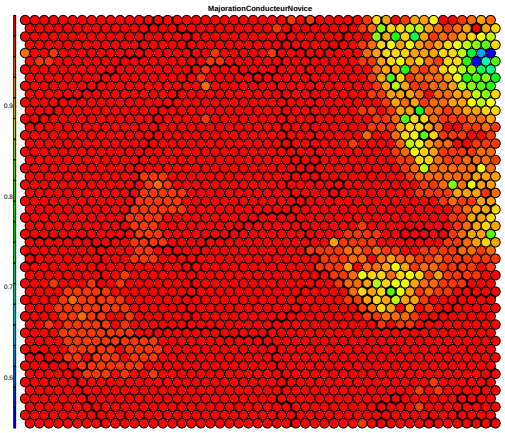


FIGURE 3.26 – Majoration Conducteur Novice

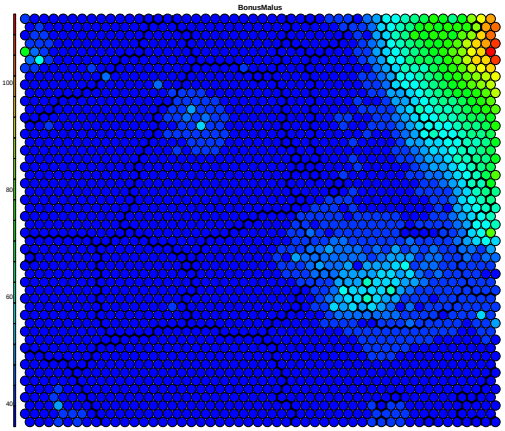


FIGURE 3.27 – Bonus-malus

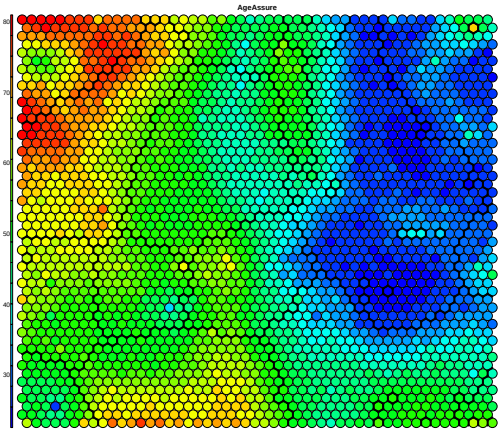


FIGURE 3.28 – Âge de l'assuré

De même, le cluster n° 11 semble être celui des assurés ayant une voiture puissante qui peut être une voiture de sport ou relativement puissante, ayant la cinquantaine, et de classe socio-professionnelle chef d'entreprise, cadre supérieur ou autre. Les cartes de la puissance du véhicule, de l'âge de l'assuré et de la classe socio-professionnelle illustrent cela.

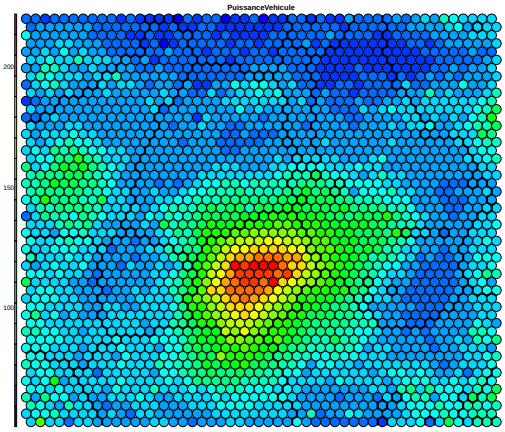
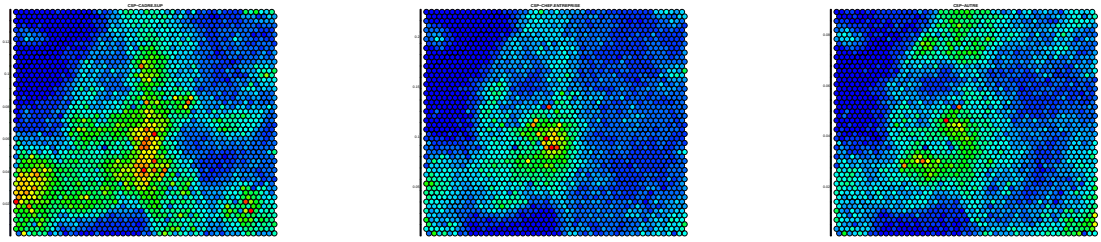


FIGURE 3.29 – Puissance du véhicule



(a) Cadre supérieur (b) Chef d'entreprise (c) Autres

FIGURE 3.30 – Classe socio-professionnelle

Le tableau 3.1 résume les interprétations des différents clusters. Par soucis de confidentialité, les tailles des différents clusters ne sont pas affichées. On notera que les clusters ont à peu près tous le même effectif, sauf le cluster n° 6 qui est deux fois plus petit que les autres et les clusters n° 1 et 8 qui sont deux fois plus grands que les autres.

Cluster	**Interprétation**	**Description détaillée**
1	Assuré auto et IARD	Marié, possesseur de l'option 2, ayant plus de contrats IARD que la moyenne, employé et d'âge environ 45 ans.
2	Jeunes conducteurs	Bonus-malus élevé, Majoration Conducteur Novice, pas marié et jeune âge
3	Clients âgés et fidèles sans réduction tarifaire	Retraité, sans réduction tarifaire, faisant un usage privé de leur véhicule, ayant leur contrat depuis un temps important.
4	Nouveaux clients retraités	Assuré âgé et retraité, faible puissance du véhicule, ayant adhéré récemment, usage privé du véhicule et conduisant plutôt des voitures diesel
5	Clients multi-équipés	Clients ayant des contrats MRH et IARD (hors auto et MRH), nombre de conducteurs élevé, lien avec d'autres assurés et une bonne partie est constituée de cadres supérieurs
6	Clients ayant un contrat prévoyance	Clients assurés en prévoyance et ayant un lien avec d'autres assurés
7	Clients ayant plusieurs contrats auto	Nombre de contrats auto élevé
8	Jeunes liés au contrat de leurs parents	Clients non mariés ayant le Bonus Famille, faisant un usage privé et professionnel de leur véhicule, ayant un lien avec d'autres assurés.
9	Clients ayant déclaré de la conduite accompagnée	Clients employés et cadres supérieurs, faisant un usage privé et professionnel de leur véhicule, ayant déclaré des conducteurs en conduite accompagnée, et ayant des contrats IARD (hors MRH et auto)
10	Clients âgés et fidèles ayant un réduction tarifaire	Clients ayant un contrat auto depuis une longue durée, faisant un usage privé de leur véhicule, ayant une réduction tarifaire, d'âge moyen environ 60 ans et ayant déclaré plusieurs conducteurs dans leur contrat
11	Cadres supérieurs	Clients d'âge moyen, chefs d'entreprises, cadres supérieurs, ayant un véhicule puissant qui peut être une voiture de sport
12	Autres profils	Clients ayant souscrit à l'option 2, d'âge moyen, souvent de sexe féminin et faisant un usage privé et professionnel de leur véhicule

Tableau 3.1 – Description des 12 clusters

3.3.3 Analyse de la résiliation dans chacun des clusters

La segmentation a permis de définir des sous-groupes homogènes mais aussi facilement interprétables grâce aux cartes auto-adaptatives. L'objectif de cette segmentation est de faciliter l'étude des comportements de résiliation.

3.3.3.1 Comparaison au taux de résiliation moyen réalisé

Il existe plusieurs façons de définir la résiliation. La convention choisie est qu'un contrat est résilié si une date de résiliation du contrat est enregistrée durant les 5 ans constituant la période d'observation et que le client n'est plus assuré auto au bout de ces 5 années (Nombre de contrats auto du client = 0). En effet, si le contrat est résilié durant les 5 ans mais que le nombre de contrats auto du client n'est pas nul, il peut avoir effectué une modification de son contrat suite à la vente de son véhicule par exemple. Bien que son contrat soit résilié, il sera donc toujours client possesseur d'un contrat auto or c'est le risque de départ du client qui intéresse l'assureur et non le risque qu'il modifie son contrat.

Le taux de résiliation entre début 2010 et fin 2014 est d'environ 20% sur le portefeuille entier. La figure 3.31 montre les taux de résiliation ventilés par cluster, avec le taux de résiliation moyen matérialisé par une ligne horizontale bleue et les barres d'erreur définissant l'intervalle de confiance à 95% du taux de résiliation :

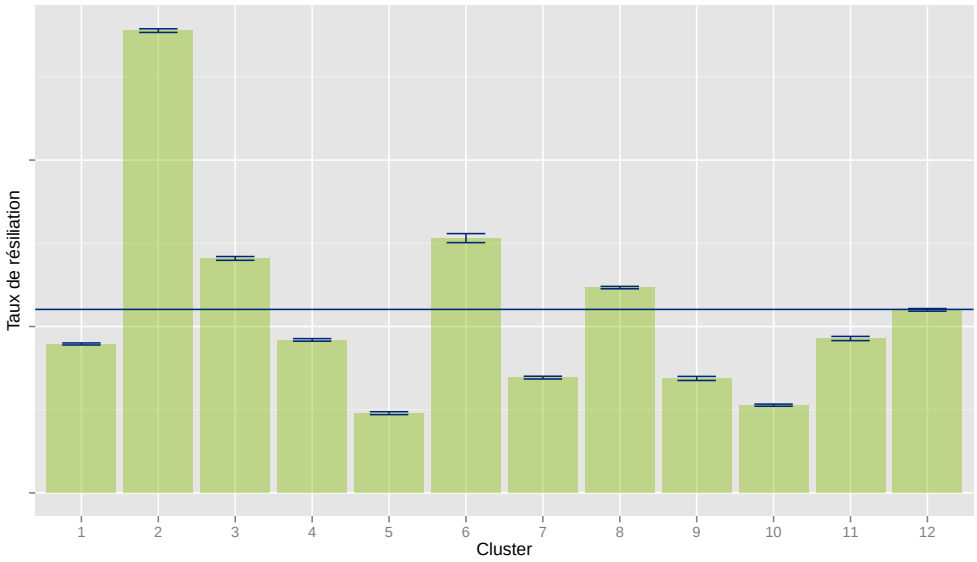


FIGURE 3.31 – Taux de résiliation entre 2010 et 2014 par cluster (échelle des taux anonymisée)

Pour obtenir un intervalle de confiance pour le taux de résiliation, l'hypothèse est faite que, dans un cluster donné, la probabilité de résiliation de chaque assuré est homogène. En notant p cette probabilité, l'estimateur du taux de résiliation dans le cluster vaut :

$$\bar{p} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

avec n le nombre d'assurés dans le cluster, $\mathbf{X}$ une variable aléatoire suivant une loi de Bernoulli $\mathcal{B}(p)$ et $(X_i)_{i\in[\![1,n]\!]}$ n réalisations indépendantes de cette variable aléatoire.

Pour obtenir un intervalle de confiance sur p, on fait l'hypothèse classique qu'une loi binomiale $\mathcal{B}(n,p)$ est environ égale en loi à une loi normale $\mathcal{N}(np, np(1-p))$ dès lors que $np > 20$. Or np est supérieur à 20 dans notre cas de figure. De plus, $\sum_{i\leq n} X_i$ suit une loi binomiale $\mathcal{B}(n,p)$. Dès lors, l'intervalle de confiance à 95% de p est classiquement :

$$\left[\bar{p} - 1.96\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}, \bar{p} + 1.96\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\right]$$

Il est indispensable d'afficher en parallèle la duration[16] moyenne début 2010 par contrat. En effet, si le taux de résiliation dans un cluster est important mais que la duration moyenne des contrats y est élevée, le cluster regroupe peut-être les contrats en fin de vie et un résiliation élevée n'apporte pas beaucoup d'information autre que celle, en faisant un parallèle avec la mortalité, que des personnes en fin de vie meurent plus en proportion que des personnes en début de vie. La duration moyenne par cluster est affichée en figure 3.32. Les intervalles de confiance ne sont pas affichés car les clusters ont plusieurs dizaines de milliers d'observations donc les intervalles de confiance sont très étroits.

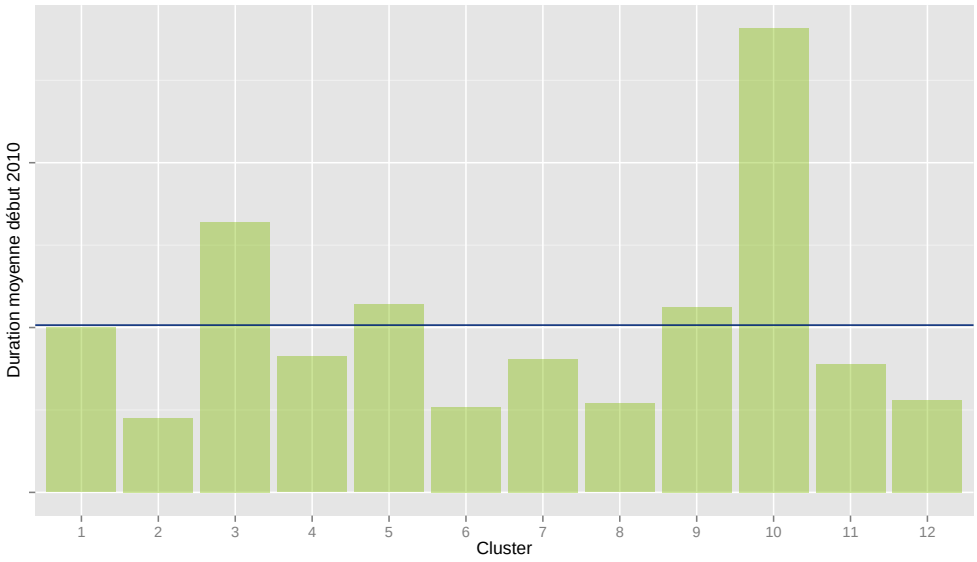


FIGURE 3.32 – Duration moyenne début 2010 par cluster (échelle des durées anonymisée)

Deux clusters semblent se démarquer des autres. Tout d'abord, le cluster n° 2, qui regroupe les jeunes conducteurs, a eu un taux de résiliation de 55,6% entre début 2010 et fin 2014, pour une duration moyenne début 2010 de 4,52 ans. Cela se compare au taux moyen d'environ 20% et à la duration moyenne d'environ 10 ans. Ce cluster semble donc être intéressant car il présente un taux de résiliation très supérieur à la moyenne.

16. Afin de ne pas faire de confusion avec la durée de vie du contrat que l'on a déjà défini précédemment et qui désignait la durée d'assurance d'un contrat une fois celui-ci résilié, on introduit le terme *duration* qui est la durée de vie du contrat depuis la date de souscription, sachant que ce dernier peut être ou non résilié.

Au contraire, le cluster n° 10 a enregistré un taux de résiliation de 10,5% entre 2010 et 2014 pour une duration moyenne de 28,2 ans. Ce cluster regroupe les clients âgés et fidèles ayant une réduction tarifaire et a un taux de résiliation très inférieur à la moyenne.

De plus, de façon moins marquée, les assurés du cluster n° 5, identifiés comme les clients multi-équipés, présente un taux de résiliation inférieur à la moyenne (9,6%) pour une duration moyenne de 11,40 ans (contre environ 10 ans en moyenne sur l'ensemble du portefeuille). Les clients multi-équipés semblent donc résilier moins que la moyenne.

Enfin, et aussi de façon moins marquée, le cluster n° 6 présente un taux de résiliation supérieur à la moyenne (30,6%) pour une duration moyenne début 2010 de 5,18 ans (contre environ 10 ans en moyenne sur l'ensemble du portefeuille). Cela tend à dire que les clients possédant par ailleurs des contrats prévoyance chez le Partenaire résilient plus que la moyenne.

3.3.3.2 Choix du cluster d'intérêt

L'objectif de la suite de l'étude est l'analyse d'un des clusters trouvés. L'apport de valeur passe par la découverte de poches clients au sein de ce cluster qui ont un taux de résiliation très éloigné du taux de résiliation du cluster. Par exemple, si le taux de résiliation est faible dans un cluster, trouver des poches clients qui regroupent les clients ayant une plus forte propension à résilier que la moyenne des clients du cluster est intéressant. En effet, ces poches clients vont à contre-courant de l'intuition selon laquelle les clients du cluster en question sont rentables car ils résilient peu. Cet apport d'information peut donc permettre de prendre la décision de ne plus chercher à prospecter des clients ayant les caractéristiques d'une poche client résilient plus que la moyenne. L'inverse est aussi vrai, des profils résiliant peu au sein d'un cluster ayant un taux de résiliation élevé présentent de l'intérêt.

La suite de l'étude doit permettre d'établir des règles définissant un profil d'assuré qui présente un intérêt en termes de marketing.

Un des facteurs déterminant le choix du cluster est le motif de résiliation des assurés au sein du cluster.

Une résiliation peut intervenir pour les motifs suivants :

— **Vente** : Résiliation suite à la vente du véhicule
— **Demande assuré** : résiliation à la demande de l'assuré
— **Augmentation Tarif/Loi Châtel** : résiliation dans le cadre de la loi Châtel ou car le client refuse une augmentation de son tarif
— **Sinistres** : résiliation car le véhicule a été détruit à la suite d'un accident
— **Non-paiement** : résiliation par l'assureur pour non-paiement des primes
— **Non utilisation** : résiliation car l'assuré n'utilise pas son véhicule
— **Départ étranger** : résiliation car l'assuré part à l'étranger
— **Initiative assureur** : résiliation à l'initiative de l'assureur
— **Autre** : autres motifs

Le cluster n° 3 est choisi pour l'étude de la résiliation. Cette décision est prise car ce segment, celui des clients âgés et fidèles sans réduction tarifaire, intéresse particulièrement le Partenaire. Le taux de résiliation au sein de ce cluster est de 28,2% pour une duration moyenne début 2010 de 16,42 ans. Il s'agit d'un groupe d'assurés ayant souscrit depuis une longue durée mais qui résilie beaucoup. Les raisons de cette résiliation sont affichées en figure 3.33. La principale cause de résiliation au sein de ce segment est la vente du véhicule.

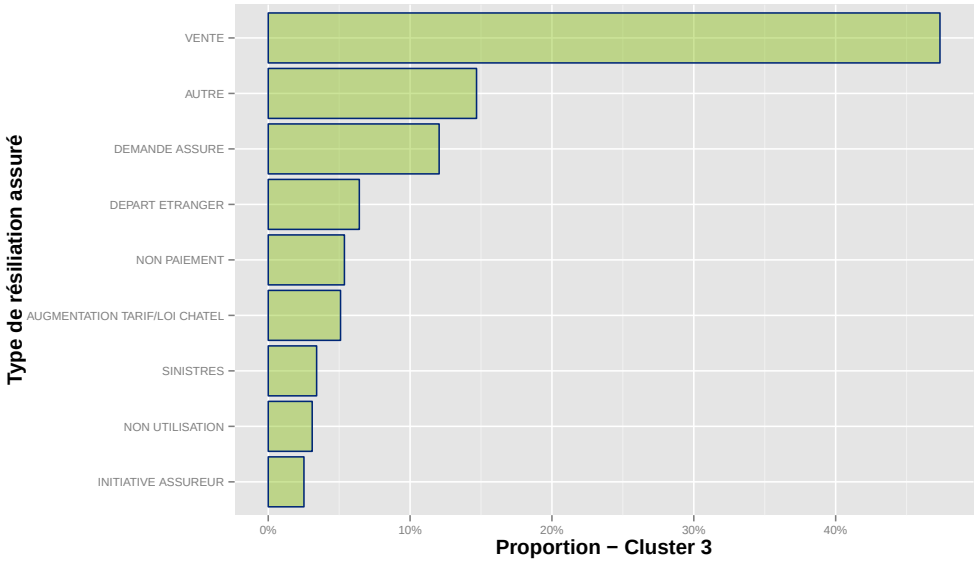


FIGURE 3.33 – Types de résiliation du contrat pour le cluster 3

La moyenne d'âge au sein de ce cluster étant élevée, le second type de résiliation, *Autre*, est à priori le décès de l'assuré. C'est un point à noter car une poche clients composée de clients très âgés n'est pas très intéressante.

3.4 Conclusion et limites sur la segmentation

Cette partie a montré la pertinence des cartes auto-adaptatives pour étudier un portefeuille d'assurance non-vie (bien qu'elles puissent aussi s'appliquer à ceux d'assurance vie). Ces techniques, qui s'écartent des méthodes classiques, réalisent un clustering en utilisant deux niveaux d'abstraction :

— Un premier niveau d'abstraction qui résume les données à des vecteurs prototypes représentés par des neurones,
— Un deuxième niveau d'abstraction qui regroupe les prototypes proches en des clusters à l'aide d'un algorithme K-means.

Les cartes auto-adaptatives permettent tout d'abord de visualiser dans un espace à deux dimensions les relations entre les variables du portefeuille. Cette visualisation a permis une interprétation aisée des différents clusters obtenus.

De plus, en réalisant une quantification vectorielle des vecteurs de l'espace d'entrée, la carte auto-adaptative résume plusieurs millions de vecteurs à quelques milliers. Ce procédé permet une élimination des valeurs très éloignées de la moyenne, notamment les valeurs aberrantes, qui auraient pu perturber un clustering appliqué directement aux vecteurs de l'espace d'entrée.

Les SOM présentent cependant plusieurs limites qui rendent leur implémentation délicate dans certains cas.

Tout d'abord, la bonne exécution d'une carte auto-adaptative nécessite la calibration des différents paramètres qui la compose. C'est ce qui caractérise les algorithmes d'apprentissage automatique : si les paramètres sont bien réglés, les performances sont très supérieures aux méthodes statistiques classiques, mais mal réglés ils font beaucoup moins bien. Comme on l'a montré, un réglage approximatif de ces paramètres aboutit à des résultats très satisfaisants, ce qui prouve la robustesse des SOM. Il existe cependant des variantes des cartes auto-adaptatives de Kohonen qui apprennent automatiquement le nombre optimal de neurones ainsi que les autres paramètres mais ces algorithmes sont plus coûteux en temps de calcul.

Une autre limite des cartes auto-adaptatives est le temps d'exécution de celles-ci sur un portefeuille de 2 millions de lignes. L'exécution de la SOM appliquée au portefeuille avait par exemple duré plus de 7 heures. Ce temps de calcul peut cependant être réduit car certaines phases de l'algorithme, comme l'attribution du **Best Matching Unit**, peuvent être parallélisées. Le temps d'exécution est la limite principale de cet algorithme.

De plus, une limite constatée est le traitement préalable des variables en vue de l'exécution d'un algorithme SOM. Ce dernier requiert en effet des variables numériques, nécessitant la création de *dummy variables*. De plus, les données ne doivent pas présenter de valeurs manquantes même si des variantes de l'algorithme présenté savent traiter les valeurs manquantes sans problème.

Enfin, on conclura sur le fait que cette segmentation n'est qu'un modèle. De nombreux biais existent et notamment le retraitement préalable des données avant l'exécution du SOM. D'autres retraitements ou même un autre algorithme auraient sans doute pu produire des résultats très différents. Il ne faut donc pas voir le clustering réalisé comme la segmentation exacte du portefeuille (ce qui n'existe bien sûr pas). Ce modèle ne constitue qu'une proposition de clustering qui a ses qualités comme ses limites.

L'étape suivante est la constitution de poches clients au sein du cluster n° 3, toujours avec les cartes auto-adaptatives, et la création d'un modèle explicatif au sein de ces poches afin de comprendre de manière très détaillée les raisons poussant un assuré d'une poche à résilier son contrat.

Chapitre 4

Analyse des comportements de résiliation au sein de poches clients

Sommaire

Cette partie vise à trouver parmi les assurés du cluster n° 3 des poches clients au profil très éloigné de la moyenne. Comme ce cluster présente un taux de résiliation élevé, le principal apport sera de mettre en évidence des poches clients ayant un taux de résiliation très inférieur au taux de résiliation du cluster, 28,2%. Ayant identifié des poches clients d'intérêt, l'approche consistera ensuite à comprendre les raisons de la résiliation au sein de ces poches de clients homogènes. Cette étude doit permettre de dégager des règles sur les caractéristiques d'assurés intéressants et à l'aide de celles-ci, de formaliser des actions opérationnelles.

4.1 Recherche de poches clients homogènes

L'objectif est de constituer des poches clients homogènes au sein du cluster n° 3. La segmentation globale du portefeuille était effectuée sur plus de 2 millions de contrats et la volonté n'était pas d'obtenir une exécution performante de la carte auto-adaptative [1]. En effet, le temps d'exécution était trop limitant et les résultats obtenus étaient facilement explicables donc satisfaisants. Pour segmenter le cluster n° 3, l'objectif est d'obtenir des poches très homogènes de clients. La carte auto-adaptative doit donc très bien converger de manière à obtenir les meilleurs résultats possibles en termes de poches clients. Cela est plus aisé que lors du clustering de tout le portefeuille pour des raisons de temps d'exécution. En effet, le cluster n° 3 ne compte qu'environ 150 000 contrats contre plus de 2 millions pour tout le portefeuille.

4.1.1 Théorie : optimisation de la carte auto-adaptative

Afin d'avoir la carte la plus détaillée possible, on choisit la heuristique maximisant le nombre de neurones sur la carte : $5 \times n^{0.54321}$. Avec $n \approx 150000$ contrats, la carte auto-adaptative optimale, selon cette heuristique, comporte 3256 neurones.

Un autre paramètre, sur lequel nous n'avons pas encore joué, est la forme de la carte : c'est-à-dire sa longueur et sa largeur. En effet, jusque-là, nous nous sommes restreints à des cartes carrées. Lorsque le nuage de points est allongé selon une direction, une carte carrée n'est pas forcément adaptée. Par exemple, si les données forment un nuage gaussien à 3 dimensions de moyenne nulle et de variances respectives sur les axes (Ox), (Oy) et (Oz) égales à a^2, b^2 et c^2 avec $a \geqslant b \geqslant c$ alors utiliser une grille telle que le rapport longueur sur largeur est égal à $\frac{a}{b}$ peut pousser la grille à se déformer prioritairement dans les deux directions (Ox) et (Oy).

Pour faire le choix de la longueur et de la largeur de la carte, on choisit d'utiliser la technique de la décomposition en valeur singulières, afin d'étendre l'exemple du nuage gaussien ainsi présenté :

Soit $\mathbf{M} = (m_{ij})_{(i,j)\in[\![1,m]\!]\times[\![1,n]\!]}$ une matrice représentant les données. Chaque ligne représente un vecteur et chaque colonne une variable. Il est possible de factoriser $\mathbf{M}$ sous la forme :

$$\mathbf{M} = \mathbf{UDV^T}$$

1. C'est-à-dire avec une erreur de quantification et une erreur topographique assez élevées.

avec :

— **V** une matrice orthogonale de taille $n \times n$, dite « matrice d'entrée ».
— **U** une matrice orthogonale de taille $m \times m$, dite « matrice de sortie »
— **D** une matrice diagonale contenant les valeurs singulières $(\sigma_i)_{i \in [\![1,n]\!]}$ de **M** classées par ordre décroissant.

Les directions associées aux valeurs singulières de **M** les plus grandes peuvent être interprétées comme les directions de plus grande variation de l'ensemble des points de données. Dans le cas du nuage gaussien, les valeurs singulières σ_1, σ_2 et σ_3 sont de l'ordre de $a\sqrt{m}$, $b\sqrt{m}$ et $c\sqrt{m}$. On retrouve les paramètres associés aux trois directions du nuage gaussien.

On réalise une décomposition en valeurs singulières de la matrice des données normalisées **M** et on obtient un ratio entre σ_2 et σ_1 égal à $\frac{\sigma_2}{\sigma_1} \approx 0,66$. En notant p le nombre de neurones, l la largeur de la carte et L sa longueur, on veut donc avoir :

$$p = l \times L = 0,66L \times L = 0,66L^2$$

soit avec $p = 3256$, on obtient que la carte optimale est de dimension 70×47.

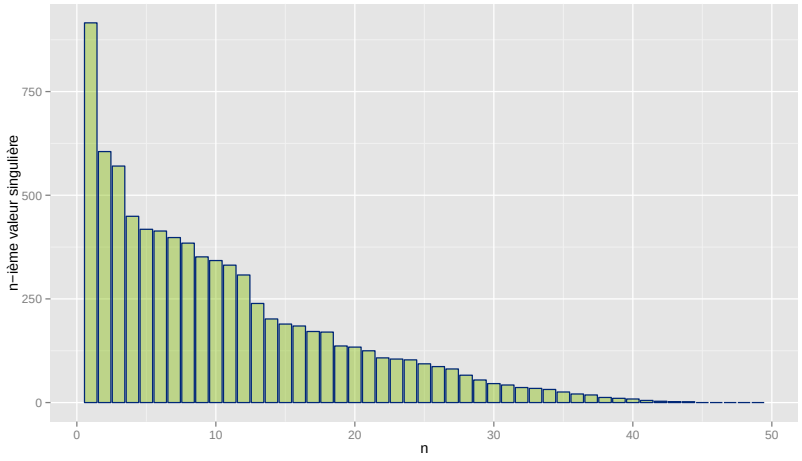


FIGURE 4.1 – Valeurs singulières de **M**

Une autre façon d'améliorer l'exécution est, plutôt que d'initialiser au hasard les vecteurs de poids, d'initialiser les vecteurs de poids dans le plan défini par les deux premières directions principales du nuage de points. Cela peut forcer la grille à s'étendre dans la bonne direction. On ne réalise pas cette opération.

4.1.2 Résultats : obtention de poches clients

Après des tests préliminaires, on utilise un taux d'apprentissage de $(\alpha, \beta) = (0.05, 0.01)$ et 100 présentations de la base complète à la carte auto-adaptative. Le résultat est affiché en figure 4.2. La convergence semble survenir pour un nombre d'itérations de 50.

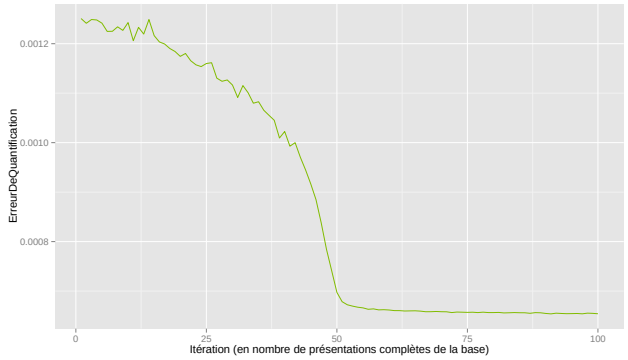


FIGURE 4.2 – Convergence de la carte auto-adaptative

L'erreur topographique est cependant élevée et vaut 0.63487. On réalise une seconde exécution avec un taux d'apprentissage plus faible ($\alpha = (0.01, 0.005)$) et 3 itérations afin de finaliser la convergence de la carte. Cela permet d'avoir une erreur topographique de 0.22511.

Le clustering de la carte auto-adaptative est réalisé à l'aide de l'algorithme des k-means et on trouve un nombre optimal de clusters pour $k = 10$ (voir figure 4.3).

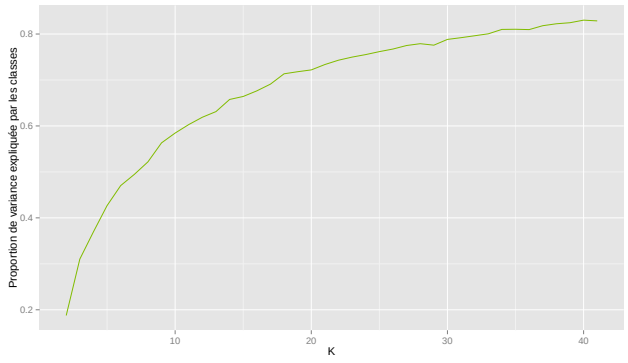


FIGURE 4.3 – Nombre optimal de clusters

Le résultat du clustering est affiché en figure 4.4. Hormis quelques neurones orphelins, la carte est visuellement de bonne qualité et cela est montré par les indicateurs de qualité présentés (erreur topographique et convergence de l'erreur moyenne de quantification).

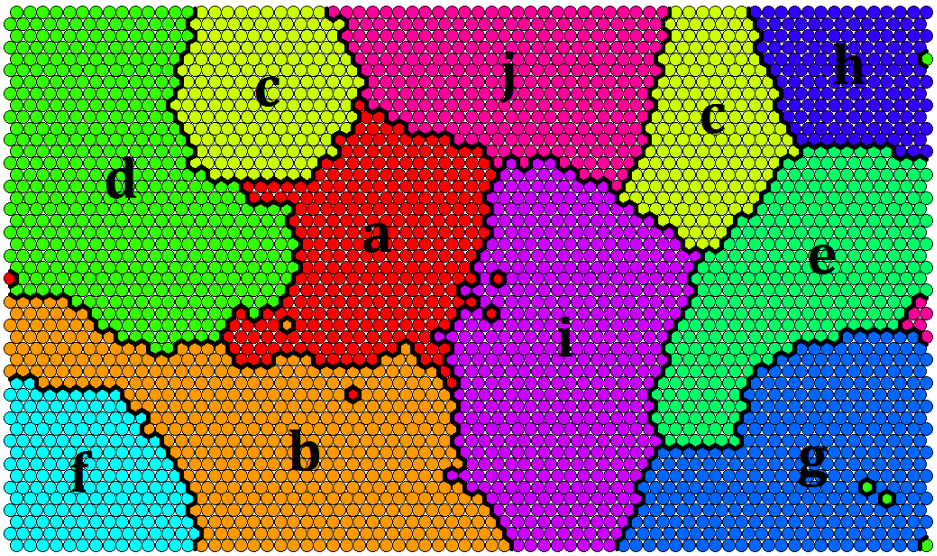


FIGURE 4.4 – Poches clients du cluster n° 3

De même que lors de l'analyse des clusters, on affiche le taux de résiliation par poche entre 2010 et 2014 en figure 4.5 et la duration moyenne par poche en 2010 en figure 4.6.

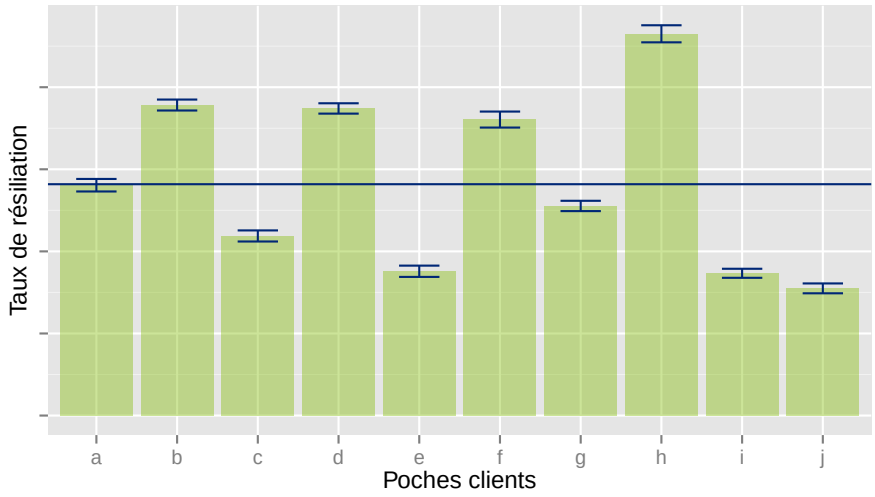


FIGURE 4.5 – Taux de résiliation pour chaque poche client (taux du cluster en bleu)

L'interprétation des poches clients est plus complexe que celle des clusters. En effet, le portefeuille entier était très hétérogène et les clusters produits sont donc plutôt clairs. Les poches clients sont obtenues à partir du cluster n° 3, qui est déjà relativement homogène. Il est donc compliqué d'exploiter les cartes auto-adaptatives obtenues lors de la constitution des poches clients pour interpréter ces dernières.

L'approche retenue afin d'attribuer un assuré âgé et fidèle sans réduction tarifaire à une poche donnée est d'utiliser une arbre de classification (voir 4.5) avec comme variable cible la poche client de l'assuré. La variable cible a 10 modalités réparties

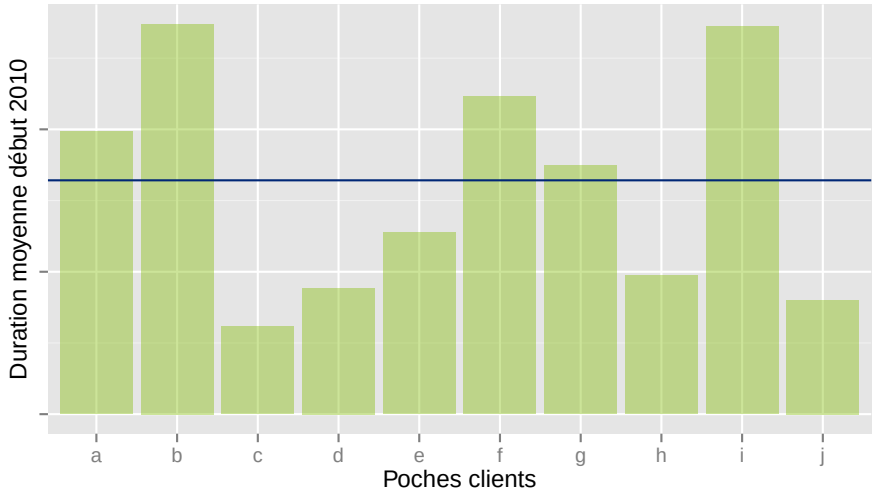


FIGURE 4.6 – Duration moyenne pour chaque poche client (duration du cluster en bleu)

dans chaque feuille en proportions $p_1, \cdots, p_{10}$ et l'arbre CART utilise le critère de Gini suivant pour effectuer les découpages :

$$I_{\text{Gini}} = 1 - \sum_{i=1}^{10} p_i^2$$

L'arbre obtenu a été réalisé sur toutes les données contenues dans le cluster n° 3. Afin de s'assurer qu'il est généralisable, l'arbre est d'abord construit sur une base d'apprentissage, par exemple 75% de la base totale et est testé sur les données restantes, la base de test. On obtient des erreurs de classification faibles sur la base de test. L'erreur globale est de 9,8% et sur certaines classes, celle-ci s'élève à environ 7% (b, d, f, h ou j) ou environ 18% (a, c, e, g ou i).

L'arbre construit sur toutes les données est affiché en figure 4.7. Pour lire l'arbre, chaque feuille contient l'information sur la classe majoritaire, sa probabilité et la proportion des données totale contenue dans la feuille en question. Par exemple, dans la feuille tout en bas à gauche, on peut lire que la classe majoritaire est la classe b, que sa probabilité est de 0.68 et que cette feuille contient 14% des données de la base. De plus, à chaque découpage, la feuille de gauche est celle qui remplissait le critère. Les paramètres de l'arbre sont un nombre minimal de données pour effectuer un découpage égal à 2000 et un effectif minimal par feuille de 2000 vecteurs.

On valide la pertinence de cet arbre à l'aide des cartes auto-adaptatives. Par exemple, la poche j est caractérisée par un nombre de contrats IARD hors auto et MRH égal à 0, une duration du contrat inférieure à 15 ans et un nombre de contrats auto au moins égale à 2. Cela se constate en figure 4.8 sur les cartes associées. Par exemple, on constate que la poche j est caractérisée par un nombre de contrats auto élevé. On retrouve cela sur la carte auto-adaptative (nuage jaune-rouge sur la carte des contrats auto à l'emplacement de la poche j) et sur l'arbre (découpage **NbContratAuto** < 1.5 dans l'arbre de classification qui caractérise la poche j).

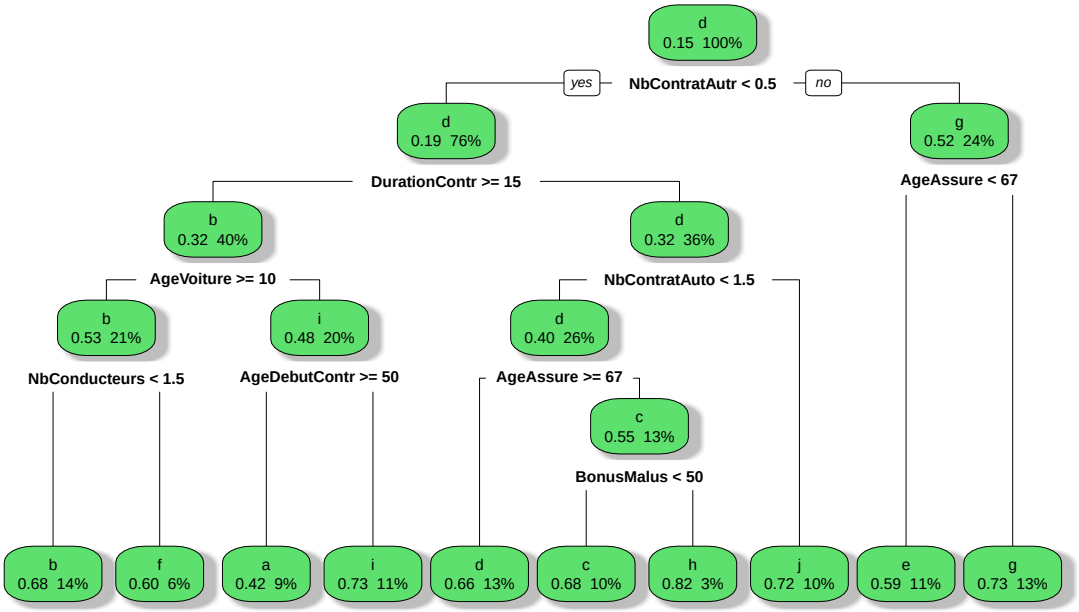
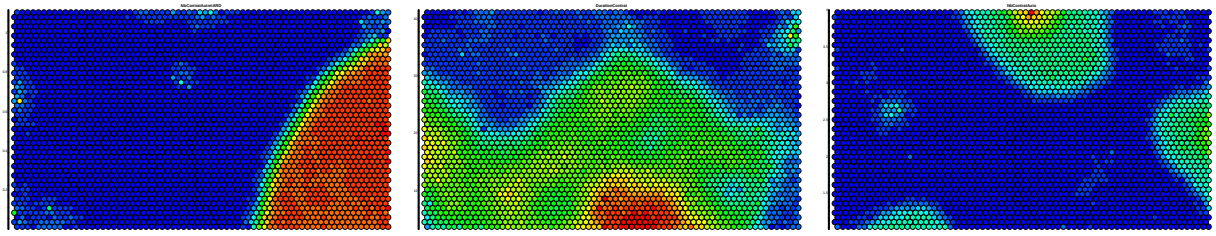


FIGURE 4.7 – Arbre définissant l'attribution d'un assuré à une poche client



(a) Contrats IARD (b) Duration du contrat (c) Contrats auto

FIGURE 4.8 – Validation des résultats de l'arbre de classification sur la SOM

Avant de mettre en place un modèle explicatif de la résiliation, il est nécessaire de choisir une poche d'intérêt afin de l'étudier plus en détail.

Tout d'abord, la poche clients i (18146 contrats) présente une duration moyenne très élevée (27,2 ans) pour un taux de résiliation faible (17,3%) alors que le taux moyen du cluster n° 3 est de 28,2% pour une duration moyenne de 16,42 ans. Cette poche représente les assurés du cluster n° 3 qui n'ont pas d'assurance IARD hors auto et MRH chez le Partenaire, ayant souscrit depuis plus de 16 ans et avant l'âge de 60 ans et avec une voiture récente (d'âge inférieur à 10 ans) [2]. On constate donc que cette poche contient des clients très âgés. La moyenne d'âge des assurés de cette poche est d'ailleurs élevée. Cette poche présente de l'intérêt car, au sein d'un cluster ayant un fort taux de résiliation, les assurés de la poche i bousculent l'intuition selon laquelle les clients âgés et fidèles mais qui n'ont pas de réduction tarifaire vont beaucoup résilier. Cette poche peut donc permettre la formalisation d'actions marketing ciblées.

2. Cette définition des contrats de la poche i découle de l'arbre de classification. L'arbre donnait à des contrats vérifiant ces conditions une probabilité de 73% d'appartenir à la poche i.

La figure 4.9 montre les types de résiliation pour les assurés de la poche client i. On remarque que les types de résiliations sont sensiblement les mêmes qu'au niveau du cluster n° 3 tout entier (voir figure 3.33). Le type de résiliation « Non-paiement » est tout de même légèrement moins représenté dans cette poche par rapport au cluster tout entier.

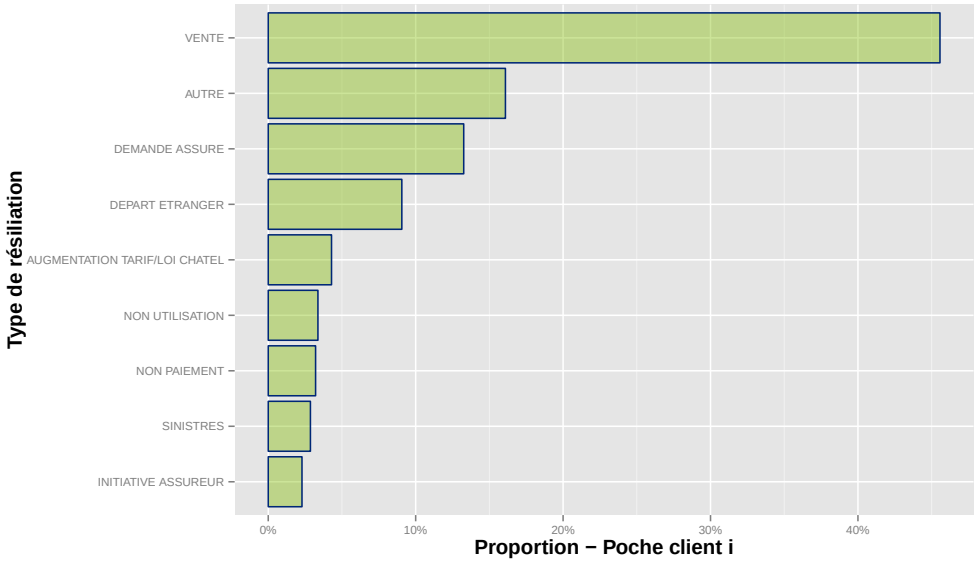


FIGURE 4.9 – Type de résiliation pour les contrats de la poche i

La poche h (8842 contrats) a une duration moyenne faible (9,72 ans) pour un taux de résiliation élevé (46,5%). Cette poche, d'après l'arbre de classification, contient des clients sans contrat IARD hors auto et MRH, ayant souscrit depuis moins de 16 ans, ayant un seul contrat auto, un âge inférieur à 67 ans et un coefficient de bonus-malus supérieur à 60. Cette poche présente de l'intérêt car elle semble constituée de clients qui ont une forte proportion à résilier. Comprendre pourquoi ceux-ci résilient peut donc être intéressant afin, soit de prendre des actions pour tenter de conserver ces clients, soit de provoquer les résiliations de ces assurés en ne tentant pas de les retenir.

Pour la poche h, les types de résiliation sont assez différents de la moyenne du cluster n° 3 (voir figure 4.10). Tout d'abord, la raison « Autre » n'est qu'en quatrième position dans cette poche par rapport au cluster n° 3 dans lequelle « Autre » se trouve en deuxième position. Cela est normal car les assurés de la poche h sont plus jeunes que la moyenne dans le cluster n° 3 (61,2 ans pour la poche h contre 70,7 ans pour le cluster n° 3). En effet, « Autre » inclut le décès de l'assuré et il est donc normal qu'on retrouve plus cette cause de rupture du contrat pour une moyenne d'âge de l'assuré plus élevée. De plus, contrairement à la moyenne du cluster n° 3, on retrouve le non-paiement en troisième position pour les types de résiliation les plus rencontrés. Les assurés de la poche h, en plus d'avoir une forte propension à résilier, semblent donc être plus enclins que la moyenne à être résiliés pour non-paiement. Cela valide l'intérêt d'analyser cette poche client qui semble être très peu rentable.

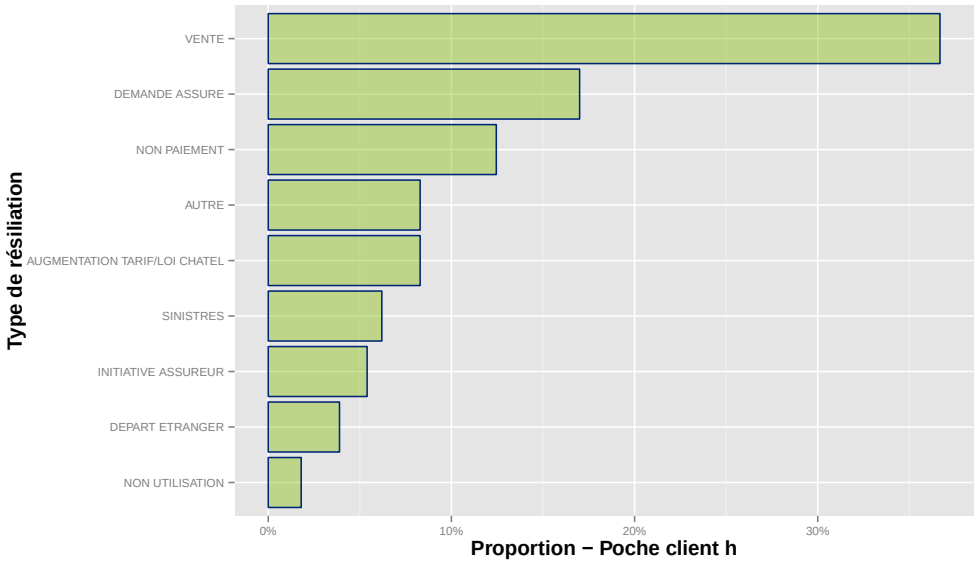


FIGURE 4.10 – Type de résiliation pour les contrats de la poche h

4.1.3 Validation : robustesse de la méthode

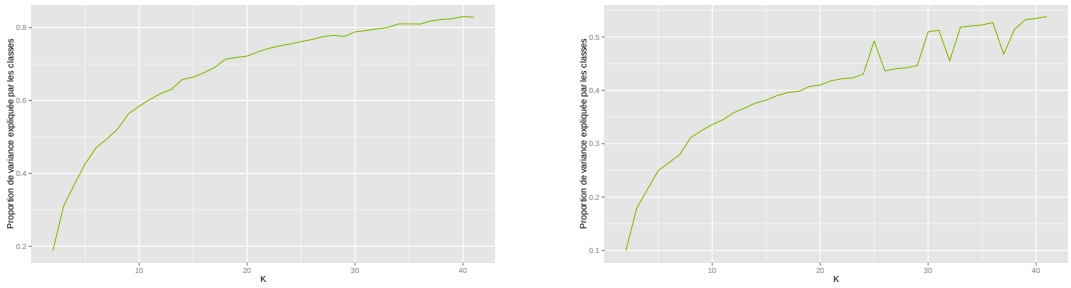
Il est nécessaire de justifier que l'algorithme employé a effectivement permis d'obtenir des groupes homogènes et de valider le fait que le cluster n° 3 est constitué d'un nombre de poches de l'ordre de 10.

Pour cela, l'approche consiste à exécuter un algorithme de clustering sur les données du cluster n° 3 et à comparer les résultats avec ceux obtenus précédemment.

Pour obtenir les poches clients, une combinaison SOM + clustering k-means a été employée. L'idée est de comparer les résultats obtenus avec l'exécution d'un k-means directement sur les données. Cette méthode basique, consistant à appliquer un k-means au cluster n° 3, doit donner des résultats semblables en terme de nombre optimal de poches mais doit aussi former des poches relativement semblables à celles obtenues avec l'approche de la carte auto-adaptative.

On compare les résultats obtenus en termes de proportion de variance expliquée par les classes. Le critère du coude s'applique dans les deux cas et donne, avec le simple clustering k-means, un nombre optimal de clusters compris entre 10 et 13 (voir figure 4.11). On remarque un comportement erratique pour la proportion de variance expliquée par les classes pour l'approche k-means pour $k > 20$. Cela est dû à la mauvaise convergence de l'algorithme pour ces valeurs de k. En effet, le nombre de vecteurs prenant part au clustering dans le cas d'un k-means appliqué directement aux données est de 151 242 contre seulement 3290 pour l'approche SOM + k-means et plus le nombre de vecteurs est grand, plus un nombre important d'itérations est nécessaire. Le nombre optimal de poches est du même ordre avec les deux algorithmes, ce qui valide en partie la pertinence des cartes auto-adaptatives pour le clustering. De plus, on remarque que, pour k égal à 10, l'approche SOM + k-means permet d'expliquer 60% de la variance contre seulement 35% pour l'approche plus simpliste. Il est cependant nécessaire de prendre en compte que le SOM a déjà fait perdre de l'information en réalisant une

quantification vectorielle. L'approche SOM + k-means permet donc seulement d'expliquer 60% des vecteurs de poids qui représentent eux-même une partie de l'information de la base de départ.



(a) Approche SOM + k-means (b) Approche k-means

FIGURE 4.11 – Comparaison des proportions de variance expliquées par les classes

L'autre critère permettant de valider la pertinence de l'approche SOM + k-means est le fait que les poches obtenues sont proches de celles qu'auraient permis d'obtenir un autre type de clustering. Pour cela, l'idée est de comparer les poches obtenues avec les 2 approches, sachant que l'on choisit un nombre de 10 poches pour l'approche k-means.

Pour cela, on utilise la notion d'indice de Rand (voir RAND W. (1971)) qui permet de calculer une distance entre deux clusterings.

Soit $\mathcal{S} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ un ensemble de vecteurs. S représente les données. Soit alors deux partitions $\mathcal{E} = \{E_1, \cdots, E_p\}$ et $\mathcal{F} = \{F_1, \cdots, F_q\}$ de l'ensemble $\mathcal{S}$, respectivement en p et q parties.

On note :

— a le nombre de paires d'éléments de $\mathcal{S}$ qui sont dans la même partie dans $\mathcal{E}$ et dans la même partie dans $\mathcal{F}$,
— b le nombre de paires d'éléments de $\mathcal{S}$ qui sont dans des parties différentes dans $\mathcal{E}$ et dans des parties différentes dans $\mathcal{F}$,
— c le nombre de paires d'éléments de $\mathcal{S}$ qui sont dans la même partie dans $\mathcal{E}$ et dans des parties différentes dans $\mathcal{F}$,
— d le nombre de paires d'éléments de $\mathcal{S}$ qui sont dans des parties différentes dans $\mathcal{E}$ et dans la même partie dans $\mathcal{F}$.

L'indice de Rand est alors défini par :

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

Si les deux clusterings sont égaux, alors toute paire d'éléments de $\mathcal{S}$ est soit dans la même partie dans $\mathcal{E}$ et dans la même partie dans $\mathcal{F}$, soit dans des parties différentes dans $\mathcal{E}$ et dans des parties différentes dans $\mathcal{F}$. Dans ce cas précis, R est alors égal à 1.

Au contraire, si les deux clusterings sont très différents, R est proche de 0.

Dans le cas des deux méthodes de clustering utilisées, l'indice de Rand vaut 0.8992. Cela signifie que les deux clusterings correspondent pour presque 90% des paires d'éléments de $\mathcal{S}$. Cela nous permet donc de valider la robustesse de l'algorithme SOM + k-means. La proportion de variance expliquée par les classes de cette méthode étant plus élevée que pour la méthode du k-means, malgré la remarque précédente sur la perte d'information induite par le SOM, on peut donc dire que le clustering à l'aide des cartes auto-adaptatives donne des groupes plus homogènes qu'une approche plus basique.

4.2 Un peu de théorie sur les modèles de classification

La suite de l'étude porte sur la compréhension des facteurs de résiliation dans les poches clients considérées. L'idée est d'utiliser un modèle de classification pour mesurer ces facteurs.

Un tel modèle permet de prédire l'appartenance à une classe donnée pour une observation.

En d'autres termes, on considère des observations (y_i, x_i) d'une base, avec $y_i \in \{1, \cdots, K\}$ et K entier naturel. Un modèle de classification permet alors d'estimer la quantité $\mathbb{P}(\mathbf{Y} = k | \mathbf{X} = \mathrm{x}_i)$ avec $k \in \{1, \cdots, K\}$.

Dans le cas de la prédiction de la résiliation, on a $y_i \in \{0, 1\}$ avec $y_i = 1$ l'évènement de résiliation.

On considère les deux algorithmes suivants pour le modèle de classification voulu :

— Random Forests (forêts aléatoires)
— Modèle linéaire généralisé

L'objectif est de comparer les résultats obtenus avec deux méthodes statistiques, l'une plus classique le modèle linéaire généralisé, et l'autre plus récente les forêts aléatoires.

Cette partie rappelle le fonctionnement de ces deux algorithmes. La partie suivante expliquera comment ceux-ci sont utilisés pour mesurer la résiliation.

4.2.1 Forêts aléatoires (Random Forests)

Les forêts aléatoires ou Random Forests (voir BREIMAN L. (2001)) sont une classe de modèles appelés méta-modèles. Les méta-modèles sont une agrégation de modèles plus simples afin de créer un modèle robuste.

4.2.1.1 Bagging

Les arbres CART sont des modèles peu robustes qui sont très instables, c'est-à-dire que la structure de l'arbre peut complètement changer si les données d'apprentissage changent légèrement. Une façon d'améliorer la robustesse de la prédiction est d'utiliser le bagging.

Le bagging repose sur la notion de boostrap. Soit $\mathcal{Z} = \{z_1, \cdots, z_n\}$ avec $z_i = (x_i, y_i)$ les données que l'on considère.

L'idée du bootstrap est de créer B ($B = 100$ par exemple) échantillons en effectuant un tirage avec remise des données dans $\mathcal{Z}$. On crée alors B jeu de données bootstrappés $Z^{*1}, \cdots, Z^{*B}$ à partir des données.

On note alors $S(\mathcal{Z})$ toute quantité calculée à partir des données $\mathcal{Z}$. L'avantage du bootstrap est de permettre l'estimation de la distribution de $S(\mathcal{Z})$ sans à priori sur le modèle. Par exemple, le bootstrap permet d'obtenir un intervalle de confiance sur n'importe quelle statistique liée aux données, comme la médiane.

On approche la quantité $S(\mathcal{Z})$ par :

$$\bar{S}^* = \frac{1}{B}\sum_{b=1}^{B} S(\mathcal{Z}^{*b})$$

et, par exemple, l'estimateur non biaisé de sa variance par :

$$\hat{\text{Var}}[S(\mathcal{Z})] = \frac{1}{B-1}\sum_{b=1}^{B}\left(S(\mathcal{Z}^{*b}) - \bar{S}^*\right)^2$$

.

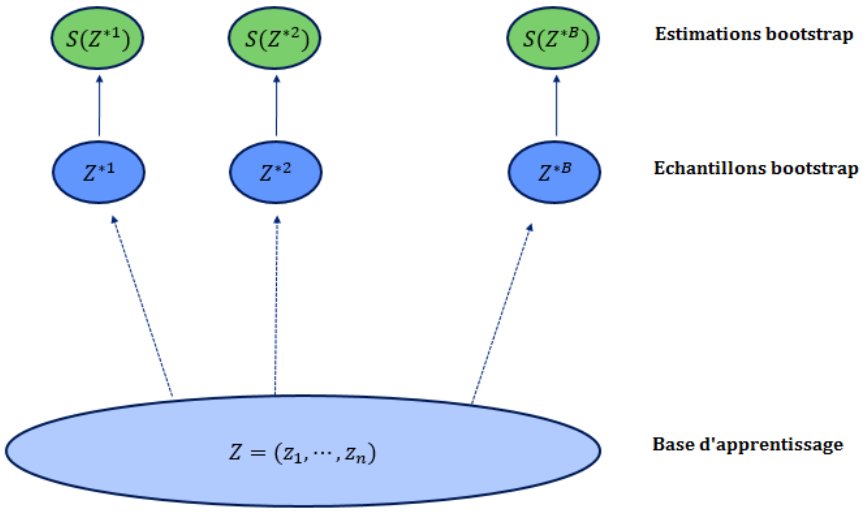


FIGURE 4.12 – Illustration du bootstrap

Dans le cadre du bagging basé sur les arbres CART, en notant ψ_{*b} la fonction de prédiction basée sur les données $\mathcal{Z}^{*b}$, l'estimateur bootstrap, destiné à rendre plus robuste les arbres CART, vaut pour $z = (x, y) \in \mathcal{Z}$:

Pour les arbres de régression ($y \in \mathbb{R}$) :

$$\bar{\psi}_*(z) = \frac{1}{B}\sum_{b=1}^{B}\psi_{*b}(z)$$

Pour les arbres de classification (Si 2 classes, $y \in \{0, 1\}$) :

$$\bar{\psi}_*(z) = \mathbb{I}\left(\frac{1}{B}\sum_{b=1}^{B}\psi_{*b}(z) > \frac{1}{2}\right)$$

On peut voir cette dernière formule comme un vote à la majorité. Si plus de $B/2$ classifieurs prédisent 1, alors c'est 1. Sinon c'est 0.

Le bagging permet de réduire l'intervalle de confiance sur la prédiction faite par les arbres. Ce procédé revient à moyenner les valeurs trouvées sur plusieurs échantillons basés sur $\mathcal{Z}$ afin de réduire l'instabilité inhérente aux arbres. Comme la plupart des algorithmes de machine learning, le bagging améliore la performance des prédictions mais n'est pas ou mal justifié de manière théorique.

On note aussi l'existence du boosting dans le cas de la classification. Dans ce cas, on effectue un tirage bootstrap sur la base d'apprentissage pour créer un classifieur faible puis à chaque itération, on crée un nouveau classifieur basé sur l'échantillon bootstrappé de telle sorte que les probabilités de tirage des données mal classées dans les échantillons bootstrappés précédents soient plus élevées que les autres probabilités. Cela permet d'augmenter le pouvoir classifiant du prédicteur final.

4.2.1.2 Principe des forêts aléatoires

Les Random Forests (voir FRIEDMAN J., HASTIE T. and TIBSHIRANI R. (2001)) sont une extension du bagging. Ce dernier permet d'améliorer les performances de prédiction, notamment des arbres CART.

Les forêts aléatoires se basent sur l'idée suivante : soient B variables aléatoires $X_1, \cdots, X_B$ de variance σ^2. Si elles sont i.i.d [3], alors la variance de $\bar{X} = \frac{1}{B}\sum_{b=1}^{B} X_b$ vaut $\frac{\sigma^2}{B}$. Si les variables sont seulement identiquement distribuées (et pas indépendantes) et de corrélation ρ deux à deux, alors la variance de $\bar{X}$ vaut :

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

Lorsque B augmente, le deuxième terme tend vers 0 mais pas le premier. En faisant l'analogie entre chaque X_i et un des arbres du bagging, on voit qu'il est nécessaire de diminuer la corrélation entre les arbres du bagging afin de diminuer la variance des prédictions. C'est cette idée qui a inspiré les Random Forests.

Pour diminuer la corrélation entre les arbres, on rajoute une part d'aléatoire dans le processus de construction de ceux-ci. A chaque nœud, parmi les p variables de découpage possibles, on tire aléatoirement avec remise $m \leq p$ variables et on recherche le critère de coupe seulement parmi ces variables. On appelle cela la méthode des sous-espaces aléatoires (*random subspaces method*).

On expose ici l'algorithme des forêts aléatoires :

Algorithme Pour b allant de 1 à B :

— On crée un échantillon bootstrap $\mathcal{Z}^*$ de taille la taille de $\mathcal{Z}$.

3. indépendantes et identiquement distribuées

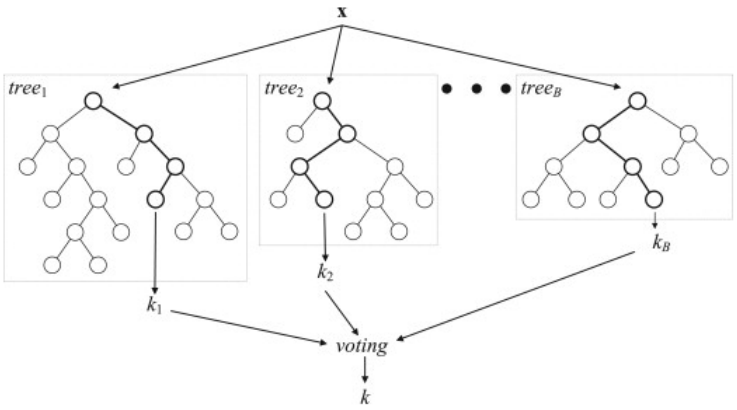


FIGURE 4.13 – Agrégation d'arbres de classification (Source : Vayatis - Cours de l'Ecole Centrale Paris)

— On construit un arbre $\mathcal{T}_b$ basé sur les données bootstrappées que l'on fait grandir tant qu'il y a plus de $n_{\min}$ données par nœud.

1. On sélectionne $m \leq p$ variables par un tirage avec remise.
2. Parmi ces m variables, on choisit celle qui permet d'obtenir le meilleur découpage.
3. On crée les deux nœuds fils.

L'algorithme fournit en sortie un ensemble d'arbres $\{\mathcal{T}_b\}_1^B$.

Pour faire une prédiction sur un vecteur x :

— **Régression** $\hat{f}_{\text{rf}}^B(x) = \frac{1}{B}\sum_{b=1}^B \mathcal{T}_b(x)$

— **Classification** Soit $\hat{C}_b(x)$ la prédiction de classe pour x avec le b-ième arbre $\mathcal{T}_b$. Alors $\hat{C}_{\text{rf}}^B(x) =$ vote à la majorité sur chaque $\hat{C}_b(x)$ pour $b = 1, \cdots, B$.

Estimation de la probabilité d'une classe Dans le cas de forêts aléatoires de classification, il est possible de donner une estimation des probabilités d'appartenance d'un vecteur x à chacune des classes prédites. La probabilité d'appartenance de x à une classe donnée est la moyenne des probabilités d'appartenance de x à la classe pour chacun des arbres. La probabilité d'appartenance de x à une classe sur un arbre de classification est le nombre d'individus appartenant à la classe dans la feuille associée à x sur le nombre total d'individus dans la feuille.

Choix des paramètres Trois paramètres permettent de régler les forêts aléatoires : la valeur de m qui est le nombre de prédicteurs pris aléatoirement parmi les prédicteurs disponibles, $n_{\min}$, le nombre minimum de données par feuille et le nombre d'arbres. On adopte souvent les paramètres suivants :

— **Régression** Pour un Random Forest basé sur des arbres de régression, on prend $m = \lfloor p/3 \rfloor$ [4] et $n_{\min} = 5$.

4. $\lfloor \cdot \rfloor$ est la fonction partie entière.

— **Classification** Pour un Random Forest basé sur des arbres de classification, on prend $m = \lfloor \sqrt{p} \rfloor$ et $n_{\min} = 1$.

Dans ces deux cas, on choisit un nombre d'arbres de l'ordre de la centaine en général.

En pratique, le choix des bons paramètres se fait au cas par cas.

Avantages et limites Les Random Forests sont de très bons prédicteurs. Ils sont très robustes et sont parmi les méthodes classiques de machine learning les plus performantes. De plus, compte tenu des performances obtenues, les Random Forests sont très peu coûteuses en temps de calcul.

4.2.1.3 Importance relative des variables

Les Random Forests permettent aussi une autre exploitation des données. Ceux-ci permettent de savoir quelles sont les variables les plus importantes et celles qui en ont moins dans un sens que l'on va donner.

Il existe deux approches pour calculer l'importance des variables.

L'approche par randomisation

Définissons tout d'abord la notion d'estimateur OOB (out-of-bag). Pour $z_i = (x_i, y_i)$, on construit l'estimateur out-of-bag de la prédiction de y_i en ne prenant en compte dans le prédicteur Random Forests que les arbres basés sur des échantillons bootstrap ne contenant pas z_i.

Le but est de quantifier l'importance de la variable X_l en terme de pouvoir explicatif. Pour chaque arbre $\mathcal{T}_b$, l'approche considère les vecteurs OOB $z_i = (x_i, y_i)$, c'est-à-dire ceux qui ne sont pas contenus dans l'échantillon $\mathcal{Z}^{*b}$ ayant servi à construire l'arbre. Pour chaque vecteur OOB, l'écart quadratique $\phi(z_i) = (y_i - \mathcal{T}_b(x_i))^2$ est calculé. La même opération est réalisée en permutant parmi les vecteurs out-of-bag la valeur du prédicteur X_l afin de voir si la capacité à prédire est diminuée si une valeur aléatoire est donnée à la variable X_l. Si l'erreur quadratique augmente notablement entre l'erreur calculée sans permutation et celle calculée après permutation, c'est que la variable X_l était à priori très explicative. Pour obtenir l'importance I_l^2 de X_l, il faut moyenner les augmentations d'erreur quadratique sur les B arbres $\{\mathcal{T}_b\}_{b \in [\![1,B]\!]}$.

On classe alors les variables par I_l^2 croissant pour savoir quelles sont les variables importantes et celles qui ne le sont pas.

L'approche du critère d'impureté

Pour un arbre $\mathcal{T}$ donné, on pose :

$$I_l^2(\mathcal{T}) = \sum_{t=1}^{J} \hat{i}_t^2 \cdot \mathbb{I}(v(t) = l)$$

comme une mesure de pertinence de chaque prédicteur X_l. On somme sur les J nœuds internes de l'arbre. A chaque nœud t, la variable $X_{v(t)}$ est utilisée pour découper les observations en deux sous-régions. Le découpage en 2 régions permet une amélioration du risque quadratique égale à $\hat{i}_t^2$. L'importance d'une variable X_l est donc définie par la somme des améliorations relatives du risque quadratique pour lesquelles la variable X_l est choisie comme critère de découpe.

Pour rappel, dans le cadre des arbres de régression, le risque quadratique est défini par $\sum_i |\mathrm{R}_i|\, Q_i$ (voir 4.5). Pour un arbre de classification en K classes, on utilise souvent le critère de Gini comme critère de pureté d'un nœud. Si on note $(p_1, \cdots, p_K)$ les proportions des K classes, l'impureté vaut $1 - \sum_{k=1}^{K} p_k^2$.

Pour un algorithme des forêts aléatoires, associé à l'ensemble d'arbres $\{\mathcal{T}_b\}_1^B$, on définit alors l'importance relative de X_l par :

$$I_l^2 = \frac{1}{B} \sum_{b=1}^{B} I_l^2(\mathcal{T}_b)$$

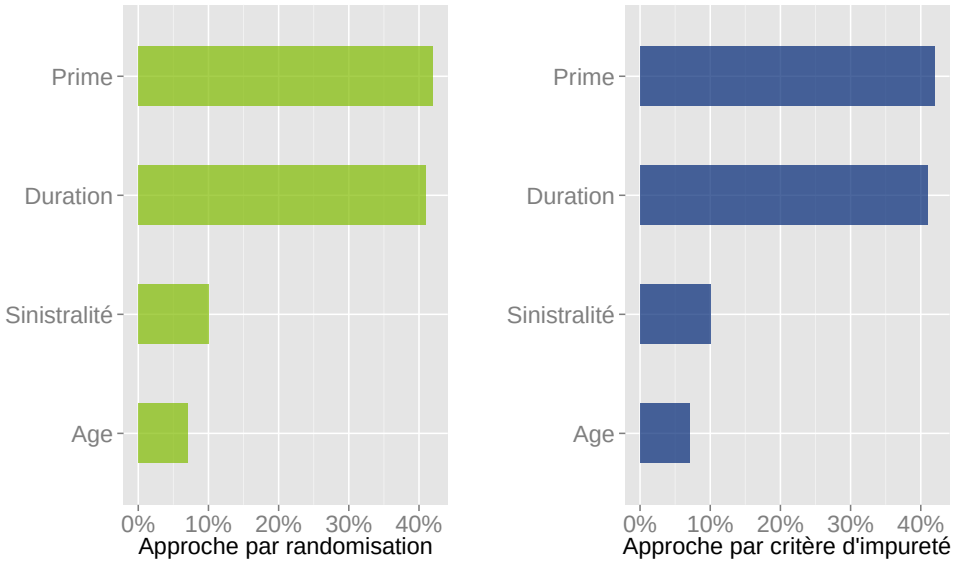


FIGURE 4.14 – Les deux méthodes de calcul de l'importance relative des variables

4.2.2 Modèle linéaire généralisé (Generalized Linear Model)

On détaille ici le fonctionnement global du modèle linéaire généralisé (GLM). On utilisera, car cela est plus pertinent pour l'étude, la régression logistique qui est un cas particulier des GLMs.

Le modèle linéaire $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ n'est pas adapté lorsque y_i prend des valeurs dans un intervalle précis. C'est typiquement le cas quand on cherche à évaluer une probabilité (donc un nombre entre 0 et 1). La régression logistique permet d'évaluer la probabilité $\mathbb{P}(Y = y_i | X = (x_{i1}, \cdots, x_{ip}))$ et est un cas particulier du modèle linéaire généralisé.

Contrairement au modèle linéaire classique, le modèle linéaire généralisé cherche à prédire linéairement $\mathbb{E}[Y = y_i | X = (x_{i1}, \cdots, x_{ip})]$ en fonction de x_i.

Le modèle linéaire généralisé s'énonce de la sorte :

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

avec $\mu_i = \mathbb{E}[y_i]$ et $\eta_i = g(\mu_i)$.

4.2.2.1 Choix de la distribution de la variable cible

Le modèle linéaire généralisé suppose que la variable cible Y a une distribution appartenant à la famille exponentielle, c'est-à-dire que :

$$f(y) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right]$$

avec $a(\cdot)$, $b(\cdot)$, $c(\cdot, \cdot)$ des fonctions et θ, ϕ des paramètres. Souvent $a(\phi) = \phi$.

Il est possible de montrer que :

— $\mathbb{E}[Y] = \mu = b'(\theta)$

— $\mathbb{V}[Y] = \sigma^2 = b''(\theta)\,\phi$

La loi normale est un exemple de loi appartenant à la famille exponentielle :

$$f(y) = \exp\left[\frac{y\theta - \frac{1}{2}\theta^2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\ln\left(2\pi\sigma^2\right)\right]$$

Distribution	θ_i	$b(\theta_i)$	ϕ	μ_i
Bernoulli $B(1,\pi_i)$	$\ln\left[\frac{\pi_i}{1-\pi_i}\right]$	$\ln(1+\exp\theta_i)$	1	$\pi_i = \frac{\exp\theta_i}{1+\exp\theta_i}$
Poisson $P(\lambda_i)$	$\ln\lambda_i$	$\exp\theta_i$	1	$\exp\theta_i = \lambda_i$
Normale $\mathcal{N}(\mu_i,\sigma^2)$	μ_i	$\frac{\theta_i^2}{2}$	σ^2	$\theta_i = \mu_i$
Gamma $\Gamma\left(\nu,\frac{\nu}{\mu_i}\right)$	$-\frac{1}{\mu_i}$	$-\ln(-\theta_i)$	$\frac{1}{\nu}$	$-\frac{1}{\theta_i} = \mu_i$
Inverse Gaussienne $IG(\mu_i,\sigma^2)$	$-\frac{1}{2\mu_i^2}$	$-(-2\theta_i)^{\frac{1}{2}}$	σ^2	$\frac{1}{(-2\theta_i)^{\frac{1}{2}}} = \mu_i$

TABLEAU 4.1 – Lois classiques de la famille exponentielle

4.2.2.2 La fonction de lien

Un autre paramètre intervenant dans le modèle linéaire généralisé est la fonction g, appelée **fonction de lien**. Comme son nom l'indique, elle détermine le lien entre l'espérance de la variable réponse et les prédicteurs. Lorsque $g(x) = x$, on retrouve le modèle linéaire classique.

Comme $\mathbb{E}[Y] = b'(\theta)$, il est normal de poser $g(x) = b'^{-1}(x)$ de telle sorte que :

$$\theta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

Dans ce cas, g est appelée fonction canonique.

Un des exemples les plus connus du modèle linéaire généralisé est obtenu pour $g(x) = \ln\left(\frac{x}{1-x}\right)$. g est alors appelée fonction logit et le modèle obtenu est appelé régression logistique. Cette dernière fonction de lien permet de modéliser une variable cible prenant des valeurs entre 0 et 1, comme une probabilité par exemple.

4.2.2.3 Calibration du modèle

L'objectif de la calibration est de trouver des estimateurs $\hat{\beta} = \left(\hat{\beta}_0, \cdots, \hat{\beta}_p\right)$ des coefficients du modèle. Une fois celle-ci effectuée, il est alors possible d'estimer μ_i de la façon suivante :

$$\mu_i = g^{-1}\left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}\right)$$

Les coefficients estimés sont ceux qui maximisent la vraisemblance de l'échantillon. En notant f la densité de Y, $\hat{\beta}$ est solution du problème :

$$\min_{\beta\in\mathbb{R}^{p+1}} \mathcal{L}(\beta) = \prod_{i=1}^{n} f(y_i|\mathrm{x}_i)$$

Ce problème dépend bien de β car $f(y_i|\mathrm{x}_i)$ dépend de μ_i dépendant lui-même de β.

Il est plus commode de résoudre le problème de maximisation de la log-vraisemblance $\sum_{i=1}^{n} \ln f(y_i|\mathrm{x}_i)$. Cette résolution, selon la fonction de lien, peut se résoudre de manière exacte ou par itérations successives jusqu'à aboutir à une convergence.

4.2.2.4 Sélection des variables

La sélection des variables est un procédé qui permet d'obtenir le meilleur modèle possible. Si le modèle a trop de prédicteurs, alors il aura une certaine tendance à générer le phénomène de sur-apprentissage. Sa capacité à réaliser des prédictions sur des données n'ayant pas servi à la calibration du modèle sera donc réduite. D'un autre côté, si le modèle a trop peu de prédicteurs, alors sa précision sera réduite. Celle-ci est aussi appelée biais. Par exemple, un modèle cherchant à prédire la prime d'un assuré peut être biaisé si seul l'âge est utilisé comme prédicteur. Il existe donc un compromis entre le biais du modèle et le nombre de prédicteurs utilisés.

Une des méthodes classiques de sélection des variables est le critère AIC [5]. Pour un modèle ayant q prédicteurs et en notant $\mathcal{L}(\cdot)$ sa vraisemblance, on a :

$$\text{AIC} = -2 \ln \mathcal{L}(\beta) + 2q$$

Au sens du critère d'Akaike, le meilleur modèle est celui pour lequel AIC est minimal.

4.2.2.5 Avantages et limites du modèle linéaire généralisé

Le modèle linéaire généralisé présente de nombreux avantages. Celui-ci a tout d'abord l'avantage d'être compréhensible facilement. Ce n'est pas une boîte noire. En fonction du signe des $\hat{\beta}_i$, sous réserve que le coefficient soit significatif [6], il est possible de savoir si l'augmentation d'une unité du prédicteur i va diminuer ou augmenter la valeur de la variable cible. De plus, le modèle linéaire généralisé est très utilisé, notamment en actuariat, et ses performances sont donc très bien connues. Les spécialistes savent donc comment rendre un tel modèle le plus performant possible alors que leurs connaissances seront plus limitées pour d'autres modèles comme les SVM. De plus, les méthodes de sélection des variables sont bien connues et justifiées. Contrairement à certains algorithmes d'apprentissage automatique, la sélection des variables se fait alors aisément.

Cependant, le modèle linéaire généralisé a ses limites. Comme son nom l'indique, il suppose une relation linéaire entre les variables. La variable cible peut dépendre du carré d'un prédicteur plutôt que de ce prédicteur lui-même. Sans transformation préalable sur les variables, et qui doit être justifiée par l'intuition ou l'observation, le modèle linéaire généralisé n'est donc pas au mieux de ses performances. Contrairement au modèle linéaire généralisé, les algorithmes d'apprentissage automatique ne font pas d'à priori sur les variables et certains captent même les non-linéarités intrinsèques aux données.

4.2.3 Évaluation des modèles de classification

Deux outils, la courbe ROC et la matrice de confusion, sont utiles pour mesurer les performances d'un modèle de classification.

5. Akaike Information Criterion, ou critère d'information d'Akaike
6. Selon un test : $H_0 : \beta_i = 0$ contre $H_1 : \beta_i \neq 0$.

4.2.3.1 Courbe ROC

Une façon d'évaluer la performance d'un modèle de classification et de comparer sa performance avec un autre modèle est d'utiliser la courbe ROC (pour *Receiver Operating Characteristic*).

Un modèle de classification construit une fonction de score $s(\cdot)$ et la prédiction entre les classes 0 et 1 est obtenue en utilisant le classifieur binaire :

$$f(\mathbf{x}) = \mathbb{I}\{s(\mathbf{x}) > S\}$$

avec S un seuil.

La courbe ROC représente le taux de vrais positifs en fonction du taux de faux positifs lorsque le seuil S varie.

L'aire sous la courbe obtenue (notée AUC pour *Area Under the Curve*) représente une mesure permettant de comparer deux modèles. Un modèle donné sera meilleur qu'un autre si son AUC est supérieure. La mesure AUC est comprise entre 0,5 et 1. Un classifieur aléatoire [7] aura une AUC de 0.5 tandis que le classifieur idéal aura une AUC égale à 1.

4.2.3.2 Matrice de confusion

La matrice de confusion permet de visualiser les performances d'un modèle de classification. Dans le cas où le modèle est appliqué à deux classes (0 ou 1), la matrice de confusion permet de mettre en regard les prédictions réalisées avec la classe réelle des observations.

Un modèle sera d'autant meilleur que les 0 sont prédits 0 et les 1 prédits 1. Dans ce cas, la matrice de confusion aura des 0 dans sa diagonale.

On donne ici un exemple de matrice de confusion ainsi que les erreurs de classification associées :

		Prédit		
		0	**1**	**Erreur**
Réel	**0**	1201	323	26,9%
	1	208	2202	9,5%

Tableau 4.2 – Exemple de matrice de confusion

Pour évaluer un modèle, il est préférable de construire la matrice de confusion sur les données d'une base de test, n'ayant pas servi à construire le modèle.

7. C'est-à-dire prédisant 0 ou 1 au hasard.

4.3 Mise en place du modèle explicatif de la résiliation

Une fois la théorie sur les modèles de classification détaillée, il est possible de passer à la modélisation à proprement parler.

L'objectif que l'on se fixe pour le modèle de classification est de savoir quels contrats vont être résiliés à horizon 1 an.

On considère un contrat souscrit avant le 1er janvier 2010. Celui-ci peut ou non être résilié au cours de la période d'observation allant de début 2010 à fin 2014.

Prenons par exemple le cas où le contrat est résilié le 30 juin 2012. Le contrat aura été observé 2 ans et demi, soit 3 ans en découpant la période d'observation en 5 années. Durant la première et la deuxième période, la résiliation du contrat n'a pas été observée. La variable cible **Résiliation** vaut donc 0. Sur la troisième période, la résiliation du contrat a été observée et la variable cible vaut donc 1.

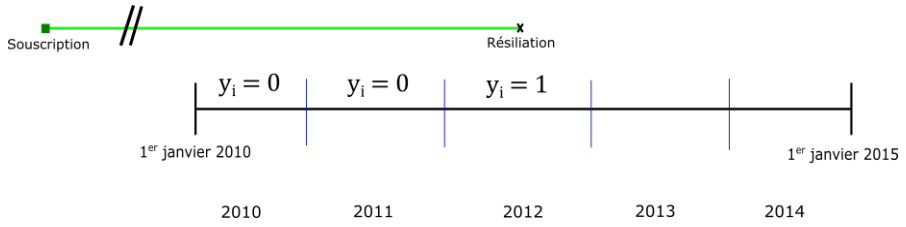


FIGURE 4.15 – Format des données appliqué au contrat pris en exemple

Pour cela, le format des observations dans la base de données doit être adapté. Avec l'exemple précédent, le contrat en question sera répliqué 3 fois dans la base de données. Chacune des 3 lignes sera composée du contrat vu à la période considérée et de la variable cible qui vaudra 0 en 2010 et 2011 mais 1 en 2012.

Un contrat non résilié durant les 5 années se retrouvera en 5 lignes dans la base de données utilisée pour le modèle. Pour chacune de ces lignes, la variable cible vaudra 0. De même, un contrat résilié dès l'année 2010 se retrouvera en un exemplaire dans la base de données et la variable cible vaudra 1.

4.3.1 Adaptation des données à un modèle de classification

Chaque contrat est dupliqué autant de fois que d'années entières durant lesquelles il est observé. Dans le pire des cas, un contrat sera donc dupliqué 5 fois.

Chacune des lignes associées à un contrat donné représente le contrat vu à l'année en question.

Certaines variables utilisées lors de la segmentation sont vues au 1er janvier 2010. C'est par exemple le cas de :

— la duration du contrat,
— l'âge de l'assuré,
— l'âge de la voiture,
— etc...

Il est essentiel, lors de la création de la base, de mettre à jour les champs évoluant au cours du temps.

De plus, afin de mesurer leur impact sur la résiliation, il est nécessaire de créer des variables liées aux contacts et à la sinistralité.

Plusieurs difficultés en découlent. Tout d'abord, la base contacts couvre seulement les années 2012 à 2014. Il n'est donc pas possible à priori de l'utiliser.

De plus, l'objectif est de disposer des variables suivantes pour la sinistralité :

— Nombre de sinistres sur la période,
— Type de sinistre constaté,
— Sinistres responsables ou non.

Si plus d'un sinistre est constaté et que ceux-ci sont de types différents, la question est de savoir ce que va contenir la variable **Type de sinistre constaté**. De même, si aucun sinistre n'est constaté, que doit contenir cette variable ?

La réponse à ces questions nécessite de faire un choix qui va créer un biais mais qui est indispensable à l'exploitation des données sur la sinistralité.

La difficulté majeure de ces approches est que la prédiction entre deux classes ayant des proportions différentes résultera en un modèle biaisé envers la classe majoritaire. En effet, dans le cas d'un arbre, pour lequel le choix entre les classes se fait par vote majoritaire, la classe la plus nombreuse sera prédite plus souvent que la classe minoritaire. Le modèle sera donc biaisé de façon optimiste pour ce qui est de prédire l'appartenance à la classe majoritaire. Dans un cas extrême, le modèle pourrait même, sur un échantillon de test ayant lui aussi des proportions très différentes entre les classes, prédire que toutes les observations appartiennent à la classe majoritaire. L'erreur de classification sur la classe minoritaire sera donc égale à 100%.

Il est donc nécessaire de modifier la structure de la base d'apprentissage, en diminuant le nombre d'observations de la classe majoritaire ou en augmentant le nombre d'observations de la classe minoritaire.

Des tests préliminaires nous font préférer la seconde solution qui donne de meilleurs résultats malgré un temps d'exécution plus important. Afin d'obtenir autant d'observations des deux classes, on tire aléatoirement avec remise parmi les observations correspondantes à des contrats résiliés et l'on concatène la base ainsi créée avec celle des contrats résiliés.

4.4 Résultats, analyses et limites du modèle explicatif

Les deux algorithmes utilisés sont comparés dans cette sous-partie afin de valider l'utilité des algorithmes de l'apprentissage automatique. L'objectif est d'expliquer les facteurs de la résiliation dans les deux poches clients choisies afin d'en déduire des règles transposables en actions marketing. La poche h regroupe des clients résiliant beaucoup tandis que la poche i est formée par des clients résiliant relativement peu.

On résume avec la figure 4.16 le déroulé de la partie qui suit :

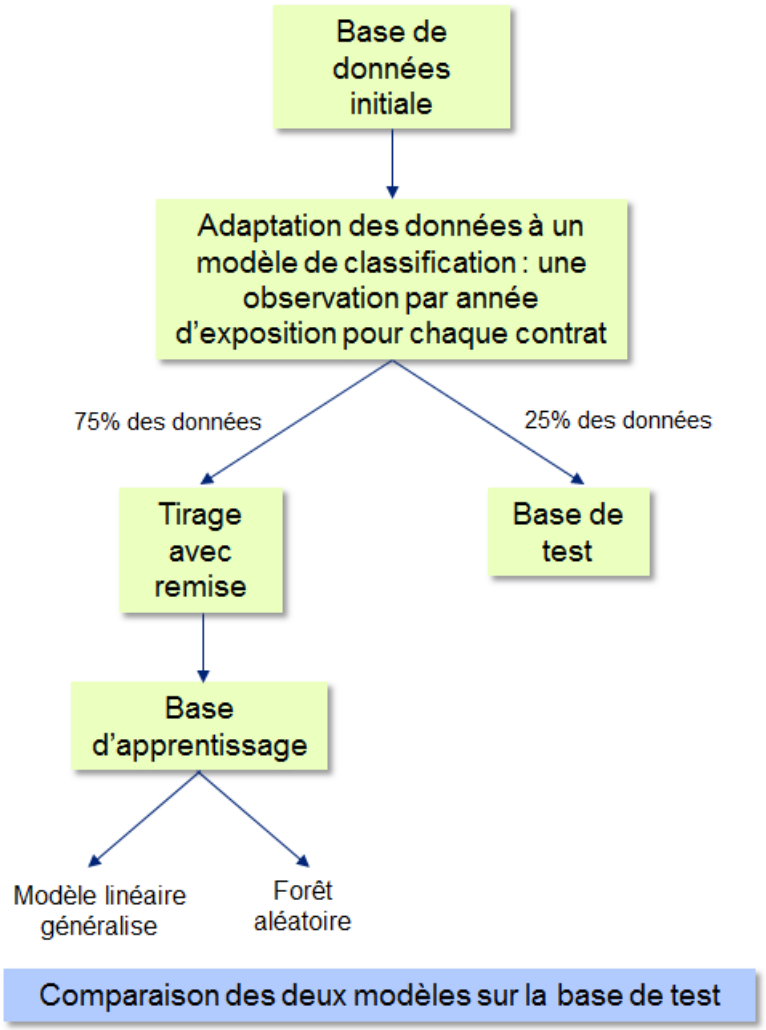


FIGURE 4.16 – Schéma explicatif du travail effectué

4.4.1 Analyse de la résiliation dans la poche *h*

La poche client *h* est constituée de clients ayant une forte propension à résilier. D'après les résultats déjà obtenus, les clients de cette poche sont des clients âgés et fidèles sans réduction tarifaire tels que :

— **NbContratAutreIARD** est égal à 0,
— **DurationContrat** est inférieure à 15 ans,
— **NbContratAuto** est supérieur ou égal à 2,
— **AgeAssure** est inférieur à 67 ans,
— Le coefficient de **BonusMalus** est supérieur à 50.

Comprendre les facteurs de la résiliation propres à ces assurés permet de savoir quelles sont les caractéristiques propres à l'assuré qui vont augmenter ses chances de résilier. Un tel modèle, comme la plupart des modèles, ne permet pas d'expliquer pourquoi un assuré a résilié car cela impliquerait d'avoir une relation de causalité entre les facteurs de la résiliation (sinistre d'un type particulier, âge, ...) et la résiliation du client. Ce que l'on pourra tirer de ce modèle, c'est l'éventuelle corrélation entre un facteur et la résiliation.

Afin de s'assurer que le modèle est apte à prédire sur de nouvelles données, on divise la base en une base d'apprentissage (75% des données) et une base de test (les 25% restants). Si l'erreur de classification entre *Résilié* et *Non résilié* est faible sur la base de test, alors le modèle est bien construit et ne sur-apprend pas. Dans ce cas, il est possible de décrire les influences relatives des différentes variables sur la résiliation à l'aide du modèle.

La poche client *h* contient 8842 contrats, ce qui donne avant adaptation des données au modèle[8], une base d'apprentissage de 6632 contrats et une base de test de 2210 contrats. La base d'apprentissage est adaptée au modèle, de telle sorte que la proportion *Résilié*/*Non résilié* soit de 50%/50%. La base de test est adaptée au modèle mais aucun tirage avec remise n'est effectué pour obtenir un équi-proportion des classes *Résilié*/*Non résilié*.

Les prédicteurs sont les variables utilisées dans la segmentation et les données sur les sinistres.

4.4.1.1 Régression logistique

Résultats La régression logistique, construite sur la base d'apprentissage, donne 34,1% d'erreur de classification sur la base d'apprentissage et 34,8% d'erreur de classification sur la base de test. Comme souvent, l'erreur est plus importante sur la base de test que sur la base d'apprentissage. Le tableau 4.3 montre la matrice de confusion sur la base test pour ce modèle de régression logistique. Les performances semblent équivalentes pour ce qui est de prédire une classe ou l'autre. Les performances de prédiction ne sont toutefois pas très bonnes car la régression logistique n'arrive pas à bien prédire l'appartenance à une classe donnée dans environ un tiers des cas.

8. Voir 4.3.1

		Prédit		
		0	**1**	**Erreur**
Réel	**0**	3824	2055	34,9%
	1	307	601	33,8%

TABLEAU 4.3 – Matrice de confusion sur la base test - Régression logistique

La figure 4.17 montre la courbe ROC obtenue sur la base test. L'AUC est égale à 0,72, ce qui doit être comparé avec l'AUC pour la forêt aléatoire afin de conclure.

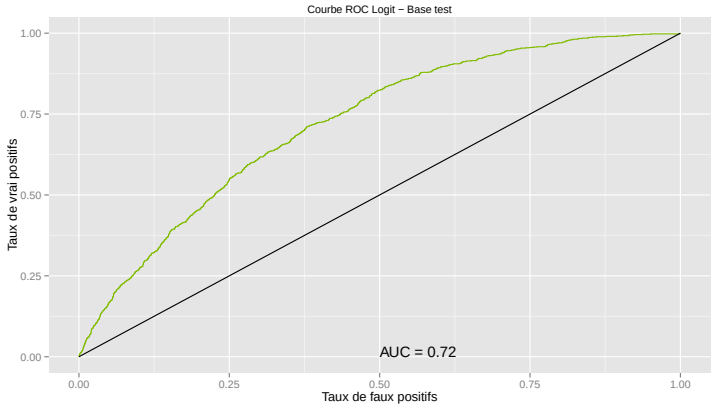


FIGURE 4.17 – Courbe ROC sur la base test - Régression logistique

On détaille, en fonction de leur p-value, les 30 variables les plus significatives[9]. Les p-values de ces variables sont toutes inférieures à 0,001.

Interprétation L'âge de la voiture semble être très significatif pour expliquer la probabilité de résiliation. D'après le signe du coefficient, plus la voiture est ancienne, plus la probabilité de résiliation est élevée. Cela est conforme à l'intuition. De plus, le **LienAutreAssure** est très significatif et a un impact négatif sur la résiliation. Le coefficient de **ReducTarif** est négatif, ce qui est conforme à l'intuition. Au contraire, on voit que si **BonusMalus** est élevé, le risque de résiliation augmente. Cela n'est pas forcément intuitif. Il convient aussi de remarquer la présence de variables liées au **TypeSinistres** par les 30 variables les plus significatives, notamment le type de sinistre « Remplacement Glace » qui tend à montrer que les assurés ayant eu dans l'année ce type de sinistre sont moins enclins à résilier qu'un assuré identique ayant un autre type de sinistre ou pas de sinistres. Il s'agit du même effet marginal pour le type de sinistre « Réparation glaces », « Choc entre deux véhicules » et « Pare-brise ». Au contraire, un assuré ayant eu un sinistre de type « Dommage aux tiers » semble plus enclin à résilier que les autres.

Les impacts marginaux des différentes variables sont assez utiles. En effet, des assurés ayant eu un certain type de sinistres ont plus de chances de résilier que des assurés ayant les mêmes caractéristiques excepté pour cette variable. Il est donc possible de

9. Les variables catégoriques étant transformées en *dummy variables*, les 30 variables contiennent plusieurs modalités d'une même variable catégorique.

Variable	Coefficient	Erreur type	z value
AgeVoiture	0.10	0.00	33.05
LienAutreAssure	-0.69	0.03	-27.58
ReducTarif	-2.56	0.10	-26.65
BonusMalus	0.02	0.00	22.38
UsageVehicule=Prive.et.Pro	0.39	0.03	13.96
AgeAdhesion	0.02	0.00	9.48
ReducFamille	0.97	0.11	8.93
PuissanceVehicule	0.00	0.00	8.78
AgeObtentionPermis	-0.02	0.00	-8.51
SitFamiliale=Marié	0.30	0.04	8.28
UsageVehicule=Professionnel	1.02	0.13	7.90
AgeAssure	0.01	0.00	6.32
CSP=RETRAITE	-0.62	0.10	-6.02
Option=6	-0.58	0.11	-5.42
NbConducteurs	0.08	0.02	5.21
MajorConducteurNovice	-0.34	0.07	-5.14
CSP=OUVRIER	-0.49	0.10	-4.88
TypeSinistre=REMPLACEMENT.GLACE	-0.75	0.16	-4.80
CSP=EMPLOYE	-0.47	0.10	-4.76
GroupeMarques=ToutPublicCourante	0.70	0.15	4.66
Coefficient constant	-4.36	0.99	-4.40
TypeSinistre=DOMMAGE.TIERS	0.82	0.19	4.32
Option=7	-0.45	0.11	-4.18
NbContratAuto	-0.15	0.04	-4.13
TauxSinResp	-0.29	0.08	-3.83
TypeSinistre=REPAR.GLACES	-0.90	0.25	-3.60
TypeSinistre=PARE.BRISE	-0.51	0.14	-3.57
Moteur=DIESEL	1.49	0.43	3.47
CSP=CADRE	-0.42	0.12	-3.46
TypeSinistre=CHOC.2.VEH	-0.40	0.12	-3.44

Tableau 4.4 – 30 variables les plus significatives

décider de ne pas chercher à retenir les premiers et préférer contacter les seconds. En pratique, le modèle donne la probabilité de résiliation d'un assuré donné, ce qui permet d'ordonner les assurés en fonction de leur risque de résiliation.

Le modèle permet alors de gagner en efficacité opérationnelle, car il permet de concentrer les actions de rétention client sur les assurés dont le risque de résiliation est le plus important. Dans le contexte de la loi Hamon, qui met une pression supplémentaire sur les assureurs, une telle connaissance est utile. Bien entendu, il convient de tester le modèle construit sur des nouvelles données afin de conclure sur la pertinence du modèle. La segmentation du portefeuille et la constitution des poches clients ayant été obtenues à partir de toutes les données pour ne pas biaiser les résultats présentés dans le cadre du mémoire, tester le modèle prédictif sur de nouvelles données n'est donc pas possible.

Comparons les résultats obtenus avec la régression logistique en implémentant le modèle utilisant les forêts aléatoires.

4.4.1.2 Forêt aléatoire

Résultats La forêt aléatoire est construite avec $m = 2$ variables tirées aléatoirement à chaque découpage (car la heuristique $m = \lfloor\sqrt{30}\rfloor = 5$ donnait des mauvais résultats) et 500 arbres.

Construite sur la base d'apprentissage, elle donne 29,7% d'erreur de classification sur la base d'apprentissage et 40,0% d'erreur de classification sur la base de test. De même que pour la régression logistique, l'erreur est plus importante sur la base de test que sur la base d'apprentissage. Le tableau 4.5 montre la matrice de confusion sur la base test pour ce modèle de forêt aléatoire. L'erreur de prédiction est plus importante pour ce qui est de prédire la non-résiliation (41,3%). Etant donné que l'on préfère un modèle qui prédit bien l'évènement de résiliation, ce modèle est satisfaisant.

		Prédit		
		0	**1**	**Erreur**
Réel	**0**	3446	2433	41,3%
	1	285	623	31,4%

TABLEAU 4.5 – Matrice de confusion sur la base test - Forêt aléatoire

La forêt aléatoire est plus performante que la régression logistique pour ce qui est de prédire la classe 1 (33,8% d'erreur contre 31,4% d'erreur pour la classe 0). L'AUC vaut 0.7 pour la forêt aléatoire, ce qui est inférieur à l'AUC de la régression logistique.

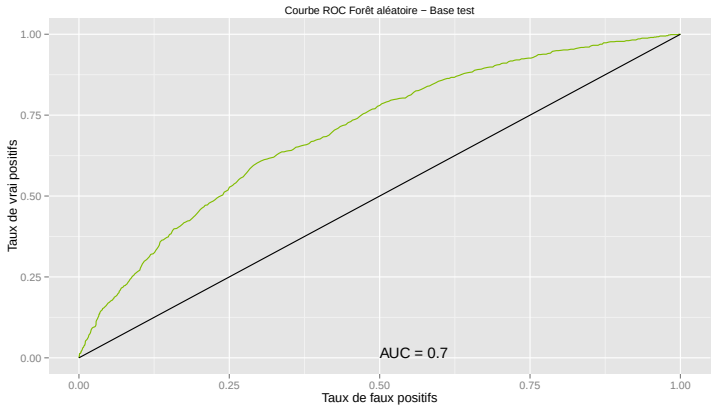


FIGURE 4.18 – Courbe ROC sur la base test - Forêt aléatoire

L'importance relative des différentes variables est affichée est figure 4.19. L'importance est calculée à l'aide des deux approches décrites dans la théorie : l'approche par randomisation et celle par critère d'impureté.

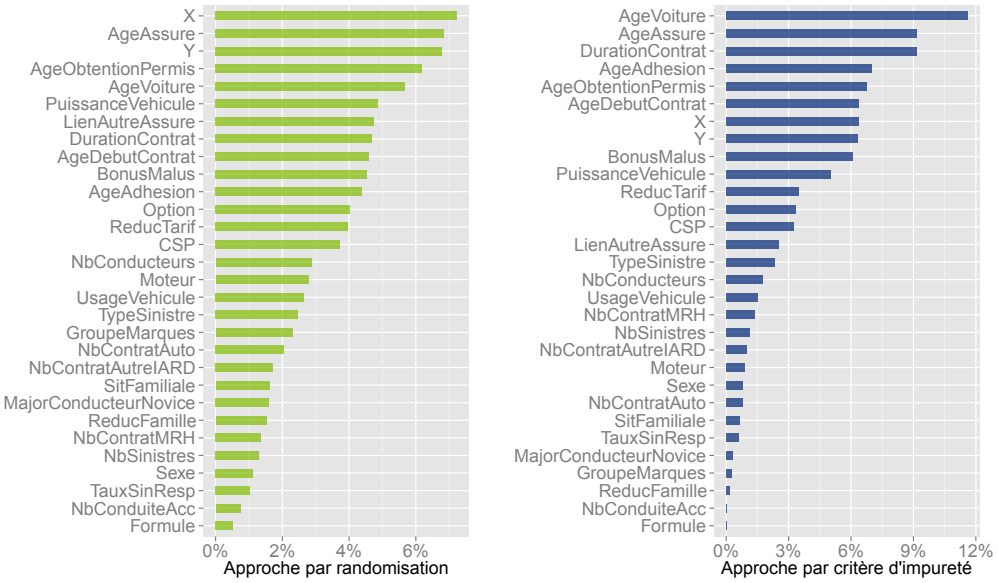


FIGURE 4.19 – Importance relative des variables

Interprétation Tout d'abord, en regardant l'approche par critère d'impureté, on trouve des résultats semblables à ceux de la régression logistique. **AgeVoiture** semble être une variable très significative pour expliquer le risque de résiliation, de même qu'**AgeObtentionPermis**. Les deux approches semblent montrer l'importance de la variable **AgeAssure**.

Plus étonnant, les variables géographiques **X** et **Y** semblent avoir un pouvoir prédictif certain, puisqu'elles sont haut placées dans l'importance des variables pour les deux approches. Ces variables n'apparaissaient cependant pas dans les 30 variables les plus significatives de la régression logistique. Cela peut être expliqué par le fait que le caractère linéaire de la régression logistique n'arrive pas à capter le pouvoir prédictif de ces variables. Il est possible, car plusieurs observations peuvent être associées à un même contrat, que les observations de la base test soient associées à des contrats de la base d'apprentissage. L'algorithme risque donc de prédire 0 pour une observation de la base de test en utilisant la connaissance apportée par une observation de la base d'apprentissage associée au même contrat. Cependant, cet effet tend à biaiser l'erreur de prédiction de la classe 0 seulement puisque, pour un contrat donné, plusieurs observations (de 1 à 5) peuvent être associées à la variable cible 0 tandis qu'il y a au maximum une seule observation avec 1 en variable cible. L'erreur de classification sur la classe 1 étant faible, on en déduit que les variables géographiques ont tout de même une importance dans la prédiction de la résiliation. Il serait intéressant d'étudier, à ce titre, les variations géographiques de la probabilité de résiliation.

Les variables **LienAutreAssure**, **BonusMalus** et **ReducTarif** semblent aussi avoir de l'importance. L'intuition peut donner le signe de l'impact marginal de ces différentes variables mais la forêt aléatoire ne donne pas naturellement l'information sur ce signe.

Comparaison des deux algorithmes Les résultats trouvés avec les deux algorithmes sont relativement semblables, ce qui tend à prouver l'importance des variables

comme l'**AgeVoiture**. La forêt aléatoire a donné de bien meilleures performances que la régression logistique sur les mêmes données. Cet algorithme a réussi à capter le pouvoir prédictif des variables **X** et **Y**. Cela permet de conclure sur l'apport des forêts aléatoires pour améliorer les modèles de classification. La prochaine étape consiste à comparer les résultats obtenus sur la poche h avec les résultats sur la poche i.

4.4.2 Analyse de la résiliation dans la poche i

La poche client i est constituée de clients ayant une faible propension à résilier (taux de résiliation de 17,3% pour une duration moyenne de 27,2 ans). D'après les résultats déjà obtenus, les clients de cette poche sont des clients âgés et fidèles sans réduction tarifaire tels que :

— **NbContratAutreIARD** est égal à 0,
— **DurationContrat** est supérieure à 15 ans,
— **AgeVoiture** est inférieur ou égal à 10,
— **AgeDebutContrat** est inférieur à 50 ans.

Au sein d'un cluster ayant un taux de résiliation important, cette poche bouscule l'intuition selon laquelle les assurés du cluster n° 3 (les clients âgés et fidèles sans réduction tarifaire) sont peu rentables car résilient beaucoup. Mieux connaître les comportements de résiliation associés à ce type de clientèle est donc plutôt intéressant.

Cette poche contient 18146 contrats. Le nombre important d'observations après adaptation des données est dû tout d'abord à la taille initiale de la poche i (deux fois plus grande que la poche h), ensuite au faible taux de résiliation dans cette poche, 17,3%, fait que le tirage avec remise effectué lors de l'adaptation des données démultiplie la taille de la base (car les observations résiliées sont tirées avec remise, jusqu'à avoir autant d'observations résiliées que celles non résiliées). Cette démultiplication de la taille de la base dans le cas d'un faible taux de résiliation est une des faiblesses du modèle.

4.4.2.1 Régression logistique

Résultats La régression logistique, construite sur la base d'apprentissage, donne 35,0% d'erreur de classification sur la base d'apprentissage et 90,2% d'erreur de classification sur la base de test. L'erreur est très importante sur la base de test et la régression logistique n'est donc pas très adaptée aux données. Le tableau 4.6 montre la matrice de confusion sur la base test pour ce modèle de régression logistique. Les performances semblent équivalentes pour ce qui est de prédire une classe ou l'autre. Les performances de prédiction ne sont toutefois pas très bonnes car la régression logistique n'arrive pas à bien prédire l'appartenance à une classe donnée dans environ un tiers des cas.

La courbe ROC obtenue sur la base test donne une AUC de 0.5. Elle est affichée en figure 4.20. La régression logistique est très peu performante dans ce cas car elle est quasiment équivalente à un classifieur attribuant aléatoirement à une observation aux classes 0 ou 1.

		Prédit		
		0	**1**	**Erreur**
Réel	**0**	946	14829	94,0%
	1	42	675	5,9%

TABLEAU 4.6 – Matrice de confusion sur la base test - Régression logistique

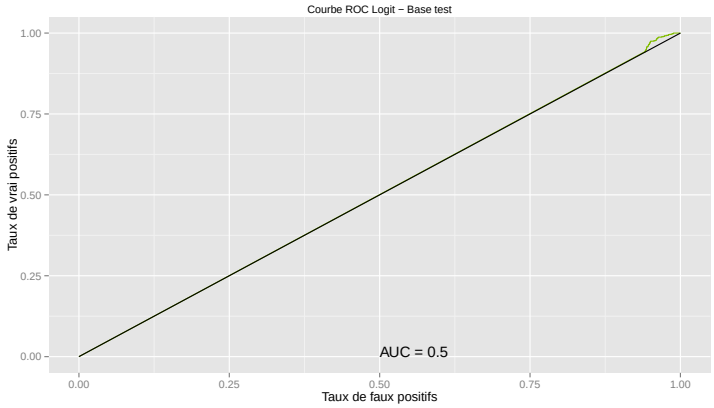


FIGURE 4.20 – Courbe ROC sur la base test - Régression logistique

Interprétation Le modèle permet cependant d'obtenir les 30 variables les plus significatives, classées par p-value. Toutes les p-values sont inférieures à 0,001. La variable la plus significative semble être **ReducTarif**. Le cluster n° 3 étant globalement un cluster d'assurés sans réduction tarifaire, le signe du coefficient nous montre qu'avoir une réduction tarifaire diminue fortement, pour ces assurés, la probabilité de résiliation. Cela est conforme à l'intuition. Cette variable était aussi très significative pour les assurés de la poche h, ce qui prouve que cette variable a un impact certain, négatif qui plus est, sur la propension à résilier. On retrouve les variables **AgeVoiture**, **BonusMalus** parmi les variables les plus significatives, comme dans l'autre poche client étudiée. La variable **NbContratAuto**, dans cette poche i, semble très significative, contrairement à la poche h dans laquelle sa significativité était moins marquée. Cela est normal car la poche i contient des contrats ayant un nombre de contrats auto supérieur à la moyenne tandis que cette variable est globalement égale à 1 dans la poche h. Dans cette poche client, on note que **NbSinistres** est significative contrairement au cas de la poche h ou elle ne l'était pas du tout (p-value de l'ordre de 0.5). La probabilité de résiliation de ce type d'assuré semble donc être diminuée si l'assuré a eu un nombre de sinistres important dans l'année. On note aussi que, contrairement aux résultats de la poche h, la variable **AgeObtentionPermis** n'explique pas la résiliation dans cette poche puisqu'elle ne fait pas partie des 30 variables les plus significatives. De même, les variables liées à la classe socio-professionnelle n'expliquent pas la résiliation dans cette poche, bien que celle-ci ne soit pas homogène en termes de classe socio-professionnelle. Ceci est un résultat pour la poche h, dans laquelle on voit alors que connaître la classe socio-professionnelle apporte de l'information sur la risque de résiliation.

4.4.2.2 Forêt aléatoire

Résultats La forêt aléatoire est construit avec $m = \lfloor\sqrt{30}\rfloor = 5$ variables tirées aléatoirement à chaque découpage.

Variable	Coefficient	Erreur type	z value
ReducTarif	-2.25	0.04	-57.03
AgeAssure	0.06	0.00	45.18
AgeVoiture	0.08	0.00	38.65
BonusMalus	0.06	0.00	33.66
NbContratAuto	-0.51	0.03	-18.93
LienAutreAssure	-0.24	0.02	-14.32
Sexe=Homme	-0.24	0.02	-13.88
UsageVehicule=Prive.et.Pro	0.34	0.03	12.79
NbContratMRH	-0.13	0.01	-9.73
NbConducteurs	-0.07	0.01	-8.08
PuissanceVehicule	0.00	0.00	7.02
UsageVehicule=Professionnel	0.87	0.13	6.93
NbSinistres	-0.48	0.07	-6.83
TypeSinistre=DOMMAGE.TIERS	1.15	0.19	6.16
TypeSinistre=STATIONNEMENT	-0.79	0.13	-6.06
TauxSinResp	-0.45	0.07	-6.03
ReducFamille	3.99	0.75	5.30
TypeSinistre=METEO.OU.INCENDIE	0.95	0.19	5.10
TypeSinistre=CARAMBOLAGE	0.76	0.15	4.94
Option=4	-0.22	0.05	-4.57
TypeSinistre=CHOC.2.VEH	-0.46	0.10	-4.45
TypeSinistre=Pas.de.Sinistre	-0.50	0.13	-3.97
AgeDebutContrat	-0.01	0.00	-3.87
TypeSinistre=REPAR.GLACES	-0.58	0.15	-3.81
Option=2	-0.13	0.04	-3.52
GroupeMarques=ToutPublicTresCourante	-0.29	0.08	-3.50
TypeSinistre=PARE.BRISE	-0.37	0.12	-3.07
TypeSinistre=REMPLACEMENT.GLACE	-0.36	0.13	-2.87
NbContratAutreIARD	-0.14	0.05	-2.60
Option=7	0.17	0.07	2.59

Tableau 4.7 – 30 variables les plus significatives

Construite sur la base d'apprentissage, elle donne 31,6% d'erreur de classification sur la base d'apprentissage et 18,3% d'erreur de classification sur la base de test. L'erreur est plus importante sur la base d'apprentissage que sur la base de test, ce qui n'est pas habituel. Le tableau suivant montre la matrice de confusion sur la base test pour ce modèle de forêt aléatoire. L'erreur de prédiction est plus importante pour ce qui est de prédire la résiliation (63,5%).

		Prédit		
		0	**1**	**Erreur**
Réel	**0**	13202	2573	16,3%
	1	456	261	63,5%

Tableau 4.8 – Matrice de confusion sur la base test - Forêt aléatoire

La courbe ROC donne une AUC égale à 0,67. Dans le cas de cette poche, le résultat est meilleur que pour la régression logistique.

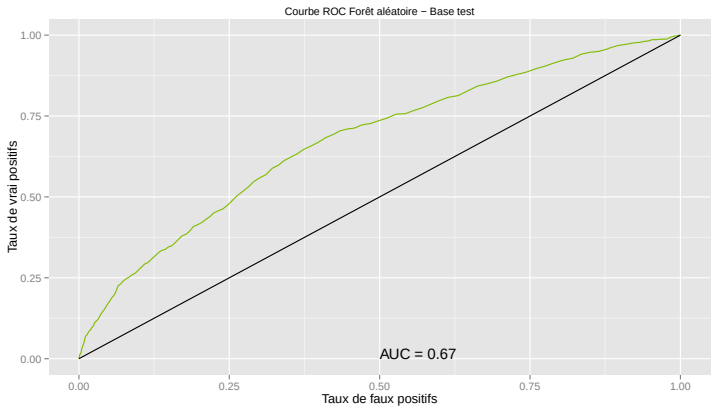


FIGURE 4.21 – Courbe ROC sur la base test - Forêt aléatoire

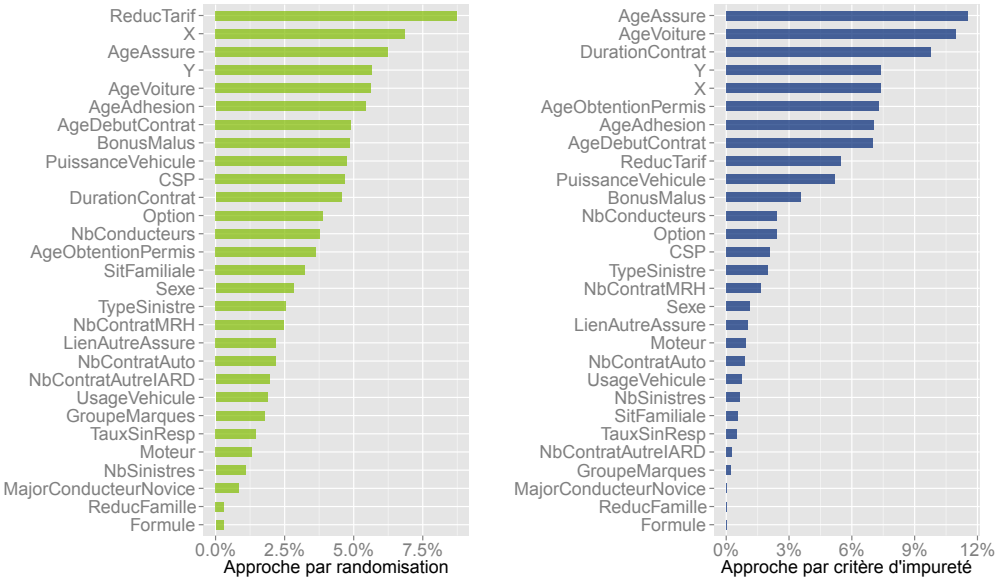


FIGURE 4.22 – Importance relative des variables

Interprétation En figure 4.22, on retrouve des résultats très similaires entre l'importance des variables par l'approche du critère d'impureté pour les deux poches i et h. Avec l'approche par randomisation, on voit que la variable **ReducTarif** a une importance certaine, que l'on retrouve de façon moins marquée avec l'approche par critère d'impureté mais très marquée avec la régression logistique. On fait globalement les mêmes interprétations que celles qui découlent de la régression logistique.

4.4.3 Comparaison des résultats dans les poches *h* et *i*

Les résultats obtenus dans les deux poches clients étudiées montrent que l'âge de la voiture est une variable déterminante pour expliquer la résiliation des assurés. Ceci s'explique très bien car un assuré va rompre son contrat auto si sa voiture devient trop ancienne pour rouler. Si ce dernier ne choisit pas de faire assurer sa nouvelle voiture (s'il en achète une nouvelle) chez le Partenaire, une résiliation sera enregistrée. C'est cet effet qui est mesuré dans les données.

De plus, la variable **ReducTarif** semble être significative dans les deux poches pour expliquer la résiliation. Un assuré ayant une réduction tarifaire va, comme le signe du coefficient dans la régression logistique tend à le montrer, avoir tendance à moins résilier que s'il n'en avait pas eu. Il est cependant nécessaire de rappeler les assurés du cluster n^o 3 sont globalement sans réduction tarifaire.

Un bonus-malus élevé tend, dans les deux poches, à augmenter le risque de résiliation. Au contraire, les assurés des deux poches ayant un lien avec d'autres assurés voient leur probabilité de résiliation réduite. Cette dernière variable est relativement significative dans les deux poches.

Une autre variable qui est significative, quel que soit le modèle employé et la poche considérée, est l'âge de l'assuré. D'après le signe du coefficient des deux régressions logistiques, la probabilité de résiliation semble augmenter d'autant plus que l'âge est grand.

La régression logistique dans la poche i nous a montré que le nombre de sinistres était significatif pour expliquer la résiliation, ce que l'on ne retrouvait pas avec la poche h. De plus, on compte parmi les 30 variables les plus significatives pour les régressions logistiques sur les deux poches 5 modalités de **TypeSinistre** dans la poche h contre 9 dans la poche i. La survenance d'un sinistre semble donc être beaucoup plus explicative de la résiliation pour les assurés de la poche i que pour les assurés de la poche h.

L'étude des facteurs de la résiliation dans les deux poches a montré que certaines variables étaient plus importantes que les autres pour expliquer l'évènement de résiliation. Pour une poche donnée, la comparaison des résultats obtenus entre régression logistique et forêt aléatoire a permis de valider les influences des différentes variables. Pour la poche i, l'algorithme de forêt aléatoire a donné des meilleures performances que la régression logistique, ce qui est un atout si l'on cherche à prédire l'évènement de résiliation plutôt que de décrire ses causes. La régression logistique semble plus utile pour un modèle descriptif que les forêts aléatoires car elle donne le signe de l'influence des variables via leurs coefficients respectifs. Dans le cadre d'un mémoire d'actuariat, la régression logistique semble donc être plus adaptée. Cependant, si l'objectif est de mettre en place un modèle prédisant la résiliation avec l'objectif de fournir une liste d'assurés risquant de résilier prochainement, alors il est préférable d'utiliser des algorithmes d'apprentissage automatique.

4.4.4 Limites des modèles explicatifs utilisés

Deux algorithmes ont été utilisés pour prédire l'évènement de résiliation. Comme on l'a vu, les performances de ces modèles sont limitées même si l'algorithme des forêts aléatoires semble être plus performant que la régression logistique qui n'arrive pas à bien capter la structure des données.

La grande limite de ces modèles est qu'un contrat peut être associé à un nombre de 1 à 5 observations. Celles-ci ne sont pas extrêmement différentes les unes des autres, sauf pour les variables qui varient au cours du temps (**DurationContrat**, **NbSinistres**, **TypeSinistre**, **TauxSinResp**, **AgeVoiture** et **AgeAssure**). Les algorithmes ont donc du mal à distinguer une observation de variable cible 0 d'une observation de variable cible 1.

De plus, la transformation effectuée sur les données fait que les observations ayant le label 1 (signifiant une résiliation) sont en faible proportion par rapport aux observations ayant le label 0 dans le cas où le taux de résiliation dans la base initiale est faible. L'astuce consistant à effectuer un tirage avec remise pour obtenir une proportion équi-répartie entre les deux classes ne compense pas totalement le problème induit par le déséquilibre de proportion entre les 0 et les 1. C'est à cause de ce phénomène que les performances de prédiction dans la poche i, au taux de résiliation de 17,3%, sont relativement faibles.

4.5 Conclusion sur l'analyse des poches clients

Cette partie a consisté en la recherche de poches clients dont le comportement de résiliation était contraire à l'intuition. Au sein d'un cluster d'assurés relativement intéressant pour le Partenaire et résiliant plus que la moyenne, l'approche a permis de trouver notamment deux poches : l'une regroupant des clients ayant une forte propension à résilier (la poche h) et l'autre constituée de clients résiliant très peu (la poche i).

La robustesse du clustering effectué a montré que ces poches étaient relativement bien définies. De plus, l'utilisation d'un arbre de classification permet de donner les caractéristiques des assurés des différentes poches. Grâce aux règles ainsi définies, il est possible d'attribuer chaque client à une poche donnée. Cela permet de savoir si un client va plutôt être de ceux qui résilient beaucoup ou de ceux qui vont peu résilier.

La modélisation qui a suivi a ensuite permis d'en apprendre plus sur les facteurs de la résiliation. Certaines variables comme l'âge de la voiture, le lien avec un autre assuré ou même la réduction tarifaire semblaient bien expliquer les différences dans la propension à résilier. Les performances des différents modèles étaient par contre perfectibles.

L'approche présentée est très utile pour améliorer la connaissance de son portefeuille et peut être appliquée dans de nombreux domaines. Elle permet ici d'améliorer la connaissance de la résiliation dans le contexte Hamon qui impose aux assureurs de mieux connaître leurs clients.

Une des limites de cette approche est qu'il est nécessaire, pour exploiter les résultats sur les poches clients, de mettre en regard taux de résiliation et durée du contrat. En effet, comme précisé, un faible taux de résiliation pour une faible durée d'assurance n'apporte pas forcément de l'information. Idem pour ce qui est d'un fort taux de résiliation couplé avec une durée d'assurance moyenne élevée. Cette analyse nécessitant de s'intéresser à ces deux indicateurs a parfois rendu l'interprétation difficile.

Le principal apport de valeur de l'approche est la découverte des poches clients au sein du cluster n° 3. Par exemple, la découverte de la poche i, ayant un taux de résiliation de 17,3% pour une durée d'assurance moyenne de 27,2 ans, permet à l'assureur de savoir qu'il doit particulièrement prendre soin de ces clients qui résilient très peu.

Conclusion générale

Tout au long de ce mémoire, nous avons étudié la résiliation en assurance non-vie à l'aide de techniques issues de l'apprentissage automatique (*machine learning*). L'approche employée nous a permis de mieux comprendre les comportements de résiliation au sein du portefeuille étudié.

Le principal apport de l'étude est l'identification de poches clients appartenant à une classe d'assurés donnée dont la propension à résilier est très différente de celle de cette classe d'assurés. La découverte de ces poches, allant contre l'intuition, permet de mettre en place des actions opérationnelles pour optimiser la gestion du portefeuille de l'assureur et ainsi augmenter sa rentabilité dans un contexte Hamon.

Les techniques employées, et notamment les cartes auto-adaptatives, permettent d'obtenir aisément des sous-groupes du portefeuille relativement homogènes en termes de comportement. L'interprétabilité des cartes auto-adaptatives a notamment permis d'améliorer la compréhension de la structure du portefeuille, tant au niveau de la répartition entre assurés que pour les comportements de résiliation au sein de chaque groupe. L'intérêt de l'utilisation des arbres de classification pour décrire les poches clients est de permettre l'attribution aisée d'un assuré à une des poches clients constituées en fonction de ses caractéristiques.

L'étude a permis de mettre en évidence, au sein du portefeuille de contrats auto du Partenaire, 12 groupes homogènes d'assurés comme par exemple les jeunes conducteurs ou les cadres supérieurs. L'analyse des taux de résiliation de ces différents groupes a poussé à l'étude du cluster des clients âgés et fidèles sans réduction tarifaire qui est un groupe résiliant plus que la moyenne du portefeuille. 10 poches clients ont été mises en évidence au sein de ce groupe, avec des taux de résiliation également hétérogènes. Deux de ces poches ont ensuite été analysées, l'une étant constituée de clients ayant une très forte propension à résilier et l'autre regroupant des clients très fidèles. L'analyse des facteurs de la résiliation dans ces deux groupes a ensuite mis en évidence plusieurs variables significatives pour expliquer la résiliation tandis que certaines autres étaient spécifiques à l'une des poches.

L'approche du mémoire présente quelques limites. Le retraitement dans les variables crée notamment un biais dans le résultat final. De plus, les cartes auto-adaptatives demandent une puissance de calcul importante. Cependant, comme démontré, elles peuvent être employées sur plus de deux millions de lignes de données. Enfin, le modèle prédictif de la résiliation présentait des performances améliorables et l'utilisation de plus de données dépendant du temps (connexions de l'assuré sur le site de l'assureur, ...) aurait pu améliorer les performances de prédiction.

Pour aller plus loin, il serait aussi intéressant de prendre en compte la charge de sinistres, variable qui n'était pas disponible dans la base fournie. Cela permettrait de mettre en place un modèle de la valeur client qui serait plus intéressant qu'une étude de la résiliation. Dans un contexte où l'assureur doit améliorer la connaissance de ses clients, il aurait aussi été intéressant de réaliser une étude de la résiliation à la maille du client plutôt qu'à celle du contrat auto comme cela a été fait. Un dernier point d'amélioration du modèle obtenu est de quantifier la probabilité d'appartenance à une

poche donnée plutôt que de savoir seulement si un assuré est dans une poche plutôt que dans toutes les autres. Cela aurait permis d'obtenir une probabilité de résiliation individuelle pour chaque assuré, en décloisonnant l'information sur la résiliation disponible dans chaque poche.

Ce mémoire d'actuariat se terminant ici, il convient de conclure par trois citations pouvant résumer tout l'apport des techniques de l'apprentissage automatique pour l'analyse de données. Celles-ci décrivent la façon dont le travail a été mené et comment on peut en interpréter les résultats.

The most exciting phrase to hear in science, the one that heralds new discoveries, is not "Eureka !" but rather "Hmm... that's funny..."

—Isaac Asimov

The first step towards predicting the future is admitting you can't

—Stephen Dubner, Freakonomics Radio, 30 mars 2011

The "prediction paradox" : The more humility we have about our ability to make predictions, the more successful we can be in planning for the future.

—Nate Silver, The Signal and the Noise : Why So Many Predictions Fail—but Some Don't

Bibliographie

BREIMAN L. (2001). Random forests. *Machine learning*, 45(1) :5–32. 92

BREIMAN L., FRIEDMAN J., OLSHEN C.-J. and STONE R.A. (1984). Classification and regression trees. V

CEIOPS (2009). CEIOPS' Advice for Level 2 Implementing Measures on Solvency II : Technical Provisions - Article 86 f Standard for Data Quality.

CHEVRIER C. (2013). Le big data, une solution miracle ? *L'Argus de l'assurance*. 12

CHOQUET C. (2011). Structuration d'une offre pour les jeunes conducteurs. *Mémoire d'actuariat*. 19

COSTES Y. (2000). Comprendre et mesurer le profil et le comportement des internautes. *Revue française du marketing*, (177) :153–168. 19

ENTORF H. and SPENGLER H. (2000). Socioeconomic and demographic factors of crime in germany : Evidence from panel data of the german states. *International review of law and economics*, 20(1) :75–106. 19

FRIEDMAN J., HASTIE T. and TIBSHIRANI R. (2001). *The elements of statistical learning*, volume 1. Springer. 94

FROIDEFOND E.A. (2014). Le big data dans l'assurance. *MBA ENASS*, page 89. 20

KOHONEN T. (1982). Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1) :59–69. 48

KOHONEN T. (1998). The self-organizing map. *Neurocomputing*, 21(1) :1–6. 52

KRIEGER J. (2014). L'exploration des données en assurance : apport de l'Analytics. *Mémoire d'actuariat*, pages 35–65. 24

LEGIFRANCE (2014). *Loi n°2014-344 du 17 mars 2014 relative à la consommation ou loi Hamon*. 7

O DUDA R., HART P. and STORK D. (2001). Pattern classification. *A Wiley-Interscience*, pages 526–527. III

PÖLZLBAUER G. (2004). *Survey and comparison of quality measures for self-organizing maps*. 53

RAND W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*. 90

REPUBLIQUE FRANCAISE (2014). Michel Sapin, Emmanuel Macron et Carole Delga saluent la publication du décret relatif à la résiliation à tout moment des contrats d'assurance. *Communiqué de presse*. 8

SCHAAL O. (2014). L'assurance des biens et des responsabilités. *Cours de l'Université Paris Dauphine*. 4

SIEGEL E. (2013). *Predictive analytics : The power to predict who will click, buy, lie, or die*. John Wiley & Sons. 13, 14, 16, I

ULTSCH A. (2003). *U*-matrix : a tool to visualize clusters in high dimensional data*. Fachbereich Mathematik und Informatik Berlin. 54

VESANTO J. and ALHONIEMI E. (2000). Clustering of the self-organizing map. *Neural Networks, IEEE Transactions on*, 11(3) :586–600. 57

WEHRENS R. and BUYDENS L. (2015). Self- and Super-organizing Maps in R : The kohonen Package. 52

YIN H. (2008). The self-organizing maps : Background, theories, extensions and applications. In *Computational intelligence : A compendium*, pages 715–762. Springer. 50

Annexes

Five effects of Prediction (en anglais)

Ces phénomènes sont décrits par Eric Siegel, ex-professeur de Machine Learning à Columbia University, dans son livre : *Predictive Analytics : The Power to Predict Who Will Click, Buy, Lie or Die* (voir SIEGEL E. (2013)). Ils sont très généraux et guident le développement d'un modèle de machine learning. Cette annexe les liste tels qu'ils sont écrits dans le livre.

1. The Prediction Effect : A little prediction goes a long way.
2. The Data Effect : Data is always predictive.
3. The Induction Effect : Art drives machine learning ; when followed by computed programs, strategies designed in part by informal human creativity succeed in developing predictive models that perform well on new cases.
4. The Ensemble Effect : When joined in an ensemble, predictive models compensate for one another's limitations, so the ensemble as a whole is more likely to predict correctly than its component models are.
5. The Persuasion Effect : Although imperceivable, the persuasion of an individual can be predicted by uplift modeling, predictively modelling across two distinct training data sets that record, respectively, the outcomes of two competing treatments.

Quantification vectorielle

Soit $\mathcal{E}$ un ensemble de données de taille et dimension quelconques. Une quantification vectorielle de $\mathcal{E}$ se définit par une fonction f et un ensemble $\mathcal{Q} \subset \mathcal{E}$ telle que :

$$\forall \mathrm{x} \in \mathcal{E}, f\,(\mathrm{x}) \in \mathcal{Q}$$

Comme on a à priori $\mathrm{Card}\, Q < \mathrm{Card}\, E$, f résume les données de $\mathcal{E}$ en un nombre de valeurs plus petit.

Exemple Si on prend $\mathcal{E} = \mathbb{R}$, $\mathcal{Q} = \mathbb{N}$ et f la fonction partie entière, alors on voit que f réalise un résumé de $\mathrm{x} \in \mathbb{R}$, l'entier tout de suite inférieur à x.

C'est ce principe consistant à résumer les données qui est utilisé dans les Self-Organizing Maps mais aussi dans l'algorithme K-means et même dans les arbres.

Algorithme des K-moyennes (K-means)

Soient n et p des entiers strictement positifs. On note $\mathcal{X} = \{\mathrm{x}_i\}_{i\in[\![1,n]\!]} \in (\mathbb{R}^p)^n$ les données. On cherche à les classer en K classes $\mathcal{S} = (S_1, \cdots, S_K)$, appelées clusters formant une partition de $\mathbb{R}^p$.

On note c_k le centre du cluster S_k défini par :

$$c_k = \frac{1}{|S_k|} \sum_{\mathrm{x}\in S_k} \mathrm{x}$$

avec $|S_k|$ le nombre d'éléments contenus dans S_k.

L'objectif du K-means (voir O DUDA R., HART P. and STORK D. (2001)) est de trouver une partition $\mathcal{S}^*$ minimisant la sommes des variances intra-classes, c'est-à-dire trouver $\mathcal{S}^*$ telle que :

$$\mathcal{S}^* = \mathrm{argmin}_{\mathcal{S}=(S_1,\cdots,S_K)} \sum_{i=1}^{K} \sum_{\mathrm{x}\in S_i} \|\mathrm{x} - c_i\|^2$$

$\sum_{\mathrm{x}\in S_i} \|\mathrm{x} - c_i\|^2$ représente la variance des distances des points de S_i à son centre c_i.

Le but du K-means est donc d'obtenir des clusters pour lesquels le centre et les points ne sont pas trop éloignés, de façon à avoir des groupes homogènes.

Ce problème de minimisation étant NP-difficile, il est nécessaire de recourir à des algorithmes heuristiques pour le résoudre. Voici l'algorithme généralement utilisé :

Algorithme

Initialisation On initialise aléatoirement K centres $\left(c_1^{(0)}, \cdots, c_K^{(0)}\right)$ pour les clusters $\left(S_1^{(0)}, \cdots, S_K^{(0)}\right)$.

Attribution aux clusters On note A la fonction telle que $A(\mathrm{x}_i) = k$ si $\mathrm{x}_i \in S_k$. A l'étape t, on attribue chacun des vecteurs au cluster $S_k^{(t)}$ le plus proche. On a :

$$\forall i \in [\![1, n]\!], A^{(t)}(\mathrm{x}_i) = \mathrm{argmin}_{k\in[\![1,K]\!]} \|\mathrm{x}_i - c_k^{(t)}\|$$

Calcul des nouveaux centroïdes Si $t = T$, fin de l'algorithme. Sinon on calcule les coordonnées des centroïdes $\left(c_1^{(t+1)}, \cdots, c_K^{(t+1)}\right)$.
On a :

$$\forall k \in [\![1, K]\!], c_k^{(t+1)} = \frac{1}{\left|S_k^{(t)}\right|} \sum_{\mathrm{x}\in S_k^{(t)}} \mathrm{x}$$

Tant que $t < T$ avec T prédéterminé, $t = t + 1$ et retour à l'étape d'attribution aux clusters.

On trace en figure 23 [10] les différentes étapes de l'algorithme de l'initialisation aléatoire des clusters à la convergence apparente de l'algorithme.

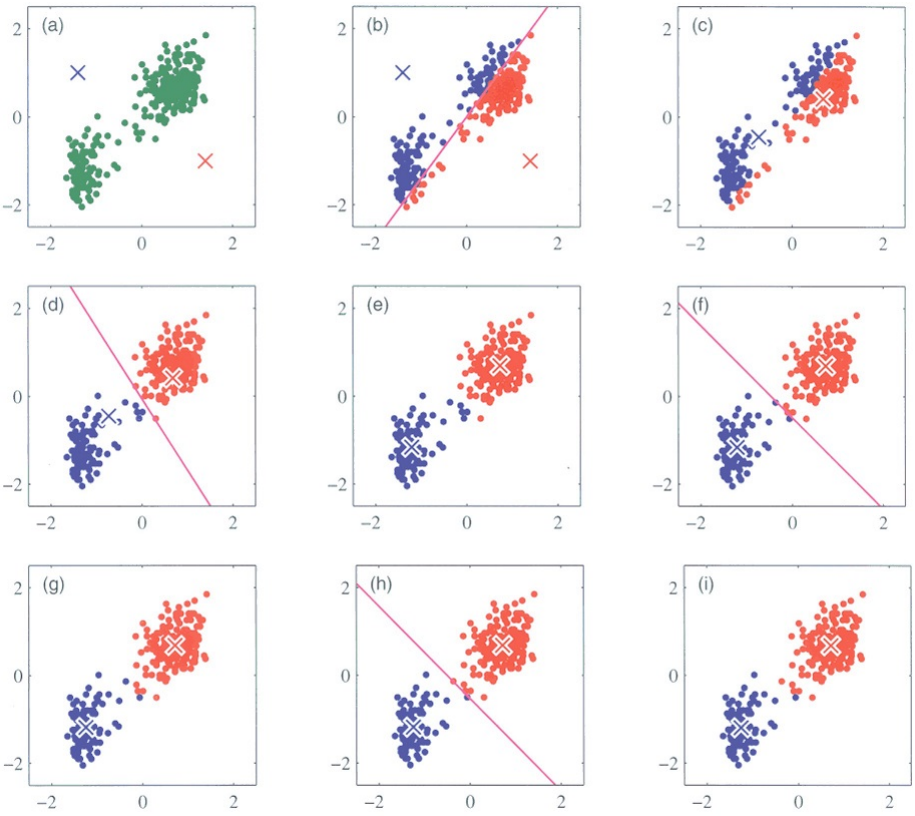


FIGURE 23 – Illustration des étapes de l'algorithme

L'algorithme converge vers une partition de $\mathbb{R}^p$ qui forme des groupes homogènes.

L'algorithme K-means a l'avantage d'être plus robuste que la classification ascendante hiérarchique.

10. Source : Fäy, Vayatis et Mougeot - Cours de l'Ecole Centrale Paris

Théorie des arbres CART

Les arbres de classification et de régression sont des méthodes qui cherchent à découper les données à l'aide de critères comme $\{\text{Variable}_1 < 12\}$, ... afin de prédire l'appartenance à une classe (pour les arbres de classification) ou une valeur (pour les arbres de régression). L'algorithme le plus connu d'arbres de classification est le CART pour *Classification and Regression Trees* (voir BREIMAN L., FRIEDMAN J., OLSHEN C.-J. and STONE R.A. (1984)).

Un arbre se présente sous la forme ci-après (voir figure 24). Il est constitué de nœuds et de feuilles. A chaque nœud, l'arbre pose une question aux données dont la réponse est oui ou non. Une feuille regroupe les données répondant aux critères (oui ou non) des questions posées dans les nœuds qui la précède.

Dans la feuille la plus à gauche, on peut lire que 76% des observations vérifient les conditions « La prime est supérieure ou égale à 1012 et la duration est inférieure à 7.5 ». Parmi ces observations, on en compte 95% avec le label « Yes » et 5% avec le label « No ».

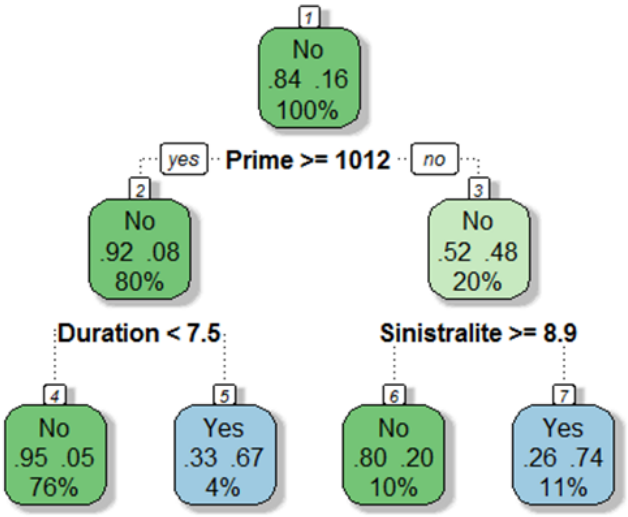


FIGURE 24 – Exemple d'arbre de classification CART

On utilisera les arbres de régression pour l'influence des variables sur la résiliation. Ceux-ci, comme les arbres de classification, reposent sur un découpage des données en régions (voir figure 25).

Soit $n \in \mathbb{N}^*$. Soit $\mathcal{Z} = \left\{\left(\mathrm{x}^{(k)}, y_k\right)\right\}_{k\in[\![1,n]\!]}$ avec $\forall k \in [\![1,n]\!], y_k \in \mathbb{R}$ et $\mathrm{x}^{(k)} = \left(x_1^{(k)}, \cdots, x_p^{(k)}\right) \in \mathbb{R}^p$ les données que l'on considère.

Pour les arbres de régression, on cherche à prédire y_k sous la forme :

$$\psi\left(\mathrm{x}^{(k)}\right) = \sum_{i=1}^{L} \bar{y}_i \cdot \mathbb{I}\left(\mathrm{x}^{(k)} \in \mathrm{R}_i\right)$$

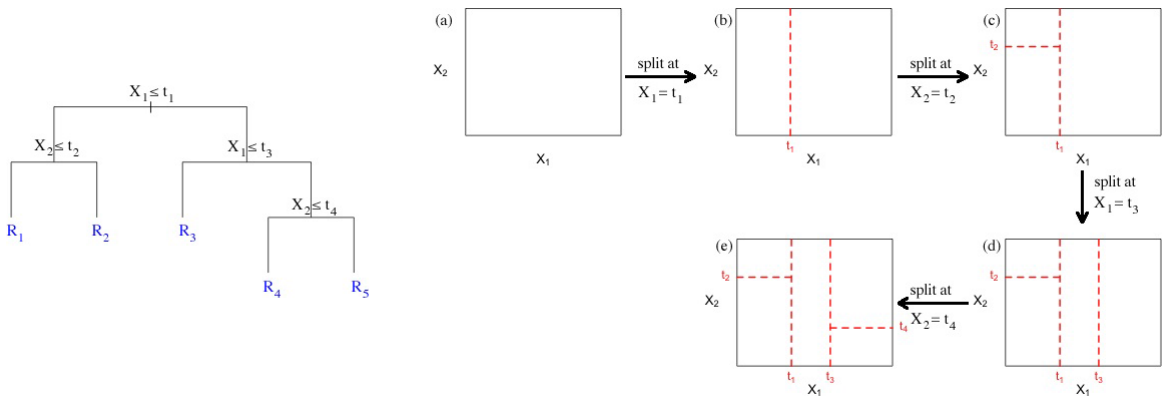


FIGURE 25 – CART et découpage en régions (Source : Fäy, Vayatis et Mougeot)

avec $(\mathrm{R}_i)_{i\in[\![1,L]\!]}$ les L régions définies par les feuilles de l'arbre et $\bar{y}_i = \frac{1}{|\mathrm{R}_i|} \sum_{(\mathrm{x},y)\in\mathcal{Z}\backslash \mathrm{x}\in\mathrm{R}_i} y$.

$\psi(\cdot)$ est donc une fonction constante par morceaux.

A chaque nœud, dans le processus de création de l'arbre, l'algorithme cherche à créer 2 nœuds fils tels que la somme des variances à l'intérieur de ces 2 nœuds fils est minimale. En notant Y_g (resp. Y_d) les valeurs des variables cibles à prédire dans le nœud fils gauche (resp. droit), le but est de trouver le critère de découpage optimal, c'est-à-dire la solution du problème :

$$\min_{x_i \leqslant x_j^R, j=1,\cdots,M} \left[P_g \operatorname{Var}(Y_g) + P_d \operatorname{Var}(Y_d)\right]$$

avec P_g (resp. P_d) l'effectif du nœud gauche (resp. droit).

Cette procédure est appliquée tout d'abord à l'ensemble des données puis à chaque nœud créé par la procédure. On s'arrête si le nœud contient moins de $n_{\min}$ observations. On a souvent $n_{\min}$ égal à 10% de la taille de la base d'apprentissage.

Une fois l'arbre noté $\mathcal{T}_0$ de taille maximale créé, il est nécessaire de l'élaguer afin d'enlever certaines feuilles pour obtenir un sous-arbre $\mathcal{T}_* \subset \mathcal{T}_0$ optimal. Cette opération permet de rendre l'arbre plus robuste dans sa capacité à prédire sur des nouvelles données. Il s'agit d'éviter le phénomène de sur-apprentissage. Cependant, un arbre qui n'a pas assez de feuilles ne permettra pas non plus une bonne prédiction de par sa faible complexité. Il est donc nécessaire d'établir une méthode pour obtenir $\mathcal{T}_*$.

On note, pour $i \in [\![1, L]\!]$:

$$Q_i = \frac{1}{|\mathrm{R}_i|} \sum_{(\mathrm{x},y)\in\mathcal{Z}\backslash \mathrm{x}\in\mathrm{R}_i} (y - \bar{y}_i)^2$$

De plus, on note $|\mathcal{T}|$ le nombre de feuilles ou nœuds terminaux de l'arbre noté $\mathcal{T}$.

Pour $\alpha \geq 0$, on pose alors :

$$C_\alpha(\mathcal{T}) = \sum_{i=1}^{|\mathcal{T}|} |\mathrm{R}_i|\, Q_i + \alpha\, |\mathcal{T}|$$

Afin de réaliser l'élagage de l'arbre, on va chercher, pour α dans un ensemble fini, le sous-arbre $\mathcal{T}_\alpha \subseteq \mathcal{T}_0$ minimisant $C_\alpha(\mathcal{T})$. Le paramètre α contrôle le compromis entre la taille de l'arbre et le bon ajustement aux données (c'est-à-dire une bonne prédiction). Pour $\alpha = 0$, le sous-arbre optimal est $\mathcal{T}_0$ et plus α augmente, plus le sous-arbre optimal est petit jusqu'à être réduit à sa racine $\mathcal{T}_\infty$.

Pour trouver le sous-arbre $\mathcal{T}_\alpha$ optimal à α donné, on fusionne tour à tour les feuilles adjacentes de l'arbre qui provoquent la plus petite augmentation dans $\sum_i |\mathrm{R}_i|\, Q_i$ jusqu'à obtenir l'arbre égal à la racine. Bien entendu, lorsqu'il n'est possible de fusionner que 2 feuilles, on les fusionne. Il est possible de montrer que parmi les arbres obtenus, on trouve $\mathcal{T}_\alpha$ qui est donc l'arbre minimisant $C_\alpha(\mathcal{T})$.

Le meilleur arbre $\mathcal{T}_{\alpha^*}$ est celui qui minimise l'erreur quadratique moyenne sur des nouvelles données, la base de test. Pour plus de robustesse, on utilise souvent la procédure suivante :

Validation croisée « k-fold » La validation croisée est une méthode d'estimation de la fiabilité d'un modèle basé sur des paramètres devant être réglés. Elle permet de trouver les paramètres qui rendent le modèle optimal. La validation croisée « k-fold fg est une des techniques de validation croisée existantes. Pour cela, les données sont tout d'abord découpées en K échantillons $\mathcal{Z}_1, \cdots, \mathcal{Z}_K$.

Obtention des valeurs du paramètre de complexité Tout d'abord, on calcule l'arbre non élagué $\mathcal{T}_0$ associé aux données. Cet arbre est de taille $|\mathcal{T}_0|$. Cet arbre permet de construire $m = |\mathcal{T}_0|$ sous-arbres différents que l'on obtient en élaguant avec différents valeurs de α. Ceci définit m intervalles pour α sur lesquels l'arbre optimal trouvé est le même pour tout $\alpha \in I_i$:

— $I_1 = [0, \alpha_1[$,
— $I_2 = [\alpha_1, \alpha_2[$,
— $\cdots$
— $I_{m-1} = [\alpha_{m-2}, \alpha_{m-1}[$,
— $I_m = [\alpha_{m-1}, +\infty[$.

On pose alors :

— $\beta_1 = 0$,
— $\beta_2 = \sqrt{\alpha_1 \alpha_2}$,
— $\cdots$
— $\beta_{m-1} = \sqrt{\alpha_{m-2} \alpha_{m-1}}$,
— $\beta_m = +\infty$.

Calcul des arbres Pour chaque $k \in [\![1, K]\!]$, on calcule les arbres optimaux $\mathcal{T}_{\beta_1}, \cdots, \mathcal{T}_{\beta_m}$ calculés avec toutes les données sauf $\mathcal{Z}_k$.

Pour chaque observation de $\mathcal{D}_k$, on calcule la valeur prédite par l'arbre $\mathcal{T}_{\beta_j}$ pour $j \in [\![1, K]\!]$. Grâce à cela, on calcule chaque $\mathcal{R}\Big(\mathcal{Z}_k, \mathcal{T}_{\beta_j}\Big) = \left[\sum_i |\mathrm{R}_i| \, Q_i\right]_{\mathcal{Z}=\mathcal{Z}_k, \mathcal{T}=\mathcal{T}_{\beta_j}}$.

Choix de l'arbre optimal On choisit alors β qui minimise la fonction suivante :

$$\mathcal{R}\Big(\mathcal{T}_{\beta_j}\Big) = \frac{1}{K} \sum_{k=1}^{K} \mathcal{R}\Big(\mathcal{Z}_k, \mathcal{T}_{\beta_j}\Big)$$

On obtient l'arbre final élagué en calculant $\mathcal{T}_\beta$ minimisant $C_\beta(\mathcal{T})$ et calculé sur toutes les données. On a souvent $K = 5$ ou $K = 10$.

Pour les arbres de classification, une approche semblable est employée.

Avantages et limites Les arbres sont facilement compréhensibles. Ce ne sont pas des algorithmes de type « boîte noire ». Un autre avantage des arbres est d'afficher clairement quelles sont les variables explicatives de la variable cible. Il est possible de raffiner les critères d'impureté pour s'adapter aux besoins de l'étude. De plus, les arbres peuvent gérer les valeurs manquantes et utilise pour cela un critère de coupe alternatif (*surrogate split*) pour les entrées dont la variable choisie pour la coupe est manquante. Cependant, ils peuvent être très instables. Si les données changeaient légèrement, l'arbre pourrait être très différent de celui trouvé avant changement. Le calcul de l'arbre optimal étant un problème NP-difficile, il est nécessaire d'utiliser les formules approchées pour construire l'arbre. L'arbre trouvé par les méthodes implémentées sur les logiciels de statistiques peut donc être sous-optimal. De plus, il est difficile de capturer dans les arbres des relations de la forme $Y = \mathbb{I}(x_1 + x_p \geq 2)$.

Distribution des variables sur la carte auto-adaptative

Par souci de concision, toutes les cartes obtenues lors de la segmentation n'ont pas été affichées dans le corps du texte. Elles sont disponibles ici pour de plus amples informations. Les valeurs sont signifiées par un dégradé du bleu au rouge en passant par le vert et le jaune. Le bleu correspondant aux petites valeurs et le rouge aux grandes. Lorsqu'une variable est catégorique, le rouge signifie 1 et le bleu 0.

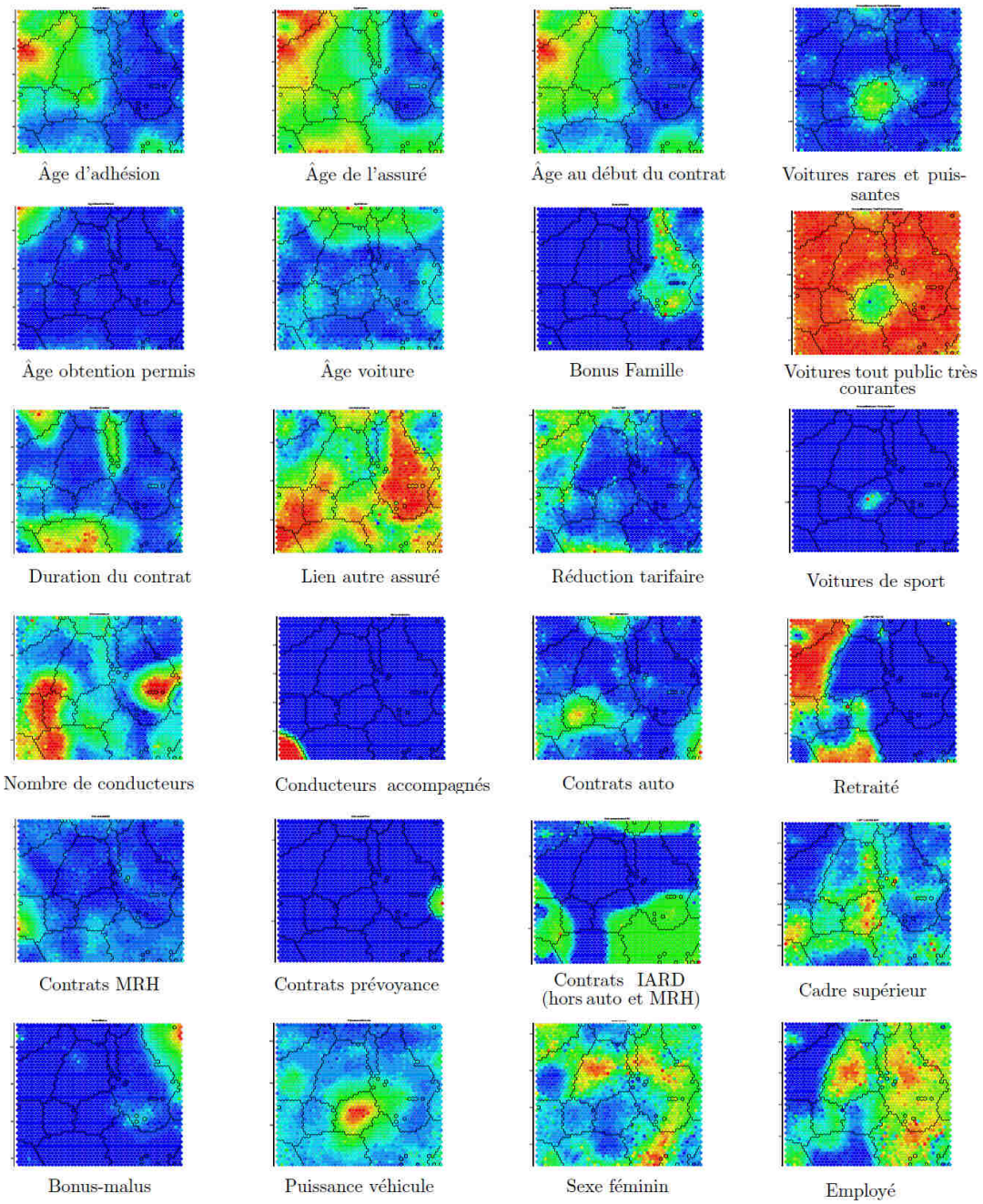


FIGURE 26 – Distribution des variables sur la carte auto-adaptative (1)

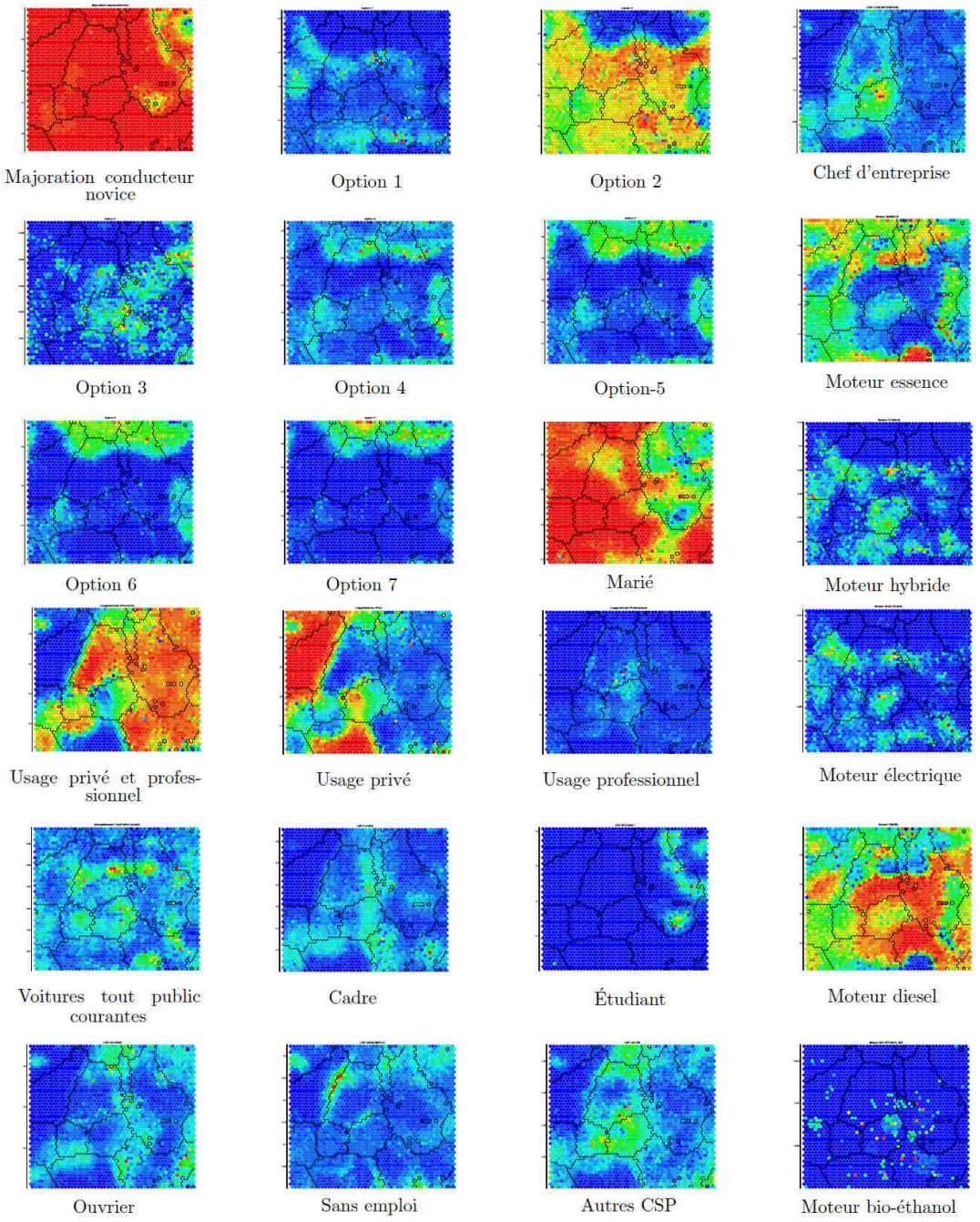


FIGURE 27 – Distribution des variables sur la carte auto-adaptative (2)

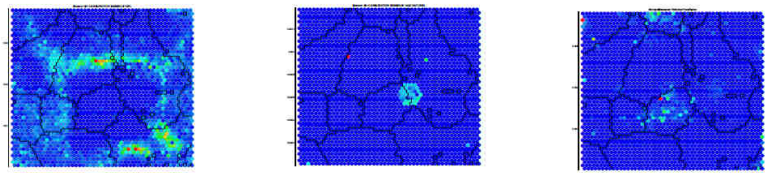


FIGURE 28 – Distribution des variables sur la carte auto-adaptative (3)